EPA's Model Averaging Methods for Dose-Response Analysis Workshop Webinar
Responses to Discussion Questions

Discussants:
Ruth Hummel, US EPA
Michael Messner, US EPA
Walter Piegorsch, University of Arizona
Woody Setzer, US EPA
Matthew Wheeler, National Institute for Occupational Safety and Health (NIOSH)

***All figures and references in Appendix**

OVERALL COMMENTS

Ruth Hummel: The workshop materials and the analyses therein (as well as the conclusions and suggestions for the future) are well thought-out and craft a very reasonable domain of model options and simulated and real data sets to represent the field – this is a great framework to build on for any follow-up comparisons and sensitivity studies. Thank you for advancing the work on dose-response Model Averaging for use in EPA chemical risk assessments. We in the program offices are very excited to improve the performance of our estimation by incorporating model uncertainty, and I am personally very hopeful that these materials can be wrapped up and distributed quickly for use at EPA. Before addressing the Discussion Questions, I have two general comments about common dose-response practices (which underpin the testing of the Model Averaging methods and therefore seem relevant here).

1. *General concern over using summarized data instead of raw data for continuous endpoints*
   I wonder at the reduction of the continuous data to means and variances for use in these models. This seems to be a standard practice throughout the dose-response literature, but I struggle to understand why. In my work analyzing data for an EPA program office, there are certainly cases when I am restricted to information summarized in a publication. This will often be summary statistics like means and variances per dose group. But in most cases, I have access to the raw underlying data (and even in the case of published work, it is often possible to request the underlying data). The means and variances per dose group are not sufficient statistics for the purpose of maximum likelihood estimation of the model parameters, so we are certainly losing important information contained in the raw data. While the performance coverage estimates presented in the workshop materials were computed only for the simulated data (which is likely less perturbed from the generating distributions than we might expect from real data), nevertheless it is likely that the model likelihoods, and therefore the model weights in the averaging, will differ when fitting for the raw data compared to fitting the summarized data. I'm curious to see this impact in the coverage rates, and I certainly recommend retaining the full data information when fitting these models in general.
2. *Modeled control mean versus observed control mean for calculation of BMR response*
   Another source of variability in the BMD(L) is the use of the modeled control mean to determine the BMR response. It would be interesting to see the output for a relative deviation from the modeled control response along with output from the observed control response (and even compared to the results from a historical control response). There are certainly theoretical advantages to comparing to the modeled control dose, but when the true model is unknown, and the data must be the driver for the modeling, it makes sense to me to use a data-driven

value for the control dose rather than relying twice (for the model shape AND for the control value) on a possibly incorrect model.

DISCUSSION QUESTION 1: Overall approach to model-averaging – Are there other model averaging methods that EPA should consider?

Ruth Hummel: No. The methods presented in the supporting document are sufficient to represent reasonable ideas and methods currently suggested in the literature. However, I recommend further investigation into candidate weighting schemes and further investigation into candidate model suites as well as adoption of a nonparametric approach to the bootstrapping.

Michael Messner: I have no suggestions.

Walter Piegorsch: The EPA report indicates that what is essentially a frequentist model averaging (FMA) approach using bootstrap resampling to calculate BMDLs with continuous dose-response data may represent the next evolution in quantitative risk assessment for addressing the clear problem of model uncertainty. They note in passing that Bayesian model averaging (BMA) could also be applied, and I would agree with this if the prior modeling and posterior calculations were done very carefully. (BMAs require proper prior distributions, and use of "vague" or diffuse priors in a BMA can detrimentally affect the posterior inferences.) Application of BMAs for continuous data has not been deeply studied, and could be worth investigation. [For BMA BMDLs with dichotomous-response data, see the article by Fang et al. (2015).] Also see item #10, below.

Woodrow Setzer: It would be good to see a comparison with AIC and especially AIC_c based weights. Burnham and Anderson (1998) were convinced of the superiority of AIC_c over BIC. Any evaluation should be over a wide range of dose-response curves and conditions, including a range of error variability. These simulations only considered a single level of variability for each error model.

Matthew Wheeler:

Model Choices- It is my experience that the models used in the proposed model average approach are often very difficult to fit using maximum likelihood estimation. This is especially true for the hill and exponential models. These two model forms frequently fail to converge, which may be problematic for a bootstrapping procedure, and may be a cause of the poor coverage.

Further, the main ideas behind model averaging is that a large suite of models, all having differing shapes be employed in the average. The EPA model suite for continuous models is especially limited, and may not be, in itself, adequate for model averaging. With this thought, I suggest the EPA add a suite of fractional polynomials (Faes et al., 2007) in the model average.

Additionally, it seems that the EPA is not considering variance models (i.e., heteroskedastic variance) to be a different model in the model average. In a continuous model, it would seem that the variance choices would also be considered as a model choice, and should be included in the model average. These additions, may provide a more robust suite of models.

Bootstrap Approach- The US EPA's proposed approach focuses on continuous models, where the data are derived from the reported sufficient statistics of a normal distribution, i.e., the mean and standard deviation. As it is a departure from other model averaging proposals, this altered focus is a significant challenge. Specifically, it provides a challenge in developing confidence intervals on the estimated statistics that are at the advertised type I error rate. As none of the methods used seem to be supported theoretically, this result is predictable.

When bootstrapping regression models, residuals of the data to the fitted model are bootstrapped, and, when the data display heteroscedasticity, more complicated approaches such as the Wild Bootstrap (Mammen, 1993; Flachaire, 2005) are employed to provide correct confidence intervals on the estimated statistic. When only the sufficient statistics are used in the regression standard

bootstrap methodologies mentioned above cannot be applied, and it is not clear that any of the methods used by the US EPA are theoretically appropriate. Though method 4b seems to be the most appropriate of all of the methods attempted, as it uses the sufficient statistics directly, it still does not produce acceptable coverage levels.

The problem is evident in the poor observed coverage in the simulations (e.g. the e1template), and it is difficult to recommend an alternate bootstrap methodology. Further, the EPAs profile likelihood approach seems equally problematic; because this method produces worse coverage than the bootstrap approach. As a possible remedy to this problem, I would suggest that the EPA investigate the Model Averaged Profile likelihood approach proposed by Fletcher and Turek (2012). This method is supported theoretically (Hjort and Claeskens, 2003) and has performed well in practice. I played with it some, and received coverage at or above method 4*b*: This method is not like the approach used by the EPA in their software. In that approach, the BMD is estimated from the averaged individual BMDLs computed at the given 100(1 – α)% lower confidence level. Instead, the model averaged BMD's has the following approximate posterior distribution given *Y*; which is

$$Pr(BMD|Y) = \sum_{i=1}^{M} w_i \, p_i(BMD|Y)$$

where *M* is the number of models considered, and w_i is the corresponding weight constructed using the *BIC*: The BMDL is then approximated by finding the value BMDL such that:

$$\int_{-\infty}^{BMDL} Pr(BMD|Y) \, dBMD = \alpha.$$

As the profile likelihood is an approximation of the integral $\int_{-\infty}^{BMDL} p_i(BMD|Y) \, dBMD$ one can use method of profile likelihood to compute (1). I believe this method should provide better coverage rates, and should be faster than the bootstrap methodology.

DISCUSSION QUESTION 2: Completeness of Suite of Models – Are there other parametric models that should be included in model-averaging?

Ruth Hummel: Maybe. I recommend further comparisons of potential model suites. For additional details, see my responses to question 5.

Michael Messner: I expect there are additional models. Other models should be sought when the likelihoods associated with the available models are all surprisingly small. Most of my dose-response modeling has been with microbial pathogens. Models for microbial dose-response include beta-Poisson, exponential, exponential with immunity, and fractional Poisson.

Walter Piegorsch: One could always think of additional forms for inclusion in the models being averaged (in various publications I've called this collection an "Uncertainty Class" of models). Based on the results seen in the EPA report with FMA for continuous data, I am led to argue for inclusion of any models the analysts or domain experts feel are pertinent for the risk assessment under study. I have no specific suggestions for other particular continuous dose-response models at this time, and would leave such decisions to the analysts.

Woodrow Setzer: The current set seems to be adequate. I have some concern that nested models are included in the set of models: linear in poly3, and exp3 in exp5. This can end up overweighting these models a bit.

