Discussion Questions for Model Averaging Workshop

The National Research Council and EPA's Science Advisory Board have, in various reports, 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 AIC. One approach is to employ model averaging.

The primary goal of this Model Averaging Workshop is to obtain expert consultation that will assist EPA in the identification of a model averaging approach for dose-response analyses that offers the greatest advantage for the development of chemical health assessments. To facilitate the workshop, a support document has been provided that documents the development and testing of a software package that implements several model averaging methods.

The methods used in the model averaging software that is being distributed to facilitate this workshop have all been proposed in the literature for dichotomous data and models, or are extensions of such methods. They are predominantly based on Bayesian statistics. However, while a full Bayesian analysis may be possible in some instances, simpler approximate methods for averaging have been presented.

The model averaging software package distributed for the workshop is intended to facilitate 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).

EPA is seeking expert input in order to identify a model averaging approach that has the greatest potential to facilitate the development of health assessments, with a focus on the following discussion questions.

1. **Overall approach to model-averaging** Are there other model averaging methods that EPA should consider?

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

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

   (a) 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)?

   (b) 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]

(c) 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?

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

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:

   (a) 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.

   (b) 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?

   (c) 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?

   (d) 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?

   (e) 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?

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?

6. **Motives for using model averaging in chemical health assessment.**
    (a) 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).
    (b) 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?

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).

    (a) 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?
    (b) 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?
    (c) 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.

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?

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? Misleading? Please elaborate.

[1] The present workshop support material and Slob and Setzer (2014) provide evidence that certain flexible 4-parameter models may perform as well as model averaging with respect to fit and coverage.

10. **Conclusions –** Do you agree with the conclusions made in Section 4.1 of the workshop support material? Please elaborate on points that you question.

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)