EPA/600/R-16/001

MEETING NOTES
Webinar Workshop on Model Averaging Methods for Dose-Response Analysis

Held on December 10-11, 2015

https://www.epa.gov/bmds/model-averaging-webinar-workshop-announcement

Final – March 2016

National Center for Environmental Assessment

US Environmental Protection Agency

Research Triangle Park, North Carolina

DISCLAIMER

The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency. It has been subjected to review by the National Center for Environmental Assessment and approved for publication. Approval does not signify that the contents reflect the views of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use. This document reflects the proceedings of a workshop, including presentations made by invited speakers, the discussions consequent to those presentations, and summaries of the individual Sessions. Any statements included in this document which were made by the presenters or by participants in the discussions or in the session summary discussions are those of the individuals and does not signify that the statements reflect the views of the US Environmental Protection Agency.

Contents

Acknowledgements

This project was only made possible by the support and assistance of a number of individuals and groups. Many of the groups and individuals involved in those groups are listed below.

EPA Core Planning Team

Jeff Gift, PhD (Team and Project Lead, NCEA-RTP)

Allen Davis, MS (Team and Project Co-Lead, NCEA-RTP)

EPA Internal Reviewers of Workshop Support Materials

Todd Blessinger, PhD (NCEA-DC)

Christine Cai, PhD (NCEA-IRIS)

John Fox, PhD (NCEA-DC)

David Farrar, PhD (NCEA-Cin)

Leonid Kopylev, PhD (NCEA-DC)

Paul Schlosser, PhD (NCEA-DC)

Maria Spassova, PhD (NCEA-DC)

Workshop Discussants

David Dunson, PhD	Duke University
Ruth Hummel, PhD	U.S. EPA, Office of Pollution Prevention and Toxics (OPPT)
Michael Messner, PhD	U.S. EPA, Office of Ground Water and Drinking Water (OGWDW)
Walter Piegorsch, PhD	University of Arizona
Woodrow Setzer, PhD	U.S. EPA, National Center for Computational Toxicology (NCCT)
Matthew Wheeler, PhD	National Institute for Occupational Safety and Health (NIOSH)

Contractor Support – The following individuals provided support for workshop planning and conduct.

Susan Blaine, BS	ICF International
Audrey Turley, MS	ICF International
William Mendez, PhD	ICF International
Pam Ross, MS	ICF International

The following individuals provided support for development of workshop support materials.

Bruce Allen, MS	Bruce Allen, Inc.
Michael Brown, MS	Lockheed Martin
Louis Olszyk, BSE	Lockheed Martin
Kan Shao, PhD	Indiana University
Cody Simmons, PhD	Lockheed Martin
Mike Uhl, MS	Lockheed Martin

Background

On December 10-11, 2015, U.S. Environmental Protection Agency's (USEPA) National Center for Environmental Assessment (NCEA) sponsored two half-day webinar workshop sessions,[1] to provide a forum for peer discussion of model averaging methods for benchmark dose (BMD) analyses. The benchmark dose (BMD) method is a key component of chemical risk assessments performed by EPA, other U.S Federal and State Agencies, and multiple international organizations (USEPA, 2012; RIVM, 2009; EFSA, 2009; JECFA, 2006). The BMD method is used to establish guideline values such as USEPA reference doses and cancer slope factors (USEPA, 2012) for chemicals expected to have ecological or human health effects.

Existing BMD methods involve the selection of a single dose-response model from among a suite of models with little biological basis, based largely on a comparison of model fits to the data. The National Research Council and EPA's Science Advisory Board, in various reports (NRC, 2009; 2011; 2013; 2014), has urged EPA (in particular, EPA's IRIS Program) to develop and apply methods that account for model uncertainty, allow for the incorporation of prior knowledge regarding a chemical's mode of action, and offer alternatives to the current approach of selecting a single 'best' model based on goodness of fit and the Akaike Information Criterion (AIC). Model averaging, one approach that has been developed to address these concerns, has been the primary focus of recent EPA research because it has been extensively vetted in the literature for these purposes. It offers potential advantages over existing approaches that rely on selection of a single model, including the ability to characterize model uncertainty and incorporate prior knowledge (e.g., biological and historical information) regarding models and parameters under consideration.

To facilitate the December 10-11, 2015 model averaging webinar workshop, EPA NCEA prepared workshop support materials,[1] including a description of methods that are consistent with published model averaging recommendations, prototype software that illustrates how those existing methods could be implemented, test results from the implementation of the prototype software and questions that EPA would like the panel to address. The software package developed by EPA for the purposes of this workshop facilitates the analysis of continuous data, i.e., dose-response data that have responses measured (and reported) on a continuous scale (e.g., body weight or serum enzyme levels). Workshop discussants were asked to assess these materials and provide input that will help EPA determine the value and optimal model averaging methods for dose-response analysis of continuous response measures, as well as the potential for extension of preferred methods to dichotomous response measures.

This document contains meeting notes documenting the proceedings of the workshop, as well as Appendix material representing written comments and presentation slides submitted by the workshop discussants and public participants. The Workshop participants were provided an opportunity to review the summary and statements attributed to them.

[1] The workshop website containing details such as the workshop agenda and workshop support materials can be viewed at https://www.epa.gov/bmds/model-averaging-webinar-workshop-announcement.

Background References

EFSA (European Food Safety Authority), 2009. Use of benchmark dose approach in risk assessment. The EFSA Journal. 1150, 1-72.

JECFA (Joint FAO/WHO Expert committee on food Additives – JECFA. (2006). Sixty-fourth meeting, WHO/IPCS Safety evaluation of certain contaminants in food. WHO Food Additives Series 55.

NRC (National Research Council). (2009). Science and decisions: Advancing risk assessment. Washington, DC: National Academies Press. http://www.nap.edu/catalog/12209.html.

NRC (National Research Council). (2011). Review of the Environmental Protection Agency's draft IRIS assessment of formaldehyde. Washington, DC: National Academies Press. http://www.nap.edu/catalog/13142.html.

NRC (National Research Council). (2013). Critical aspects of EPA's IRIS assessment of inorganic arsenic: Interim report. Washington, D.C: The National Academies Press.

NRC (National Research Council). (2014). Review of EPA's Integrated Risk Information System (IRIS) process. Washington, DC: The National Academies Press. http://www.nap.edu/catalog.php?record_id=18764.

RIVM (National Institute for Public Health and the Environment). (2009). PROAST: Software for dose-response modeling and benchmark dose analysis. http://www.rivm.nl/en/foodnutritionandwater/foodsafety/proast.jsp

U.S. EPA (U.S. Environmental Protection Agency). (2012). Benchmark dose technical guidance. (EPA/100/R-12/001). Washington, DC: Risk Assessment Forum. http://www.epa.gov/raf/publications/pdfs/benchmark_dose_guidance.pdf.

U.S. Environmental Protection Agency
Workshop on Model Averaging Methods for Dose-Response Analysis
December 10–11, 2015

Meeting Notes

Day 1: Thursday, December 10, 2015

Welcoming Remarks

Jeff Gift | *U.S. EPA, National Center for Environmental Assessment (NCEA)*

- Background
  - National Research Council (NRC) has provided recommendations to EPA for conducting dose-response assessments to better understand model uncertainty and to possibly incorporate prior information such as mode-of-action (MOA) into dose-response evaluations.
  - Model averaging is a well-published topic, and the workshop is an extension of current EPA publishing.
  - Methods for model averaging of dichotomous endpoints was published several years ago by Wheeler and Bailer (2007).
  - EPA wanted to focus on continuous models using those available from U.S. EPA's Benchmark Dose Software (BMDS) and build on existing models.
  - EPA tested five distinct endpoints, and each approach has various options within them.
  - As outlined in the workshop materials, a report was produced and software was prepared to test the five approaches and reviewed by the panel discussants.
- Acknowledgements
  - Thanks to the very important contributions of:
    - Kan Shao (Assistant Professor at Indiana University) – implementation and programming;
    - Bruce Allen (independent consultant) – implementation;
    - Louise Olzyk and Cody Simmons (Lockheed Martin) – testing and programming; and
    - EPA Statistical Working Group (SWG) including Leonid Kopylev, John Fox, David Farrar – critical review and comments.
- Key questions
  - Does the workshop adequately address NRC concerns?
  - Are maximum likelihood estimates (MLE) and bootstrap methods sufficient? Should Bayesian modeling approach be investigated?
  - How close are we to using continuous endpoints in model averaging?
  - What are lessons learned for investigations to improve existing dichotomous endpoints?
- EPA is not promoting modeling averaging as the only approach to address NRC concerns; rather, the workshop outcomes will be used as a guide for future assessments.

Day 1: Thursday, December 10, 2015

Public Presentation

John French | *National Institute for Environmental Health Sciences (NIEHS)*

See Appendix B for presentation slides

- Presentation outlines considerations of the shape of the dose response curves based on genetically diverse population models (results presented at recent NRC conference on interindividual variability).
- Models that are population drivers are genetically diverse but in the past, we primarily have used homozygous mouse models as surrogates for humans.
- Proof of concept: using a population-based approach can help to refine the shape of dose-response curves, aid in hazard identification (eliminate potential false negatives, false positives), and prepare relevant data for benchmark dose modeling to derive points of departure (POD).
- Project description:
    - Published in French et al. (2015)[2].
    - Tested benzene exposure (chemical with sufficient information on toxicity) on a diversity outbred mouse population; used experimental protocols previously used on homozygous inbred mice.
        - Outcomes tested were % reticulocytes and micronucleated reticulocytes (MN RET) in peripheral blood and bone marrow.
    - Results
        - Panel A – Pre-exposure peripheral blood measurements (MN RET).
            - No significant differences among group means, but background levels are variable.
        - Panel B – Post-exposure peripheral blood measurements (MN RET).
            - Significantly increased response at 100 ppm exposure compared to control.
        - Panel C – Post-exposure bone marrow measurements (MN RET).
            - Dose-response shift to the right; significantly increased response at 1, 10, and 100 ppm compared to control.
        - Data fit to continuous models in EPA's benchmark dose software (BMDS).
            - Exp4 as best fit of the data of all data with benchmark responses (BMRs) of 1 standard deviation and 10% relative deviation; data at 100 ppm exposure skewed the data.
            - Removed highest dose, and fit was improved.
        - Compared to results in Farris et al. (1996)[3] (B6C3F1 inbred strain), the genetically diverse outbred model gave a robust response and provided an 18-fold difference from the inbred strain.
- Conclusions
    - Using a diversity outbred population model would be a better model of the human population to increase response to toxicity; can also include mouse strains with additional single nucleotide polymorphisms (SNPs) that could better represent a genetically-diverse population.

[2] French, J. E., et al. (2015). "Diversity Outbred Mice Identify Population-Based Exposure Thresholds and Genetic Factors that Influence Benzene-Induced Genotoxicity." Environ Health Perspect 123(3): 237-245.

[3] Farris, G. M., et al. (1996). "Benzene-induced micronuclei in erythrocytes: an inhalation concentration-response study in B6C3F1 mice." Mutagenesis 11(5): 455-462.

Day 1: Thursday, December 10, 2015

Discussion Question 1: Overall approach to model-averaging

Are there other model averaging methods that EPA should consider?

Matt Wheeler | *NIOSH*

- Disappointed with the results of coverage when models were on the edge of the model space; there are other possible methods to consider.
- Use of the Bayesian Information Criterion (BIC) as the weighting criteria for the models should be examined more closely.
    - Based on preliminary comparisons, Laplace approximation to the weights provided a better approach than BIC.
- Unsure whether bootstrapping is appropriate.

David Dunson | *Duke University*

- Broader context/theory behind Bayesian model averaging.
    - BIC weights v. Laplace approximation – BIC was derived by Schwarz based on taking the Laplace approximation and simplifying assumptions.
        - Uses "unit information coefficient" (prior) in each of the models.
        - Unit information prior is a very high-variation prior and puts higher weight on lower-parameter models (linear v. quadratic).
        - Bayesian modeling assumes that one of the models in the list is true; however, it is more realistic that none of the models is true, and may want to instead choose best linear combination of models instead of choosing one model.
        - Less of an issue in small sample sizes, but large sample sizes require choosing one model anyway; "ensembling" approaches won't do that.
    - Preliminary comparisons by Matt Wheeler suggest that nonparametric or flexible models seem to perform better than the default model averaging approach provided in the workshop material.

- Background on Laplace approximation (from Matt Wheeler):
    - Given a Maximum Likelihood Estimate (MLE) or mode of the distribution, assume some normality constraint at that mode.
        - Approximate integral assuming normal distribution; assume that there are enough data around a normal approximation, and integral to find the probability is relatively well-controlled at the mode and appropriate to approximate based on normal approximation.
- Follow-up: thought behind model averaging is that you give weights on an initial list of models, and the weights on the models are posterior probabilities of the models (assuming one of the models is true); posterior probabilities are proportional to marginal likelihood which takes likelihood of data under that model and integrates out parameters over the prior distribution.
    - In practice, the integral may be intractable; therefore, may run sampling algorithms or analytical approximation instead (such as Laplace approximation).
    - Laplace approximation will be increasingly accurate as more data is added (because posterior distribution will be closer to a normal distribution, and Laplace will be more accurate); does reasonably well with small sample sizes; expression of the central limit theorem.

Day 1: Thursday, December 10, 2015

Ruth Hummel | *U.S. EPA, OPPT*

- Framework is great, no comment on general methods but there is some concern for weighting, coverage probability, and edge cases and when true model is not included.
- Different prior weights seem to force methods into one model; room for counter ideas for prior weighting schemes, expanding model suites, and testing.
- Comment on continuous dose response modeling in general:
    - Unsure why literature only uses mean and standard deviation (SD) rather than the actual data (what is the benefit to using only the mean and SD? Can bootstrap from empirical distribution; coverage can be improved).
    - For the coverage probabilities, look at how these are changed by looking at control means based on the data rather than modeled control means.

Woody Setzer | *U.S. EPA, NCCT*

- To address R. Hummel's question: why not use individual data? Integrated Risk Information System (IRIS) program does not have that data available; only have the summary data and often cannot find the original data; use if available, but not common.
- Unless there is a theoretical issue with the method, would like to have seen model averaging done with other weighting sources such as corrected Akaike Information Criterion (AIC) in addition to BIC; might have given a better result.
- Useful to consider Slob and Setzer paper[4].
    - Better to get an idea of the shape of the universe of the dose response curves that we are targeting with model uncertainty evaluation (e.g., 4-parameter sigmoid model).

Walt Piegorsch | *University of Arizona*

- Are there other model averaging methods that EPA should consider?
    - Two fundamental approaches:
        - Bayesian – incorporates prior information; and
        - Frequentist – does not attempt to apply Bayesian thinking because needed information may not be available or there may not be a use for prior information.
- Do we want to use a Bayesian pathway or stay with frequentist model averaging methods?
    - EPA report seemed to be frequentist model averaging.
    - Only Bayesian part was the prior information on which model would have more or less weight; does not use a fully Bayesian approach.
- Do we want to use the model averaging route or semi- or non-parametric route? Does the panel see a separation on this?
    - If there is prior information available, then build the full model and use the Bayesian method, but frequentist can be effective; use the method that makes sense for the appropriate situation.

[4] Slob, W. and R. W. Setzer (2014). "Shape and steepness of toxicological dose-response relationships of continuous endpoints." Crit Rev Toxicol 44(3): 270-297.