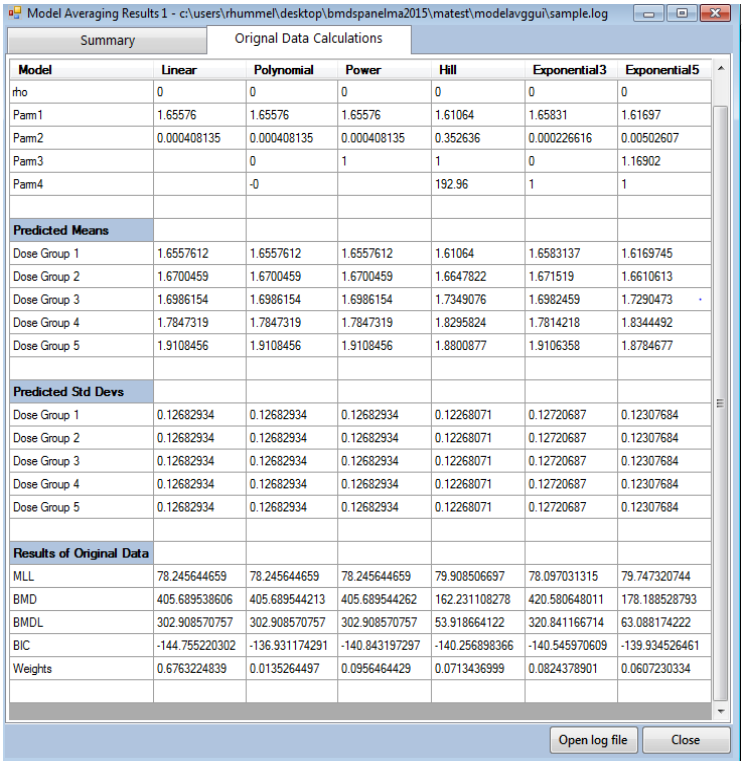
Matthew Wheeler: As recommended above, the EPA should investigate fractional polynomial models for continuous data. This provides a large suite of possible model shapes, which are easy to fit using maximum likelihood or least squares. Their use would expand the model space and provide very little added overhead in computational costs. I also question the use of the Hill model and the Exponential

model in the model average. These are very complicated models to fit, and it is not clear if larger more robust model suite is created if these models are necessary.

DISCUSSION QUESTION 3: Implementation of Methods – Do you agree with the approaches used to implement the methods reviewed in the workshop support material? In particular:

Ruth Hummel: Possible error (or a misunderstanding on my part):

These are the results from running 1,000 simulations (which comes up as the default, rather than 10,000 as stated in the Quick Start document, just FYI) on the sample data set included in the GUI folder. I'm looking at the Linear and Polynomial "Results of Original Data." It seems, based on the MLL, BMD, and BMDL being identical, that the polynomial model defaulted to the linear fit. Yet the BICs are not identical. I thought, from the first bullet point on page 8 in section 2.2.1, that the likelihood for the polynomial model would be set equal to that of the linear model. Then the BIC should be the same (or possibly off by 2, if you were still considering the Poly3 model to have three free parameters?), yet it is not. Is this displaying the calculated value before the substitution, or is the substitution not functioning properly?


Model Averaging Results 1 - c:\users\rhummel\desktop\bmdspanelma2015\matest\modelavggui\sample.log

Summary | Orignal Data Calculations

Model	Linear	Polynomial	Power	Hill	Exponential3	Exponential5
rho	0	0	0	0	0	0
Parm1	1.65576	1.65576	1.65576	1.61064	1.65831	1.61697
Parm2	0.000408135	0.000408135	0.000408135	0.352636	0.000226616	0.00502607
Parm3		0	1	1	0	1.16902
Parm4		-0		192.96	1	1
Predicted Means						
Dose Group 1	1.6557612	1.6557612	1.6557612	1.61064	1.6583137	1.6169745
Dose Group 2	1.6700459	1.6700459	1.6700459	1.6647822	1.671519	1.6610613
Dose Group 3	1.6986154	1.6986154	1.6986154	1.7349076	1.6982459	1.7290473
Dose Group 4	1.7847319	1.7847319	1.7847319	1.8295824	1.7814218	1.8344492
Dose Group 5	1.9108456	1.9108456	1.9108456	1.8800877	1.9106358	1.8784677
Predicted Std Devs						
Dose Group 1	0.12682934	0.12682934	0.12682934	0.12268071	0.12720687	0.12307684
Dose Group 2	0.12682934	0.12682934	0.12682934	0.12268071	0.12720687	0.12307684
Dose Group 3	0.12682934	0.12682934	0.12682934	0.12268071	0.12720687	0.12307684
Dose Group 4	0.12682934	0.12682934	0.12682934	0.12268071	0.12720687	0.12307684
Dose Group 5	0.12682934	0.12682934	0.12682934	0.12268071	0.12720687	0.12307684
Results of Original Data						
MLL	78.245644659	78.245644659	78.245644659	79.908506697	78.097031315	79.747320744
BMD	405.689538606	405.689544213	405.689544262	162.231108278	420.580648011	178.188528793
BMDL	302.908570757	302.908570757	302.908570757	53.918664122	320.841166714	63.088174222
BIC	-144.755220302	-136.931174291	-140.843197297	-140.256898366	-140.545970609	-139.934526461
Weights	0.6763224839	0.0135264497	0.0956464429	0.0713436999	0.0824378901	0.0607230334

Open log file | Close

3a) What is the viability of the alternative approach described in Section 4.2 for generating bootstrap samples called for in Methods 3 and 5 (i.e., treating the saturated model as another model that gets considered for use in generating the bootstrap sample)?

Ruth Hummel: I may be misunderstanding the purpose of the Method 3 and 5 approach: is there a theoretical benefit, e.g., capturing more of the variability in both the weighting metric (the likelihood) and the estimates as a result of varying the model from which the bootstrap data are generated? If so, then I think it might be worth pursuing the idea in bullet point 7 on the bottom of page 24 in Section 4.2. (If there is no theoretical advantage – and since the simulations tested so far show no advantage in practice– I recommend against using the Methods 3 and 5.)

That said, if the raw data are available, I would prefer to empirically bootstrap from the data rather than generating semi-parametric or parametric bootstrap samples. In my work we often do not know the true biological dose-response relationship. I would like to see the general methods developed for purely data-driven analysis, where prior knowledge (including for the bootstrap sampling) can be built in for use in cases where there is additional information beyond the data.

I recommend further testing of the impact of empirical versus semi-parametric (as in Methods 2 and 4) versus parametric (as in Methods 3 and 5) versus hybrid (as suggested in the bullet point) bootstrapping.

Michael Messner: It seems like this is "using the data twice" and could lead to false or over-confidence.

Walter Piegorsch: I was generally comfortable with use of bootstrapping to find the BMDLs, where necessary. I was, however, less excited by the call in Methods 3 and 5 for parametrically structured bootstraps, especially since the specific parametric model is chosen randomly (with, admittedly, use of information in the BIC weights). In most cases, my general predilection for employing bootstrap resampling leans toward nonparametric or if needed, semi-parametric bootstraps, since this better avoids problems with model uncertainty to which the fully parametric bootstrap can remain hostage.

Woodrow Setzer: It makes sense, but you'd want to run more simulations as a check. It would be important to ask, first, what problem with the current methods this solves. Right now, the biggest problem I can see with the MA methods tested here is the likelihood of very low coverages for some templates. This seems to be a problem related to being able to track some DR curves, not variance (see my **Figure 1**).

Matthew Wheeler: As stated before, I question the implementation of all of the Bootstrap methods. The approach that I lean on the most is 4b. As it uses the sufficient statistic, it is essentially the most non-parametric approach. One could compute a "group residual' from the fit to form mew sufficient statistics. That is, given the fit $\mu_j(d)$, for model j; one has the residual $\epsilon_{ji} = \bar{Y}_i - \mu_j(d_i)$. One could then sample these residuals with probability proportional to the weights $w_1, \ldots, w_M$, and form a new sufficient statistic $\bar{Y}_k = \mu_j(d_k) + \epsilon_{ji}$. This might perform better than 4b.

3b) What is the viability of the alternative approach described in Section 4.2 for modeling variance (i.e., fit a saturated variance model that allows each dose group variance to be estimated independently)? Would it be reasonable to use only a model for variance as a power of the mean with power = 0 as a boundary case (constant variance)? [This question is intended to apply only when variance is a nuisance parameter, i.e., when it is not part of the BMR]

Ruth Hummel: This seems like an important feature to add in the future. There will of course be a trade-off between the flexibility of allowing individual estimates (and the corresponding addition in complexity of estimation and increased model parameters) versus forcing constant variance or constraining the variances to follow a power model. From my experience analyzing dose-response data, I don't recall seeing many cases where the variances were significantly non-constant but followed a power model. My inclination is to recommend that this option be built in at some point, but with the option still available to try to fit each group's variance separately, and that some Lack of Fit test be provided to help the user decide if the power model is sufficient.

