SUMMARY REPORT OF THE WORKSHOP TO PEER REVIEW THE BENCHMARK DOSE TECHNICAL GUIDANCE DOCUMENT

U.S. Environmental Protection Agency
Washington, D.C.
December 7–8, 2000

Prepared for:

Risk Assessment Forum
U.S. Environmental Protection Agency
Washington, D.C. 20460

Prepared by:
Eastern Research Group, Inc.
110 Hartwell Avenue
Lexington, MA 02421

March 12, 2001

NOTICE

The statements in this report reflect the views and opinions of the workshop experts. They do not represent analyses or positions of the Risk Assessment Forum or the U.S. Environmental Protection Agency (EPA).

This report was prepared by Eastern Research Group, Inc., an EPA contractor, as a general record of discussion held during the Workshop to Peer Review the Benchmark Dose Technical Guidance Document (December 7–8, 2000). As EPA requested, this report captures the main points and highlights of the meeting. It is not a complete record of all details discussed, nor does it embellish, interpret, or enlarge upon matters that were incomplete or unclear.

CONTENTS

LIST OF ACRONYMS iii

EXECUTIVE SUMMARY iv

1. INTRODUCTION 1-1
   1.1 Workshop Purpose 1-1
   1.2 Workshop Participants 1-1
   1.3 Charge to the Panel and Premeeting Comments 1-1
   1.4 Agenda 1-1
   1.5 Workshop Summary 1-2

2. SUMMARY OF OPENING PRESENTATION AND REMARKS 2-1
   2.1 Welcome and Introductions 2-1
   2.2 Background on the Benchmark Dose Technical Guidance Document 2-1
   2.3 Review of Workshop Format and Charge 2-3
   2.4 Opening Reviewer Comments 2-4

3. PREPARATION FOR COMPUTING A BENCHMARK DOSE: SELECTING DATA AND AN APPROPRIATE BENCHMARK RESPONSE LEVEL . . 3-1
   3.1 Chair's Discussion Summary 3-1
   3.2 Discussion of Question 1 3-5
   3.3 Discussion of Question 2 3-7
   3.4 Discussion of Question 3 3-7
   3.5 Observer Comment 3-15

4. MODELING TO COMPUTE A BENCHMARK DOSE: MODEL SELECTION, FITTING, AND CONFIDENCE LIMITS 4-1
   4.1 Chair's Discussion Summary 4-1
   4.2 Discussion of Question 4a 4-7
   4.3 Discussion of Question 4b 4-10
   4.4 Discussion of Question 4c 4-14
   4.5 Discussion of Question 5 4-16
   4.6 Discussion of Question 6 4-19

5. INTERPRETATION AND USE OF THE BENCHMARK DOSE 5-1
   5.1 Chair's Discussion Summary 5-1
   5.2 Discussion of Question 7 5-4
   5.3 Discussion of Question 8 5-6
   5.4 Observer Comment and Subsequent Discussion 5-9

6. FINAL WORKSHOP SUMMARY 6-1

APPENDIX A: LIST OF REVIEWERS A-1

APPENDIX B: LIST OF OBSERVERS B-1

APPENDIX C: WORKSHOP CHARGE C-1

APPENDIX D: REVIEWER PREMEETING COMMENTS D-1

APPENDIX E: WORKSHOP AGENDA E-1

APPENDIX F: OVERHEADS F-1
Overheads From Dr. Setzer's Presentation F-3
Overheads Used by Dr. Frederick F-9

APPENDIX G: ADDITIONAL REFERENCES SUGGESTED BY REVIEWERS G-1

LIST OF ACRONYMS

AIC = Akaike Information Criteria
BMD = benchmark dose
BMDL = A lower one-sided confidence limit on the BMD
BMDS = EPA's benchmark dose software
BMR = benchmark response
ED = effective dose
EPA = U.S. Environmental Protection Agency
GEE = generalized estimating equations
GOF = goodness of fit
HEC = human equivalent concentration
HED = human equivalent dose
IRIS = EPA's Integrated Risk Information System
LCL = lower bound (or lower confidence limit) on the effective dose
LED = lowest effective dose
LOAEL = lowest-observed-adverse-effect level
MOE = margin of exposure
NIEHS = National Institute of Environmental Health Sciences
NOAEL = no-observed-adverse-effect level
NTP = National Toxicology Program
OEHHA = Office of Environmental Health Hazard Assessment
PBPK = physiologically based pharmacokinetic
POD = point of departure
RfC = reference concentration
RfD = reference dose

EXECUTIVE SUMMARY
Prepared by Colin Park, Workshop Chair

On December 7–8, 2000, Eastern Research Group, Inc. (ERG), a contractor to the U.S. Environmental Protection Agency (EPA), held a workshop in Washington, D.C., to externally peer review the October 2000 draft of the "Benchmark Dose Technical Guidance Document." Reviewers consisted of 15 scientists representing academia, state government, and the private sector, whose expertise included toxicology, health risk assessment, statistics, and physiologically based pharmacokinetic modeling.

During the 1½-day workshop, reviewers responded to specific charge questions concerning computation of the benchmark dose (BMD), modeling to compute a benchmark dose, and interpretation and use of the benchmark dose. The proceedings of this workshop are summarized in this report. Key conclusions and recommendations developed by reviewers include the following.

Reviewers were very supportive of the concept of the benchmark dose as a more quantitative alternative to the no-observed-adverse-effect level (NOAEL) approach for risk assessment. They were also supportive of the Benchmark Dose Technical Support document produced by EPA, and of the related software that has been developed. Reviewers provided detailed comments on how to improve the focus and usability of the technical document, realizing that the intended audience may be broad. Reviewers spent most of their time at the workshop discussing issues for which there was a divergence of opinion. They also provided some relatively minor suggestions for modifying the document, and they made a few recommendations regarding concepts that reviewers felt EPA should not include in the document.

There was considerable discussion on the slope of the technical support document for calculating the BMD and the use of the BMD in regulatory applications. Since EPA will be producing other documents on how to apply the BMD, reviewers focused on providing suggestions for where the guidance document should make linkages to the regulatory process.

Some reviewers questioned why the document included a section on the advantages and disadvantages of the BMD as compared to the NOAEL procedure. After some discussion they recognized that this language was necessary, as a historical perspective, to support the recommended use of this procedure as a potential alternative. Reviewers suggested that this rationale should be more clearly spelled out, and that it should discuss, for example, the reasons why the Agency is recommending the newer BMD approach over the historically used NOAEL concept.

A continuing theme throughout the discussions was that the document should be more user-friendly, for example by including tables of defaults and options and step-by-step examples, and by using language and terminology consistently. Reviewers noted that EPA has collected and presented operating characteristic data in other documents, but suggested that the operating characteristics of the procedure should be summarized here. Reviewers pointed out that the target audience may be both toxicologists and statisticians.

Reviewers were concerned about the potential for model shopping, and suggested that EPA establish a hierarchy of models. Reasons for this concern included the volume of work entailed in having unlimited models, the potential for ever-decreasing BMDs and BMDLs as different models and families of models are used, and the sheer volume of the outcomes and documentation. Some reviewers felt there should be a strong hierarchy of models, such as used by the Office of Environmental Health Hazard Assessment in

California, while others felt that a less prescriptive hierarchy should be suggested. However, reviewers generally agreed that some sort of hierarchy for models should be presented so that fitting all models to all endpoints would not become the general procedure.

The document suggests that, as a guideline, all endpoints with a NOAEL within 10-fold of the lowest NOAEL should have a BMD calculated. While reviewers recognized that some screening is advisable, there was concern that 10-fold may not be sufficient considering that different endpoints may have different cross-species dosimetry. Also, the selection of the specific BMD for regulatory purposes may depend heavily upon the adversity of the endpoint. Therefore, endpoints occurring at higher administered dose levels may need to be considered, since the resulting BMD may be at a lower dose level after dosimetric conversion. Related to this was the recommendation from a dosimetry working group that the BMD modeling be done on animal doses, then converted to human dosimetry, rather than converting to human effective doses before the modeling. It was pointed out that the selection of which BMD to use among a set of BMDs is not completely within the scope of this document. For example, the question of severity of response or adverse versus adaptive responses is not within this discussion.

An issue of some discussion was the calculation and use of different benchmark response (BMR) levels in calculating the BMD. A recommendation was made that, for quantal data, the 10% response always be calculated for comparison. Different higher or lower responses may then be calculated, as appropriate, depending on the robustness of the data. If a BMR different from 10% is used in regulatory applications, that difference should be considered when selecting uncertainty factors and modifying factors.

Reviewers expressed discomfort with the decision rule that results in the selection of the lowest BMDL among a group of BMDLs that differ by more than three-fold. Reviewers felt that this situation was likely to occur frequently, particularly if many models are used. They felt that the decision rule was not generally appropriate. This concern relates back to the issue of hierarchy of models and to the issue of model shopping, as well as the question of the bounds of this document.

Reviewers suggested that EPA should consider the use of bootstrap confidence intervals, as they may be more accurate in their nominal coverage.

As expected, there was a significant divergence of opinion on whether the point of departure for regulatory purposes should be from the BMD or the BMDL. The discussion revolved around simplicity of application versus statistical accuracy. It was also mentioned that the choice of which of these two to use interacts with the uncertainty factors and/or modifying factors to be used in regulatory application.

In the discussion of model-fitting procedures and criteria, reviewers suggested that there should be more explicit recommendations on dropping results from higher doses so that models fit better at the lower doses of interest. This recommendation should be consistent with the modeling recommendations in EPA's cancer guidelines.

Reviewers were highly supportive of the use of the BMD methodology for epidemiology data, but felt that the appropriate applications, and the pitfalls of those applications, should be more fully discussed in the technical document. The BMD process appears to have significant advantages over the NOAEL concept for characterizing epidemiological data, since the NOAEL has inherent arbitrariness, and therefore biases, in the discrete clumping of exposure data. The particular discrete intervals used can influence the NOAEL, but not the BMD, when the BMD is calculated using individual exposure levels.

Reviewers made numerous references to the inherent conceptual problems in applying the methodology to continuous data. They were highly supportive of using the BMD concept for continuous data, and eventually became comfortable with the recommended procedures. Reviewers strongly preferred hybrid models over dichotomization models. They felt that "dynamic dose range" is a concept that is not sufficiently well developed to be included in the document at present.

Future research topics for application of the BMD process would include use of covariates, repeated measures, nested data, multivariate data, Bayesian approaches, and the joint likelihood of numerous BMDLs. Reviewers recommended that these subjects should be acknowledged in the document as potential future applications.

Reviewers pointed out that additional risk and extra risk, as measures of risk, are both useful, but when extra risk is calculated, the results are very often applied incorrectly to the human population. For extrapolations to humans, reviewers suggested, additional risk is more likely to be appropriately applied to existing populations.

Although it was not directly relevant to technical guidance, reviewers discussed the fact that the BMD process, when used as a supplementary approach to the traditional NOAEL/LOAEL approach, can reduce or increase use of experimental animals. Under the NOAEL/LOAEL methodology, studies without a NOAEL are often repeated to determine where the bottom end of the dose-response occurs, thus more animals are used.