Day 1: Thursday, December 10, 2015

Mike Messner | *U.S. EPA, OGWDW*

- Concerned with how bootstrapping is implemented; resort to bootstrap if nothing else is better.
- Wants to better-understand how bootstrapping is done without the raw data, using reported means and standard deviations, how to factor in the sample size (will get a better feeling of the modeling averaging once this question is answered).
- Prefers to use raw data in a fully Bayesian framework.

Question from Bill Hirzy (American Environmental Health Studies Project): RE: Ruth's "use all the data, not just means and S.D.s" for animal studies, where dose can be closely controlled, this may be OK, but how can it work for epidemiological data, where doses are almost impossible to be precisely quantified?

- Discussants' response: Not possible to use all data for epidemiologic data analysis.

Matt Wheeler | *NIOSH*

- A fully Bayesian approach is preferable but would be difficult to do Markov Chain Monte Carlo (MCMC) sampling due to computational complexity of the models and heavy parameterization (e.g., nonlinearity of Hill, exponential models).

Walt Piegorsch | *University of Arizona*

- Agrees with M. Wheeler and interested in how we could deal with heavily-parameterized models via a Bayesian approach; needs further exploration Another thought: when building in high-variance, vague, less informative priors into model averaging approach, the model averaging could be detrimental – anyone with more experience on this?

David Dunson | *Duke University*

- Regarding M. Wheeler's comment, when running MCMC for complicated nonlinear models, it tends to be very unstable, asymptotic, and confidence intervals tend to be unstable from a frequentist perspective; rather would implement "full approximate Bayesian" (e.g., Laplace approximate) to approximate weights specific to each model, and method would be better in general; main point is to have good full practical performance.
- Regarding the use of diffuse/high-variance priors: if there is a list of models, cannot choose high variance priors for model-specific parameters because of the Bartlett-Lindley paradox.
    - As you choose higher variance prior for model-specific parameters, it will put higher posterior probability weights on the smallest model (true with nested or non-nested models).
    - Rich literature on default type priors (e.g., mixtures of g priors are a nice default approach – can be applied in general to different types of models; however, a highly nonlinear model is an issue; better to address nonlinearity through some appropriate basis expansion rather than addressed as a parameter).

Question from Ted Simon (Ted Simon, LLC): Regarding Matt Wheeler's question--if MCMC is used to obtain many parameters in PBPK models why not the three parameters in the Hill model? This indicates the use of MCMC in PBPK models is incorrect.

David Dunson | *Duke University*

- Models that are nonlinear or Hill models tend to be over-parameterized, and aberrant behavior creates

Day 1: Thursday, December 10, 2015

problems for estimating models, but Bayesian approach could be better because priors account for likelihood function instability; for a fast approach of model averaging, could use Laplace approximation after parameterization of nonlinear model, but MCMC might not be recommended.

Comment from Jeff Gift (U.S. EPA): Bayesian model averaging approach from Dr. Kan Shao (available on www.benchmarkdose.com) has similar features to what M. Wheeler was describing. Dr. Shao will be available for questions.

Additional attendee comments:

- Regarding availability of raw data – in the applicant submission documents, documents are in PDFs and have the raw data that could be used (R. Hummel generally works with this data).
- Better to have more doses and fewer animals per dose (W. Piegorsch– echoes the comment, should start thinking about sacrificing the number of observations per dose for more doses; recognizes that everything has an associated cost).
    - OECD has ongoing project to revisit guideline recommendations to increase number of dose groups for the explicit goal of dose-response estimation.
    - Dose spacing is often politically motivated, fewer animals may not be appropriate for dichotomous endpoints.
    - Many frustrations with the quality of available toxicological data.

Discussion Question 2, 5: Completeness of suite of models

Are there other parametric models that should be included in model-averaging? If so, do you recommend testing performance of the new suite?

David Dunson | ***Duke University***

- Prefers parametric models that are flexible but can capture nonlinearity (expanding in an appropriate basis function).
- Also prefers to move away from parametric models to non- or semi-parametric model that have shape restrictions but otherwise relatively nonparametric (even if not formally so).
- Prefers methods that do not restrict the dose response to be strictly monotone (methods that allow for U shape at low doses or downturn at high doses; would capture model shapes seen in practice).
- Flexible parametric models with restrictions would be useful and could get away from issues of model averaging; could have one model with restrictions, centered on a good parametric model but allowed capturing more shapes than averaging over simple parametric models.

Question from the audience: Would you still be able to use a Bayesian approach to make use of prior information with a nonparametric model?

David Dunson | ***Duke University***

- Yes, can use Bayesian approaches centered on a simple or existing model but allowed for flexibility where model may not be exactly correct; common case is Gaussian process or process of emulation (examples outside of toxicology/risk assessment include some physical models such as modeling volcano eruptions).

Day 1: Thursday, December 10, 2015

- Suggest centering Gaussian process on parametric model or combination of models and then apply an order restriction.

Ruth Hummel | *U.S. EPA, OPPT*

- Important to expand suite of models to anything we find could potentially work.
  - Also include flexible models of the suite with fewer restrictions – no specific extra recommendations but we should explore further.
- Is staying monotonic as important if we are model averaging? Important that models are data-driven.

Woody Setzer | *U.S. EPA, NCCT*

- Agrees with D. Dunson that we should be thinking of alternative semi-parametric approaches, but if thinking purely about model averaging, the disappointing results may have been due to the constraints on model parameters (e.g., power parameters in Exp5 and Hill models); relaxing parameters could give more flexibility in the model shapes.
- Seems that incorporating nested models in suite of parametric models can modestly overweight the models in the overall model average; could simplify suite of models and perhaps get a better result if only the most complex, parameterized form of the model were included—is this true?

Bruce Allen | *Consultant*

- Nested models included in the workshop case study were ones that were on the boundary of the more complex version.
  - Simpler models were nested, and they were nested with specific values for parameters that were on the boundaries of the parameter space in the more complex model version.
  - Assumption is that to give the simpler model some weight, it should be explicitly included due to the boundary of parameter space concern.

Walt Piegorsch | *University of Arizona*

- Response to B. Allen: that should include simpler models that are nested (refer to Occam's window, strategy of shrinking model space).
- Echoes R. Hummel's comments, need to include as many models as possible, and model average approach would be better; need to be careful if Bayesian approach is used because of the computational complexity.
- If using frequentist model averaging approach, would seem appropriate to use more models not less.
- Also should consider non- and semi-parametric methods that do not require a model averaging approach.

Question from Bill Mendez (ICF International): For Dunson, primarily: BMDS-type model forms are not biologically (mechanistically) based. We'd love it if we were in a situation where mechanistic data could be incorporated in a sensible way into a "semi-parametric" model. In addition to volcanoes, etc., can he provide examples of such approaches in biological science? Do other panelists have any suggestions?

David Dunson | *Duke University*

- An example would be a biologically-based model for muscle twitch dynamics based on a differential equation (mechanistic Gaussian process that is a nonparametric Bayesian approach that would mimic

Day 1: Thursday, December 10, 2015

behavior of biologically-based differential equation model but allow deviations from it).

- Frequentist approach from Jim Ramsay: take biological information or chemical dynamics information and put them in as a penalty in frequentist way to estimate a regression function.

Matt Wheeler | *NIOSH*

- Interested in quantifying risk and not identifying the right model; nonparametric methods are superior at doing this.
- However, with a fractional polynomial model (suite of models with reasonable flexibility), parameters can be bounded in such a way to induce monotonicity.
  - Has fewer parameters compared to Exp5 or Hill model; changing number of parameters would change the weight.
  - Fractional polynomial would cover different shapes and allow for a larger number of models compared to a structured Hill model that has so many parameters that it is down-weighted.

David Dunson | *Duke University*

- Regarding relative weights for small versus big models, the BIC-based weights use "unit information prior" (has high variance).
  - The higher the variance prior, the more that a smaller model is favored.
  - Using a default distribution for the model-specific parameter rather than unit information prior might be important not to overly favor small models.
- Fractional polynomial could be a better alternative for the less stable parametric models.

Mike Messner | *U.S. EPA, OGWDW*

- For application to microbial dose response, suite of models would need to be somewhat different to ensure risk is not overestimated and also need to account for uncertainty in dose metric/quantification.

Walt Piegorsch | *University of Arizona*

- Need to address uncertainty in exposure/dose metric.

Comment from Jeff Gift (U.S. EPA): EPA is addressing exposure uncertainty in the arsenic assessment.

Question from Michael Schümann (BfR, RKI & UBA Germany): What about models describing more than one influence factor (e.g. response function in relation (or conditional) to the stage of development or aging)? Often seen in epidemiology.

Walt Piegorsch | *University of Arizona*

- Difficulty is that the problem gets squared, adds another dimension.
  - Only some research has been done on this for continuous and dichotomous endpoints and is currently being studied (some work with Roland Deutsch).
  - Did not attempt model averaging with the problem.

Day 1: Thursday, December 10, 2015

Discussion Question 3: Implementation of methods

Do you agree with the approaches used to implement the methods reviewed in the workshop support material?

Ruth Hummel | *U.S. EPA, OPPT*

- Overall, framework is great; sets stage for follow-up.
    - Some questions about software: when one model is defaulting to its simpler, nested version, what is actually happening in the software? Workshop materials indicate that the weight is set equal to the simpler model.
        - May not the best method to do this; may be better to collapse model to simpler model.
    - Biggest concern is weighting; prior weights are very impactful so sensitivity analysis may be needed; can bias be reduced? Is there a new way to correct for parameters besides BIC, AIC specific to type of data seen in this context?

Woody Setzer | *U.S. EPA, NCCT*

- Major concern is the overall summary of coverages; surprised at how bad the coverages were for continuous models in the workshop materials.
    - Dichotomous models observed in other publications seem to have better coverages, so something must be wrong but not known why it is wrong.
- General advice for a way forward: as EPA explores different methodologies, need to continue simulating results, know what the goals are, and makes sure methods are achieving the goals.
- Specific questions about dealing with variance (Sections A and B of the workshop materials) in simulation results – interesting questions but only edge of the problems and do not deal with the big effects.

Walt Piegorsch | *University of Arizona*

- Bootstrapping is heavily used in the approaches and unavoidable; need to be careful with bootstrapping due to a lack of better alternatives.
- Comfortable with methods but an operable, non-Monte Carlo alternative would be preferable; has worked with dichotomous data to work on alternative to bootstrap.
- Alternatives to BIC weights – strongly in favor of considering alternatives.
    - AIC may produce more stable results than BIC (for their research questions).
    - May be other information criteria such as Kashyap's information criterion (KIC) or corrected Akaike Information Criterion (AICc) that could be used (need something more than just the BIC).
- How to deal with experiments less than four dose groups? Build a better design and get more dose groups. (Recognizes that cost/resource constraints make this hard -- but, yes, it's a hard problem.)
- Add more implementations and other approaches; still need to study it more.

Mike Messner | *U.S. EPA, OGWDW*

- Comment on page 4 under Method 1 of the workshop materials.
- Uncomfortable with idea of calculating weights and using those to produce weighted sum of BMDLs; not

Day 1: Thursday, December 10, 2015

aware of any statistical justification for validity of bootstrap-averaged confidence limits.

- Another out-of-the-box thought on 3b of the workshop materials, question about modeling variance (e.g., variance seen in instrument response, calibrating an analytical instrument).
    - Often there is a constant variance, and variance is proportional to dose; wondering if models that include that combination of variances would be something that has been encountered.

Question from Ted Simon (Ted Simon LLC): Why not use the simplest method? There seems to be very little difference between the results of the various averaging methods and the RfD will be expressed with a single significant digit only?

Matt Wheeler | *NIOSH*

- Response to T. Simon: not sure about what is meant by "simplest method".
- Instead of doing a full Bayesian MCMC, find mode and do Laplace approximation.
- With regard to Hill model and certain parameters of the Hill model – could add prior information (penalized-type likelihood approach) to account for the parameters.
    - Add prior information, then do approach and Laplace approximation and could get a different weighting strategy that is more authentic to Bayesian model averaging and a different weight outside of AIC/BIC framework.
- Not comfortable with bootstrap method because the residuals of regression function are usually available, and the residuals can be bootstrapped, and if there was heteroscedasticity, would use wild bootstrap; however, the workshop material examples, only deal with the sufficient statistic; have not seen a bootstrap only relating to sufficient statistic.
    - Paper by Fletcher and Turek[5]; simulation and approach is to integrate out model average (integrate to find the BMD, but this is not the approach used by EPA).
        - Not able to fully implement this method, but as an alternative, took some of the results and put them into R; BMDLs were more conservative, might produce better coverage.
- Model suite expansion or something else is needed; maybe change model parameter bounding.
- Nonparametric approaches may be better for situations with less data; absent sufficient data, nonparametric approaches prefer a linear relationship, which is the more conservative approach.

David Dunson | *Duke University*

- Agrees with everyone; to follow up on M. Wheeler's point, nonparametric would only go to a line if there was a nonparametric model or prior for a dose response curve that was concentrated around the line and would allow the model to move away from the line if there was more information available; might not necessarily collapse on the line.
- In terms of the undercoverage issue, not surprising; more data will put stronger weight on one model, and Bayesian model averaging would tend to undercover in general unless one of the models was particularly accurate; weights not concentrating because not enough information available.
- Flexible parametric model with some restrictions would be a good alternative.

[5] Fletcher, D. and D. Turek (2011). "Model-Averaged Profile Likelihood Intervals." Journal of Agricultural, Biological, and Environmental Statistics 17(1): 38-51.

Day 1: Thursday, December 10, 2015

- Model averaging (whether in the Bayesian way or not) would putting in AIC weight would be ad hoc.
    - In general, use Laplace approximation in priors that were not concentrated or overly favored simplified models.
- Depends on what is accomplished by model averaging.
    - For quantitative risk assessment to estimate response at low dose and get benchmark doses, the weights on the models are going to depend on where the data are; might be putting a high weight on a model that fits data in a region that you are not interested in.
        - For example, linear model might fit poorly in the low doses but fit well in the higher doses; so if most of the data are in the high dose, there would be higher weight on this model; but the overall intent is to look at the low dose where there is less data.

Comment from Alan Feiveson (NASA JSC): With two predictors, say X1 and X2 where X1 is dose and X2 could be something like diet, the analog of benchmarking would be to define a region in (X1, X2) space such that the predicted response is unacceptable. This is sort of like upside-down response surface analysis.

Walt Piegorsch | *University of Arizona*

- W. Piegorsch published with Deutch[6] to define a two-dimensional BMD, called the benchmark profile; region in (X1,X2) space where risk is acceptable below points of concern or unacceptable above points of concern.
- Create dose-response surface and invert surface into (X1, X2) plane.
- But from a model averaging perspective, still a wide open question.

Question from Ted W. Simon (Ted Simon LLC): How do the discussants feel about dropping models with big spreads between the BMD and BMDL from the group being averaged? For example, some fits may have a BMD=3 and BMDL =5E-08.

Woody Setzer | *U.S. EPA, NCCT*

- Many situations where there is not enough information about data near the BMD (and thus this fit situation may occur).
- Reluctant to drop models only because they have a big gap between BMD and BMDL.
- Data configurations where response plateaus before getting to the first dose (Hill or Exp5 may fit the data) are consistent with a wide range of benchmark doses, so there should be a big gap between BMD and BMDL; excluding these models may be causing confidence in data where there is no confidence/information.
- W. Piegorsch follow-up: Agrees with W. Setzer's comment that the model should not be excluded and that the spread between BMD and BMDL should be further investigated.
- R. Hummel follow-up: The point of doing model averaging is that we do not know which model is correct, so even if there is a model with a large spread, hopefully the correct weighting will capture that at the right level.