Michael Messner: Consider adding a model with two variance terms: one constant (additive) and one multiplicative. The proposed "independent variances" seems to loose, given the nature and small numbers of animals/subjects per dose-group.

Walter Piegorsch: This suggestion for fitting models that relax the assumption on the variance seemed novel. At a first glance, my biggest concern would be whether larger sample sizes would be required for the operation and if these would be available in practice.

Woodrow Setzer: Fitting a separate variance for each dose group Is certainly feasible, though will be somewhat inefficient if it is unwarranted. This could be included as an option to use when variances seem heterogeneous, but without pattern. If I understand it, the second part of this question suggests JUST modeling the variance, allowing the boundary case as a possibility. This gets around the problem of somehow evaluating variance heterogeneity first. This seems like a good idea.

Matthew Wheeler: I think the alternative variance models should be part of the model average. There is a model selection component that is being ignored if the different variance models are not included in the model average.

3c) Is an investigation of alternatives to the BIC-based weights warranted (see last bullet in Section 4.2)? What is your opinion about weights based on information criteria in general? Which approach best approximates Bayesian model averaging?

Ruth Hummel: The weighting of the models is an area that needs further research. The weighting is really the major point of this entire endeavor: we will certainly have different results and coverage rates for various possible weighting schemes. As was noted on page 19 under case 3 of 3.2.2, "enough weight was still given to the other models to "degrade" the performance of the averaging." Can we correct this? In West et al. (2012) there was a clear preference for lower-order models by use of the AIC for the weighting. Can the -2k penalty be bias-corrected specific to this dose-response context (with similar experimental designs and sample sizes and parameters)? Can we use other features of the model fit, features that EPA's BMDS Guidance already incorporates into the choice of a single model e.g., BMD/BMDL ratio, fit at the dose(s) nearest the BMR, etc., to contribute to the weighting values? I would like to see, following or in tandem with additional testing on various suites of models (because of the interrelationship between the suite of models and the selected weights) (and possibly in tandem with a more empirical treatment of the data by bootstrapping from the raw data and using the raw data for fitting the models), a comparison of coverage performance for candidate weight criteria. The weight criteria seem so fundamental to the selection and use of a model averaging (or BMD(L) averaging) technique that I would consider this an essential and immediate research need.

That said, as I will describe in question 9, I still recommend going forward with this current advancement for use now, since it will certainly improve on the single-best-model approach currently in use. Delaying use of an improvement while we work out even more room for improvement is not practical, given the typical pace of innovation in government.

One final comment on the weighting: according to the workshop materials, when one model is a limiting case of another model and the more complicated model defaults to the simpler model, the more complicated model results are set equal to those of the simpler model. It probably makes sense to do this as described in the workshop materials in order to keep the method consistent and the results comparable for any data set over the full model suite, but I am curious how the essentially double-weighting of a single model would compare to, say, removing the failed model from the pool (as in Piegorsch 2014 section 4.1, referencing Wheeler and Bailer (2009)) rather than reducing it to the limiting model.

Michael Messner: Not sure. I would like to see some outcomes treated different ways (BIC weights vs alternatives)

Walter Piegorsch: Yes, other alternative ICs should be investigated. For instance, we (Piegorsch et al., 2013) studied a non- bootstrap form of FMA BMDLs with dichotomous data. They found that the AIC actually worked better than the BIC (and, than similar alternatives such as AICc or KIC) when employed in Akaike-style weights, as in the EPA report. Based on our results, I would strongly argue for further study of other ICs.

Woodrow Setzer: Investigation of a variety of information criteria-based weights is warranted. While weights based on information criteria have shown themselves to be practical, the literature seems pretty murky about whether one of the criteria is clearly superior to the others. My recollection is that which information criterion best approximates Bayes factors depends on the priors used. However, that is not the right criterion to use. What you want is the best estimate available of true BMDs, and the best quantitation of uncertainty available. Again, while Bayes MA is reasonable, it is not at all clear it is globally optimal.

Matthew Wheeler: Alternative weighting strategies should be investigated only if a viable bootstrap procedure is identified. If a viable Bootstrap is identified, then the AIC and possibly the FIC should be

investigated. The BIC is the only procedure that approximates Bayesian model averaging. The EPA should also consider the Laplace approximation to the weights. This approximation is more accurate than the BIC approximation, and this approximation may not suffer from the BIC's problem of picking models that are too small.

3d) What options would you recommend for dealing with experiments having fewer than four positive dose groups plus a control?

Ruth Hummel: I have no recommendations at this time.

Michael Messner: I'm also concerned with the numbers of animals/subjects per dose-group. I need to better understand how the bootstrap is employed when each dose-group results are summarized by statistics (mean and standard deviation).

Walter Piegorsch: Not to seem facetious, but in short I'd tell the investigators to acquire more data. Our team's experience with fewer than 4 groups (and, even with only four or five groups) has shown that such limited dose-response information makes estimation of the dose-response relationship and of complicated dose-related quantities such as the BMD – and its BMDL – extremely difficult. This essentially undermines the goal of making careful, accurate inferences on the BMD. We have argued that moves to upwards of 6–8 groups, incl. the control, are necessary if truly accurate estimation of BMD is the study's primary goal.

Woodrow Setzer: That is a hard question. One approach, which we suggested in our paper, would be to use Bayesian methods and informative priors. With enough information about plausible values for the parameters in question, that should make the model parameters identifiable in studies with fewer than three positive dose groups plus control (assuming this question has misspoken – there is no problem fitting models with four parameters to datasets with four positive dose groups plus control).

Matthew Wheeler: As long as large model space is identified, one could conceivably do model averaging for a control plus two positive dose groups, assuming the Hill and Exponential models are removed. Below that, a line should be fit and the BMDL calculated from this line.

DISCUSSION QUESTION 4: Testing Approach – Should additional testing be performed to identify a model averaging approach for dose-response analyses that offers the greatest advantage for the development of chemical health assessments? For example:

4a) Should additional dose-response patterns be tested? For instance, the workshop support material suggests that the Exp4 and Exp2 models could be added because they are bounding cases for models already considered.

Ruth Hummel: Not rigorously at this time. (I am interpreting this question to mean dose-response patterns for generating the simulated data. I certainly do recommend testing additional models in the suite of models, as discussed in question 5.)

Michael Messner: This may be a bit off target for these questions, but this is where it occurred to me: Can't the likelihood, itself, indicate when the best model is still poor? Take the max likelihood parameter values for the best-fitting model and repeatedly simulate a study of the same design, each time, observing the likelihood. If the great bulk of those likelihoods are greater than the likelihood of the actual data, then we would know that even the best fitting of the available models is very poor and suggests that either the data are bad or some other model is needed.

Walter Piegorsch: Yes, consider a wider variety of dose-response patterns.

Woodrow Setzer: Yes. You only have 16 dose-response curves in your test set. It would be good to extensively expand that set to look more like the distribution of dose-response shapes that are out

there. In addition, it would be useful to be sure that known or suspected problem cases are included In the test set. You also need to consider different levels of variability, relative to the dynamic range of the dose-response curves in the test set.

Matthew Wheeler: To fully test the methodology, more bounding models should be considered, and the EPA should attempt to find an approach that reaches nominal coverage (or at least is better than the Hill and Exponential5 model).

4b) Would testing of additional relative risk BMR values (e.g., 1% and 5%) provide additional information that could impact EPA's decision regarding the identification of a model averaging approach for dose-response analyses that is best suited for the development of chemical health assessments?

Ruth Hummel: Not rigorously at this time. Such a comparison is not critical for the purpose of identifying a model averaging approach, although some small comparison of a subset of case studies would provide a useful approximation of the sensitivity of the results to this factor. Additional testing to compare results across BMRs would be very useful as a follow-up simulation, once the methods are fully vetted, for two purposes: (1) identifying any different behavior of the model averaging method(s) in performance at various BMR values that might indicate a weakness in the methodology to apply to any low-dose BMR, and (2) studying the behavior (coefficient of variation, ratio of BMD to BMDL, etc.) of the model results at various BMRs in order to inform agency understanding about variability in these values.

Michael Messner: Maybe.

Walter Piegorsch: Yes, clearly.

Woodrow Setzer: Your testing needs to attempt to include all reasonable situations the software could be asked to cover. Certainly, you need to consider 1% and 5% BMR values. If BMDs based on control standard deviations are going to be used, you need to test those as well.

Matthew Wheeler: BMRs of 1% should be investigated as well. This will provide evidence of how well the model average is doing across a spectrum of risk levels.

4c) Should additional testing be performed to determine the extent to which the constraints placed on model parameters impacted the test results? If so, what additional testing would you recommend?