The BMD methodology can, in theory, increase animal use in two ways. One is that using more animals generally results in shorter confidence intervals, which results in higher regulatory levels. The other potential for increased use can occur when there is no significant dose-response for calculating a BMD, invalidating the calculation of a BMD. But in this case the highest NOAEL could be used as a point of departure. This comment emphasizes the point that either approach can be used, depending on the available data. Reviewers felt that the animal use factor should be mentioned when discussing the advantages of the BMD process.

1. INTRODUCTION

1.1 Workshop Purpose

Since 1990, the U.S. Environmental Protection Agency's (EPA's) Risk Assessment Forum has been promoting research and discussion on benchmark dose (BMD) issues. Over the past decade, the Forum sponsored several workshops and forums on the BMD approach, and began development of the "Benchmark Dose Technical Guidance Document." This document was first externally reviewed in 1996 and subsequently revised based on reviewer comment. On December 7–8, 2000, the EPA sponsored a workshop in Washington, D.C., to externally peer review the December 2000 draft of the "Benchmark Dose Technical Guidance Document." The results of this workshop are summarized in this report. Following this review, the technical panel developing the document will further revise the document and then submit it for final Agency review.

1.2 Workshop Participants

Panelists at the workshop consisted of 15 scientists whose expertise included toxicology, health risk assessment, statistics, and physiologically based pharmacokinetic (PBPK) modeling. Reviewers represented state government, academia, and industry. A number of observers also attended the workshop. Reviewers and observers are listed in Appendices A and B, respectively.

1.3 Charge to the Panel and Premeeting Comments

Before the workshop, reviewers were sent the review document and a charge listing several specific questions to which reviewers were asked to respond. These questions covered such areas as selecting data and an appropriate response level; model selection, fitting, and confidence limits; and interpretation and use of BMD. Appendix C provides the complete charge to participants. Working individually, reviewers prepared and submitted premeeting comments, which were compiled into a booklet. The booklet was distributed to reviewers approximately a week before the workshop, and was made available to observers at the workshop. Appendix D provides the reviewers' premeeting comments.

1.4 Agenda

Reviewers convened in Washington, D.C., on December 7–8, 2000, for 1½ days of discussion on the draft BMD methodology guidance document. The workshop agenda is provided in Appendix E. The workshop began with opening remarks and reviewer introductions. An EPA scientist then made a presentation to provide the reviewers with background and context on the development of the guidance document. Next, the workshop chair reviewed the charge to reviewers and summarized the reviewers' premeeting comments. After this, three discussion sessions were held to address charge questions on preparation for computing BMD (Session 1), modeling to compute a BMD (Session 2), and interpreting and using BMD (Session 3). A brief period of observer comment was held on each day of the workshop. The workshop concluded with a review and summary of key points made during the workshop.

1.5 Workshop Summary

This report summarizes the workshop presentations and discussions and is organized as follows:

- Section 2 of this report summarizes the opening presentation and remarks. Overheads used in the presentation are provided in Appendix F.

- Sections 3, 4, and 5 summarize the discussions and observer comment in Sessions 1, 2, and 3, respectively. Appendix G includes a list of additional references suggested by reviewers.

- Section 6 summarizes the final discussion, during which reviewers reviewed and summarized key points.

2. SUMMARY OF OPENING PRESENTATION AND REMARKS

2.1 Welcome and Introductions

Kate Schalk, representing Eastern Research Group, Inc., the EPA contractor that had organized the meeting, opened the workshop and welcomed all reviewers. She introduced Colin Park as the workshop chair and turned the meeting over to him. Dr. Park welcomed participants and reminded them of the overall purpose of the workshop, which was to review EPA's October 2000 draft of the "Benchmark Technical Guidance Document." The peer reviewers introduced themselves. Dr. Park then gave the floor to Woodrow Setzer (of EPA's National Health and Environmental Effects Research Laboratory), who provided background on the document.

2.2 Background on the Benchmark Dose Technical Guidance Document

Woodrow Setzer, EPA National Health and Environmental Effects Research Laboratory

Dr. Setzer began his presentation by reviewing some of the historical activity in the development of the BMD approach. (Dr. Setzer's overheads are provided in Appendix F.) He said that Dr. Kenneth Crump introduced the topic in 1984, and there have been many technical papers on applications of BMD and other BMD topics since then. There also have been many public meetings sponsored by various organizations to discuss the BMD concept and its applications in risk assessment. In 1993, the Risk Science Institute sponsored a workshop at which the concept of BMD was discussed. Soon after that, EPA convened a technical panel to write guidance on use of the BMD in Agency risk assessments. This panel, originally chaired by Carole Kimmel of EPA, produced a draft guidance document that was discussed (though not reviewed) in a 1996 workshop. The technical panel currently consists of five members (four EPA and one former member of the U.S. Food and Drug Administration) and is chaired by Dr. Setzer.

The draft BMD guidance document that is the subject of this review was written to help EPA risk assessors and their technical (primarily statistical) support apply BMD methodology to real dose-response situations that arise in Agency practice. Thus, the document aims to achieve several goals: (1) to provide basic instruction in application of BMD methodology in routine risk assessment situations, (2) to facilitate decision-making regarding the use of BMD methodology in nonroutine risk assessment situations, and (3) to provide guidance for reviewing other applications of the BMD methodology. The document also establishes a system of consistent defaults to help foster the consistent application of BMD across EPA offices and applications. In developing the document, the authors always kept in mind that from a statistical standpoint the core of BMD application is essentially nonlinear modeling of a relatively diverse range of data sets. Dr. Setzer acknowledged that it is impossible for a guidance document to cover everything related to BMD methodology, so there may be many situations in which statistically trained professionals will be needed to assist in application of BMD methodology.

Dr. Setzer said that, because BMD and risk assessment in general are continually evolving, the guidance document could be considered a "work in eternal progress." He also emphasized it was important to remember the broader context of other risk assessment–related guidance that EPA has developed or is developing. When the BMD guidance is applied to Agency risk assessments, it will be applied in this broader context, not in isolation. Other related Agency guidance that already exists, or is under development, includes guidelines for endpoint-specific toxicities (such as the guidelines for developmental toxicity, reproductive toxicity, and neurotoxicity risk assessments); a framework for harmonizing approaches to cancer and noncancer risk assessment; a revision of the cancer guidelines; a review of the

overall reference dose (RfD) process in the Agency; an effort to develop on the selection of benchmark response (BMR) levels; and development of guidance on uncertainty factor selection in the RfD process. The BMD guidance document does not provide substantial guidance in these areas, since they are already covered by other existing or planned guidance documents.

In the context of risk assessment, the BMD has been developed as a tool for quantification of the dose-response relationship to enable comparisons across studies (in which the same chemical was used) and endpoints, and for extrapolation purposes. The original and still dominant approach for nonlinear dose-response in noncancer and some cancer assessments is called the no-observed-adverse-effect level (NOAEL). In general terms, the NOAEL is the highest dose in a study for which there is insufficient evidence to infer that the response differs from the control response. This measure has some undesirable properties. For example, the fact that a NOAEL for a study must be one of the doses used in that study makes it difficult to compare across studies. When studies use different dose levels, selection of a dose level confounds the comparison of toxicities across studies. Dr. Setzer pointed out that when there is a population threshold in the dose area being studied, then as sample size increases, the NOAELs tend to converge to the highest dose in the study that is below the threshold. With finite samples, generally an investigator can cause a NOAEL to increase simply by decreasing sample size. Thus, a less thorough study could yield larger BMDs than a better study. This bias in the NOAEL is not a good thing.

Although the NOAEL is a statistic, Dr. Setzer pointed out, it is difficult to quantify its uncertainty. Some work was done in the past to quantify this uncertainty for certain types of studies, but that work cannot be applied to the general situation. It is important to remember that NOAEL means no *observed* response, not no response. In other words, a response could exist that falls below the study's limits of detectability. Because of the way the NOAEL is constructed, there is no way to recover this information.

Dr. Crump proposed the BMD approach in 1984. This approach involves fitting the dose to the data to interpolate a dose that is expected to give a pre-specified response. Since the resulting dose is an interpolation of the doses in the study, it is not restricted to being one of the doses in the study. Thus the BMD approach enables one to circumvent the type of problem that arises with NOAEL, and at least a portion of the uncertainty of this estimate can be estimated using standard statistical methodology. Dr. Crump proposed using the estimate of that uncertainty (i.e., the lower confidence limit, or LCL) as the starting point for further dose-response assessment exercises, such as determining the RfD or reference concentration (RfC). It is important to remember that BMD is not a NOAEL or lowest-observed-adverse-effect level (LOAEL). A BMD can be calibrated so that it falls in the same general range as NOAELs might fall into for a range of study types, but NOAELS are different and must be thought about distinctly in risk assessment. The goal in assessing the BMD for a given response level is to get the best available estimate of the dose that should yield that response level on average and to quantify the uncertainty of that estimate.

Dr. Setzer reviewed the data context in which the BMD methodology is applied. One element of this context is conformity with EPA testing guidelines. These guidelines were not written from the standpoint of estimating dose-response, though they were written from the standpoint of trying to evaluate toxicity over a range of endpoints for a broadly focused purpose. The level of detail of the available data ranges from full access to individual data to only summary data in which standard deviations are reported to one significant digit, if at all. Different EPA offices have different abilities to require further studies when data are lacking. Some offices may request studies, others may not (in which case they are restricted to data they can find in the literature). The kinds of data available range from epidemiological data to data

from bioassays involving invertebrates. One of the major changes in toxicology practice that the BMD approach probably helps foster is focusing bioassays on the problem of estimating dose-response over a range of doses, rather than trying to determine a dose that is identified with a real no-effect level. Past studies' experimental designs are sometimes not optimal for estimating dose-response. In the future, we will need to rethink our experimental designs so they are more appropriate for estimating dose-response.

2.3 Review of Workshop Format and Charge

Dr. Park reviewed the format of the workshop. He said that he would begin discussions by asking each peer reviewer to make opening remarks. The discussions themselves would be organized into three areas, based on the three areas indicated in the charge to peer reviewers. Times were set aside for observer comment on both days of the workshop.

Dr. Park reminded the peer reviewers that they were at the workshop to peer review the draft document, not to tell EPA how to use it; however, it would be appropriate to comment on what steps they feel the Agency may need to take to ensure that the guidance is used properly. Dr. Park emphasized that reviewers should make sure to express their full range of opinions. It was not necessary for reviewers to agree on topics or issues. He also asked reviewers to remember that the document is intended as a tool for toxicologists and that there were a number of toxicologists in the audience—to the extent possible, reviewers should aim to make their discussions understandable to them.

Dr. Park then briefly reviewed the charge questions (see Appendix C) and summarized the reviewers' premeeting comments as follows:

- Most reviewers thought the review draft was generally a good document, and that it supported the concept of BMD.

- A primary issue raised by reviewers was the problem of how the BMD would be used in the regulatory process. For example, one needs to know how an endpoint is going to be used in order to determine the right response level. Users of the BMD methodology will need to understand what BMR was used and what the implications are of using different BMRs.