[6] Deutsch, Roland C., and Walter W. Piegorsch. "Benchmark Dose Profiles for Joint-Action Quantal Data in Quantitative Risk Assessment." Biometrics 68.4 (2012): 1313–1322. PMC3539281.

Day 1: Thursday, December 10, 2015

Question from Michael Schümann (BfR, RKI & UBA Germany): Which of the approaches takes the population heterogeneity (inter-individual variance e.g. for susceptibility) best into account? Method 3? The selection of the 5th percentile of the weighted BMDs over all iterations is expected (intuitively) to be "best" for it. Do you agree?

Walt Piegorsch | *University of Arizona*

- Does not prefer Method 3 (described in the workshop materials) relative to other methods when there is significant population heterogeneity.
- M. Wheeler follow-up: In frequentist model averaging paper[7], an argument in section 10 of the paper states that bootstrapping is not valid if true model is not in the model space, so can be troublesome to do so.

Woody Setzer | *U.S. EPA, NCCT*

- None of these models take very explicitly into account population heterogeneity but rather incorporate it as a noise distribution.
- Attempts at using population heterogeneity to set benchmark doses use a quantile of the background distribution.
- Tend to target dose response curve at some point in the median of the distribution but could do quantile regression of the data and models that look at 95th percentile rather than the median (Dr. Wheeler had done this but not in the context of model averaging).

Question from Joakim Ringblom (Karolinska Institute): In the article "Comparing model averaging with other model selection strategies for benchmark dose estimation" Wheeler and Bailer[8] suggested to restrict the models averaged to those with a Pearson X2 goodness-of-fit statistic with a P-value of 0.1 or greater. Have anyone tried such restriction in the model averaging for continuous data? What are your thoughts about doing it?

Matt Wheeler | *NIOSH*

- M. Wheeler's thinking has changed since the paper had been published.
- If a model has a poor p-value, then it will get a bad weight (weight will be practically nothing); now, not a proponent of removing model if it has a low p-value.
- Some ideas with Occam's window if there is a very large model space.
- W. Setzer follow-up: if all the models in the suite have low p-value for goodness-of-fit, then no models are appropriate.
    - M. Wheeler response: Can get a bootstrap p-value for the dichotomous data; if you can get a bootstrap p-value that reasonably describes data, modeling averaging is reasonable, then continue; however, if all models fit poorly, then averaging bad models is not going to produce a good result.

Woody Setzer | *U.S. EPA, NCCT*

Regarding comment on dealing with experiments having fewer than four dose groups: if a dataset has fewer than

[7] Hjort, N. L., & Claeskens, G. (2003). "Frequentist Model Average Estimators." *Journal of the American Statistical Association*, 98(464), 879–899. R

[8] Wheeler, M. & Bailer, A. (2009). "Comparing model averaging with other model selection strategies for benchmark dose estimation" *Environ Ecol Stat.* 16(1): 37-51.

Day 1: Thursday, December 10, 2015

four dose groups, could possibly use penalized likelihood or a prior in a Bayesian setting to estimate models even in deficient designs.

Ruth Hummel | *U.S. EPA, OPPT*

- Comment regarding goodness-of-fit: when model averaging, we should not be as concerned in the end that the bootstrapped model may not be fitting well to the data; if we have a model suite that captures all of the possible ideas about fitting a model to the data, then it is the best available data and model suite.
- Still can move forward with a decision even if the goodness-of-fit tests are not as good.

Question from Bruce Allen (Consultant): Could you get good out of all bad if some over-estimate and some under-estimate the BMD?

Matt Wheeler | *NIOSH*

- It is possible if the value is reasonable but not a great estimate.
- Have an idea that model average values are reasonable but not guaranteed.
- R. Hummel follow-up: intent is not to undo one bad thing with one good thing; rather we are using the data to drive a decision when we have nothing better.
- W. Setzer follow-up: sympathized with R. Hummel's point with wanting to move forward when fit is not perfect, and it will be important to convey quality of fit to the data (and associated uncertainty) to management in such cases.
    - R. Hummel agrees with follow-up point.

Question from Keith O'Rourke (Health Canada): Would the data anomalies not give rise to inappropriate weights for different models?

Matt Wheeler | *NIOSH*

- Currently writing a paper with D. Dunson describing nonparametric method that addresses irregularities like high-dose downturn.
    - Fitting the suite of models currently in the workshop materials would not be appropriate.
    - Continually dropping the high dose would not be appropriate.
    - Assumption that downturn occurs after the last dose before the high dose, and if that is the case, risk is underestimated by that assumption; hesitant to use model averaging for such data.

Question from Ted W. Simon (Ted Simon LLC): Muri et al (2009)[9] (food chem toxicol 47:2906) indicate the use of the BMD/BMDL ratio of <2 as a measure of good fit? Are they wrong?

Woody Setzer | *U.S. EPA, NCCT*

- Without any context, yes this is wrong; depends on the question and if the data are noisy.

(NOTE: There is more elaborate discussion on this topic in the notes summary to the earlier question posed by Ted Simon: "How do the discussants feel about dropping models with big spreads between the BMD and BMDL from

[9] Muri, S. D., et al. (2009). "The benchmark dose approach in food risk assessment: is it applicable and worthwhile?" Food Chem Toxicol 47(12): 2906-2925.

Day 1: Thursday, December 10, 2015

the group being averaged, for example, some fits may have a BMD=3 and BMDL =5E-08?")

Comment from Jeff Gift (U.S. EPA): In response to those who say that a response within an individual has a threshold or is nonlinear, NRC has said that on a population level, due to human heterogeneity (e.g., genetic variability), the response becomes linear. Can model averaging be used to simulate or approximate this phenomenon [e.g., by giving prior weights to linear models based on the percentages of the population we think might be sensitive (or insensitive to the endpoint of concern)]?

Walt Piegorsch | *University of Arizona*

- Possible if you can build a large enough suite or class of models to overcome uncertainty (the models could incorporate the features).
- With Bayesian model averaging, you can incorporate informative priors to incorporate the features.

Woody Setzer | *U.S. EPA, NCCT*

- The way model averaging would handle this, the suite of all models all have the feature of low-dose linear, then can build into the models that you use.

Walt Piegorsch | *University of Arizona*

- Can make sure some of the models have that feature, others that don't; push away non-useful models; build in careful parameterization or priors.

Woody Setzer | *U.S. EPA, NCCT*

- Caveat to this is you need to be modeling in the species where you care about low-dose linearity (e.g., modeling human data and not single strain rat data) and need doses that go down to that low dose range because model averaging will not pick it up anyway.

David Dunson | *Duke University*

- People may confuse mixture models and model averaging.
    - Heterogeneous population and different dose response models applied to different members of the population; doubtful that model averaging would characterize this structure well; increased sample size model averaging will fixate on one of the models, and not necessarily characterize that mixture structure.
        - In a mixed membership model, each person, rat, etc. is in one of the models whereas the greater membership model says none of the models is exactly right but each individual has a probability weight on each of the models.
        - Gets at the aspects of mixture models better than model averaging does.
        - In machine learning, this is called mixtures of expert models.

Comment from Bill Hirzy American Environmental Health Studies Project: If that last deals with the SDWA requirement to protect sensitive subpopulations, then simply playing with models is not likely to account for population variance in iodine levels, or arsenic or lead exposures, etc.

Day 1: Thursday, December 10, 2015

Woody Setzer | *U.S. EPA, NCCT*

- In a case like this with a lot of information, may not be fair to characterize as simply playing with models.
- There is much information on population variation in iodine levels or consequences of arsenic and lead exposure; can use this information in models to the extent that we can understand biological consequences of exposures to arsenic, etc.; ideally refine estimates for fluoride based on this information.
- B. Hirzy follow-up: Issue is that in a large population iodine, arsenic, and lead exposures are one thing but then there are people who are unable to clear fluoride from their bodies due to kidney issues and increased fluoride retention has the possibility to cause adverse effects on the brain.
    - Unsure how "tinkering" with models to account for population-level variability is going to be a very realistic way of performing health risk assessments.
    - Rather, need to make policy decisions on what an adequate margin of safety is and take a reasonable benchmark dose dataset and apply the adequate margin of safety.
- W. Setzer follow-up: Ideally, can start with conventional modeling and treat population as a subpopulation, identify fraction that falls within that population; use combination of pharmacokinetic analysis and blood fluoride levels of toxicity to develop an intraspecies/intrahuman uncertainty factor.
- B. Hirzy follow-up: Agrees that a reasonable approach would be 1 order of magnitude or half order of magnitude uncertainty factor with regard to adverse effect levels as a reasonable policy decision to cover sensitive subpopulations.
- W. Setzer follow-up: Although sensitive subpopulations is a problem, not likely to be a problem that model averaging will be able to deal with on a population level.
- B. Hirzy follow-up: Agrees there are simply too many variables that come into play such as genetic, environmental, nutritional factors, etc.; testing needs to take into account complex applied/internal dose and species differences.
- R. Hummel follow-up: agrees with Bill Hirzy; many complicating factors that could be treated as covariates; current model averaging discussion is much simpler (assumes that there are no other complications; other factors such as susceptible populations would then be added on in a risk assessment setting).

Question from Randall Smith (NIOSH): Question about comment on not needing to achieve adequate model fit to use BMD/BMDL estimate; in that case, wouldn't the data fail to meet minimum criteria for BMD estimation, and it may be preferable to use another effect level estimate such as NOAEL or LOAEL (despite the limitations in those estimates)?

Woody Setzer | *U.S. EPA, NCCT*

- No advantage to use NOEL/LOEL over benchmark doses; NOAEL/LOAEL more familiar but approach is not better at describing data than benchmark dose modeling.

Walt Piegorsch | *University of Arizona*

- Caution against relying on NOEL/LOEL almost always.
- Find ways to make nonparametric, semiparametric, or modeling average approaches to use the dose-response data to better inform on the lower limits.

Comment from Bill Hirzy: Currently have a paper under review to use epidemiology data to derive reference dose for

Day 1: Thursday, December 10, 2015

fluoride to protect against the loss of IQ; we used both methods (NOEL/LOEL and BMD), policy implications were not much different.

Continued discussion to question from Keith O'Rourke (Health Canada): Would the data anomalies not give rise to inappropriate weights for different models? Especially if the irregularities are at high doses?

Woody Setzer | *U.S. EPA, NCCT*

- Earlier in this workshop, discussants originally discussed this question in terms of systematic deviations of our models from the monotonic models from the suite but another point of discussion was that initial investigations found surprising prevalence of additional dose-group level variability (likely arising due to something like failures of randomization or issues of caging); did not trust the probability models being fitted; if underlying probability model is part of the fit and parameters are weighted incorrectly, it could give wrong answers.

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?

Woody Setzer | *U.S. EPA, NCCT*

- Recommendations were addressed in written comments but yes, additional testing should be performed; highlights of comments include:
  - Test additional dose-response patterns;
  - Test curves that better represent what are seen in practice;
  - Test additional/different BMR values, e.g., 1%;
  - For constraints question, before additional testing do analysis of what shapes model suite can achieve relative to test curves;
  - Should not constrain power parameter to 1, try 0.5 or 0.25; and
  - Might want to add some arithmetically-spaced dosing.

Walt Piegorsch | *University of Arizona*

- Yes, we will learn more about how the methods work if we do things like add additional dose response patterns, more BMR values (though not highest priority), etc.; dose scaling also seems reasonable.
- Regarding experimental design, more doses are better; need better experimental designs.
- Additional analyses or methods are needed; interested in excess risk at a given dose (e.g., at the BMDL).

Mike Messner | *U.S. EPA, OGWDW*

- Concerned with bootstrapping; recommend bootstrapping approach when using summary statistics (maybe compare results for simulated data vs summary statistics).
- For constraints, report and examine posterior densities; look for clustering of bootstrapping results for parameter estimates.

Day 1: Thursday, December 10, 2015

Matt Wheeler | *NIOSH*

- Agreed with previous test suggestions.
- Would nice to see a test suite to come in to the future; make it available so that anything in the future could be verified against the test suite.

David Dunson | *Duke University*

- Agree with previous suggestions; good to conduct additional testing beyond current model averaging approach; run tests on other methods and determine if alternatives perform better.

Walt Piegorsch | *University of Arizona*

- If we all could come up with set of standards that we can get behind, then we can evaluate and test them.
    - D. Dunson comment: Run realistic test on suite of methods to see if alternatives do better than current model averaging approach.
    - W. Piegorsch comment: Want to see what exactly is working where in comparing alternative methods; interested in model averaging to get guidance on what a better BMD would be.
    - D. Dunson comment: Also would want to introduce nonparametric approaches as an alternative to model averaging.
    - M. Wheeler comment: Agrees but one drawback is that some people are fixated on parametric model being the "right" model; big challenge is to get people from the "right" model to the correct analysis.
    - R. Hummel: Agrees with other discussants; as user of EPA methods, would like to advance something practical as soon as possible. Follow-up with major fixes into the suite of models then move quickly to improve the model averaging approach/coverage, complete the testing, and distribute model average software, given that once the issues are resolved it will be a significant improvement on what we have at EPA. Also would like to know methods in a data-rich environment when individual data are available.

Question from Michael Schümann (BfR, RKI & UBA Germany): Up to now testing has been performed to identify the model averaging approach for dose-response analyze given nearly appropriate data. Would it be possible to test the approach with intentional "inappropriate" data (e.g. nonmonotonic dose-response, threshold models, mixture models, over proportional increasing response variance over dose)?

- D. Dunson comment: Downturns at high doses are not uncommon and models need to be tested against such dose-response data; at least one model in the suite should be flexible enough to capture these other shapes.
- W. Setzer comment: one of the problems with the initial testing for BMDS was that relatively clean dose response data was used; later problems can be avoided by early testing of unusual data sets.

Question from Ted W. Simon (Ted Simon LLC): Since the CV values in Table 6a are all 1% at the central value of the BMDL, does it really make a difference what MA method is used?

- Discussants' response: In reference to page 20 of 23, W. Setzer's Figure 1 – boxplots of coverage values of range of model averaging approaches plus Exp5 and Hill model. If perfect, the values would be around 0.95, but the methodologies vary greatly in their coverage; means that the variability in the estimate is bigger than we think it should be. So it does matter which we pick. The fact that none is concentrated around 0.95

Day 1: Thursday, December 10, 2015

means we have not found the best methodology.

- W. Setzer comment: also look at the BMD/BMDL ratios for the various methods as an important factor.

Question from Bill Hirzy (American Environmental Health Studies Project): Wouldn't such testing have to look at (animal) data in testing separately for threshold and non-threshold endpoints, some of which might be multi-stage, e.g. initiation/promotion/proliferation, (each of which might have unique dose-response behavior) phenomena?

Matt Wheeler | *NIOSH*

- Leery of statistical testing for threshold because you can get arbitrarily close with a non-threshold model and it be indistinguishable. If there was some model that included nonmonotonicity and those possibilities, then you are covered.

Walt Piegorsch | *University of Arizona*

- Get models that get the quality model estimation that you want; Echo M. Wheeler's idea about threshold modeling.