Ruth Hummel: Not rigorously at this time. Again, I suggest that this comparison is not critical for the purpose of identifying a model averaging approach but would have some value as a small-scale sensitivity analysis or as a post hoc simulation study.

Michael Messner: Plots of posterior density plots (for model parameters) can reveal when constraints are influential. Scatterplots for paired parameters can sometimes reveal issues with constraints that aren't obvious in density plots.

Walter Piegorsch: I have no strong opinion on this issue, but would not expect that additional testing would be detrimental.

Woodrow Setzer: Probably, yes. However, the results of the current testing should be more thoroughly analyzed before any further testing is carried out. How well do each of the model average component models fit the test curves, and what are their corresponding BMDs? Are they biased high because of parameter constraints? If so, would the bias be decreased if the constraints were relaxed? Slob and Setzer (2014) found that the average value of 'd' in the exponential 5 model was about 1, which means some individual estimates would have to fall less than 1. That is impossible with the standard parameter constraints.

Matthew Wheeler: The constraint on the power model should be tested, and allowed to go below 1.

4d) Should additional testing be performed to determine the extent to which dose scaling impacted the test results? If so, what additional testing would you recommend?

Ruth Hummel: Not rigorously at this time. Again, I suggest conducting a post hoc sensitivity study.

Michael Messner: ?

Walter Piegorsch: I have no strong opinion on this issue, but would not expect that additional testing would be detrimental.

Woodrow Setzer: I would be surprised if dose-scaling was an issue. This is a computational, numerical issue. Models that raise dose to an arbitrary power should probably scale dose internally to the interval 0-1. I would have expected those sorts of problems would have turned up while testing the individual models. So, no, I don't think additional testing related to dose-scaling is warranted.

Matthew Wheeler: I do not believe any additional tests on dose scaling need to be performed.

4e) The experimental designs considered so far have log-spaced doses and one of two patterns of group-specific sample sizes. Should additional experimental designs be considered as part of the process of identifying a model averaging approach for dose-response analysis? In general, can you recommend any additional tests or analyses of the methods that would facilitate selection of a recommended method?

Ruth Hummel: Not rigorously at this time. Again, I recommend a post hoc sensitivity study once the more critical research areas (weighting criteria, use of full raw data, and comparison of performance on different model suites) have been investigated further. I have not scoped the frequency with which my EPA program office receives designs that differ substantially from the designs used in these workshop materials. If it is the case that EPA is regularly reviewing studies with very different designs, then other designs should be considered, at least for a small-scale sensitivity study and perhaps for a post hoc simulation study.

OECD's test guidelines for chronic toxicity studies recommend using at least 20 animals per sex group for each dose level (or a minimum of 4 per sex per group for non-rodents), with at least three dose levels in addition to the control (http://www.oecd-ilibrary.org/environment/test-no-452-chronic-toxicity-studies_9789264071209-en;jsessionid=10j0v3ev92qmk.x-oecd-live-02). EPA's Health Effect Test Guidelines for pesticides and toxics can be found at: http://www2.epa.gov/test-guidelines-pesticides-and-toxic-substances/series-870-health-effects-test-guidelines. Presented as an example are the following details from EPA's subchronic "Repeated Dose 28-Day Oral Toxicity Study in Rodents (July 2000)" guidelines:

> "At least 10 animals (five female and five male) should be used at each dose level. If interim kills are planned, the number should be increased by the number of animals scheduled to be killed before the completion of the study. Consideration should be given to an additional satellite group of 10 animals (five per sex) in the control and in the top dose group for observation of reversibility, persistence, or delayed occurrence of toxic effects, for at least 14 days post treatment." "Generally, at least three test groups and a control group should be used." "Dose levels should be selected taking into account any existing toxicity and (toxico-) kinetic data available for the test compound or related materials. The highest dose level should be chosen with the aim of inducing toxic effects but not death or severe suffering. Thereafter, a descending sequence of dose levels should be selected with a view to demonstrating any dosage related response and NOAEL at the lowest dose level. Two to four fold intervals are frequently optimal for setting the descending dose levels and addition of a fourth test group is often preferable to using very large intervals (e.g. more than a factor of 10) between dosages."

I am curious to see the change in performance of the various methods depending on the sample size of the dose groups. I think this is a less pressing concern than others, but it would be an interesting follow-up simulation, once the methods are fully vetted.

Given that many rodent studies require a certain number of animals per sex, it would also be nice to build in to the modeling the ability to include the sex strata effect (which, if nonsignificant for a particular endpoint, could justify the use of a combined analysis which would give more power and smaller CI).

I recommend:

1. further investigation into candidate weighting schemes
2. further investigation into candidate model suites
3. retaining the full data information (rather than substituting means and variances) when fitting these models
4. adoption of a nonparametric approach to the bootstrapping
5. comparison of best MA approach with simply using a four-parameter Hill or exponential model

Michael Messner: I would like to see flexible experimental designs that adapt, based on data, to aid in both model selection and parameter estimation. Rather than sharply define all the dose levels before-hand, decide on the second dose level after observing results from the first, decide on the third after observing the second, and so on.

Walter Piegorsch: For a greater understanding of how the proposed FMA approach operates, I think it is incumbent upon the research community to study a wider variety of experimental designs here. I admit a lack of expertise on the sorts of designs to consider with continuous data, but can certainly suggest a variety of possibilities for the dichotomous data setting.

As for additional tests of the methods, I am generally satisfied with careful examination of BMDL coverage. One possible extension could be the approach taken in Piegorsch et al. (2013), where we studied both BMDL coverage and also investigated via simulation exactly what values of extra risks were achieved at our FMA BMDLs (these should have been slightly below the target BMR, but for some cases we found substantial departures). We argued that in the end, maintaining control of the target extra risk value was a fundamental component of the risk-analytic decision process here.

Woodrow Setzer: The designs you have used are reasonable caricatures of typical experimental designs for tox. dose-response. Unless a particular design is thought to be particularly problematic, what you have should be adequate. However, you might try a few tests with arithmetically spaced dosing, since that turns up occasionally.

Matthew Wheeler: Additional tests on different benchmark dose definitions should be performed. Given the variety of BMD definitions available for continuous models, testing should be performed on each definition the EPA plans to implement in the final version of the software.

DISCUSSION QUESTION 5: Contingency of Results Upon Including the True Model in the Set of Averaged Models – Section 4.1 (first bullet) notes that best performance of model averaging occurs when the model generating the data is a member of the suite of averaged models. West et al. (2012) also noted this. They also warned that expanding the suite of models (see Section 4.1, first bullet) may increase the risk of selecting an inappropriate model and an incorrect BMDL. (a) Would you recommend increasing the suite of models or changing it in some way? If so, do you recommend testing performance of the new suite?

Ruth Hummel: Yes – at least additional testing of potential model suites and in combination with testing for other weighting schemes. Based on model performance (coverage) shown in these workshop

materials, as well as evidence in West et al. (2012), it seems important to include as diverse a set of model options as possible in the hope of always including the true model. On the other hand, it may be the case that the models currently proposed for inclusion in the model suite are sufficient when the data resampling for the iterations is empirical rather than semi- or fully parametric. I wouldn't discount the possibility of improvement from a simple change to use of the full data. The disappointing coverages for the cases where the true model was not included (Case 2), as well as the weak performance for the cases where the true model is a bounding case of an included model (Case 3), surely need to be improved if at all possible. At minimum, a larger investigation seems warranted. I suspect that some of this performance may also be redeemed, especially in Case 3, by an improvement in the weighting scheme.

Some potential additions to the pool of candidate models are the other exponential models (with the caveats that some research (see last paragraph of Ritz et al. (2013)) shows that there may be an issue with including nested models and research may be needed to determine whether higher-order models that converge to their limiting cases should be excluded from the weighting), splines or isotonic regression (which have the benefit of being strongly data-driven but have some drawbacks (see "Splines as dose-response models" in Slob and Setzer (2014))), and fractional polynomials (as in Ritz et al. (2013)).

Michael Messner: Include models for dichotomous (binary) data.

Walter Piegorsch: Yes and yes. BTW: In West et al. (2012) we warned against including additional models in a model ***selection*** effort. That article did not address model averaging in any depth; that was left to the Piegorsch et al. (2013) article. Based on the results in the EPA report, I would not at this time make any such cautionary warning regarding an expansion of the suite of models.