- Another issue raised by reviewers was the central estimate of the BMD versus the lower 95% confidence limit.

- Many reviewers raised the issue of different BMRs; in other words, how is a BMD level determined when there are multiple responses and multiple endpoints and therefore a distribution of BMD levels?

- Some reviewers wanted to see more guidance on default models, rather than a free-for-all on any or all models.

- Some reviewers pointed out that, even though one purpose of BMD is to encourage better experimentation, there are disincentives for better experimentation in some of the methodologies.

- Reviewers emphasized that the BMD procedure makes perfect sense for quantal applications, but is much more difficult to apply in continuous applications. They felt BMD should be used for continuous applications, and wanted more guidance on how to do this.

- One commenter pointed out that one standard deviation above the control may be a very significant clinical issue for some endpoints but not for others.

- Reviewers commented that the document should reference other documents that discuss and define the term adverse effect.

- The method for determining relative confidence intervals was brought up by a number of commenters. One of the charge questions asks whether there is sufficient documentation for deviating from the 95% confidence interval. A number of reviewers were not sure what documentation the question was referring to.

2.4 Opening Reviewer Comments

Dr. Park then asked each reviewer to make opening comments about the document. Reviewers responded as follows (each bullet represents the comments of a separate reviewer):

- A particular concern is how to combine data from different studies when using a BMD approach. When there is a good data set, a NOAEL or BMD approach can be used fairly readily. However, when data points are few and far between (as is the case for many compounds) and when the limited data comprise both human and animal data, the BMD approach can be much more challenging. The statistical validity of combining different types of data should be discussed.

- The document should clarify that the BMD methodology is a work in progress, and should indicate what elements of the methodology are still evolving and are not fully understood. For example, it may not be possible to find comparability between quantal and continuous data in terms of choice of the BMR. Also, the issue of clinical significance will be challenging. The document should be better structured to help the user clearly navigate the BMD methodology and understand what decisions must be made at various points. The document also should provide a rank ordering that shows which models EPA considers preferable, and should clarify whether an implicit or explicit dichotomization approach is preferred.

- Again, the document should clearly distinguish between which aspects of the BMD methodology are known and which are being researched or under development. The document should also provide guidance on other competing or complementary methods. For example, it should provide more information on how to handle multiple outcomes and continuous outcomes. The examples in the document should be revised to be more operationally useful. For example, they should better address the issue of adjusting for covariance and other complex modeling situations.

- The use of PBPK modeling and dosimetric adjustments, like deposited dose, is a very significant area that is not covered in the document. A treatment of this area should be added.

- The document should be made more user-friendly to better accommodate the broad range of anticipated users. Currently, advanced users must wade through much information to locate the

portions of the document that may be of interest to them. These users would be better served if the document clearly indicates to the reader what the key issues are and where they are addressed. Also, tables that summarize and compare information about the various models, including the advantages and disadvantages of each model, the key assumptions, and model fit would help the high-end user. The guide should provide better linkage between models and the text. As other reviewers have mentioned, the document should clearly indicate what is known and not known with regard to BMD methodology. Finally, there is a concern about the potential for model shopping; this needs to be addressed.

C The document currently appears to be written for statisticians, yet it is toxicologists who most often perform risk assessments, and the document and software likely will be disseminated to the toxicological community in states and the private sector. Therefore, the target audience should be defined as toxicologists interested in dose-response evaluation and the document should be modified to be more understandable to them. For example, the discussion in example 3 refers to the Hill model, yet few toxicologists have experience with this model. This discussion is written at too complex a level for toxicologists and could be handled by providing more straightforward guidance on the use of this methodology. The discussion of figures A-4.1 and A-4.2 in example 4 also would be difficult for most toxicologists to understand.

C The overall technical quality of the document is good, but changes are needed (as mentioned by previous reviewers) to improve the readability and overall user-friendliness of the document. The examples should be clearer and should be linked more intimately to the points made in the rest of the document.

C When moving toward harmonized approaches, two distinctions must be made:

— First, in risk assessment applications, one must consider whether one is trying to feed into a standardized semiregulatory process (such as calculating an RfD) or whether one is trying to calculate a point of departure (POD) for high-dose to low-dose risk projection. These are different.

— Second, distinctions are needed according to the mechanistic theory that appears to be underlying particular kinds of responses. So, for example, one must do something different in a tolerance distribution situation than one would in a situation governed by stochastic processes such as the multiple mutation processes of cancer complicated by nonlinear influences of pharmacokinetics and toxic responses. Thus, the document needs to remind the user that just because a commonality of approach is desirable, one should not choose a final model form without considering other information and the application of the risk assessment.

Another document is needed that will investigate the BMD methodology in an empirical way. One approach to this would be to take stratified random selections from EPA's Integrated Risk Information System (IRIS) and apply the BMD methodology to these data to see what problems may arise and whether some of the cases that are theoretically possible arise in actual experience.

C As others have commented, the document is a good start. It is important to carefully consider what a BMD really is. Is it defined by a process, or does it have an independent meaning—is the

process just a way to get to this meaning? If the former, BMDs are relatively easy to do: just use the process and the result is a BMD. However, that puts a lot of weight on questions like defaults, alternative models, and model shopping. It also makes it difficult to know whether revisions to the methodology produce better BMDs, since there is no standard against which to apply revised BMDs. Defining a BMD in terms of what it is intended to do, on the other hand, provides a way of judging whether a methodology is succeeding and whether an alternative methodology could do better. This type of definition is important but difficult, since statistical issues (e.g., detectability) and biological issues (e.g., what a BMD is supposed to mean in terms of the process being characterized must be considered. On another point, some elements in the document contradict other currently available EPA guidance on dose-response analysis. If this is intentional, EPA should make it clear that the new guidance in the BMD document supersedes previous guidance.

C Several points:

— First, discussion is needed on the issue of broader characterization of the dose-response data than a single number. If we are going to focus on the BMD, what is the best characterization of BMD for the purpose of comparison and for low-dose extrapolation? In his presentation, Dr. Setzer said the task was to focus on identifying the dose corresponding to a pre-specified risk level—i.e., the best estimate of the dose with a particular BMR. That sounds like a focus on BMD rather than BMDL, since we do not know what the risk level is at the BMDL but we have an idea of what it is supposed to be at the BMD.

— Second, earlier commenters focused on adversity. Adversity is an issue not only for continuous data, but also for quantal data. For continuous data, there is a lot of confusion regarding what is unusual versus what is adverse. This needs discussion.

— Third, the document defines everything in terms of *extra* risk rather than *added* risk. This needs discussion.

— Fourth, we should not be too restrictive on models in light of the evolution of mechanistic modeling, particularly with respect to categorical data as well as continuous data. Dr. Setzer talked about quantifying uncertainty in BMD estimation. This provides opportunities to use bootstrap procedures to characterize uncertainty.

— Finally, we need to focus on special features concerning modeling of epidemiological data.

C This a good document that includes some real gems, such as how to handle U-shaped dose-response curves and the concern about the loss of precision when data are moved from a continuous to a quantal form. The document was hard to follow in terms of the sequence of procedures. As others have stated, it needs to be more user-friendly, with clearer examples that are more easily followed.

C The document does a good job of outlining the BMD for binary responses, but the approach described for continuous responses is too simplistic. In practice, complications arise with continuous responses. The document needs to acknowledge and address them. Also, there are serious concerns about the examples.

- This document represents a big step forward in BMD methodology, but many issues remain open. The document should indicate what issues still need research, and it should distinguish between what is known and not known regarding BMD methodology. The document should put greater emphasis on data quality and should indicate that data quality is an open challenge. A yet unpublished analysis of the National Toxicology Program's (NTP's) developmental toxicity database shows that about 40% of NTP data do not yield any model that converges, either because the data are out of the range of interest or because there is no toxic effect. The document should better describe how continuous responses can be utilized. Essentially, this is risk assessment at a population level, not necessarily an individual level. According to the document, when a consensus criterion exists that clearly identifies what type of measurement range indicates an abnormality, this is essentially an explicit dichotomization, or an individual-level assessment. However, when such a criterion is lacking, then the change in distribution of the exposed population compared to the control is usually examined; this is essentially a population-based assessment of exposure effects. The examples provided in the document include mostly cases covered by the hybrid and implicit dichotomization, and a more general framework would be helpful; the explicit dichotomization is a special case of population-level assessment in which some shape change occurs in the distribution (e.g., mean shifting). The document seems to suggest that the Akaike Information Criteria (AIC) can always be used, when in fact this is not the case. It would be useful to discuss what statistical criteria reviewers agree with and what they do not agree with.

- Many reviewers said this document is a "good start," but we should be past a "good start": we should encourage EPA to move forward. One issue is whether to use the LCL rather than the central estimate as the POD for risk assessment. The reasons for using LCL are largely predicated on the fact that using LCL accounts for variability or lack in resolution of the study design. Much time has been spent in other meetings talking about study design—making sure it is standardized, has resolution, and so on. Hopefully, study design issues are pertinent to this discussion, in that when the study design is rigorous, a central estimate of the dose-response from those studies would be acceptable. Perhaps less-than-adequate studies could be viewed as the exception rather than the rule in risk assessment. Another issue is floating levels of response, about which there are two concerns. First, they seem to defeat the purpose of having a BMD as means of comparing data. One reason for the BMD is harmonization among endpoints in risk assessment. But how can this be done effectively unless common levels of response across endpoints are considered? It almost seems as if the concept of floating levels of response is driven by different levels of resolution of test designs. This issue should be addressed, but is not a reason to punt on the BMD. Rather, it is a reason to make sure the data generated are of sufficient resolution to use for risk assessment. Finally, the BMD has a number of advantages and adds an important tool to the toolkit. One real advantage is to make more use of the available experimental data in setting a POD for a risk assessment. Perhaps additional experience with the BMD will encourage researchers to design studies to be more flexible (and thus make more efficient use of experimental animals in establishing the dose-response curve), rather than repeat studies to come up with NOAELs. This can be an important advantage of the BMD approach.

Following these opening remarks, Dr. Park turned the floor over to the discussion leaders, who facilitated peer reviewer discussion in response to the charge questions (see charge in Appendix C). The discussion was divided into three sessions:

- Session 1 covered the questions in charge area 1: "Preparation for Computing a Benchmark Dose: Selecting Data and an Appropriate Benchmark Response Level." This session was facilitated by George Alexeeff.

- Session 2 covered the questions in charge area 2: "Modeling to Compute a Benchmark Dose: Model Selection, Fitting, and Confidence Limits." It was facilitated by Lynne Haber.

- Session 3 covered the questions in charge area 3: "Interpretation and Use of the Benchmark Dose." This session was facilitated by Lorenz Rhomberg.

The discussions during sessions 1, 2, and 3 are covered in sections 3, 4, and 5 of this report, respectively.

3. PREPARATION FOR COMPUTING A BENCHMARK DOSE: SELECTING DATA AND AN APPROPRIATE BENCHMARK RESPONSE LEVEL

Discussion Leader: George Alexeeff

The first discussion session covered "Preparation for Computing a Benchmark Dose: Selecting Data and an Appropriate Benchmark Response Level." This included the following questions, as set forth in the charge (see Appendix C):

- Question 1: What concepts and terms related to the benchmark dose (BMD), if any, do we need to define more clearly?