Woody Setzer | *U.S. EPA, NCCT*

- Misunderstanding exists between mathematicians and biologist/toxicologists; to a mathematician, the threshold is the dose where the response changes from exactly the background response. Below the threshold, the effect is 0. To a biologist/toxicologist, the threshold is the dose where the response changes trivially or at unimportant magnitude, not that the response is always exactly the background response,and below the change point, the effects are pretty small, and above that, increase is rapid; getting close to a threshold model is probably good enough.

Comment from Jeff Gift (U.S. EPA): I suggest that discussants view Bayesian modeling approach on Dr. Shao's benchmarkdose.com website.

Kan Shao | *Indiana University*

- Website is based on PyStan core for MCMC analysis; front end and back end on Python.
- Website is capable of fitting seven dose response models for dichotomous data.
- Can specify different prior model weights for model selected and analyzed.
- In the near future, will implement feature to let user specify parameter prior for the models.

Model averaging on Dr. Shao's website is different than materials in workshop:

- Methods based on Wasserman 2000 paper.[10]
- Two levels of model weight calculations:
    - First level based on data; fit model to data, then website will produce model weight for each model selected; model weight at this stage is then calculated based on the likelihood estimate, number of parameters in model, and number of data points in the input (similar to the BIC method), slightly different because in the Bayesian setting;
    - Second level, when calculating BMD from each model, then can specify the prior model weight for

[10] Wasserman, L. (2000). "Bayesian Model Selection and Model Averaging." J Math Psychol 44(1): 92-107.

Day 1: Thursday, December 10, 2015

different model, then get model average BMD estimate; at this level, user-specified model weight and the weight calculated based on the pure data fitting, both will play a role in calculating the final posterior model weight.

- For the model average BMD calculation, only one method: for each model, we calculate BMD; 30K posterior sample for each model, then each 30K sample will time the model weight and add the final weight to get final average BMD distribution.

Comment from Ted Simon: I think what Woody said about what S. Sand was trying to do by pegging the BMR=21% as a measure of the start of the rising phase of a Hill model, about thresholds paper in Tox Sci 2006[11]. He indicated the Hill coeff=1 occurred at BMR = 21%.

- Comment from W. Setzer: Depending on the shape of the power parameter in the Hill model, the rising phase could be all over the place; lacking context, unsure where 21% BMR is required.
- Ted Simon will send the paper.

Comment from Keith O'Rourke (Health Canada): I would look forward to a wider discussion of methods and relative merits anticipating Matthew's suite of tests to assess these. Perhaps similar to what is outlined in section 6 of this preprint: https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf.

- W. Piegorsch comment: Agrees that a wider discussion is needed on methods and merits of how to assess evolution of BMD and BMDLs.
- M. Wheeler comment: Clarify that earlier discussion was about a hypothetical test suite, not something that actually exists.

Comment from Bill Mendez (ICF International): There is an alternative to BMDL for sigmoidal dose response in Sand et al. (2012) paper[12]; where the curve change point where it turns most sharply upward (Sand paper); place where the 2nd derivative is taken.

Walt Piegorsch | ***University of Arizona***

- Taking a BMD is just a point estimate. BMDL provides a measure of the variability of the point estimate; you need to get the measure of uncertainty such as a lower confidence limit. Sand has profile likelihood method for estimating the confidence interval.

WRAP-UP DAY 1

---

[11] Sand, S., et al. (2006). "Identification of a critical dose level for risk assessment: developments in benchmark dose analysis of continuous endpoints." Toxicol Sci 90(1): 241-251.

[12] Sand, S., et al. (2012). "The point of transition on the dose-effect curve as a reference point in the evaluation of in vitro toxicity data." J Appl Toxicol 32(10): 843-849.

Day 2: Friday, December 11, 2015

Discussion Question 6: Motives for using model averaging in health assessments

a. How useful is model averaging for characterizing model uncertainty? (MU)

b. How useful is model averaging for incorporating prior information?

Discussion Question 7: Should alternatives or complements to model averaging be investigated?

Examples: Isotonic regression; Non-parametric and semi-parametric (Bayesian and frequentist) modeling; Full Bayesian model averaging; Flexible parametric models

Walt Piegorsch | ***University of Arizona***

- Start by examining the motives for model averaging. Papers by West et al. (2012)[13] and Ringblom et al. (2014)[14] have shown that for both dichotomous and continuous models, unadjusted simple model selection is not a good strategy, especially when model uncertainty is present. Using the unadjusted method means the true BMD and BMDL could be missed.
- Model averaging is a viable tool, and can be a powerful and useful approach for addressing the model uncertainty issue.
- Should alternatives or complements to model averaging be investigated?
    - Yes, it is completely necessary; model uncertainty is a big problem, or we wouldn't be talking about it. Just trying something simple can bomb frequently, so need something more.
    - Possible alternative approaches include nonparametric regression, full Bayesian, and parametric models.
    - It is evident that these methods are valuable and appear to be answers to the model uncertainty problem.
- Dr. Piegorsch has done more work in the frequentist area of model averaging, but has also looked at the Bayesian area too.
- Nonparametric methods might be longer term, but there are existing approaches that can be used now. More work needs to be done to expand those methods.

Mike Messner | ***U.S. EPA, OGWDW***

- In favor of all things Bayesian, and model averaging seems to be right direction.
- Subpart of Question 6 regarding inferences below the range of observations; there is no good solution for this. Model averaging can fail at this just like anything else. At low doses, nothing may work when extrapolating beyond established data.
- For the long term, could use model averaging to estimate BMDLs for numerous contaminants. As that happens, it will make sense to take what has been learned, and distributions of parameter values, and begin to build useful priors on parameters, as well as priors on models; priors and equal weights among models are good defaults.

[13] 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

[14] 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. http://dx.doi.org/10.1016/j.yrtph.2014.03.004

Day 2: Friday, December 11, 2015

Matt Wheeler | *NIOSH*

- Should also explore the concept of how model averaging could be tested.
- Regarding how model averaging could be used in health risk assessment and how useful is it for characterizing model uncertainty, M. Wheeler proposed databases. (See Appendix C for slides)
    - Build a database with the various curve shapes observed in prior analyses of chemicals (start with IRIS chemicals, for example) and incorporate that prior information into model averaging to estimate what kind of shapes may be observed for chemicals tested in the future.
    - Add different papers and places in the literature to find databases, and have a good amount of shapes, both continuous and dichotomous.
    - Gave examples of possible database sources to create dose-response curve test suite (See Appendix C).
- Regarding methods, and model averaging, W. Piegorsch has published good things for nonparametric methods showing that they are better than current practice. In the short term, there are lots of methods out there that have been proven to be better than current practice; they are not perfect, but having 85% coverage is better than 50%, even if it's not 95%.
- Could create some test suites, and use **decision table** (See Appendix C) to choose between different methods for a certain data situation.
- **Risk Definition/ Data Type**: Go through for a given risk definition and data type, for example, dichotomous added risk has been studied well.
- **Benchmark References (BMRs)**: Could then start to study given BMRs, for example starting with 10%. Might be interested in other levels as well, but start with it since it has been studied.
    - Can ask if it is better than current practice (in terms of coverage, bias, alpha, etc.).
    - In Dr. Wheeler's opinion, if it is biased to one side it is okay, but a higher bias not so much.
    - Using a BMR 10%, if a method (nonparametric or model averaging) was developed and established that coverage, then it's better than current practice and should be used. Then go to 5% and 1%, etc. and compare.
- **Point of Departure Linearization (POD):** Dr. Wheeler disagrees with the comment that the low dose is outside the range of data; recent paper (Wheeler et al. 2013)[15] suggests something different.
    - When accounting for model uncertainty, also accounting for model uncertainty in the low dose regions that have not been studied.
    - Right now estimate 5 or 10% BMR as the POD, and linearize down to risks of interest.
    - As a decision rule, there is almost no difference in estimation if linearizing from 10% vs. computing directly at 0.1%.
    - One can have cases where this approach (e.g., using model averaging or nonparametric modeling to compute directly at the risk of interest, e.g., 0.1% used by NIOSH), is actually more conservative (determines the chemical to be more toxic) than the EPA current cancer approach of linearizing from 10% using a single model; even when there is essentially a quantal-linear model in a dichotomous space.

[15] Wheeler, M.W. and Bailer, A.J., 2013. An empirical comparison of low-dose extrapolation from points of departure (PoD) compared to extrapolations based upon methods that account for model uncertainty. Regulatory Toxicology and Pharmacology, 67(1), pp.75-82.

Day 2: Friday, December 11, 2015

  - Don't always have to consider POD linearization all the way down; if using a nonparametric method, might be doing no different than current practice. With nonparametric or model averaging, it may be better not to linearize at 10%.
  - If studying a chemical for thresholds, if there's some level where it doesn't increase by much, then all of a sudden it does, and a data set representative of that; the POD linearization method is overly conservative by several orders of magnitude.
  - Using a model averaging approach, can get numbers 30-fold greater than what would be in published risk assessments; one can feel confident that they're being public health protective.
- **Non-Monotone**: Can develop nonparametric methods for nonmonotonicity with "goofy" data sets.
- All of this information provides a **decision table** to determine whether the method (model averaging, nonparametric, etc.) in question is better than current practice, and if it should be used in this data situation.

David Dunson | *Duke University*

- Agreed with and enjoyed M. Wheeler's presentation; suggested that width of confidence intervals is another important thing to look at, on top of coverage.
- In terms of model averaging, it's very clear it's better than current practice.
  - Short term, within a year, seems to be a good option to switch to using model averaging over selecting one model.
- Regarding its usefulness for capturing model uncertainty; it is useful, but with caveats that can hopefully be addressed in the long term.
  - Suggested "competitions" to come up with best way to estimate BMDs. This idea was seconded and endorsed by many discussants.
  - NIEHS had a similar exercise for mixtures recently.
- Regarding model uncertainty and model averaging, one limitation, compared to good nonparametric methods, is capturing uncertainty when going beyond the range of available doses, or even between wide-spaced doses in a study.
  - Parametric model averaging will confidently extrapolate beyond the available range of doses; to estimate parameters precisely and decide which models to weigh, then it will be interpolating beyond the range of data without appropriately accounting for uncertainty in that range (e.g., low dose).
  - As M. Wheeler and others say, there is no great solution; however appropriate nonparametric methods will inflate uncertainty as they go to regions of dose without direct data or observations.
- In summary, model averaging is a great step forward in the short term.
- In the medium term we could be doing things like:
  - The exercise M. Wheeler suggested, to see which other methods can beat simple model averaging (parametric models in simple way).
  - For nonparametric methods, the Bayesian methods tend to have some advantages in terms of characterizing model uncertainty and getting good interval methods, particularly at low doses; it is hard to get accurate estimates using frequentist methods.
    - Good when putting in prior information about shape restrictions; those will do better in terms of M. Wheeler's exercise than short term model averaging approaches.

Day 2: Friday, December 11, 2015

- Quite practical to very quickly try nonparametric methods more often; probably going to be Bayesian with shape restrictions (not necessarily monotonic, allows for deviation from monotonicity in the low and high dose), maybe something that allows flexibility. This will be a better approach in general moving forward.

Ruth Hummel | *U.S. EPA, OPPT*

- From the perspective of R. Hummel and OPPT, model averaging is better than what is currently in use, so it should be put into play as soon as possible.
- In the short term, should commit to frequentist model averaging as a new method, knowing that other methods may be better in the long term.
    - Should do more research into the weighting scheme, as there is room for improvement there, and also consider the candidate model for the suite of models being used; maybe throw in flexible fractional polynomial model.
    - Remove method one, and perhaps methods three and five, as they are not as "data driven".
    - Add in an alternative for using raw data when available.
- Once the best method has been determined a few months from now, do some small scale sensitivity analyses on certain features; look at additional data dose-response patterns as an extra level of testing, and then put that out as soon as possible, with the caveat that we are working towards improving the method's coverage and ability to obtain the smallest confidence intervals possible.
- In the longer term, should develop a test suite, distribute it, maybe at a competition, get others to compare what they have done with others (maybe splines, isotonic regression, Bayesian, etc.)
- Should get frequentist model averaging in the program offices as soon as possible.

Woody Setzer | *U.S. EPA, NCCT*

- Model averaging is useful for characterizing model uncertainty.
- Use of prior information is an excellent idea, but we need transparency (so as not to appear to make a selection just to get the answer desired). There needs to be legwork (guidance, documentation, and justification) on how to use prior information.
- In terms of coverage, the best approaches (most consistently to 95%), were the two simple models (Hill and Exp5).
- Perception of uncertainty in models of continuous data is due to the data having small number of doses. Need to distinguish real model uncertainty from apparent uncertainty from over-parameterization.
    - Have characterized a wide range of dose-response datasets with the Exp5 model. In this trial, Hill and Exp5 outperformed model averaging approaches in terms of coverage.
    - Range of model uncertainty may not be as large as thought.
- In long term, learn more about the range of dose-response shapes available; then come back to model averaging formulation, but more carefully informed by actual problems.
    - Nothing currently addresses the various ways nonmonotonicity might come into datasets (low or high end), e.g. changes in cytotoxicity at high dose can cause data to move back to control in the high end. A more complex model is needed to take that into account.
- In the mid-term, refine the model averaging approach to improve coverage and evaluate the width of confidence intervals they're producing.

Day 2: Friday, December 11, 2015

- In the very short term, the two 4-parameter models (Hill and Exp5) worked well; can create a penalized likelihood scheme to help fit them in datasets with small number of parameters.

Matt Wheeler | *NIOSH*

Response to W. Setzer

- Exp5 and Hill on average did better, but on the template p1 (normal, chronic), the best coverage rate reported for Exp5 was 68%. Ran a nonparametric method for the same dataset and got 98% coverage.
- For Hill/Exp5; one of the problems seen in risk assessments is that the models had problems converging to find BMD/BMDL.

Woody Setzer | *U.S. EPA, NCCT*

Response to M. Wheeler

- There are two issues here: 1) the types of dose-response shapes that exist, 2) parameter estimations for these models. This requires good design, and reasonably placed doses with respect to the dose-response shape.
    - Due to the nature of nonparametric methods, the estimates would likely be more stable.
- W. Setzer suggested un-constraining the power parameter in Hill and Exp5. Relaxing those constraints might, in general, address some of the coverage issues that have been identified in the supporting material.

David Dunson | *Duke University*

Response to W. Setzer

- Relaxing constraints makes it harder to estimate parameters.
- Could have historical database, perhaps with some standardized doses, and estimated parameters; there would be empirical distributions of parameters across database, useful for choosing a prior for parameters. That should greatly stabilize the estimation.
- Other issue is Hill by itself may seem better, but it doesn't get appropriately high weight compared to a simpler model, which might be due to the weights being used.
    - Using a Bayesian approach with high variance on parameters, which is implicit in BIC weights, would over-penalize the larger models; would weight highly the overly simple models, get worst coverage in certain cases.
    - D. Dunson endorses looking at other weighting schemes.
- In the short term, nonparametric methods can be used if implemented with reliable code; should have people try it out.

Bruce Allen | *Consultant*

Clarification to all discussants regarding the Hill/Exp5 in the current application.

- In the software distributed and the runs done on the test datasets, set it up so that for both models (Hill and Exp5) they ran a simpler nested model (power for Hill, Exp3 for Exp5), and did a very simple comparison.
- Since the nested model should never have max log-likelihood greater than the more complicated model, they made a substitution whenever they saw that the opposite was the case; that was to just use the predictions (BMD estimates) from the simpler models and the log likelihoods from those simpler models in the averaging and definition of the weights.