Woodrow Setzer: I do not agree that this is what your data show. While none of the polynomial models have a close match in the set of models used in model averaging, model averaging only performed badly for templates p1 and p3 in terms of coverage (see **Figure 2**). An alternative to consider is that these templates and the four ex templates challenged the MA methods you used here because of parameter constraints. A quick check of how well the individual models can reproduce the template curves, and how close their corresponding BMDs come to the template BMDs could address that question, and, if it seems to be true, then additional simulations should be undertaken to see if unrestricting constraints, particularly on power parameters, improves the performance of model averaging over this set of restrictions.

(a) It would probably have only a small effect, but try removing the linear and exp3 models from the current suite, and adding more 4-parameter models. Basically, you can generate 4- (and higher) parameter models by taking the hill model as a template, and replacing the term $x^d / (k^d + x^d)$ with any cumulative distribution function with support on the non-negative real numbers.

Matthew Wheeler: (a) The study by West et al. (2012) does not say anything about model averaging. This paper discusses model choice, and, in this context, adding additional models will be deleterious to the overall performance. The same is not true with model averaging. However, when the true model is not in the model suite it is theoretically justified (see Hjort and Claeskens (2003) section 10) that the bootstrap will have problems. I think expanding the model suite above will improve the performance of the model average. These models should not be complex, instead the fractional polynomial approach, or some other suite of linear models where the MLE is easily found, may be preferred. (b) The testing of this new suite should be done as above.

DISCUSSION QUESTION 6: Motives for using model averaging in chemical health assessment.

6a) Please comment on the use of model averaging versus other approaches to account for model uncertainty. It is important to distinguish between two cases, (a) inference within or at the margins of the range of observed responses and doses and (b) inference for responses below the range of observations. See for example West et al. (2012).

Ruth Hummel: (a) I am comfortable following a traditional statistical single-best model method for interpolation within the range or at the margins of observed responses and doses. Model uncertainty will have a much smaller (and I believe trivial, with respect to other larger sources of uncertainty) effect in this region. We can apply Model Averaging in this region, but I see less need for this more sophisticated method where data exist to inform a good single model choice.

(b) For low-dose extrapolation, on the other hand, there is clear and sometimes very large variability in BMD(L) estimates due to the choice of model. If we could provide good evidence that a single model (such as the Hill or exponential) performs well for low-dose extrapolation (in coverage and size-of-error compared to known true values), then I would see no need to capture additional model uncertainty through MA. However, in the absence of compelling evidence for a single model family, the MA concept is the most promising method for capturing model variability that is available and has traction in the literature.

Michael Messner: I think Bayesian model averaging is superior for (a) and everything is risky for (b), which could perhaps be avoided by having flexible experimental designs that allow the researcher to choose better dose levels to ensure the dose range is sufficiently wide.

Walter Piegorsch: The results of West et al. and of Ringblom et al. (2014) clearly show that unadjusted model selection is essentially an unwise strategy when calculating BMDs and BMDLs. Instead, FMA (or carefully performed BMA) methodology appears to be the best option for addressing model uncertainty at this time. (One could try some form of adjusted model selection and account for the selection step statistically, but that would be a far more complex operation than simply applying a properly constructed FMA calculation.)

I think this issue is now substantial enough that distinguishing between cases near the observed doses with those away from the dose range is a lesser concern: until we find a better way to address model uncertainty, we should look to make model averaging the default choice for BMDL calculations.

Woodrow Setzer: In the light of Slob and Setzer (2014), and your own simulation results, in which the exp5 and hill models arguably outperformed all the model averaging methods, you need to make the case that model averaging methods are needed, that there is a problem that MA solves. That is, relative to inference within or just beyond the margin of observed doses and responses. For inference much beyond the range of doses, you need to use methods that bring in more biology. It is unlikely that model averaging approaches applied to conventional dose-response data will fully capture the uncertainty of inferences made well below the dose-response range.

Matthew Wheeler: Model averaging is a reasonable approach to account for model uncertainty. Within the margin of the observed data these approaches are very reasonable to use, and, typically perform better than current practice. Though my experience is primarily with dichotomous models, this observation should transfer to continuous data when there are enough models included in the model average that have sufficient flexibility. When dealing with extrapolating below the observed data, I can only speak in the case of dichotomous data. Here Wheeler and Bailer (2013) showed that the results were effectively no different for quantal linear data, and significantly closer to the true risk when a sub-linear model was concerned. I have personally seen cases where the use of model averaging would increase the hazard of the compound. For example, investigate the IRIS analysis of Dichloroacetic Acid, and compare it to a model average or semiparametric approach, both methods produce lower BMDLs

(by an order of magnitude) at the specified risk level than the best model+ POD + linearization approach. I have also seen cases where it decreased the hazard of the compound, and talking to toxicologists, this was probably reasonable. For example, see the IRIS analysis on Analine, where there is an order of magnitude shift in the other direction.

I do think further research is needed to extrapolate this argument to continuous data, but I see no reason this argument should not hold. I again caution the interpretation of West et al. (2012), with model averaging as it deals with picking the best model. In this situation, extrapolation below the point of departure is much different. See the argument in Wheeler et al. (2015) for more information.

6b) Another motivation for using model averaging is that it is a way to apply weights based on prior information or beliefs (e.g., about mechanisms) and historical information (e.g., about model families that fit data well). What is your opinion on this use of model averaging versus alternative approaches for using prior information and data?

Ruth Hummel: I would like the ability to use prior/historical information in the modeling. This could be very valuable for demonstrating the effect of expert knowledge in determination of the POD, which could help us quantify previously unquantifiable information and could provide helpful information for decision-making.

Michael Messner: Use of informed priors will require strong justification. Sensitivity analysis can show when priors are having large influence.

Walter Piegorsch: Obviously, where prior information exists it should be incorporated into any statistical calculations. As presented, the FMA approach dose this in a simple fashion via the prior terms in the Akaike weights; however, more-complex BMA operations could also be applied (Fang et al. 2015) if done carefully.

Woodrow Setzer: Any use of such prior weighting needs to be developed very carefully and transparently. It makes sense, but it also seems like it will be difficult to justify a set of such weights. Our dose-response models really have little biological content, and there is rarely any real reason to prefer one over another on mechanistic grounds. There is a better chance that historical information will lead to a viable set of such weights, but it will require a lot of work, and datasets with an unusually large number of dose groups, so that dose-group-level variation can be separated from variation in dose-response shape.

Matthew Wheeler: Prior information, when used correctly, can make the analysis more robust. I would recommend only using prior information in terms of historical controls. Here methods exist to add this information. In terms of prior-weighting of the dose response curves, to my knowledge, there have been no methodologies that have been developed to accurately include prior information on the form of the dose-response curve. I would argue that any attempt to weight a given model would add possibly undue bias to the analysis, if this weighting scheme was not well vetted.

DISCUSSION QUESTION 7: Should alternatives or complements to model averaging be investigated? Piegorsch (2014) and West et al (2012) suggested that further research is needed before the performance of model averaging and other approaches are understood well enough to be applied in risk assessment. Alternative approaches include isotonic regression, non-parametric and semi-parametric (Bayesian and frequentist) modeling, fully Bayesian model averaging, and use of flexible parametric models (Piegorsch 2014; Ritz et al. 2013; Slob and Setzer 2014).

Michael Messner: Alternatives.

Woodrow Setzer: It is better to say there are two options available for dealing with model uncertainty: flexible models, whether it be 4-parameter sigmoid models like the hill or exp5, or various semiparametric approaches, such as splines or Gaussian processes.; and model averaging.

7a) Should EPA be concerned that other approaches may provide better goodness of fit or coverage closer to that intended, at least under some conditions (e.g., for data sets with special characteristics, such as more than 5 doses, or no doses in the response (BMR) range of interest)? If so, how do you recommend EPA explore these alternatives?

Ruth Hummel: To a limited extent. I recommend that MA (or an alternate approach such as consistent application of the four-parameter Exponential or Hill model) be advanced as a unifying approach for general dose-response modeling and determination of a POD as soon as some minimal additional research (into alternative weighting schemes, testing of a few larger suites of models, and comparison with a single four-parameter exponential or Hill model, with comparisons on coverage rates and size of the error (as a ratio) of the BMDL estimate versus the known true value) is completed, and with ongoing research into these other potential methods and special cases.

As described in Slob and Setzer (2014), GoF tests should probably be used with caution, given the influence of non-random sources of variation (litter effects, effects from non-random application of the study protocol, etc.) which are generally not accounted for in dose-response modeling but can certainly affect the fit of a statistical model. I am most concerned with seeing the performance of coverage and minimum error in the estimates rather than GoF.