- Question 2: The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?

- Question 3: What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?

Section 3.1 provides a summary of this discussion, prepared by the discussion chair, Dr. Alexeeff. Sections 3.2 through 3.4 provide a detailed record of the discussions. Section 3.5 summarizes the observer comment made during this discussion.

3.1 Chair's Discussion Summary

Overall, reviewers felt that the guidance document is a good and useful guide to the resolution of statistical issues involved in calculating BMDs. The document's authors deserve a great deal of praise for putting together this much-needed technical guidance document. Most of the comments concerned fine-tuning and focusing rather than reinventing. One issue raised by reviewers was the appropriate audience for the document. The document is uneven, with a mixture of very basic information and rather technical material. This approach leaves middle-of-the-road users somewhat unaddressed. The EPA states that the target audience is Agency risk assessors and their statistical support. To be useful for risk assessment, the document should address both sets of users and should bridge the interface between the toxicologist and the statistician. In general, this can be done by adding some summary tables and by providing additional cross-referencing of other relevant documents. A table, provided as a user's "troubleshooting" guide, could help advanced users to rapidly go to specifics. Clearly specifying what points are to be illustrated by each example would help address this issue.

The document mentions related documents and programs either developed or under development. EPA staff members have indicated that the guidance document is one of a group of documents. Others include endpoint-specific guidelines, a framework for harmonizing approaches to cancer and noncancer risk assessment, a review of the RfD process, guidance on selecting uncertainty factors, and a discussion of situations in which BMD is to be used. The technical document probably needs more information and specifics about how all these materials will connect, and about the larger picture for benchmark development. The document also should specifically indicate what areas of assessment it does not discuss.

Question 1: What concepts and terms related to the benchmark dose (BMD), if any, do we need to define more clearly?

- Most reviewers indicated that the document did a good job in describing BMD methodology and defining terms in the text, such as the difference between BMD and BMDL. The suggestions below are not intended to be a roadblock to proceeding, but are given in the spirit of providing clarity.
  - Concepts and terms suggested for clarification included practical threshold, RfD/RfC, lowest effective dose (LED), critical value, and additional risk.
  - For quantal data, a subscript was recommended for the BMR (e.g., $BMDL_{10}$).
  - For continuous data, an approach needs to be identified to designate the response level by using either a subscript (e.g., $BMDL_{1D}$) or a parenthetical (e.g., BMD[1SD]).
- It was suggested that the difference between the dose-response relationship (response incidence) and the dose-effect relationship (response changes in terms of severity of appearance or additional effects produced) should be clarified. The discussion here should be consistent with the use of multiple endpoint BMD approaches.
- Several reviewers suggested that the units be defined.
  - Some comments indicated that several equations should be added, including the Michaelis-Menten model, the gamma distribution, and the Hill equation.
- One reviewer indicated that the glossary of terms, while a good idea, was not of the same quality as the text. He suggested that the terms should be defined by meaning and use.
- It was suggested that the relationship between BMDL, NOAEL, LOAEL, and POD be more clearly described and contrasted.
- One reviewer questioned whether the term BMD should be used for central estimate and the procedure.

Question 2: The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?

Reviewers indicated that, for the most part, the references are appropriate and the text is generally appropriately referenced. The literature review dealing with the BMD approach is impressive in its breadth. The suggestions below are not intended to be a roadblock to proceeding, but are meant to supplement the review of the literature in the document. Reviewers cautioned that EPA should not try to make the document a compendium of information on related topics; rather, the document should provide references for this information.

- One concern generally raised was that the review does not place enough emphasis on what is not yet known about the probable operating characteristics of the BMD approach (labeled

"Properties of the BMD" in the document) in actual practice. The limits of that understanding should be recognized.

- It was suggested that the review also provide more discussion on what other approaches exist in the literature for evaluating the data, especially continuous data.

- Reviewers suggested that the document should discuss some of the biological, dosimetric, and mechanistic considerations that could impact the modeling or modeling choices (e.g., a risk factor approach).

- A suggestion was made to include a discussion of the distributional characterization of the BMD in the document.

Suggestions for additional references are provided below in Appendix G. These suggestions were made by individual reviewers in their premeeting comments.

Question 3: What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?

- Reviewers generally agreed that the document should be made more user-friendly, especially for the two major types of users: toxicologists and statisticians.

  — Most reviewers indicated that graphical depictions, tables, examples, and descriptive figures would help clarify a number of issues discussed in the review. For example, a table could be added to summarize the main issues discussed in the text.

  — The decision flow diagram mentioned in the document could be used as an example of a way to clarify issues discussed in the review.

- Several reviewers indicated that the examples in the appendix should be clearly linked with the text. One reviewer suggested that the objectives of these examples should be clearly articulated.

  — It was suggested that an example of a data set that fails to meet the model requirements would be instructive and should be added.

- Several reviewers indicated that previous related documents should be referenced to point the reader to more introductory material.

- The document succinctly describes the value or use of the BMD approach in contrast to the NOAEL approach. Reviewers generally agreed that this description was appropriate. Various reviewers suggested the following:

  — It would be useful to provide a little more discussion regarding when one procedure is more likely to be used than the other, since much of the data—because of their poor quality—could not be modeled.

- A paragraph should describe the limitations of the data to satisfy the requirements of the BMD approach.

- There should be a very clear statement that the BMD_x or $BMDL_x$ is not to be validated by its closeness to the NOAEL or LOAEL, but is being used in this context as a common point of comparison.

- Reviewers indicated that the proposed 10-fold rule, while a useful rule of thumb, is likely beyond the scope of the document. Also, as indicated by various reviewers, comparisons depend on the scaling used, mechanistic considerations, and endpoints of concern. Without taking such issues into account, the difference may be much greater than 10-fold.

C Reviewers agreed that the BMR should be viewed as a generic neutral value that is not necessarily an adverse level, but rather is a common point of comparison.

C Reviewers generally agreed that selection of a BMR can depend on a number of factors, including, for example, the application of the BMD (e.g., comparison purposes, POD for linear extrapolation, POD for margin of exposure, POD for RfD calculation), the severity of the response, and the endpoint under consideration. For this reason, reviewers thought that the document should more explicitly distinguish the BMD_x or $BMDL_x$ and the POD.

C Reviewers agreed that, for quantal data, a default BMR of 10% appears to be a reasonable yardstick, since a default BMR's purpose is to identify a measure of dose-response at the low end of the curve that could be used for comparison across data sets for a particular chemical. Related points include:

- The default BMR should not be considered a measure of toxicity.

- Preferably, 10% should be within the range of the data.

- If extrapolation is required, there should be some indication if the extrapolation to 10% is occurring upward or downward.

- Various circumstances may suggest a BMR smaller (e.g., 1%, 5%) or greater (e.g., 20%, 50%) than 10%.

C Reviewers agreed that, for continuous data, a default BMR of one standard deviation from the mean appears to be a reasonable yardstick. Related points include:

- The default BMR should not be considered a measure of toxicity.

- The response at one standard deviation from the mean may or may not be adverse.

- The definition of the BMR depends on the application of the BMD and the endpoint.

- The absence of a mean shift does not negate the possibility of toxicity. (Reviewers were asked to submit examples of this concept if they have them. Examples may include

effects on heart rate variability or blood pressure that may not be adverse, but may be correlated with increases in myocardial infarction.)

— A reviewer suggested that to solve the "mean shift" problem, one may need to consider other endpoints in the data set or develop a quantal approach to the continuous data.

C For continuous data, reviewers generally agreed, the standard deviation for the BMR is preferably generated from the study data, as is indicated in the document. The reviewers also discussed the fact that, when the standard deviation is not available from the dose group data, a standard deviation from historical controls or from study controls may be appropriate. It was not clear if this concept was adequately explained in the document.

C Reviewers generally agreed that the document should more clearly discuss and clarify the relationship between the BMR_{10} and the BMR_{1SD}. It was suggested that a table or figure would help define the relationship.

C A reviewer recommended that the document should more clearly define the term "dichotomize," both in the text and the glossary.

C One reviewer mentioned that the document should indicate that dichotomization of continuous data may result in a loss of power of the study.

C One reviewer indicated that, for clustered data or multivariate analysis, response variability within the group may be needed in addition to individual level.

C Several reviewers suggested that requiring a positive trend test may preclude useful epidemiological data. While reviewers appeared to agree, they mentioned two important caveats: the endpoint and dose-response relationship must be validated by other studies, and the possibility for exposure misclassification should be taken into account.

C At least one reviewer indicated that more information on combining endpoints within a study or across studies would be helpful. It was suggested that other dose-response models may be appropriate. There was also some interest in the literature available on distributional analysis, but some reviewers cautioned that the methodology was still in development.

C Several reviewers indicated a need to clarify in the document that there may be changes in severity in the dose-response curve.

3.2 Discussion of Question 1

Dr. Alexeeff kicked off the discussion by summarizing the premeeting comments in response to charge question 1: *What concepts and terms related to the benchmark dose (BMD), if any, do we need to define more clearly?*

C Most reviewers indicated that the BMD Technical Guidance Document did a good job in describing BMD methodology and in defining terms, such as the difference between BMD and BMDL.

- However, a number of concepts and terms were suggested for clarification, such as POD, practical threshold, multinomial modeling, RfD/RfC, LED, critical value, bootstrap, extra risk, additional risk.

- A subscript was recommended for the BMR (e.g., $BMDL_{10}$).

- It was suggested that the difference between the dose-response relationship and the dose-effect relationship be clarified.

- Several reviewers suggested that the units be defined.

- Some comments indicated that several equations should be added (e.g., for the Michaelis-Menten model, the gamma distribution, and the Hill equation).

- One reviewer indicated that the glossary of terms, while a good idea, was not at the same quality level as the text. Also the terms should be defined by meaning and use.

- It was suggested that the relationship between BMDL, NOAEL, LOAEL, and POD be more clearly described and contrasted.

- One reviewer questioned whether the term "BMD" should be used for the central estimate and the procedure.

Reviewers then began their discussion. The first discussion topic concerned the confusion that can arise regarding terminology. Reviewers agreed there can be confusion, particularly for continuous endpoints. They discussed several options for obviating confusion. One reviewer suggested using some system that would let reviewers readily see when a BMD is for a specified effect level and when it is for a nonspecified effect level. For example, a suffix could be used to indicate a BMD that is at a specific level but nonspecified; no suffix could be used to indicate a BMD or BMR that is being referred to in a general context. Another reviewer disagreed, since a numerical suffix can have different meanings and therefore can generate more confusion than it remedies. For example, 10 means something different for continuous versus quantal endpoints. Also, does 05 mean 5% of the mean, a 5% change in dynamic range, or a 5% change in population? Another source of confusion is whether a BMD applies to a particular data set or to a chemical as a whole (considering multiple data sets). The first reviewer responded that perhaps the reviewers could simply recommend using some clearly defined consistent designation to indicate how the BMD was being used. Three reviewers voiced support for using some form of subscript. One of them suggested using a subscript like "c" for a continuous BMD, or a parenthetical statement to explain that it is a continuous variable and what it means. Another felt that a subscript for the BMR percentage would be useful and that, for continuous endpoints, some sort of definition would be useful (such as a parenthesis based on one standard deviation); this reviewer did, however, doubt that a subscript could capture all the multivariants of the continuous descriptors. Two reviewers suggested the document could be improved by omitting reference to the term "practical threshold," since it is confusing. A reviewer mentioned that several definitions in the glossary were not correct.