Day 2: Friday, December 11, 2015

- When that happened, they kept the parameters at four and the BIC calculated based on four parameters, even though they used likelihood obtained from simpler model.
- Simple method of correcting for potential convergence issues that M. Wheeler referred to.

- Does that information revise any of the thinking about the superiority of the Hill/Exp5 model?

Woody Setzer | *U.S. EPA, NCCT*

Response to B. Allen

- W. Setzer felt like B. Allen had acknowledged the fact that full Hill/Exp5 models are more difficult to use on many datasets, and came up with a step down-like procedure to get an estimate that is consistent with Hill/Exp5 model. If all four parameters can't be estimated, one or more will be fixed to a default value and the remaining will be estimated; for purposes of weighting all will be treated as if they estimated them.
- It's conceivable that they could come up with better methodologies, but this method is reasonable and does not change W. Setzer's impression/feeling about using Hill/Exp5 model.
  - W. Setzer assumed that if he looked at individual model results they would fit the same way.

Bruce Allen | *Consultant*

- B. Allen confirmed that W. Setzer's assumption was correct.
- We've been hearing a consensus that simple model selection is almost certainly not the right way to go. Is that sort of defaulting appropriate and different from the simple selection scheme that we've been using historically? And does that make the comparison of the averaging results to individual models, and our thinking about that, any different?

Woody Setzer | *U.S. EPA, NCCT*

Response to B. Allen

- This is different than the standard model selection scheme, where choosing the best invites overfitting; the confidence interval is too short, which understates the degree of uncertainty.
- This approach involves fitting the biggest model, if it can't, fit next one down; use the comparison of likelihoods as a test to see if it converged or not. Not picking smaller model because it fits better, it can't fit better than the full model.
- They've done the simulation testing, can actually see the coverage; would like to see wider range of dose-response shapes. Would endorse suggestions already made to put together a library of such shapes.

Bruce Allen | *Consultant*

- Emphasized that Hill/Exp5 have the back-up for default option; this is something important to keep in mind when reviewing those models' performance, as that feature may easily be overlooked.

Matt Wheeler | *NIOSH*

- There are better optimization methods out there that have been developed since the original publication of the BMDS software. He prefers NLopt: http://ab-initio.mit.edu/wiki/index.php/NLopt.
- Recommends NLopt-R, which is able to find a better optimum than the optimization routine, especially when relaxing constraints.

Day 2: Friday, December 11, 2015

- Some of the results are due to the convergence problem; wondered if that's just the optimizer itself not being able to find the maximum because there is such a flat likelihood to optimize.

Bruce Allen | *Consultant*

Response to M. Wheeler

- For Hill/Exp5, if there's not a distinct plateauing of response, it is hard to fit those models. The optimizer picked larger and larger k half saturation parameter, at some point it stops when it should go to infinity. Thus they chose the power model as a fallback. Even the best optimizer is going to stop somewhere when it should go to infinity.
- While interested in other optimizers for the work on BMDS in general, he thinks we would still have this issue, when we automate these runs and do our model comparisons.

Matt Wheeler | *NIOSH*

Response to B. Allen

- Putting in priors is the equivalent to putting in penalized likelihood; need to have a way to regulate that estimator.

Comment from Jeff Gift (*U.S. EPA):* EPA uses one of the optimizers from Dr. Spelucci (MIT); for the most part we have not altered the optimizers that have been used for years in the BMDS models.

Woody Setzer | *U.S. EPA, NCCT*

Response to J. Gift

- Dr. Spelucci refined his license for the software to the point where the BMDS team was uncomfortable with using recent versions, for fear of violating his license requirements.
- Used Dr. Spelucci's optimizer because it allows for optimization subject to non-linear constraints, which one has to do with polynomial models in BMDS; still have to solve that problem.
- Can get other models directly from the optimizer with bound constraints.

Matt Wheeler | *NIOSH*

Response to W. Setzer

- NLopt will do linear constraints, it has a whole suite of algorithms; he used it to do the profile likelihood for the Hill model given the linear constraints.
    - It has linear, nonlinear, inequality constraints, etc. in the algorithms section.

Comment from Ted Simon (Ted Simon LLC): M. Wheeler's statement about estimating a BMD/BMDL below the range of observation, i.e. the low dose region, is at odds with EPA's 2005 Cancer Guidelines: Section 3.2.4 (pg. 3-16) http://www.epa.gov/sites/production/files/2013-09/documents/cancer_guidelines_final_3-25-05.pdf

Matt Wheeler | *NIOSH*

- Yes, but guidelines can be updated.

Day 2: Friday, December 11, 2015

- When starting to use nonparametric methods that appropriately deal with model uncertainty and regions between doses, it is not necessarily as important. It is almost so unimportant that in situations where one would expect linear extrapolation is necessary, get identical estimates (within a factor of 1-2).
- But if on the other hand there is something going on and linearization is not appropriate, would get a different answer.

Comment from Jeff Gift (*U.S. EPA):* The guidelines say that the lowest POD that is adequately supported by the data is used. Many of you may be familiar with instances where EPA has used other BMRs besides the standard 10% to establish a cancer POD.

Comment from Ted Simon (Ted Simon LLC): But the cancer guidelines are talking about the region between the lowest dose and zero. Linear extrapolation is not the same as estimation of the POD as the BMD/BMDL.

Matt Wheeler | *NIOSH*

- For the BMDL, estimate a POD, say with a BMR 10%, calculate BMDL, which is then the "risk estimate accounting for statistical uncertainty." Take that BMDL and do the linearization from that to background.
- M. Wheeler is proposing, if one computed BMDL using a nonparametric method specifically, but could get similar result with model averaging, at least with dichotomous data, that BMDL going all the way down to background is in many situations very similar to the POD approach that's suggested at 10%, but it can be different when there are more threshold-like effects.

Question from Bill Hirzy (American Environmental Health Studies Project): Would Michael please discuss how his office used BMD in its non-cancer risk assessment that used dental fluorosis as the end-point? E.g., was model averaging used? How was the difference between fluoride levels in drinking water and actual doses ingested that resulted in observed fluorosis rates addressed in the modeling? Were levels of calcium as well as fluoride addressed in the modelling, etc.?

Refers to EPA OW's response to the NRC Committee on Fluoride Toxicity's recommendation to do a new risk assessment for fluoride. EPA's response was published in 2010. Reference follows: U.S. EPA fluoride dose-response analysis for non-cancer effects. Health and Ecological Effects Division, Office of Water. Dec. 2010. available at http://water.USEPA.gov/action/advisories/drinking/upload/Fluoride_dose_response.pdf

- Mike Messner (U.S. EPA, OGWDW) did not have any knowledge of that.

Question from Mehdi Razzaghi (Bloomsburg University): Could the panelists please comment on the use of BMA (Note: assumed to stand for Bayesian model averaging) for developmental toxicity dose-response modeling and risk assessment? I would be interested in their comments for both dichotomous and continuous responses.

- Several discussants felt this could be a long-term question to explore, given how developmental toxicity would add to the complexity of statistical problems.

David Dunson | *Duke University*

- There is really no conceptual or technical barrier to doing model averaging in the developmental toxicity arena.
- Could use Bayesian model averaging and obtain some approximation of the marginal likelihood, for example using Laplace, in a hierarchal random effect model.

Day 2: Friday, December 11, 2015

- Longer term goal because of the time it would take to develop it.

- There is literature based on developing dose-response models for developmental toxicity. Since models are available, it shouldn't be too difficult conceptually to average them.

Woody Setzer | *U.S. EPA, NCCT*

Response to D. Dunson

- There are two or three currently used general approaches for developmental data for quantal endpoints: 1.) beta binomial likelihood, and 2.) generalized estimating equations.
- The complicating factor is that developmental models also include a covariate (litter size, or other determinant, such as implantation), and it is unsure how often that is an important parameter. This complicates the fitting.

Walt Piegorsch | *University of Arizona*

- There is a fair amount of work for BMD calculations for developmental data. Although D. Dunson commented that it would not be difficult to brainstorm hierarchical constructions to make the models make sense, no one has gotten to it yet.

Comment from John Fox (U.S. EPA): Three examples of Bayesian models for developmental data:

1. A hierarchical [Bayes] modeling approach for risk assessment in developmental toxicity studies Christel Faes, Helena Geys, Marc Aerts, Geert Molenberghs Computational Statistics & Data Analysis 51 (2006) 1848 – 1861.
2. Bayesian Semiparametric Analysis of Developmental Toxicology Data Francesca Dominici and Giovanni Parmigiani Biometrics, Vol. 57, No. 1 (Mar., 2001), pp. 150-157.
3. Association Models for Clustered Data with Binary and Continuous Responses Lanjia Lin, Dipankar Bandyopadhyay, Stuart R. Lipsitz, and Debajyoti Sinha Biometrics. 2010 March; 66(1): 287–293. doi:10.1111/j.1541-0420.2008.01232.x.

Question from Keith O'Rourke (Health Canada): Would the cost and likely benefit of the Bayesian Model Combinations/Ensemble methods be short or long term?

- No responses.

Question from YS Lin (U.S. EPA): How to incorporate data uncertainty into model uncertainty issue (e.g., model averaging)? Or we just assume there is no error (e.g., measurement error or systematic error) in the data used for modeling efforts?

Woody Setzer | *U.S. EPA, NCCT*

- All methods attempt to account for uncertainty in the estimates due to variability in the data; that is a fundamental aspect of fitting single models to data, and one of the goals of doing the estimation process.
- The confidence interval seen on the BMDs that come out of a model averaging effort will incorporate uncertainty from the data and the model; nothing extra is needed.

Day 2: Friday, December 11, 2015

Mike Messner | *U.S. EPA, OGWDW*

Response to W. Setzer

- There is no accounting for exposure/dose uncertainty in these models, which is an issue with microbial dose-response assessment.

Woody Setzer | *U.S. EPA, NCCT*

Response to M. Messner

- Models are really set up for designed experiments, where the administered dose is the covariate and is known with certainty.
- An error-in-input modeling approach should be considered for that issue, which is easy to set up in a Bayesian context. That would account for dose uncertainty.

David Dunson | *Duke University*

- Can use the Laplace approximation to get weights if there are a bunch of developed models.
- There is nothing conceptually new that needs to be developed for each new case (i.e., developmental toxicity), but it is important to vet those methods for regulation, and make sure they work well in specific contexts.

Question from John Fox (U.S. EPA, co-chair of the NCEA Statistical Work Group): There seemed to be consensus for using MA short-term. However, Woody's graph of achieved coverage is disturbing. Should EPA explore alternatives like changing the shape constraints and using existing nonparametric models before applying these MA programs for health risk assessment?

Adding to this, J. Gift (U.S. EPA): Can nonparametric modeling advise efforts to improve the coverage of model averaging methods in the short-term?

David Dunson | *Duke University*

- It could be used in short-term. There is code developed already, at least in simple cases without complications (ex: no hierarchy, not developmental toxicity), so it is easy to apply to standard type studies.

Walt Piegorsch | *University of Arizona*

- The available approaches aren't old, so the vetting has been only within the study.
- Exploring how these different strategies will work when compared (model averaging and nonparametric BMD calculations) is closer to short term than long term.

Ruth Hummel | *U.S. EPA, OPPT*

- Want to improve coverage as much as possible as fast as possible, so need to decide what the highest priority is. There is plenty to explore long term; for the short term just want to get something out.

Matt Wheeler | *NIOSH*

- The code runs in R, 6-10 secs per dataset. Could run it against every one of the simulated datasets.

Day 2: Friday, December 11, 2015

Woody Setzer | *U.S. EPA, NCCT*

- Want to relax the constraints on the models used in the model averaging suite and have a competition among different methods, looking at coverage and confidence interval width for 1% and 10% BMDs. That is possible in the next six months, can then decide which way to go short-term.

Comment from Jeff Gift (*U.S. EPA):* One goal is improved coverage, but also trying to respond to the National Research Council's (NRC) concerns regarding model uncertainty and incorporating mode of action/prior information. While nonparametric methods are out there that could be implemented short term for fitting observed data and improving coverage, not so certain with respect to the other concerns.

Comment from Michael Schümann (BfR, RKI & UBA Germany): "Model averaging" seems to be one of the best ways to find out (a) which is the "residual uncertainty" testing or aggregating a variety of models (and parameters) and (b) which consequence the choice of a selected model will have on the BMD/L estimate. But, it is not always clear to me, which would be the best descriptive for "model and parameter uncertainty" for each of the models in contrast to the averaged model. I would suggest, that EPA might develop a recommendation for one index (together with restrictions) to support the standardization of evaluation of uncertainty.

David Dunson | *Duke University*

- General idea with model averaging; get some models, if have one model then have those parameters, if those parameters have some biologic/toxicological interpretation, that would be useful.
- In this setting, it might not be the case; fine to just get a good estimate of the dose-response curve and uncertainty in that estimate.
    - Don't rely on interpretation of parameters in individual models, which addresses Michael Schümann's concern.
- Then can use model averaging, nonparametric models, without losing interpretation about key quantities of interest.

Comment from Joakim Ringblom (Karolinska Institute): Regarding when the log-likelihood of e3 is higher than e5, the solution is to fit e3 and e5 sequentially and using the parameters from e3 as start values for e5. The starting values for the c-parameter could be set to an arbitrarily high value. Thereby the log-likelihood of e5 will start at a log-likelihood that are almost the same as the log-likelihood of e3 and you will almost never end up with the log-likelihood being higher for e3 than for e5.

Woody Setzer | *U.S. EPA, NCCT*

- The problem is that as one increases the number of parameters, especially for small numbers of doses; have to specify starting values for nonlinear optimization methods. The starting value has to be close to optimum or do not get convergence.
- J. Ringblom is suggesting a step-up approach using starting values from simpler models; a way of walking up the tree of nested models to maximize the chance of optimization.
    - This might be a useful diagnostic; might have convergence more often for more complex models, but on the other hand this is happening largely when more complex models should not converge.

Day 2: Friday, December 11, 2015

Bruce Allen | *Consultant*

- The current convergence problems are not ones with bad starting values; for the Hill model, they're cases which have parameter identifiable as the half-saturation point; if don't see saturation, have very little or no information about where that is. To the extent that it wants to go as high as possible, but can't get to infinity by any of these algorithms.
- For the x5 and x3 the issue is different; x3 is nested in x5, the parameter space is "funky" on that to find that nesting.
    - For any optimization routine out there, if start with x3, not going to get a full x5 model and vice versa. That has to do with a whole set of parameter estimates where likelihood will be exactly the same. Can't get from one to the other using optimization.
    - Wanted to explicitly include both to have a separate estimate for x3 and x5. Sometimes they're different, with x5 giving the higher likelihood, but other times not; that's more of an algorithm problem.

Matt Wheeler | *NIOSH*

Response to B. Allen

- Would it be possible to use a derivative-free random genetic optimizer to get away from the multiple modes? That might be a superior way to optimize it, and find starting values.
- Should be able to find the optimum or close to the optimum, take those parameter values and plug it in to an optimizer that uses a derivative to get a best estimate.

Bruce Allen | *Consultant*

- That is correct, an evolutionary algorithm might be a way to solve that problem.

David Dunson | *Duke University*

- These models are intrinsically unstable, often there are very flat regions of the likelihood surface; doing traditional things (like finding the mode and looking at the curvature around it) doesn't work very well. Can get similar behavior to these models with more stable models.
    - They should be discarded in favor of models that can fit same types of shapes with more stable parameter estimations.
- Hill can be viewed as a richly parameterized model. There are other richly parameterized models, like semi-parametric models, that could characterize the same shapes without the intrinsic instability of this model.