Walter Piegorsch: Absolutely: EPA must study other approaches for addressing the model uncertainty issue. Further comparisons are needed with isotonic/nonparametric methods for finding BMDLs (Piegorsch et al., 2012; Guha et al., 2013; Piegorsch et al., 2014; Lin et al., 2015) and semiparametric techniques (Wheeler and Bailer, 2012, Ref. 15). Bagging for the BMDL might also be included (Bornkamp, 2015). More development is needed in all these areas.

Woodrow Setzer: Well, yes. The two un-averaged, flexible models considered here outperformed model averaging in terms of coverage, over the conditions in this study. Maybe that superiority would be reversed over a broader set of conditions, but this is what we have to go on right now. EPA should try to better understand why model averaging generally performed so poorly in this analysis, identify particular performance goals for its modeling functions, and continue to explore both flexible models and model averaging. In the process, they need to take into account the degree of model uncertainty that likely really exists, considering Slob and Setzer (2014), and the degree to which dose-group-level variability is a factor in decisions about model uncertainty.

Matthew Wheeler: Model Averaging is one possible approach that can be used to account for uncertainty in the shape of the dose response function. Though it has been the most studied methodology for quantitative risk assessment in the past 10 years, it has actually been studied less than nonparametric and semiparametric approaches. For example, isotonic regression applied to benchmark dose estimation is recent in the literature Lin et al. (2014) and Piegorsch et al. (2014). However, isotonic regression has a history in the statistics literature dating back 50 years. Similarly, penalized spline based methodologies have been studied for 40 years, and much more is known about their performance. It is my belief that these approaches offer far more capabilities for continuous data than model averaged approaches.

For example, consider a monotonic Bayesian spline solution using monotone M-Splines Ramsay (1988) that are penalized using a prior similar to the auto-regressive prior used by Lang and Brezger (2004) that is applied to the simulated data templates provided. Specifically, we look at the situations where model averaging failed to achieve nominal coverage. In all cases, the BMD estimate was closer to the actual BMD, and in all but one case, the coverage was at or above nominal levels. This result is

striking given some of the model averaged estimates provided poor coverage. **Figures (1) - (3)** show three particular examples of this method applied to the simulated data sets. Each one is far superior to all of the proposed model average methods, and the computation time is similar (typically between 6 to 10 seconds per fit). I recommend the EPA investigate such methods as a possible alternative to model average.

7b) Do you wish to comment on specific situations, defined in terms of modeling options, endpoints, etc., where model averaging could be particularly valuable and might be implemented initially?

Ruth Hummel: In my work in an EPA program office (OSCPP/OPPT) we frequently analyze dose-response data for selection of a POD for risk management decision-making. For all these data sets (typical NTP studies, summarized data from published research, test data submitted according to test rule specifications and as confidential business information through our New Chemicals PMN Program, ecotoxicity and aquatic toxicity studies, etc.), we need the best methods CURRENTLY (or very soon) available that are performing with coverage at the level claimed and with estimates that are as accurate as possible. For my office, the priorities would be in this order: (1) coverage as claimed, (2) as soon as possible, and (3) minimizing the size of the confidence interval (by, for example, following the advice of Slob and Setzer (2014) and including data from other endpoints and studies on a similar endpoint or including historical information from other studies on, say, analogous chemicals) and tightening the distribution of error of the estimated BMDL compared to a true BMD (when know from simulation study).

Walter Piegorsch: Specific cases could certainly be delineated, but my concern is that model uncertainty exists at far greater levels than is explicitly acknowledged in modern BMD estimation. And, we are only now understanding how badly the established, single-model, parametric estimators perform in the presence of such uncertainty. At present, I feel we should be applying FMA (or if done carefully, BMA), or another model-robust method such as isotonic regression, to all BMDL calculations whenever any possibility exists of uncertainty in the dose-response model specification.

Woodrow Setzer: I really cannot say.

Matthew Wheeler: Initially, I believe Model averaging should be implemented using dichotomous data. In this case, there have been numerous simulation studies showing that the results it provides are superior to traditional approaches. Further, there is evidence that the estimates are similar to those provided by different methods to account for model uncertainty (Wheeler and Bailer, 2013).

7c) Do you think that the model-averaging approach is preferable to using the Hill or Exponential model as suggested by Slob and Setzer (2014)[1]. If so, please explain.

Ruth Hummel: Not necessarily. Slob and Setzer (2014) has me intrigued about the possible application of a single model for all generic dose-response analysis. I would love to see a direct comparison of a best version of Model Averaging (after working out a few more of the potential sources of poor performance for the true-model-not-included cases) with a simple single-model (exponential or Hill) approach over a range of simulated data (and using the raw data). Slob and Setzer (2014) present an approach that is very appealing in its simplicity and they present results that seem to fit the real datasets rather well; however, coverage rates of this method are unknown and are of great interest when selecting a method to develop PODs that achieve in practice what they by definition promise. I suspect that the MA will outperform this simple method, but it seems a valuable comparison in light of the advantages laid out in Slob and Setzer (2014).

Walter Piegorsch: Yes. See comment 7(b), above.

Woodrow Setzer: At the moment, no. Slob and Setzer argue that toxicological dose-responses are quite homogeneous in shape, and that the exp5 and hill models are quite good at capturing the shapes of those datasets. Some model uncertainty may well remain, but it needs to be captured with a much more restricted set of models. The results of this study indicate that the hill and exp5 models behave relatively much better than the model-averaging approaches explored here (see **Figure 1**) in terms of coverage. Further analysis needs to be done to see if this holds up in terms of bias and overall length of confidence intervals.

Matthew Wheeler: Though I agree that one may be able to find a single flexible model that will accurately describe the data and estimate a BMD. I do not agree that the model will be able to provide a BMDL at the nominal level. This is evident by the simulations, where the Hill and the Exponential models frequently failed to provide a BMDL at the specified rate. The question should be is anything lost by using an appropriate model averaging technique, and my answer to that is no.

DISCUSSION QUESTION 8: Describe any major concerns for the application of methods described in this report to dichotomous data. How do the results of the present background paper on models for continuous data compare to published work on model averaging for dichotomous models?

Ruth Hummel: I do not have an opinion on this at this time. I defer to the other discussants who have worked more extensively on this.

Michael Messner: For microbial dose-response, the models should honor single-hit theory. Probability of infection shouldn't exceed probability of exposure!

Walter Piegorsch: In general, I believe many of the larger conclusions presented in the EPA report would also apply to dichotomous data after further/pertinent investigation. My own work in this area has led to the realization in comment 7(b), above: model uncertainty exists at far greater levels than is explicitly acknowledged in modern BMD estimation and we are only now understanding how badly the established, single-model, parametric estimators perform in the presence of such uncertainty. At present, I feel we should be applying some form of model-robust estimation to all BMDL calculations whenever any possibility exists of uncertainty in the dose-response model specification.

Perhaps not surprisingly, I favor the FMA approach developed in Piegorsch et al. (2013), which we found to possess (i) a defensible theoretical/asymptotic justification; (ii) fewer computational requirements for practical implementation (no bootstraps, no Monte Carlo approximations); and (iii) very stable performance – if slightly conservative – across a variety of dichotomous dose-response patterns. An alternative would be the model-robust, non-parametric method we derived in Piegorsch et al. (2014). If substantive, informative prior information is available, one could also apply the BMA approach we developed in Fang et al. (2015).

Woodrow Setzer: You should be cautious extrapolating the current work to the domain of dichotomous models. Some of the lessons may carry over, like not bounding some of the parameters and making sure the right models are in the model averaging mix. Dichotomous models have a richer literature so far than do continuous models, and it is likely that you will be able to make an easier decision about using model averaging results about dichotomous data with a combination of literature review and judicious simulations to test any developed software. While continuous data seem to follow very similar dose-response shapes, this does not seem to be as true for dichotomous data, so some way to address model uncertainty seems necessary at this point.

Matthew Wheeler: As I recommend above, I believe that model averaging is currently most appropriate in the dichotomous data setting. Every study suggests that picking a single model is fraught with problems, and that model averaging is superior. Again, I will note that the West et al. (2012) study does not say anything on model averaging, but speaks to the practice of picking one model for risk

assessment. It is my experience for dichotomous data that even when model averaging fails (i.e., models at the edge of the space), it performs better than current practice.

DISCUSSION QUESTION 9: Is model averaging as implemented in the workshop support material suitable for use in chemical health assessments, possibly with some reservations or precautions? Can you identify circumstances when model averaging may be helpful and informative? Misleading? Please elaborate.

Ruth Hummel: The MA methods presented here take the single-model information (which is currently used for EPA's chemical risk assessments) and improve upon it by providing some distributional grounding for the range of reasonable modeled results. I highly recommend that EPA risk assessors begin including these model averaging (and/or results-averaging, depending on the conclusions from this workshop) methods in dose-response modeling supporting chemical health and risk assessments, with concurrent development of appropriate guidance materials and precautions. I continue to be concerned about the impact of the distribution assumptions for the bootstrapping (normal or lognormal from the mean and standard deviation, rather than empirically based on the raw continuous data) on the model results.