The group then discussed dose-response versus dose-effect relationships. Dr. Alexeeff explained that "dose-effect relationship" refers to the difference in magnitude, severity, or nature of effects that may occur as dose increases, while "dose-response" describes how the incidence of a particular effect

increases as dose increases. A reviewer emphasized that this distinction is important, though not yet standard, and suggested the document should not only recognize this distinction, but advocate it. Another reviewer pointed out that this distinction is important when one is combining endpoints, since a different approach must be used to combine different endpoints as opposed to the same endpoints happening at increased incidence. The discussion about dose-effect relationships should therefore be consistent with the use of multiple endpoint BMD approaches.

3.3 Discussion of Question 2

Dr. Alexeeff summarized the reviewers' premeeting comments in response to charge question 2: *The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?*

- Generally, the comments indicated that the literature review was very thorough and fairly complete.

- A number of additional references were suggested to supplement those in the document and to provide additional examples of concepts discussed.

- It was suggested that the review also provide more discussion on what is not yet known about the operating characteristics of the BMD approach.

- The document should discuss some of the biological, dosimetric, and mechanistic considerations that could affect modeling or modeling choices.

- One reviewer suggested the document should discuss the distributional characterization of BMD.

Reviewers briefly discussed additional areas they felt the document should cover. One reviewer suggested that the document should better address modeling, including the fit of the models and comparisons between models. Another reviewer suggested that the document should mention and provide references for other approaches that exist in the literature, especially approaches for continuous outcomes. A third reviewer mentioned that the risk factor approach is one example of a different approach that could be referenced in the document.

3.4 Discussion of Question 3

Dr. Alexeeff summarized the reviewers' premeeting comments in response to charge question 3: *What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?*

- Graphical depictions, examples, and descriptive figures would help clarify a number of issues discussed in the review.

- The value or use of the BMD approach in contrast to the NOAEL approach appears overstated or defensive. It would be useful to provide more discussion regarding when one procedure is more likely to be used than the other.

- The derivation and application of the 10-fold rule needs to be clarified.

- The document should provide additional discussion of circumstances that may suggest a BMR smaller or greater than 10%.

- Clarification is needed regarding changes in severity in the dose-response curve.

- The use of a standard deviation from historical controls instead of the study controls, or the control standard deviation for the dosed standard deviation, requires more discussion in the document.

- Individual-level data requirements for clustered data or multivariate analysis may be greater than response variability per dose group.

- Requiring a positive trend test may preclude useful epidemiological data.

- "Ideal" study descriptions may not be practical or ultimately useful.

- The selection process may identify an excessive number of data sets or no practical data sets.

- How does adversity of the effect fit in?

- The definition of the BMR for quantal data may be problematic. The document lacks a definition of a BMR for continuous data.

- More information on combining endpoints within a study or across studies would be helpful.

- One reviewer commented that distributional modeling may provide additional insight.

Reviewers began their discussion by suggesting a number of ways in which EPA could make the document more user-friendly. These included:

- Adding clarifying examples in the text.

- Providing references in the text wherever appropriate to clearly direct the reader to the clarifying examples that are provided at the end of the document.

- Providing a list of objectives for each example to make it clear what EPA intends the example to illustrate.

- Providing tables that clearly list the assumptions, fit, and problems associated with each model.

- Including references to existing documents to which users can refer for additional explanation and understanding of topics and concepts discussed in the document.

Reviewers talked briefly about the value of the BMD approach versus the NOAEL approach. They agreed that EPA needs to explain why they advocate the BMD approach, but should be careful to avoid giving the impression that the NOAEL approach should no longer be used. Several reviewers pointed out that there are many situations in which the available data do not enable a BMD approach, so the NOAEL will remain an essential and valid tool.

A reviewer said he felt there was a philosophical shift in this document as compared to previous BMD documents, in that EPA now appears to be saying that the BMD is a new and better endpoint than the NOAEL, not an attempt to mimic the NOAEL. Another reviewer agreed, and felt it important that readers of the document not think that the BMD and NOAEL need to be matched or that a justification of the BMD procedure is that it gives something comparable to the NOAEL.

Reviewers discussed the fact that the BMD approach is limited when the data are inadequate. Some reviewers suggested discussing the relationship of the BMD to the NOAEL and clarifying when one approach might be more suitable than the other. They pointed out that, while the BMD approach may be limited by the quality and quantity of available data, this is not a limitation of the approach itself, but rather a reflection of data limitations. Two reviewers suggested that it would be useful for the document to include examples of data sets that are not sufficiently robust for the BMD approach. Another reviewer pointed out that the NOAEL approach is also limited by poor data. He recommended that this point be made explicitly in the document. He suggested that whichever approach is used, the assessment should include language indicating that a poor data set was used when such is the case. Another reviewer suggested that the document should clearly explain the concept of degrees of freedom so that users can understand that concept's importance and restrict the range of models used to those that are appropriate based on the degrees of freedom. For example, when there are only two data points, it is unwise to apply a model with three free parameters.

Reviewers then discussed the 10-fold rule. Dr. Alexeeff mentioned that, in the premeeting comments, reviewers expressed confusion about where this rule came from, how this rule should be applied within a study and between studies, how to normalize the comparisons, and how accurate one needs to be when doing the comparisons. A reviewer pointed out the need to distinguish between safety assessments, in which the goal is to determine a safe exposure level, and risk assessments. For safety assessments for chemicals producing different endpoints by different modes of action, it is reasonable to focus on the most sensitive endpoint: protecting against the most sensitive effect presumably will also protect against the other effects. So, in these cases, the 10-fold rule may well be appropriate. However, for quantitative risk assessments, separate effects should be carried through separately as part of an overall analysis of the consequences of different population exposures to the chemical. Several other reviewers agreed with this point. A number of reviewers questioned whether the 10-fold rule should be covered in the document, since the focus of the document is how to do BMD analysis on data sets, and not on how to apply the BMD as calculated. Based on these considerations, some reviewers retracted the support for the 10-fold approach they had indicated in their premeeting comments. One reviewer pointed out that every endpoint could have different cross-species dosimetry, so even an effect level that is 100-fold higher could represent the critical effect when cross-species scaling is taken into account. Another reviewer agreed and suggested the document should indicate that effects that appear to be occurring at higher doses should be considered when one is determining a BMD. Another issue, a reviewer responded, is how to handle effects that increase in severity with increasing dose (e.g., from slight irritation, to severe irritation, to ulceration). Should these effects be bundled, or should the effect at the lowest dose be selected? Another reviewer agreed with the previous comments and cautioned against overemphasizing the most sensitive study and endpoints. He stressed the need to run multiple endpoints and to explicitly present the results, including the results from running different models, different parameterizations of the models, and different assumptions about the data. One reviewer suggested that the document could simply mention that 10-fold has historically been used as a rule of thumb, but it does not adequately deal with differences in dosimetry. A reviewer pointed out that one must consider dosimetry issues before conducting modeling in order to determine which are truly the most sensitive endpoints. Another reviewer said that even when

there is a focus on a particular model and endpoint, a BMD should be calculated for each study that provides dose-response data for that endpoint. On a final note for this portion of the discussion, a reviewer pointed out that the dosimetry issue exists for all risk assessment approaches, not just the BMD approach.

Reviewers next held a long discussion of issues concerning BMR. A reviewer began the discussion by pointing out that the selection of a BMR is a function of the intended application of the BMD modeling. One possible application is use of the BMD or BMDL as the POD for linear extrapolation, which will result in roughly the same risk estimate whether the BMR is set at 1%, 5%, or 10%. In this case, the BMR should be selected to obtain a BMD near the lower end of the experimental data. Another application is the use of the BMD or BMDL as the POD for the calculating a margin of exposure (MOE) under the new cancer guidelines, in which case the BMR needs to be relatively consistent across chemicals to avoid confusing risk managers regarding the level of protection provided. The new cancer guidelines indicate that the BMDL, rather than the BMD, should be used for both these applications, which is consistent with the past practice of using the upper bound on the cancer risk estimate. They also suggest the use of a 10% BMR, although a 1% BMR may be preferable for epidemiological studies. A third application is the use of the BMD or BMDL as the POD in noncancer risk assessment. This application is much more problematic, because the selection of the BMR is a function of the nature of the endpoint being modeled and interacts with the selection of uncertainty factors. The fourth application is the use of the BMD or BMDL as the basis for comparing toxicity across chemicals. In this case, consistency in the BMD approach across chemicals is particularly crucial, and comparisons should probably be made on the basis of the BMDs rather than the BMDLs. This last application is the one that is discussed most frequently in the draft guidance, which suggests the use of a 10% BMR for quantal data and a one standard deviation change in the mean for continuous data. At any rate, it is important for the guidance to distinguish these different applications when discussing the selection of BMRs and the use of BMDs versus BMDLs.

Dr. Alexeeff reminded reviewers that there also is the issue of continuous versus quantal approaches for BMR. He said that, in their premeeting comments, reviewers seemed to indicate general but not unanimous support for the idea that, for quantal data, a 10% response rate is a reasonable default to report on a regular basis. A reviewer pointed out that in a quantal approach, lower PODs can be estimated with some data sets. If different response levels are used, this must be taken into account when setting uncertainty factors. Also, the use of different response levels can be a disincentive to good experimentation. Another reviewer agreed that the guidance should encourage users to calculate a lower BMR if the data allow them to do that. He was concerned that setting a level of 10% might drive researchers to focus on the 10% response level, even when they might be able to get results at lower levels. When comparing data sets for a chemical, said another reviewer, it probably is best to use the same type of BMR. A reviewer suggested including an explicit statement that selection of the BMR should depend on the application. She cautioned against being overly prescriptive, since the guidance should be relevant not only to current uses of the BMD, but also to future applications.

Dr. Alexeeff asked reviewers to comment on whether they thought that 10% response was a reasonable default level to use as a yardstick to compare response rates across data sets for a chemical. A reviewer said he thought that 10% for quantal animal data is a good yardstick for most but not all endpoints. Another reviewer emphasized that the guidelines made it clear that they were not meant to cover selection of BMRs, which is an extremely complicated issue. This reviewer said the guidelines should indicate that selection of the BMR depends on the application. However, they could suggest calculating a benchmark simply as a point of comparison—while making it clear that this point of comparison is not intended as a measure of toxicity or critical effect level. This benchmark should be something that can be

calculated for most data sets, such as a 10% response for quantal data and one standard deviation for continuous data. A reviewer agreed with this idea, and also suggested including an ED_{50} when it is within the data range because, as the median of the animal data, the ED_{50} does not implicitly incorporate animal interindividual variability. The animal ED_{50} could be used to derive the human ED_{50}, and then human data on interindividual variability could be used to project risks at low doses. A reviewer recommended that if a 10% default is used, it should be made clear for each application whether it was within the range of the data or was calculated based on an extrapolation (which would be the case for epidemiologic data and some animal data). A reviewer said the document's introduction should include a statement making it clear that different applications have very different implications for the use of BMR.