Discussion Question 8: Dichotomous Data

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?

Mike Messner | *U.S. EPA, OGWDW*

- Recommended a book edited by Roger Cooke.[16]

[16] Roger Cooke (ed), Uncertainty Modeling in Dose Response. Wiley, 2009.

Day 2: Friday, December 11, 2015

- For microbial dose-response, the response is infection or illness, which is dichotomous. The other special issue is that the dose is discrete, as there are a countable number of organisms.
    - It's not worth putting effort into adapting the software for modeling microbial dose response because there aren't enough data.

Matt Wheeler | *NIOSH*

- Dichotomous data have been better studied than continuous.
- Less concerned with dichotomous data results. There is less information per observation than a continuous observation, which might be why model averaging has tended to perform well for dichotomous endpoints.
- Regardless of why it's performing well, it's definitely better than current best practice. The results may not be good in terms of the continuous model, but are very different in the dichotomous case.
- No reservations for dichotomous model averaging, nonparametric, or semiparametric modeling.
- It might be better to move to models that are linear in their parameters; as a suggestion, a mixture of g-priors.

David Dunson | *Duke University*

- With dichotomous data, the weights tend to be more spread out, as opposed to continuous modeling where there is more concentration (higher weight) on an individual model. In that sense, maybe under-estimating uncertainty.
- Computation is a different problem using dichotomous data, depending on the kind of model. Often if models are linear in the parameters, can choose default conjugate priors that don't include any substantive prior information in them, and get closed forms for the marginal likelihood and not have these issues.
- In that sense continuous data is easier. Those advantages don't hold in models that have highly non-linear parameters.
    - Can have a non-linear model, in terms of characterizing very flexible non-linear dose-response curves; but it's still linear in the parameters, making the estimation stable.
    - In terms of model averaging, can get the exact marginal likelihoods and weights on different models using very established default approaches (mixtures, g priors, etc.).
- Don't have that type of simplicity in the dichotomous case.

Ruth Hummel | *U.S. EPA, OPPT*

- There is more in the literature for dichotomous than continuous; EPA should look into this as a recommended method.

Comment from Jeff Gift (*U.S. EPA):* To clarify why the EPA looked at continuous data first; there was already literature on dichotomous model averaging, and we saw the Wheeler and Bailer software we liked. Do the discussants have any tweaks to the Wheeler and Bailer software?

Matt Wheeler | *NIOSH*

- Given the discussion today, the software is antiquated in terms of what could and should be done, but it is a good starting point.
- A software package for dichotomous model averaging (with GUI) plus semi-parametric modeling is available here:

Day 2: Friday, December 11, 2015

https://drive.google.com/folderview?id=0B_A_n8NRzYVFMnFpSUNIaUx2ZVk&usp=sharing. It has not yet been tested in a huge suite, such as BMDS software.

- Caution: Downloading the software is a multi-step process. It's a MatLab program compiled using the MatLab Compiler Library; have to go through several steps to get there. Have to install 3 programs before the actual program can be run.

Woody Setzer | *U.S. EPA, NCCT*

- Used M. Wheeler's software on 10-20 chemicals in an ILSI Europe project looking at dose-response for risk assessment for genetox carcinogens. The software performed well, although unsure of the coverage properties.
- Persuaded the European toxicologists that this was a reasonable thing to do, even though historically they have been slower to adopt dose-response modeling.
- Close to being able to use model averaging for dichotomous models; need to review what the state is now and make decisions about methods.
- Regarding NRC's mandate that dose-response modeling take into account adverse outcome pathway (AOP) or mode of action (MOA) information; he doesn't see how to do that with empirical dose-response models in any sort of substantive way.
  - It might be relevant for dichotomous models because there is a simple approach to divide the world into 1) chemicals that act through tolerance distribution mechanisms (for which there might be some threshold-like behavior), and 2) chemicals that work through an internal stochastic process (like mutagenesis, for which there might be low-dose linear behavior).
  - For continuous models, you're not able to get that far. While it would be nice for modeling to have mechanistic undertones, it can't be done with the kind of data toxicologists generate.

Walt Piegorsch | *University of Arizona*

- Frequentist and even Bayesian model averaging for dichotomous data, done carefully, is ready now (his team recently tried Bayesian model averaging, in paper coming out soon).[17] Should compare all of these together to determine the best practice.

Comment from Ted Simon (Ted Simon LLC): Agree 100% with Woody that empirical modeling has no means of incorporating biological knowledge.

Matt Wheeler | *NIOSH*

- Possibility of doing something in the middle; empirical modeling that is nonparametric but based on biological assumptions.
- Knowledge based on a biological system can be incorporated in a lot of different ways, and still come up with nonparametric estimates using Bayesian information.

Woody Setzer | *U.S. EPA, NCCT*

Response to M. Wheeler

[17] Fang, Q., Piegorsch, W. W., Simmons, S. J., Li, X., C, C., and Wang, Y. (2015). Bayesian model-averaged benchmark dose analysis via reparameterized quantal-response models. Biometrics (in press), doi:10.1111/biom.12340

Day 2: Friday, December 11, 2015

- It ought to be possible for nonparametric or semiparametric models that through a set of constraints/priors, could make general statements about the expected degree of smoothness.
    - How specific can you get with incorporating biological information?

Matt Wheeler | *NIOSH*

Response to W. Setzer

- Creating a basis upon a differential equation that creates a function that should resemble that differential equation with greater probability than a normal nonparametric method.
- Simulate a draw from a random process to see what that looks like, and actually get a function that looks like what's expected.
- For example, in a PBPK model, to take into account elimination/uptake, could do first order differential-equation first order approximation; would be a Gaussian process method but the prior would resemble after the PBPK draw, can do it for second order differential-equation as well. Did that in the paper.[18]

Question from Ted Simon (Ted Simon LLC): How detailed? Can you incorporate details of a cytotoxicity/proliferation cancer MOA into one of these models?

Matt Wheeler | *NIOSH*

- Depends upon how that information is formalized scientifically. Possibly; could find some of those Gaussian process emulators that could do a reasonable job.

Comment from Jeff Gift (*U.S. EPA):* Asked Dr. Setzer to expand upon this idea that parameters in frequentist models don't have biological meaning; distinguish that from being able to use prior information to inform a parameter.

Woody Setzer | *U.S. EPA, NCCT*

- Example of Hill-shaped parameter (D); after looking at datasets, the value of D in *in vivo* datasets varied precisely around value of 1. Use that value to construct a prior for future datasets; not going to constrain D to be 1, but center it at 1. That's how prior information could be used to fit that model.
    - There is no mechanistic biological information that went into it; D might be around 1 for lots of different mechanistic reasons. Didn't need mechanistic information to get to that.
    - Same for C; for a given endpoint it tended to be constant.
- The model choice itself might incorporate some level of mechanistic understanding; if it is believed that MOA operates in a particular way, the expectation would be one kind of dose-response shape, which would be biased towards a particular model.
    - Can't do that very well. There may be special cases in which parameters can be identified in simple empirical models, but it will be case by case.

[18] Wheeler, M.W., Dunson, D.B., Pandalai, S.P., Baker, B.A. and Herring, A.H., 2014. Mechanistic Hierarchical Gaussian Processes. Journal of the American Statistical Association, 109(507), pp.894-904.

Day 2: Friday, December 11, 2015

Comment from Jeff Gift (*U.S. EPA):* Agrees with W. Setzer that we rarely understand the mechanistic underpinning to inform curve shapes, but (as he pointed out in his example) we can use historical dose-response information as prior information for a modeling analysis.

Woody Setzer | *U.S. EPA, NCCT*

- Possibly. Even with semiparametric, should be able to build into the constraints to incorporate a plausible range of shapes for that biological endpoint.

Comment from Bruce Allen (Consultant): Rather than priors on parameters, could you see MOA info informing the overall shapes that are more likely -- like a "threshold" type shape vs low-dose linear?

Woody Setzer | *U.S. EPA, NCCT*

- Can see the argument for dividing chemicals (threshold-like versus chemicals stochasticity (e.g., genotoxicity).
- Don't see how it could be done with continuous endpoints, but others may have a different idea.

Discussion Question 9: Is Model Averaging Ready for Use in Chemical Health Assessment?

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 or misleading?

Matt Wheeler | *NIOSH*

- Yes, model averaging as implement in the support materials is suitable for use in chemical health assessments; however, model averaging for continuous endpoints still needs development.
    - Three specific areas for that development: 1) investigation of constraints and their impact, 2) use of penalized likelihoods, and 3) expanding to a larger model space.

David Dunson | *Duke University*

- Nothing to add, everything was covered in prior discussion.

Ruth Hummel | *U.S. EPA, OPPT*

- Emphasized that the sooner something is sanctioned for use at EPA, the better.

Woody Setzer | *U.S. EPA, NCCT*

- Model averaging not ready to be put into BMDS "this week," but in principal it is a promising and reasonable approach, after a short amount of tweaking and exploration.

Walt Piegorsch | *University of Arizona*

- Should move forward!

Day 2: Friday, December 11, 2015

Mike Messner | *U.S. EPA, OGWDW*

- Model averaging is an improvement over the current practice (pick one model and run with it).
- Some features that could be added to the software:
  - Saving the bootstrap sample
  - Added GUI features
  - Easy way to plot scatter plots of paired parameters
  - Visual diagnostics, graphics
- Echoed prior concerns about using summarized data rather than raw data

Comment from Jeff Gift (*U.S. EPA):* Assuming convergence can be improved using suggestions that have been made, 4b (a model averaging approach) might be the approach closest to the best. Does anyone disagree? Would add a feature for graphical representations of the averaged model. Any other suggestions for graphical or textual features for the output?

Matt Wheeler | *NIOSH*

Response to J. Gift

- Would rather see profile likelihood rather than bootstrapping (Fletcher et al. (2012**)** suggests the profile likelihood interval).[19]
- Wants to regularize models, and add the Laplace approximation to make things much faster.
- Wants to investigate and probably implement other weighting schemes.
  - Using that approach, the entire weight settings would switch to models that were more indicative of the actual curve.

Comment from Jeff Gift (*U.S. EPA):* Comments on use of other AIC (Akaike Information Criteria)? Corrected AIC (AICC)?

Walt Piegorsch | *University of Arizona*

Response to J. Gift

- AICC essentially the same as AIC when sample sizes are large (>200 observations, about the size of a chronic design) and the number of parameters is small.
  - AICC only useful with small sample sizes and moderate parameters.
- Response to B. Allen's question regarding other information criteria that might focus weights based on fit in the area of greatest concern: he has looked at the KIC in the dichotomous case but it didn't work any better. Currently doing work on the FIC (focused information criteria); finding it is competitive, but not substantially better.
- Response to J. Gift's question of whether the FIC is something that could be tuned to pick a portion of the curve: he has no knowledge of continuous case, and for dichotomous case the jury is still out.

[19] Fletcher, D. and Turek, D., 2012. Model-averaged profile likelihood intervals. Journal of agricultural, biological, and environmental statistics, 17(1), pp.38-51.

Day 2: Friday, December 11, 2015

Jeff Gift (U.S. EPA) and Ruth Hummel (U.S. EPA) both said methods that allowed them to pick models that did the best at the low end of the dose-response curve would be intriguing to their offices at the EPA.

Woody Setzer | *U.S. EPA, NCCT*

- Concerned about any method that puts more weight on one part of the dose-response curve (e.g., one that focuses on region near the BMD) because it might not fully take into account the information provided on dose-response shape by other observations.
- At least two problems: 1) in practice don't know the true model (and therefore may not know where the BMD is), and 2) irregularities in the data may be more likely to influence fits and therefore weights ("focused too much on noise").

Discussion Question 10: Conclusions

Do you agree with the conclusions made in Section 4.1 of the workshop support material?

David Dunson | *Duke University*

- No comments.

Ruth Hummel | *U.S. EPA, OPPT*

- A few things have been discussed that should be looked at, but in general on board.

Woody Setzer | *U.S. EPA, NCCT*

- Don't entirely agree with conclusions; more nuance is needed. Refer to written comments for full response (See Appendix A).
- Bullet 1: Make sure to include models that are data-generating models. Problem of constrained parameters also possible.
- Bullet 3: Difference between lognormal and normal distribution. The coefficient of variation of the lognormal distribution used was relatively small (10%), so not much difference between lognormal and normal; this could be different for datasets with more variability. This might make the effect of assuming the wrong data distribution become apparent.
- Bullet 5: Including variance model with constant variance.
- Bullet 6: Experimental design, sample size; more simulations would sort out difference between design and sample size.

Walt Piegorsch | *University of Arizona*

- Wherever possible include as many models in the suite as you can; when closest possible model isn't in the collection (uncertainty suite) that creates a lot more trouble.
- Do not worry about having too many or nested models – Occam's razor (as applied by Wheeler and Bailer) would be a way to handle those issues.
- Certainly try extending other forms of BMR.
    - Same with non-constant variance models.

Day 2: Friday, December 11, 2015

- Expand experimental designs; run into trouble when trying to parametrically (or non-parametrically) fit a complicated dose-response with only a few doses, in order to estimate something as complicated as the BMD.
    - Cannot rely on past designs that were not intended to estimate the BMD. Should start thinking about what kind of designs to use when the BMD is an important target parameter.
- Caution when extrapolating above or below doses.
- Need to keep in mind the users of this software/technique; should keep it easy, practical, and transparent.

Mike Messner | *U.S. EPA, OGWDW*

- No issues with conclusions and observations.
- Would like software to be available in an R package.
- One note about dose levels; to increase the number of dose levels, the number of subjects per dose level would have to decrease (cost, experimental constraints). Thus some other type of bootstrapping would be necessary if the number of subjects gets below five in each dose.

Matt Wheeler | *NIOSH*

- Use of these tools has to become increasingly transparent and user-friendly. Need results to be reproducible.

Woody Setzer | *U.S. EPA, NCCT*

Response to M. Messner

- Decreasing sample size per dose level as the number of dose levels increases; bootstrapping the residuals rather than the observations themselves may be a way to address this concern.

Question from Keith O'Rourke (Health Canada): Perhaps rejection sample from log-normal to match closely the means, SDs and ns in your summarized test data sets (maybe keeping 10 to 100 sets) and re-run? (Addresses raw data and log-normal assumption issues.)

- No responses.

Comment from Ted Simon (Ted Simon LLC): There's data out there done in rainbow trout with thousands of animals per dose and data in rats with 5 - 50. These might be good real datasets to explore some of the sample design issues.

Matt Wheeler | *NIOSH*

- Has used that study, called it the mega-trout study. Used it to start the argument of low-dose extrapolation with methods that account for uncertainty, in Wheeler and Bailer (2013).[20]

Comment from Jeff Gift (*U.S. EPA):* Comments on Dr. Shao's website (tool for Bayesian model averaging)? Is it something the agency should pursue?

[20] Wheeler, M.W. and Bailer, A.J., 2013. An empirical comparison of low-dose extrapolation from points of departure (PoD) compared to extrapolations based upon methods that account for model uncertainty. Regulatory Toxicology and Pharmacology, 67(1), pp.75-82.

Day 2: Friday, December 11, 2015

Woody Setzer | *U.S. EPA, NCCT*

- Missing software documentation. (J. Gift: approx. one week away at time of workshop)
- Enjoyed the interface.
- Uneasy with the idea of putting potentially sensitive or internal data in the cloud. Individuals may statutorily not be allowed to submit some data to the cloud, making the modeling platform useless for them.