Before a MA method is finalized and sanctioned for use (but as quickly as possible, in order to get any better method out for use right away) I would like to see some additional development of candidate weighting schemes and candidate model suites, plus use of the raw data, and comparisons of the coverage of these developments with what is currently presented and with the use of a single four-parameter exponential or Hill model.

Michael Messner: Yes – I think model averaging is ready, but with close control and checking. I don't think it is ready for high-throughput, turn the crank processing.

Walter Piegorsch: Yes, I think FMA and even careful use of BMA is "ready for prime-time" in BMD/BMDL calculation with Chemical Health (and other) Risk Assessments. As noted above, model uncertainty exists at far greater levels than is explicitly acknowledged in modern BMD estimation and we are only now understanding how badly the established, single-model, parametric estimators perform in the presence of such uncertainty. At present, I feel we should be applying some form of model-robust estimation to all BMDL calculations whenever any possibility exists of uncertainty in the dose-response model specification.

I think further study is necessary to determine if some of the non-bootstrap FMA methods that have been proposed could be viable – and more practicable – alternatives to the bootstrap; this is not an issue, however, that should delay promulgation of the general MA strategy.

Woodrow Setzer: I'd say "No" for continuous endpoints at this point. The high frequency of your test conditions for which all the model averaging results yielded quite low coverages (I'd suspect because the BMD estimates were biased high) will be a major problem for use.

Matthew Wheeler: Model Averaging is better than current practice, and it is often markedly better. Consequently, I would argue it is ready for chemical risk assessment. Further, NIOSH has already used it as a basis of a risk assessment, and will use it in future risk assessments (Wheeler et al., 2015). With this I urge caution not to assuming that model averaging is the best possible approach to the model uncertainty problem. I believe there are many approaches that will be shown to be superior to model averaging. Consequently, I do believe that model averaging should be considered along with other methods that account for uncertainty, and it should not be seen as the single best approach.

DISCUSSION QUESTION 10: Do you agree with the conclusions made in Section 4.1 of the workshop support material? Please elaborate on points that you question.

Ruth Hummel: Yes, I agree with the conclusions in Section 4.1, with a small minor concern that the last paragraph of the second bullet point is overlooking the possibility of removing the failed model from the pool (as in Piegorsch 2014 section 4.1, referencing Wheeler and Bailer (2009)) rather than reducing it to the limiting model (as I discussed at the end of my response to question 3c).

Michael Messner: Yes. I have no issues with the conclusions in 4.1.

Walter Piegorsch:
Points with which I agree:

- inclusion of as many reasonable dose-response relationships as possible into the uncertainty class used for the model averaging;
- (following up on the previous point) even include models that can be written as nested/sub-models within a larger class (e.g., Michaelis-Menten and Hill); apply "Occam's Razor" to expunge models when likelihood or other information indicates lower importance/impact for a larger model (see Wheeler and Bailer, 2009);
- extend the current study to other forms of BMR, such as relative to the standard deviation, or "hybrid" definitions;
- "...one probably ought to fit non-constant variance models as a matter of course for model averaging" with continuous data;
- Consider experimental designs with more than 4 (or even more than 5) dose groups – possibly reduce per-group sample sizes where necessary;
- always operate with caution when extrapolating far below (or above) tested doses and their responses.

Issues with which I am uncomfortable:

- I'm afraid I cannot support use of "Method 1" in any of its forms (Method 2a, Method 3a). To my knowledge, no probability statement exists that validates construction of a confidence limit by averaging a series of other confidence limits. Thus the Method 1/Method 2a/Method 3a quantity cannot be described as a true BMDL. (Some might argue that via simulation study the method(s) appear to operate acceptably, at least in some cases. Wheeler and Bailer's (2009) simulations showed highly varied coverage patterns, however, many with badly suboptimal under-coverage. And more generally, simulations can only be used as validations, not proof, of a proposed confidence proceedure.) I could imagine an average of BMDLs being used as an initial estimator in some sort of iterative or hierarchical estimation schema – or in some similar, informal fashion – but not for use as a final BMDL.

Woodrow Setzer:

Bullet 1: I don't think you have adequately demonstrated the point made in the first paragraph of this bullet, and I have discussed my reservations in an earlier answer. It may well be that problems are due to using models with constrained parameters. The characterization of Slob and Setzer (2014) is a bit misleading. The point of that paper is that you can adequately fit continuous dose-response data with a four-parameter hill or exp5 model. The fact that a nested approach was taken to the fitting is irrelevant, and has more to do with difficulties in estimating all the parameters for these four parameter models in some datasets than with the need for more shapes. Your own simulations suggest that coverage of these two models are at least better than that of the model-averaging approaches you used. Finally, I remain unconvinced that mechanistic biological considerations can have much of an impact on model choice, though empirical observations about the range of models that are required to fit the universe of dose-response datasets should be useful.
Bullet 2: You need to distinguish between the range of models that is needed to characterize the variety of real dose-response shapes and the difficulty of fitting some of those models to inadequate datasets.

Unfortunately, we are stuck using data from study designs that date back to the era of NOAELs (or even NOELs) as points of departure, when it was unimportant to characterize the dose-response more quantitatively (apparently). The proper approach is to first, figure out what the variability is, and next, figure out how to fit those models. For instance, using Bayesian methods with informative priors on the model parameters, based on an observations that the power parameter for the exp5 model tends to be tightly clustered around 1, and the upper or lower bound on the dose-response tends to be similar across studies for the same endpoint, allows all the parameters to be identifiable in a Bayesian analysis. I do not see that adding restricted forms of the full models will help you. If you are getting poor coverage with the full model, it is unlikely that restricting the parameters of that model further will make things better. Instead, try relaxing the constraints on the model parameters.

Bullet 3: This result is expected from Shao et al (2013). The lognormal distributions used a relatively small CV (around 14%), for which the difference between normal and lognormal is relatively small. Also, maybe the relatively small log-scale dynamic range of the dose-response models in the test set minimized the degree to which variance changed with mean. You may well find datasets where the misspecification of the error model has a bigger effect.

Bullet 5: If your variance model includes constant variance, do not include that as a separate model. I agree that it would probably be OK to just use the modelled variance versions of the models if constant variance is a possibility. Perhaps alternatively, switch to using lognormal errors as the default, since there is evidence that CVs tend to be constant across dose groups.

Bullet 6: More simulations would help sort out the difference between design and sample size. Try k keeping sample size constant and varying the design, for instance.

Bullet 7: This really looks like a problem caused by constraining power parameters to be not less than 1.

Matthew Wheeler: I agree with all of the comments, except there is no difference between the model average bootstrap methods. Method 4b performs similarly when they all work, but is far superior (even though not at the advertised rate) to the other methods when it fails.