Dr. Alexeeff then asked reviewers to discuss the issue of the benchmark for continuous data. A reviewer responded that there is a published method (in a 1995 paper authored by Dr. Crump) of comparing continous and quantal data, which suggests that one standard deviation is good point of comparison for most studies. He thought that one standard deviation is suitably valid: it is consistent with the 10% level for quantal data and, in any case, the BMR is just being used as a point of comparison. Another reviewer disagreed, saying that one standard deviation shift in a mean is not an appropriate measure for adversity. The first reviewer responded that the point of comparison is based on the data and is not associated with adversity, so adversity should not be a consideration in the discussion. The issue of adversity is handled during risk assessment. A third reviewer suggested the document could simply indicate that the definition of the BMR for continuous data depends on the application and the endpoint being evaluated, and that as a general rule, one standard deviation may serve as a good comparison point. A fourth reviewer expressed concern that a number defined as a reference number for activity could by default become a definition for adversity. A reviewer pointed out that for a number of endpoints (such as fetal weight) for which the degree of adversity has not been determined, this is currently the de facto situation. If these types of endpoints can be detected statistically, they are often considered critical endpoints that can be used in a risk assessment, even when there is no sound biological basis for doing so. Another reviewer expressed concern that use of just one number might imply greater importance for that number than was intended. For this reason, he advocated using at least two standard numbers to alert users there is a choice to be made.

A reviewer pointed out that the term "benchmark" is being used in different ways. The document provides guidance on fitting dose-response curves, characterizing dose-response in the observable range, and then pulling a particular point off the dose-response curve somewhere above the bottom of the observable range. The same general methodology can be used to estimate different points that are used for different applications, including risk assessment. The word "benchmark" was originally chosen to mean a point that is comparable across cases and that does not have any implications regarding adversity. The reviewer suggested that the term "BMD" be used to mean only the point of comparison, and that the document indicate that (1) BMD is not the same thing as the POD for risk assessment and other applications and (2) BMD is not a point of adversity comparison. Then, added another reviewer, PODs can be calculated at other response levels as well, depending on the application and what the data allow. A number of reviewers agreed with the suggestion that different terms be used to distinguish these meanings and added that the document should not provide guidance on selecting the POD. One reviewer disagreed, saying the BMD should be not only a general point of comparison, but also a point of comparison across endpoints to ensure that when an RfD is calculated, it will protect against each of the potential toxicities—in that sense, the BMD should be the POD. Reviewers discussed possible terms for the points selected for application, including effective dose and lowest effective dose (LED). A reviewer also suggested using a suffix on BMD to indicate the POD (e.g., BMD_5), but other reviewers were concerned this would not clearly separate the BMD from adversity. Another reviewer mentioned that

when the goal is to compare lots of chemicals, the central value and the confidence limits are of major interest.

At this point, reviewers suspended their discussion for a few minutes to listen to an observer comment (see Section 3.5). When they resumed, they focused on the issue of whether it is appropriate to use a standard deviation from historical controls instead of the study controls. The guidance suggests that one can use the standard deviation determined for historical controls as a substitute for the measure of spread in the dose groups, said one reviewer. Further, it appears to suggest that one might even be able to use the standard deviation exhibited by historical controls as a substitute. Yet it is not clear whether the historical control's standard deviation will be a valid substitute. Even less clear is the validity of using the control standard deviation (whether it be determined from historical or current controls) as a substitute for the standard deviation in dosed groups. The variance (indicated by a standard deviation) is quite conceivably a function of the biological distress invoked by the dose, and therefore quite possibly dose related. A strong preference should be stated for using the dosed group standard deviation for dosed groups. Where impossible, the next best option is most generally to use the standard deviation obtained from current controls (in preference to historical controls). However, in cases where the current control group is suspect (atypical), it may be appropriate to use the historical standard deviation.

Dr. Alexeeff asked the group whether they felt it would be appropriate to use a standard deviation from historical controls to estimate the benchmark when the data are continuous. A reviewer responded that data interpretation can be informed by use of the historical range (e.g., to determine what is "normal"). This has traditionally been done for clinical chemistry endpoints, but in theory could be done for other neurological endpoints. A reviewer suggested that any uncertainty regarding the standard deviation should be reflected in the confidence interval. Another reviewer agreed, emphasizing that one should take the uncertainty into account when computing BMRs for continuous data. He said there is a growing body of literature on the use of historical controls in dose-response analysis, and suggested citing that literature in the document.

A reviewer said he was troubled by the suggestion in the guidance document that one should always use the mean response observed from the study and not defer to the historical control mean. This may not be appropriate when the mean observed in the bioassay is very atypical. In such a case, it may be appropriate to defer to the historical controls.

Dr. Setzer clarified that the EPA's benchmark software (BMDS) is designed to do a maximum likelihood estimate when a standard deviation exists for the study group. EPA did not intend the user to simply plug a standard deviation from a historical control study into the standard deviation column of the data set. However, it should be possible to use historical control data for dose-response analysis if this is done carefully, with proper consideration of the issues surrounding the use of this type of data. Dr. Setzer said that the EPA authors were trying to allude to that in the document, but they did not intend to tell readers how to use a control standard deviation. Rather, the authors intended to address situations in which no standard deviation is reported for a data set, so the choice is to throw out the data set or to do something with historical control data. EPA merely wanted to indicate that use of historical control data was a possibility in these situations. Given the confusion about this intent, EPA will need to make the document clearer. Dr. Setzer also said that the methodology in the BMDS software and the guidance document includes a way to recognize the uncertainty in the standard deviation.

A reviewer pointed out that use of historical data is discussed in the Agency's endpoint-specific risk assessment guidance. The BMD document should reference and be consistent with the existing Agency

guidance on this topic. Another reviewer supported the idea that the first preference is to use the standard deviations and means for the data set, since they will be internally consistent with the data. The standard deviation can have a significant impact on the outcome, so use of a standard deviation from historical controls can affect the result. Another reviewer pointed out that the inputs to the BMDS software reflect the dose levels used, the number of subjects, the mean, and the standard deviation, so use of a standard deviation (such as a standard deviation calculated from historical data) that is not aligned with the number of subjects would violate the mathematical assumptions built into the software. A historical standard deviation could be used to identify the critical level or level of interest, said another reviewer, but it would not be part of a maximum likelihood estimation.

Dr. Alexeeff then mentioned the point, made in the premeeting comments, that the individual-level data requirements for clustered data or multivariate analysis may be greater than the response variability per dose group. A reviewer clarified this point: the document specifies that dose-level data are required for dose-level analysis. However, for multivariate or clustered data, a finer level of reporting may be needed than aggregate dose summaries. This would apply to any data for which one wants to do a rigorous multivariate analysis, including clustered data in teratology. Another reviewer pointed out that individual data are also important for time-to-response analyses.

Reviewers then discussed the comment that requiring a positive trend test may preclude useful epidemiological data. In their premeeting comments, some reviewers indicated that a benchmark analysis for epidemiological data may sometimes be useful, even when there is no positive trend test. A reviewer mentioned that a benchmark analysis had been done for both manganese and methyl mercury because there were equivalent studies demonstrating an effect. Since the cause-effect relationship had been established for the endpoints being measured, it was deemed reasonable to perform this analysis. The reviewer emphasized that it would be inappropriate to try to estimate a NOAEL when other studies had not amply demonstrated the dose-response causality. This an important caveat, agreed another reviewer. A third reviewer pointed out that, with epidemiological data, there often is much greater uncertainty about the actual doses received than in experimental studies. Therefore, the failure to find a trend test is less of a point against the data for epidemiological studies than it would be for animal studies. This means that the possibility for exposure misclassification should be taken into account when one conducts a benchmark analysis for epidemiological data without a trend test, added another reviewer.

Dr. Alexeeff then moved on to the premeeting comment that “ideal” study descriptions may not be practical or ultimately useful. He said that, in their premeeting comments, reviewers were concerned about the implication that studies could be ideal and that more experimentation should be done, when in fact simpler measures such as specialized dosing may be sufficient. This was largely a point of clarification, he explained, and suggested that EPA refer to the premeeting comments for more detail.

Dr. Alexeeff then brought up the next three issues he had listed when reviewing the premeeting comments to kick off the discussion. He said that the issue of selection process involving an excessive number of data sets was no longer an issue, since reviewers had determined that the 10-fold rule was beyond the scope of the document. The issues of adversity of effect and the definition of BMRs for quantal and continuous data had been discussed earlier. He suggested that the group move on to discuss the second-to-last issue—that more information on combining endpoints within a study or across studies would be helpful. A reviewer pointed out that a number of premeeting comments made suggestions about dose-response models for fitting multiple data sets, for example, where there might be a common slope

across data sets that have variable intercepts or some other commonality of the models that can be fit across multiple data sets. He wanted to make sure these ideas were captured.

Recalling the earlier discussion on adversity, another reviewer mentioned that for continuous endpoints there are a number of ways of describing a BMR. For example, one way mentioned in an example in the document was a percentage of the so-called dynamic range of the variable. The reviewer wanted to make sure that some time was spent discussing the different ways of describing a BMR for continuous data, none of which are compatible with quantal data. Dr. Park suggested that this topic would be appropriate as part of the discussion of charge question 2.

The reviewers then went on to discuss issues related to distributional modeling. A reviewer pointed out that the document steers the reader toward focusing on a particular data set. However, the distributional modeling currently available enables one to look at all studies that relate to an endpoint, and to perform an analysis that encompasses the universe of relevant data sets. Uncertainty in dose-response used to be strictly characterized by an LCL computed on the minimum data set, the reviewer said, because distributions were simply too difficult. Now, however, distributional characterizations of dose, uncertainty factors, and so on are more routinely possible, so it is reasonable to incorporate uncertainty or variability when characterizing dose-response. Distributional modeling could be used to characterize the variability and uncertainty of a specific endpoint from study to study or model to model, rather than simply the variability within a particular study and a particular set of assumptions. Another reviewer agreed it is not necessary to focus on a single data set, and suggested that the document discuss hierarchical Bayesian approaches that can be used for a more comprehensive analysis. A third reviewer generally agreed with the potential value of a more comprehensive analysis, but suggested that the document simply indicate that distributional modeling can be used to combine endpoints and studies, and then should provide appropriate references. She cautioned against advocating any particular modeling approaches or being overly prescriptive, since many of these approaches are under development and there are many choices. The first reviewer agreed, saying he simply wanted the document to recognize the distributional approach.