Comment from Jeff Gift (*U.S. EPA):* Not suggesting a web-based approach, looking more at what Dr. Shao has done as being very preliminary alpha versions of these approaches for testing. Would ultimately code it on internal EPA servers if decided to pursue it.

Kan Shao | *Indiana University*

- Goal of developing the website is to make it functional and then perfect. By 2016 it should be capable of dealing with both dichotomous and continuous data (currently just dichotomous). Also by March 2016, should be able to handle both summarized and individual data.
- The tool is developed for Bayesian model averaging, not for nonparametric modeling or semi-parametric modeling; however both may be incorporated in the future.
- The structure of this application is a web-based and local combined application; if the user doesn't have web access, can still use local version to run the application, but online can update with the latest features.
- Documentation: not a lot of time to do it, but will be done as soon as possible; quickstart guide for each webpage to show how to proceed and how to use the website done in one week

Bruce Allen | *Consultant*

Summary of the major items to look into for getting MA ready for use (in no particular order):

- Consider another optimizer (NLOpt)
- Remove or adjust parameter constraints
- Consider different basis for model weights (AIC or other IC that may give more weight to larger models).
- Probably take out Methods 1, 2a, 3a (and maybe 3 and 5 if their bootstrap methods are deemed to be too much dependent on modeling and not enough on data)
- Investigate and implement other bootstrapping methods (especially if can use individual data)
- Use approximate methods (like Laplace approximation) instead of bootstrap

Jeff Gift | *U.S. EPA*

- Thank you so much to discussants, and ICF International.
- There have been a lot of suggestions, we'll look at them all, there will be a number of comparative tests done, sensitivity analyses; this is one consideration of many.

WORKSHOP ADJOURNS

Appendix A

Workshop Discussants' Preliminary Written Responses to Discussion Questions

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.

The five methods all involve bootstrap sampling. When raw data (result for individual animals) are available, I would recommend Markov chain Monte Carlo sampling, rather than bootstrap. The bootstrap seems to be required because raw data are unavailable, so I wonder if sampling across all groups (rather than within groups) would be best when the numbers of animals per group is small and the number of groups is large.

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.

The suite of models should be expanded to include nonparametric models and models for dichotomous data. Consider including artificial neural network models, such as those described in Herbie Lee's thin book "Bayesian Nonparametrics via Neural Networks" (ASA/SIAM, 2004). Earlier, I suggested adding models for microbial dose-response. Now I believe the value of including them may not be worth the effort. Those models would need to account for uncertain discrete doses and they would need to satisfy single-hit theory – especially the notion that probability of infection can't exceed probability of exposure.

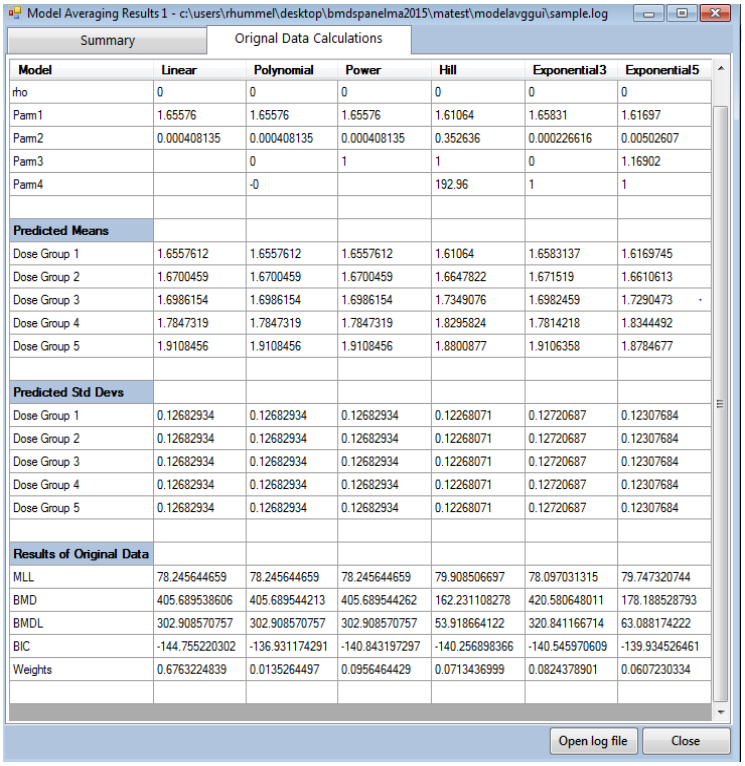
Walter Piegorsch: One could always think of additional forms for inclusion in the models being averaged (in various publications I've called this collection an "Uncertainty Class" of models). Based on the results seen in the EPA report with FMA for continuous data, I am led to argue for inclusion of 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):


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

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?

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). I would expect the bootstrap to perform poorly when the number of subjects in a group is small, say less than 5.

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.

BMRs based on 1% will be highly uncertain. Additional tests at 5% and 20% might be informative.

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.

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 and test them in a similar fashion.

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. To increase the utility of the software, I recommend (a) adding options for saving the MCMC or bootstrap samples, together with weights so that the end user can explore other options to more fully account for and communicate uncertainty over the full dose range and (b) including options for plotting posterior or bootstrap parameter estimates in pairs and for plotting the parameter estimate distributions / posterior densities. These plots could reveal highly correlated parameters that may need rethinking/reparameterization, undue influence of constraints/priors, and issues with bootstrap/MCMC sampling.

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

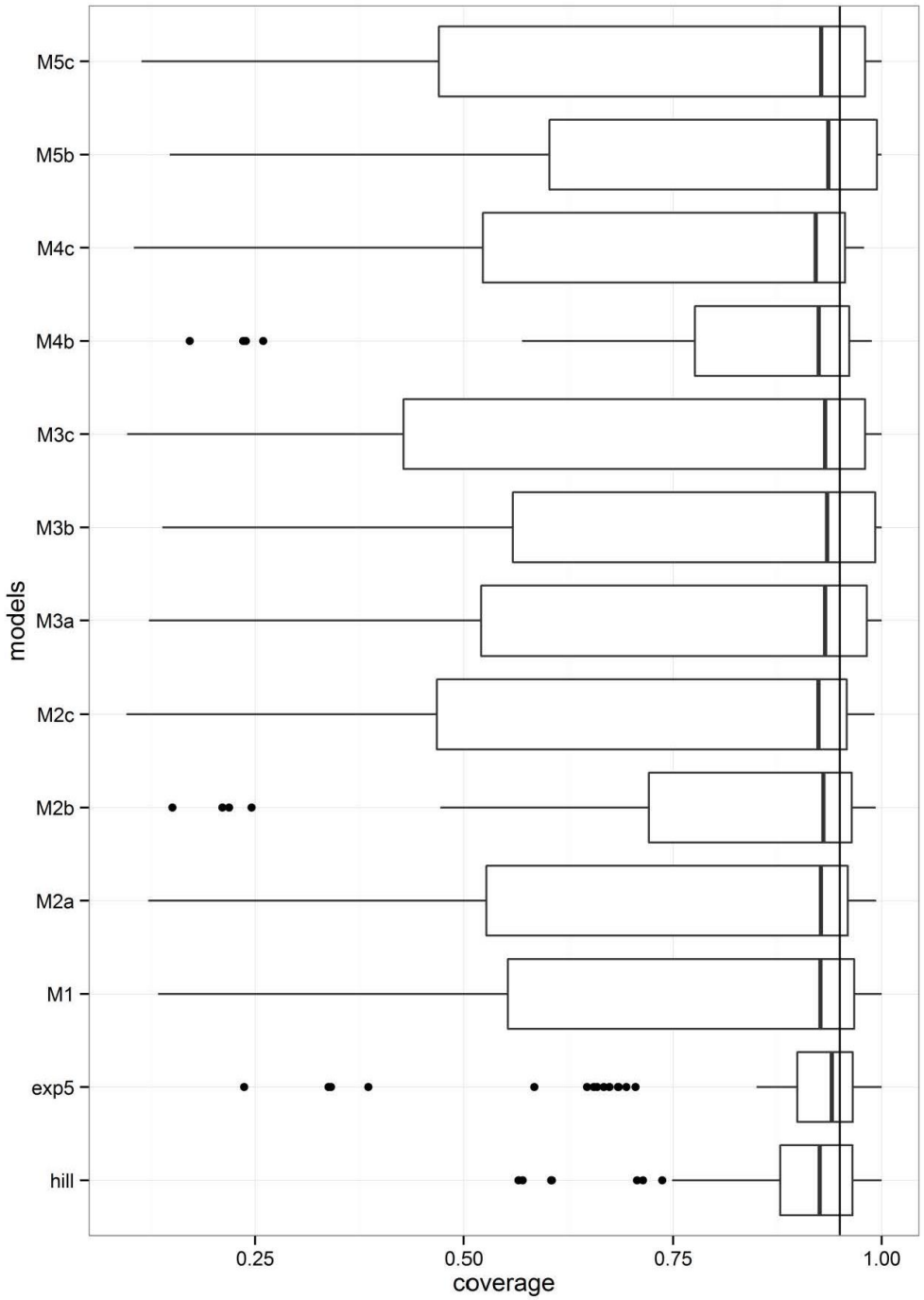


Figure 1. Distribution of coverage estimates over the model averaging methods and two simple models, over all templates, error models, both designs, and both approaches to modeling the variance.

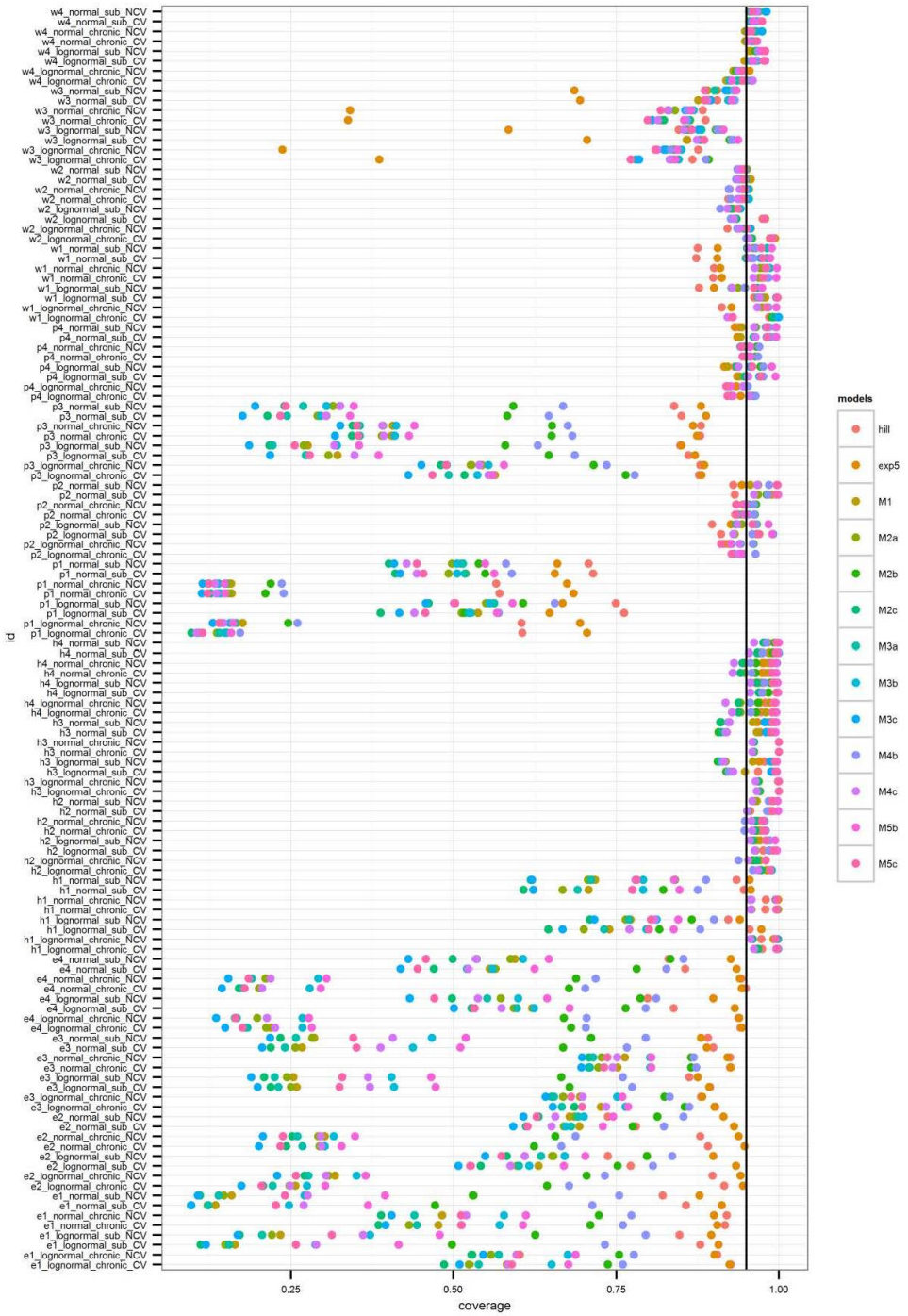


Figure 2. All coverage estimates, stratified by template, etc. and color-coded by method.

Matthew Wheeler's Figures

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

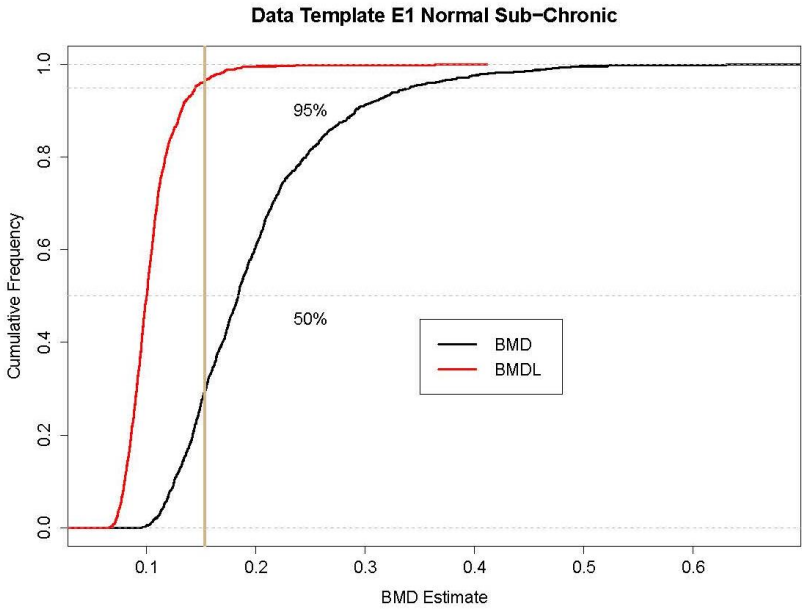


Figure 2: BMD results for data template E1 assuming Normal errors and a chronic study design.