APPENDIX

Woodrow Setzer's Figures

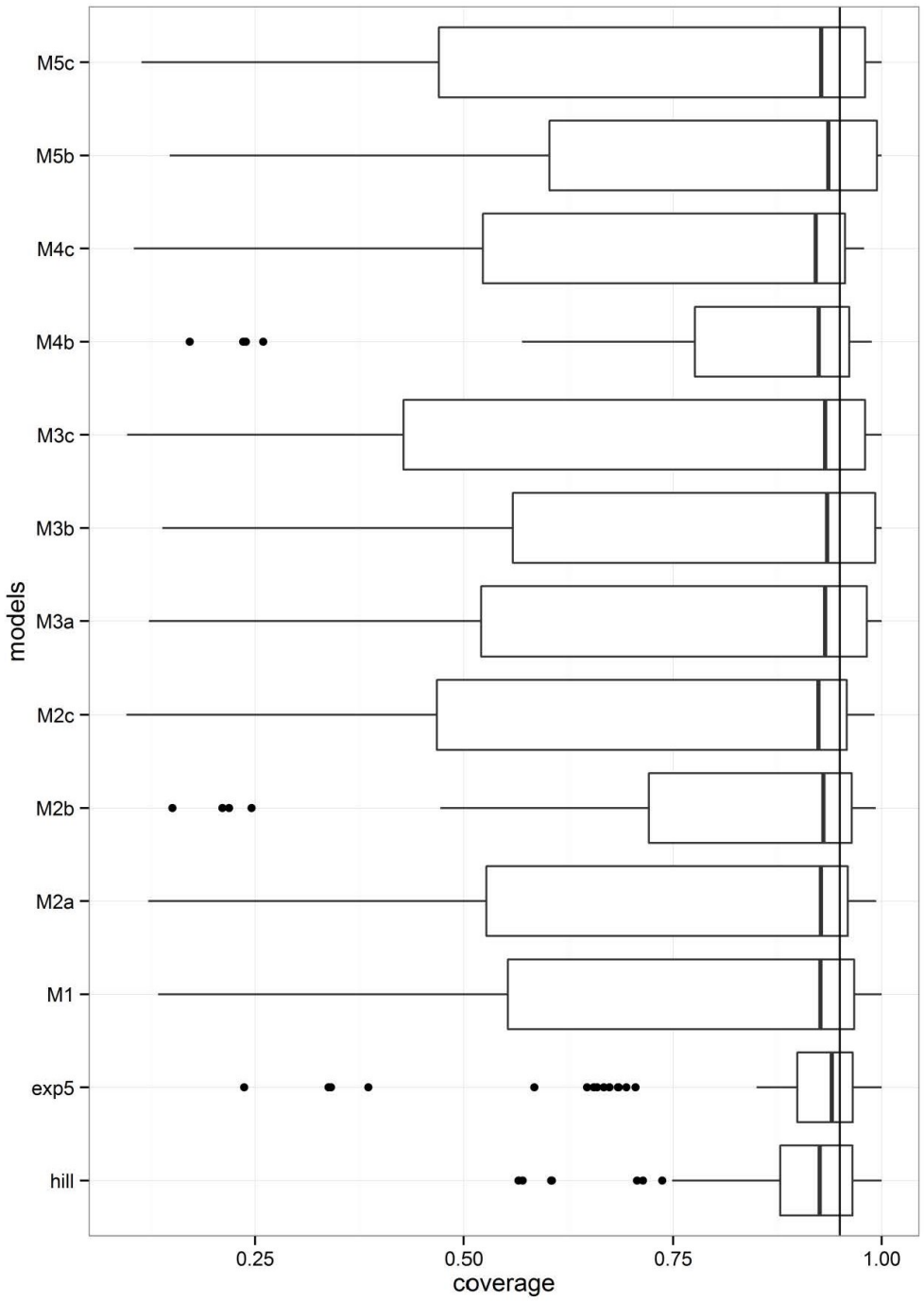


Figure 1. Distribution of coverage estimates over the model averaging methods and two simple models, over all templates, error models, both designs, and both approaches to modeling the variance.

Figure 2. All coverage estimates, stratified by template, etc. and color-coded by method.

Matthew Wheeler's Figures

Figure 1: BMD results for data template E1 assuming Normal errors and a sub chronic study design.

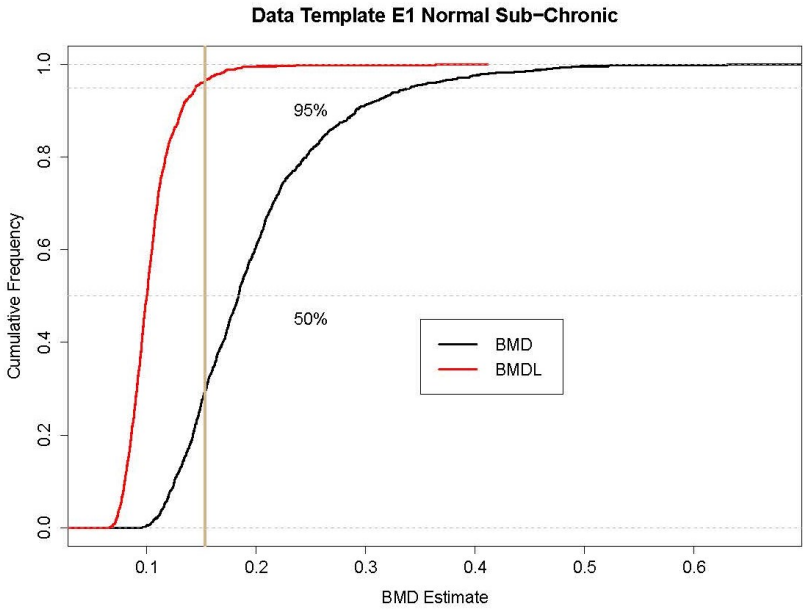


Figure 2: BMD results for data template E1 assuming Normal errors and a chronic study design.

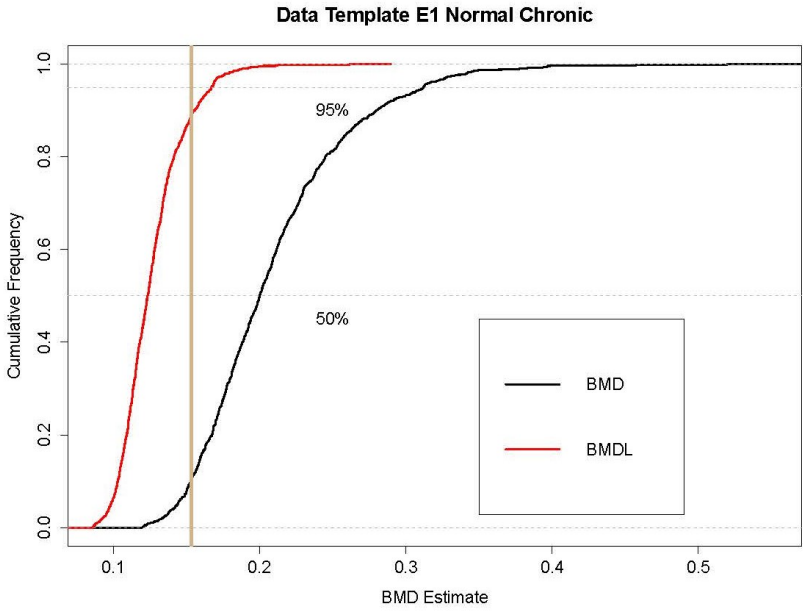
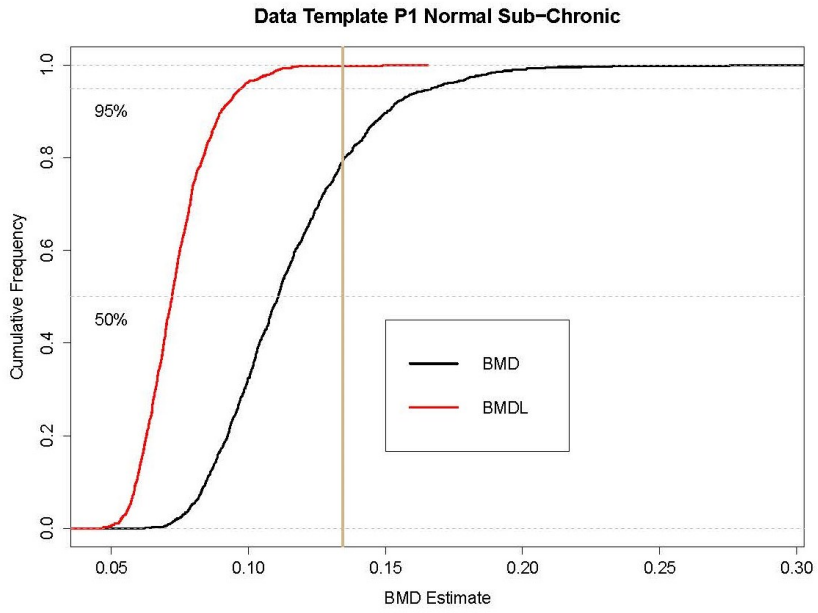


Figure 3: BMD results for data template P1 assuming Normal errors and a sub chronic study design



REFERENCES CITED

Piegorsch, WW (2014) Model uncertainty in environmental dose-response risk analysis. Statistics and Public Policy 1:78-85 (http://dx.doi.org/10.1080/2330443X.2014.937021)

West RW, Piegorsch WW, Peña EA, An L, Wu W, Wickens AA, Xiong H, Chen W. (2012) The Impact of Model Uncertainty on Benchmark Dose Estimation. Environmetrics 23(8):706-716. (http://onlinelibrary.wiley.com/doi/10.1002/env.2180/epdf)

Ritz,C, Gerhard,D & Hothorn, LA (2013) A Unified Framework for Benchmark Dose Estimation Applied to Mixed Models and Model Averaging. Statistics in Biopharmaceutical Research 5:79-90 (http://www.tandfonline.com/doi/abs/10.1080/19466315.2012.757559)

Slob W, Setzer RW (2014) Shape and steepness of toxicological dose-response relationships of continuous endpoints. Critical Reviews in Toxicology 44(3):270-97. (doi: 10.3109/10408444.2013.853726)