A reviewer then returned to an earlier topic: using one standard deviation from the mean as a reasonable default BMR. He agreed it was appropriate to use one standard deviation as the comparison point for continuous data, but only when the toxic effects represent a mean shift. He suggested that the document point out the limitation that, for some data sets, the BMR indicated by a one standard deviation shift of the mean may not be defined, but that does not mean there are no toxic effects. If one standard deviation is chosen as the default comparison method, the document should point out the limitation that toxic effects may manifest differently than mean shifting. The reviewer suggested the document (1) allow for the possibility that the BMR may need to be reported in some other way, and (2) encourage people to look at different manifestations of toxic effects. However, since many other ways are possible, he did not think the document needs to prescribe what the alternatives might be. Another reviewer agreed that the BMR cannot be calculated when something causes a dispersion of the outcome rather than a shifting of the mean, and suggested the guidelines should mention this. One solution would be to redefine the effect, said another reviewer, so the parameter being studied is the variance. In other words, if the thesis is that the chemical has damaged a control system, then some function of the homeostatic control can be defined as the object of study.

Dr. Setzer asked reviewers if they had examples of such data sets. One reviewer responded that in an aging study he currently is working on, the mean shifting is not significant, but the population being studied includes one group who have resisted memory loss on the average, but whose memory tests indicated a

significant instability (larger variation). Another reviewer mentioned the recent finding that the standard deviation of heart rates in the human population is a good predictor of cardiovascular mortality, in that people with lower variability in heart rates have more risk of heart attack (probably because they have lost some of their reserve capacity). In this case, the pathological case is lower variability rather than higher variability.[1] A reviewer suggested that the document should mention the example of using change in variability as an alternative to change in the mean. Another reviewer mentioned the example of lead, for which where there is a correlation between blood pressure shifts and myocardial infarction. The changes in myocardial infarction are too small to measure directly.

Dave Gaylor (currently with Sciences International, Inc.), who was involved in the development of the document, clarified the origin of the use of one standard deviation in the document. He said the goal was to have the same risk level associated with continuous and quantal data. With a normal distribution, if the 98.5th percentile is defined as "abnormal" compared to the rest of the population, then a one standard deviation shift in the mean shifts the tail of the distribution by about 10%. A few reviewers pointed out that other percentiles give significantly different BMRs. For example, a 95th percentile will give you a BMR of 22%, so a very small change in the percentile can cause a big difference in excess risk. Some reviewers suggested it might useful to have a figure in the document illustrating this point.

A reviewer pointed out that in some situations, both the lower and upper percentile can be regarded as representing adverse effects. The document can acknowledge this by pointing out that one standard deviation from the mean implies a one-sided endpoint, and if someone wishes to choose a BMR based on a two-ended adverse effect, that should be correspondingly adjusted. Another reviewer pointed out that sanctioning the division of a continuous distribution into arbitrary dichotomous categories of "abnormal" and "normal" is an artificial intellectual exercise. The document should acknowledge that dichotomization of continuous data may result in a loss of power of the study, said another reviewer.

At this point, reviewers agreed they had sufficiently discussed the first charge question. They moved on to discussing charge question 4 (see Section 4).

3.5 Observer Comment

During the discussion of question 1, the meeting was opened for observer comment. One observer commented.

Elizabeth Margosches, EPA: Recently, Dr. Margosches has been helping with acute oral toxicity designs. Up to now, she said, these studies have essentially looked at deaths of animals and tried to use that information for classification or, in some cases, for risk assessment. Until recently, on an international level, these studies were largely standard multidose studies with a fixed number of animals at several doses. There is a general move to redesign these studies, and their redesign will be different depending on whether the goal is to use the data for safety assessment or risk characterization. The

---

[1]H.V. Huikuri, T. Makikallio, J. Airaksinen, R. Mitrani, A. Castellanos, and R.J. Myerburg. 1999. Measurement of heart rate variability: A clinical tool or a research toy? J Am Coll Cardiol 34:1878-1883.

H.V. Huikuri, T.H. Makikallio, K.E.J. Airaksinen, T. Seppanen, P. Puukka, I.J. Haiha, and L.B. Sourander. 1998. Power-law relationship of heart rate variability as a predictor of mortality in the elderly. Circulation 97:2031-2036.

BMD methodology is a way to try to characterize the data one has. The guidance document tries to get across how well a data set is understood, given that the Agency will be using the data in a particular way when it is doing a risk assessment to find a regulatory level.

The section on reporting requirements toward the end of the document, said Dr. Margosches, may get at some of the things the reviewers have been discussing regarding the use of data sets in comparisons. No study design can give an ED_{50} and an estimate of variability with just a few animals. However, confidence intervals can be useful to characterize what the available data do encompass. This is one reason why one does not take an ED without an LED, or a BMD without a BMDL. In considering changes to the document, said Dr. Margosches, the reviewers may want to consider whether EPA could use language from one section of the document in other sections, rather than simply adding new language. Also, the reviewers may want to consider whether any new language they recommend could be taken from their premeeting comments.

4. MODELING TO COMPUTE A BENCHMARK DOSE: MODEL SELECTION, FITTING, AND CONFIDENCE LIMITS

Discussion Leader: Lynne Haber

The second discussion session covered "Modeling to Compute a Benchmark Dose: Model Selection, Fitting, and Confidence Limits." This included the following questions, as set forth in the charge (see Appendix C):

- Question 4a: *What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?*
- Question 4b: *Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?*
- Question 4c: *What are the advantages/strengths of using the methods described to select among "equally" fitting models? What other methods should be considered in making a selection?*
- Question 5: *Please comment on the approaches described to compute confidence limits.*
- Question 6: *What additional concepts, if any, should be illustrated by an example?*

Section 4.1 provides a summary of this discussion, prepared by the discussion chair, Dr. Haber. Sections 4.2 through 4.6 provide a detailed record of the discussions.

4.1 Chair's Discussion Summary

Question 4a: What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?

Reviewers felt it would be very useful to state several things more clearly: what the default parameters are for each model (this could be stated in tabular form), the assumptions inherent in each model choice, and the appropriate applications for each specific model.

Several reviewers saw a need for more strongly wording the guidance to constrain power parameters to be no less than unity, although some reviewers did not consider such constraints appropriate and felt that use of unconstrained power parameters would be beneficial in certain cases. To some extent, the recommendation to constrain the power parameters was in response to the example in the document that started out with an unconstrained power parameter (i.e., it is not clear until the reader is well into the example that this is an example of what not to do).

Some reviewers asked that the guidance not recommend limiting parameters to non-negative values. Allowing negative values can be useful in fitting U-shaped or inverse-U dose-response curves.

Several reviewers recommended that biological plausibility be a consideration in the choice of models. Although there was not overt disagreement on this point, there was some disagreement on the degree of preference and the degree of caveats that should be expressed. Some reviewers noted that there may be empirical information on the shape of the underlying distribution, which can inform the model choice.

Others considered biological models to be the preferred approach, particularly as better biology allows one to go further down the dose-response curve. However, this latter group felt that the models are generally too close to distinguish on a biological basis, and that models with an apparently biological basis (e.g., the Hill model) are still just curve-fitting in most applications. Concern was also expressed that it is harder to determine confidence limits for biological models (e.g., the Michaelis-Menten model). It was also noted that the models are not used to extrapolate much below the data, removing some of the problems of model-dependency seen with other applications.

Reviewers noted that the threshold parameter could have multiple interpretations, and there could be a time "threshold," as in a time-to-tumor model.

There was a general consensus that there should be a preferred hierarchy of models, rather than a recommendation to run every model for every situation. Reviewers did, however, express a range of opinions on the recommended strength of that hierarchy. Some reviewers use one preferred model for consistency, and then may compare others to the "standard" model to get some indication of model dependence (e.g., as is done by the California EPA's Office of Environmental Health Hazard Assessment, or OEHHA). Others wanted a small number of models in the initial phase of the hierarchy, to aid in considering model-dependence. Under this approach, the modeler would consider more complex models only if there were modeling problems that could not be addressed by the initial suite.

A few reviewers asked for guidance on how to handle the control group if one is log-transforming data. One approach mentioned was to add a constant to the data. Another approach was not to do log transforms, especially in light of concerns about distorting the dose-response curve. It was noted that a logarithmic *exposure* distribution is not a reason for doing a log-transformation, although a log-transform could be useful for certain applications and/or data sets. Another alternative mentioned was to weight the data points.

Reviewers asked that the discussion of covariates clearly state that there are methodological issues regarding how to include covariates in the model, and that this is an area in which research is needed (both from a methodological viewpoint and in terms of what should be used as a covariate). This issue can also be considered from a mechanistic viewpoint. Reviewers were asked to submit relevant references if they have them. A recent National Institute of Environmental Health Sciences (NIEHS) meeting on endocrine disruptors discussed pitfalls in this area, and could be cited. Reviewers also requested guidance on how to compute BMDs using nonlinear models that include covariates, and on how to interpret these BMDs. It was noted that some of the available commercial benchmark dose software can handle covariates.

One reviewer asked for guidance on how to select the threshold parameter, if one is used. A recent publication cited in the guidance (Haber et al., 1998) provides an example in which the use of the threshold term markedly improved the fit when there were several exposure levels with no response above background. Any discussion of the use of the threshold parameter should make it clear that this is a parameter determined by model fitting, and that there is no associated implication of a biological threshold at that level. Concern was also expressed that a threshold may be apparent visually in log-transformed data, but that if one looks at that data on a linear scale, there is no evidence for a threshold.

One reviewer addressed the use of dosimetric information in benchmark modeling. As noted by this reviewer (and as others on the panel agreed), it is generally considered preferable to conduct quantitative risk assessments by relating toxic responses to measures of internal exposure (e.g., target-tissue dose

metrics) rather than to administered dose or exposure concentration (Andersen et al., 1995).[2] Dosimetric approaches for estimating internal exposure range from the calculation of regional deposited dose for inhaled particulate (USEPA, 1994) to the prediction of target-tissue concentration profiles with physiologically based pharmacokinetic (PBPK) models (Clewell and Andersen, 1985). In the case where some form of dosimetric calculation can be performed, the resulting dose metrics for each individual or exposure group are used in the dose-response model in place of the corresponding administered doses or exposure concentrations (Clewell et al., 2001). When this is done, the results of the dose-response modeling will also be in units of the dose metric, rather than in units of an administered dose or exposure concentration. To convert the resulting benchmark dose metric to the equivalent benchmark exposure concentration or administered dose, the appropriate dosimetric conversion must be performed. In the case of a PBPK model, the model must be run repeatedly, varying the exposure concentration or administered dose until the benchmark dose metric value is obtained. To obtain the human equivalent concentration or dose (HEC or HED), this conversion is done using the dosimetric parameters for humans. Alternatively, the dosimetric parameters for the test species can be used to obtain the results of the benchmark modeling in terms of the animal exposure conditions (e.g., for comparisons with NOAELs in other studies). In the past, the HECs or HEDs for each animal exposure group have sometimes been calculated and used in the dose-response modeling. This practice is acceptable only if the same dosimetric adjustment is appropriate for all exposure groups. If the adjustment is a function of factors that vary across the exposure groups, then the dose-response modeling should be performed on the adjusted animal dose metrics. Examples of the use of PBPK-based dose metrics in BMD modeling can be found in the literature (Clewell et al., 1997; Barton and Clewell, 2000).