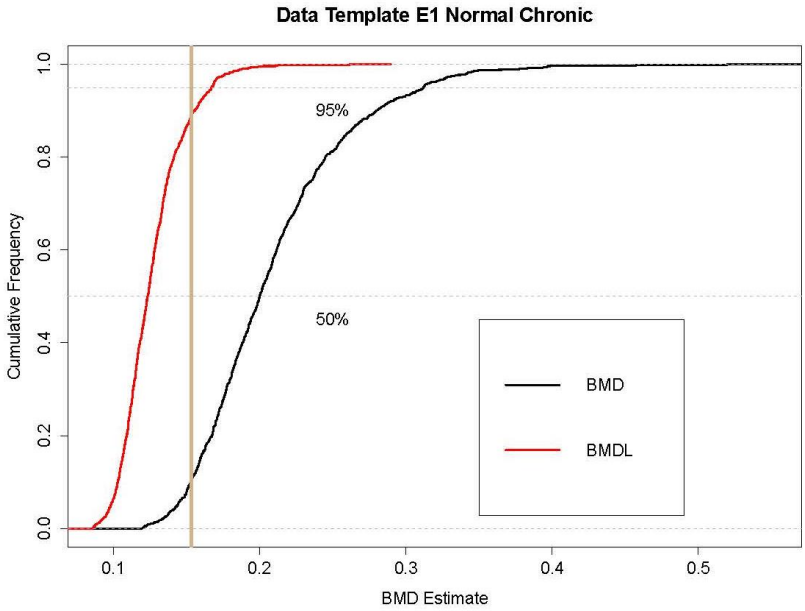
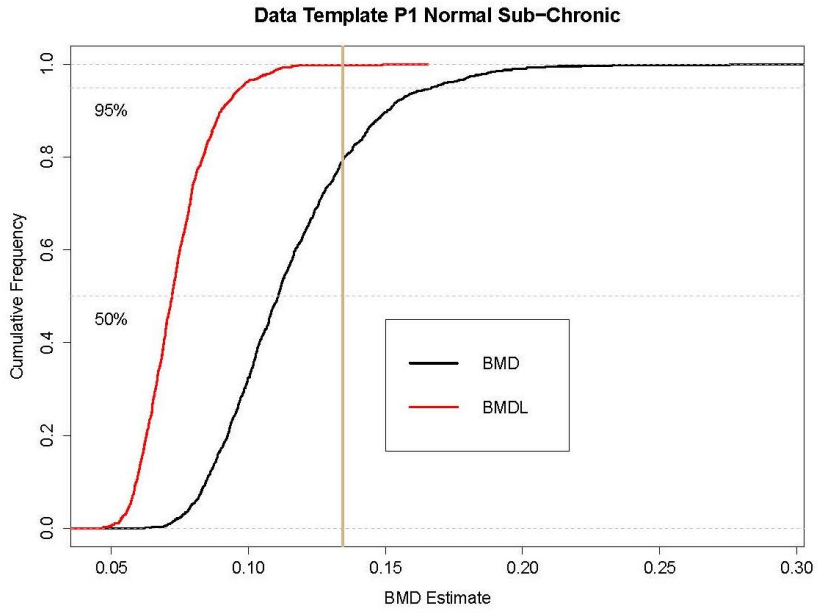


Figure 3: BMD results for data template P1 assuming Normal errors and a sub chronic study design



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)

Appendix B
Public Presentation Slides

National Institute of Environmental Health Sciences
Your Environment. Your Health.

Dose response measures of benzene toxicity in Diversity Outbred Mice: A population based model for investigation of inter-individual variation

***John E. (Jef) French, Ph.D.**

***Adjunct Professor, Center for Pharmacogenomics & Individualized Therapy and Department of Nutrition at University of North Carolina at Chapel Hill**

1) I have no conflict of interest in the issues to be discussed in this workshop and

2) (2) The comments and opinions expressed in this presentation are mine and should not be construed to reflect the opinion or policies of my former or present employer.

National Toxicology Program
U.S. Department of Health and Human Services

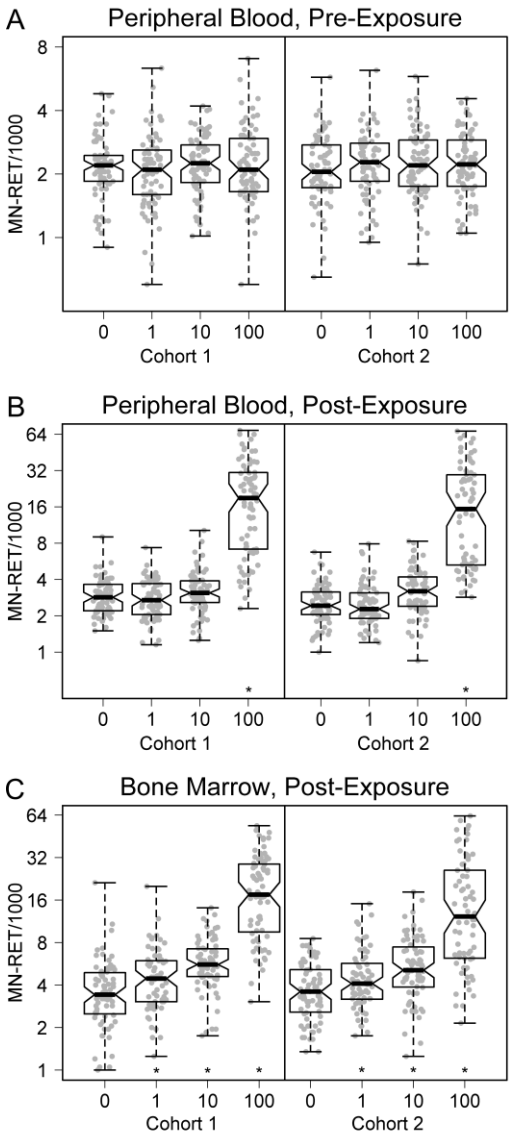
Benzene 28 Day Inhalation Exposure: **Proof of Concept**

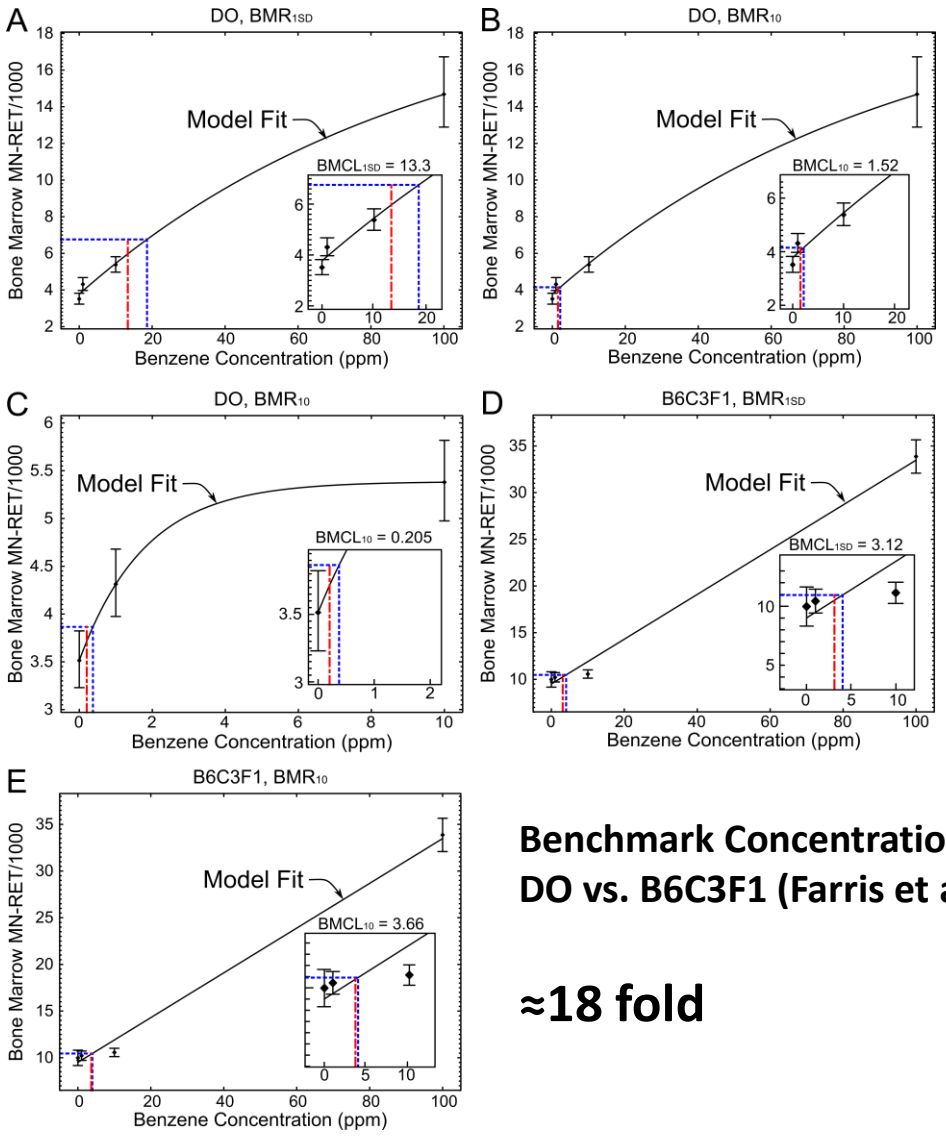
- Diversity outbred (J:DO) male mice: 7 & 8th randomly outbred generations; selected from 175 breeding pairs
- Randomly assigned to exposure groups by weight
- Dose levels: 0, 1, 10, 100 ppm benzene, 28 days, 6 hr/day
- 75 male mice per exposure group, 300 mice/study
- 2 independent cohorts to assess reproducibility (600 mice total)
- Endpoints for hematotoxicity and genetic damage
  - % reticulocytes and micronucleated reticulocytes in peripheral blood and bone marrow
  - Mouse Universal Genotyping Array (9K SNPs)
  - Linkage mapping analysis (DO QTLRel)

French et al. EHP (2015) 123:237-245.
(Advance online – 6 November 2014)

2 of 7

The results….


A
Peripheral Blood, Pre-Exposure
MN-RET/1000
8
4
2
1
0
1
10
100
Cohort 1
Cohort 2
B
Peripheral Blood, Post-Exposure
64
32
16
8
4
2
1
C
Bone Marrow, Post-Exposure

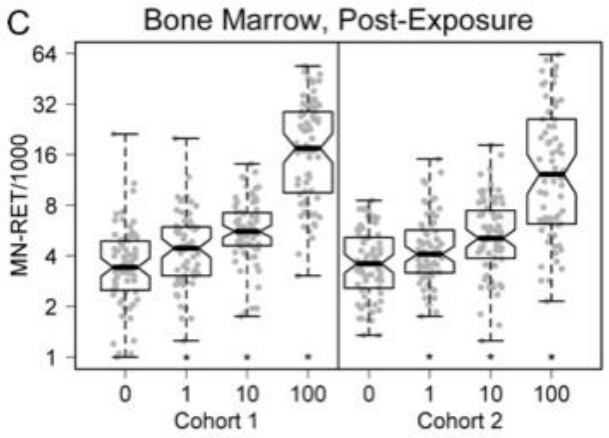


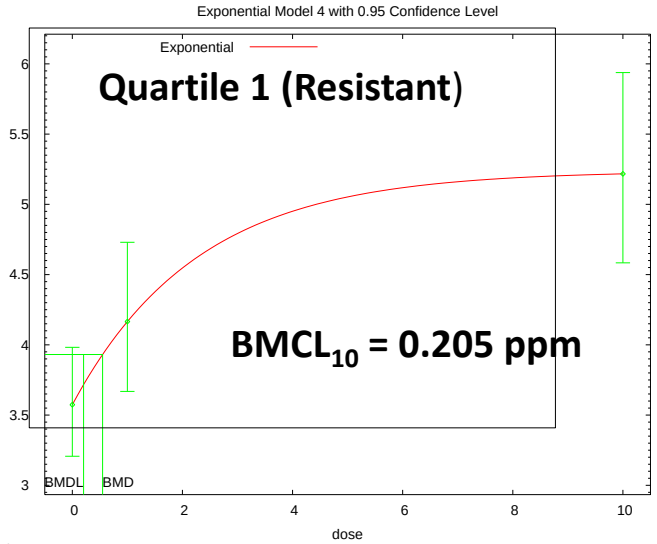
Benchmark Concentration Models
DO vs. B6C3F1 (Farris et al. 1996)

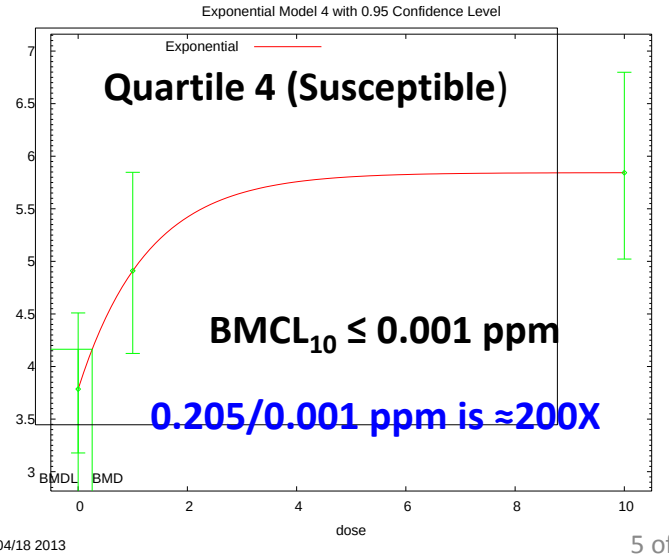
≈18 fold

4 of 7

Quartile	BMC	$BMCL_{10}$
All quartiles	0.367	0.200
Q1	0.550	0.205
Q2+Q3	0.315	0.117
Q2+Q3+Q4	0.275	0.130
Q4	0.257	<0.001







5 of 7

Benchmark Concentration Model (BMC)

EPA Risk Assessment 1988 (Updated & revised 2003)

BMCL* = 8.2 mg/m^3 (2.6 ppm); UF = 300; MF = 1

RfC = 3 x 10^{-2} mg/m^3 (9.4 ppb)**

*Decreased lymphocyte count in 44 individuals **(Rothman et al., Am. J. Ind. Med. 29:236,1996**). No metric provide (i.e. 1SD or 10% above the control mean).

***No adverse effect level peripheral blood HSC counts determined in occupational exposures to benzene (Lan et al. Science 306:1774, 2004 & McHale et al. 33:240, 2012)**

6 of 7

Appendix C
Matthew Wheeler's Workshop Presentation Slides

Hypothetical Decision Table

			BMR = 10%		BMR = 5%		BMR=1%		POD Linearization	Non-Monotone
Method	**Risk Definition**	**Data Type**	**Bias**	**Coverage α=0.05**	**Bias**	**Coverage α=0.05**	**Bias**	**Coverage α=0.05**		
Current Practice	Added	Dichotomous	0.03	89%	0.05	79%	0.1	69%	Y	N
Method A	Added	Dichotomous	-0.03	92%	-0.06	95%	-0.1	99%	N	Y
Method B	Added	Dichotomous	0.001	96%	0.02	93%	0.01	79%	Y	N

M. Wheeler (NIOSH) Day 2 Slides 1 of 2

Possible Databases Sources to create dose-response curve test suite:

Dichotomous:

Nitcheva, D.K., Piegorsch, W.W. and West, R.W., 2007. On use of the multistage dose–response model for assessing laboratory animal carcinogenicity. *Regulatory Toxicology and Pharmacology*, *48*(2), pp.135-147.

Continuous:

Slob, W. and Setzer, R.W., 2014. Shape and steepness of toxicological dose–response relationships of continuous endpoints. *Critical reviews in toxicology*, *44*(3), pp.270-297.