Reviewers recommend that the guidance document put more emphasis on the hybrid continuous model (termed “implicitly dichotomizing” in the guidance). They noted that some of the loss of emphasis was probably unintentional, resulting merely from the order in which the approaches were discussed. Overall,

---

[2]Andersen M.E., Clewell H.J., and Krishnan K. 1995. Tissue dosimetry, pharmacokinetic modeling, and interspecies scaling factors. Risk Anal 15:533-537.

Barton, H.A., and Clewell, H.J., III. 2000. Evaluating noncancer effects of trichloroethylene: dosimetry, mode of action, and risk assessment. Environmental Health Perspectives, 108(suppl 2):323-334.

Clewell H.J., Andersen M.E. 1985. Risk assessment extrapolations and physiological modeling. Toxicol Ind Health 1:111-131.

Clewell, H.J., Andersen, M.E., and Barton, H.A. 2001. A consistent approach for the application of pharmacokinetic modeling in cancer and noncancer risk assessment. Environmental Health Perspectives (submitted).

Clewell H.J., Gentry P.R., and Gearhart J.M. 1997. Investigation of the potential impact of benchmark dose and pharmacokinetic modeling in noncancer risk assessment. J Toxicol Environ Health 52:475-515.

U.S. Environmental Protection Agency (USEPA). 1994. Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry. EPA/600/8-90/066F, USEPA Office of Health and Environmental Assessment,Washington, D.C.

reviewers agreed that the guidance should state more clearly that implicit dichotomization is preferred over explicit dichotomization, due to the loss of information in the latter approach. A general preference for the use of the hybrid model over other descriptions of the BMR was also expressed.

Question 4b: Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?

Reviewers requested more graphical presentations and more explicit guidance on how to evaluate the graphical data. One approach mentioned by reviewers involved looking for systematic patterns of departure from the best fit line (e.g., having several data points on one side of the line, rather than the line falling between data points). This was expressed mathematically as balancing the residuals; some reviewers requested more information on what one should be looking for in evaluating residuals. Reviewers suggested emphasizing the need to focus on the region of the BMR. For example, they noted the issue of acceptable overall fit, but poor fit in the region of BMD. The inverse issue (acceptable fit in the region of the BMD but poor overall fit) was also noted. This latter issue can often be addressed by dropping a high dose, although reviewers noted that one high-dose group should be included. It was also noted that a Michaelis-Menten model is an alternative to dropping high-dose groups when a response plateau is observed.

Some reviewers asked for biological plausibility to be highlighted as a criterion for evaluating fit.

Several reviewers considered that the explanation in the text of model-fitting methods and criteria for evaluation may be overwhelming for the typical practitioner. This text may be at too high a level (i.e., have not enough explanation) for most toxicologists, while providing unnecessary information for statisticians. The text could be better aimed at the document's target audience.

Reviewers generally agreed with the use of $p > 0.1$ as a cutoff for the goodness-of-fit (GOF) statistic. Some concern was expressed that this criterion would be too restrictive, particularly for developmental toxicity data sets. This cutoff may also be problematic for non-monotonic data sets. However, the developmental toxicologists among the reviewers stated that the data sets that they have dealt with could be fit adequately using this criterion. In light of the broad target audience, reviewers recommended that the decision rule be stated explicitly (i.e., that a GOF $p < 0.1$ is rejected). It was also noted that it may not be possible to estimate a GOF p value for some methods, such as some applications of GEE (generalized estimating equations).

4c. What are the advantages/strengths of using the methods described to select among "equally" fitting models? What other methods should be considered in making a selection?

There was considerable discussion on how much emphasis should be placed on parsimony and how one should determine the model of choice for a data set. Several reviewers expressed a concern that the draft guidance puts too much emphasis on parsimony. They felt that good fit to the means should be emphasized over parsimony. Others liked the tighter confidence intervals obtained with simpler models. Overall, it was generally agreed that the use of a hierarchy of models (as discussed above under question 4a) would aid significantly, and would reduce reliance on the AIC in choosing among similar models. Reviewers felt that the AIC is useful in considering whether the improved fit obtained by adding a parameter is sufficient, in light of the loss of a degree of freedom. However, the AIC should be considered informative, rather than prescriptive. Some of the reviewers noted that the use of a small variety of models gives some indication of model sensitivity, although several others preferred to choose

one model a priori and then compare others to that chosen model. There was, however, general agreement that one should not use the lowest BMDL, since choosing the lowest of several lower bounds means that the confidence interval is no longer appropriately characterized. Overall, reviewers considered the model choice to be case-specific, not something to be prescribed in detail, and usually not something that has a large impact on the resulting value. Reviewers also noted that the issue of when and whether to throw out an outlier BMDL is removed if one is only running a few models. For example, with three models, an outlier cannot be defined and the issue becomes one of choosing the most appropriate model.

A number of reviewers expressed concern about "model shopping" and associated pressure to evaluate every available model in order to avoid the accusations of model shopping. A recommended hierarchy of models was considered a useful way to address this issue.

Question 5: Please comment on the approaches described to compute confidence limits.

Reviewers were divided on whether BMD or BMDL should be used for regulatory or pseudo-regulatory purposes (e.g., development of RfDs/RfCs). However, the actual use of BMD or BMDL was beyond the scope of the peer review. Despite these limitations and disagreements (discussed in more detail in the following paragraphs), there was agreement on several points. Reviewers agreed the guidance should note that there are several approaches to calculating confidence limits. Specific mention was made of the bootstrap method, and reviewers were requested to submit relevant citations (some of which were provided in the premeeting comments). Even those who preferred the use of the central estimate over the BMDL thought it would be of interest to report confidence limits to characterize the associated uncertainty, sometimes including both upper and lower confidence limits. Conversely, even those favoring BMDL for risk assessment purposes (e.g., RfD development) tended to agree that BMD would be preferred for comparisons of potencies or other comparisons across endpoints. Reviewers also generally agreed that the BMD guidance should document the reason for the choice made (BMD versus BMDL), and should acknowledge the interplay between that choice and the choice of uncertainty factors (and perhaps modifying factors). ("Interplay" refers to the fact that a response is observed at the BMD, while the BMDL is a lower bound on that response, and the choice of uncertainty factors will be different if one uses a BMD versus a BMDL.) Finally, reviewers agreed that confidence limits below zero are meaningless.

A number of arguments were made for using BMD, rather than BMDL, as a POD. The first argument has to do with the amount of statistical complexity introduced by BMDL. BMD is a simple tool that has significant advantages over the previously used NOAEL/LOAEL process. The use of lower confidence intervals, however, introduces many statistical complexities, including the methodology for computing the confidence interval, lack of convergence for the complex models (this is less of an issue for simpler models), and interpretation of the ensuing intervals. Reviewers also noted that, while it is theoretically true that using a BMDL rewards better experimentation, that argument has little practical value. Current experiments have highly prescribed minimum designs, and it would be rare for a sponsoring company to use more animals or a more complex design simply to reduce the size of the confidence intervals. A simpler approach would be to calculate a central estimate of the BMD with simple, intrinsically linear models, then apply a modifying factor to the central estimate, depending upon the quality of the data. Other reviewers noted the advantage that BMD provides an estimate of response that can be combined with probability distributions for estimating low-dose response, while no such estimate of response is possible when one uses the BMDL as the POD. Finally, reviewers expressed a preference for making

conservatism explicit (e.g., through the use of uncertainty or modifying factors), rather than automatic and unavoidable (by use of BMDL).

Other reviewers preferred that risk assessment applications (e.g., RfD development) use BMDL as the POD. While the use of BMDL may not affect study design, it does capture the uncertainty inherent in the data that are being used, since risk assessors often must develop values without having test guideline-compliant studies. Reviewers also suggested that BMDL may be more stable than BMD for model choice. They noted that confidence limit calculations interact with the number of degrees of freedom, since increasing the number of parameters could lead to larger differences between the central estimate (BMD) and the lower bound (BMDL). (This could be an argument for using BMDL or BMD, depending on which shows more stability. Reviewers also noted that this is less of an issue when calculating confidence limits by the bootstrap approach.)

Reviewers noted that the discussion of confidence limits in the guidance is highly technical compared to the rest of the document. Since such discussion is probably not useful for the typical practitioner, it may be best to put it in an appendix. It would be useful, however, to provide a plain-language description of what the confidence interval is.

Question 6: What additional concepts, if any, should be illustrated by an example?

Reviewers agreed that a number of examples should be added to the guidance. For each example, the point to be illustrated should be clearly stated up front. The reviewers recommended giving examples of basic applications to different types of data, rather than the current focus on more obscure issues. There was also an acknowledgment of the usefulness of examples addressing some more complex issues (some of which could be addressed with on-line examples). It would be useful if these examples were step-by-step illustrations, related directly to the decision tree provided in the main document, and if the supporting data (output, input) were supplied. These examples should illustrate the choices made in the process and the rationale for those choices. Basic applications and examples that should be discussed include:

- Modeling of quantal and continuous animal data. Continuous data should be modeled using the hybrid continuous model. These examples should also illustrate the point of dropping the high dose to improve fit when there is a response plateau.

- Epidemiology data for cancer and noncancer endpoints. Due to the complexity of this modeling, reviewers did not consider it useful to fully "walk through" an epidemiology example in detail. (One reviewer thought the issues were so significant that the epidemiology example should be omitted.) Instead, reference to well-done epidemiology applications, and a brief discussion of key issues, would be useful. Reviewers recommended that the text note the distinct advantage of the BMD approach over the NOAEL/LOAEL approach for continuous epidemiology data. This advantage arises from the fact that both exposure and response are continuous for such data, and that discrete exposure groups do not exist. The NOAEL/LOAEL approach requires that such data be artificially "binned" (e.g., that exposure be classified as 0 to 300 ppm-years, 300 to 600 ppm-years, etc), and the resulting NOAEL is directly related to the artificial binning. Thus, a different NOAEL would result if the data were binned as 0 to 300 ppm-years than if they were binned as 0 to 200 ppm-years. The actual individual exposure levels can be evaluated using BMD modeling, removing the need for artificial binning. The BMD approach also removes the need for an external comparison group, removing some problems related to the healthy worker effect.

Reviewers also suggested that these advantages of the BMD approach for epidemiological data be mentioned in the introduction of the guidance document.

- An example using the bootstrap method.

- An example of application to human data using the risk-factor approach (e.g., cardiovascular or sperm count data, where the latter can be projected to male infertility).

- An example illustrating the use of covariates.

Reviewers made a number of specific comments on the examples in the premeeting comments; these comments should be considered (to the degree that the current examples are retained). Primary among these comments was a general agreement that the dynamic range example is not generally accepted as the basis for the BMR, and should not be emphasized by using it as an example.

A cancer example was considered valuable, but it was recommended that the text not refer to "the cancer model." Some additional brief examples in the text were considered worthwhile, such as an example of a case in which BMD modeling would fail and should not be used.

4.2 Discussion of Question 4a

Dr. Haber kicked off the discussion by listing some of the key comments reviewers had made in their premeeting comments in response to charge question 4a: *What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?* These comments included:

- The need for constraining power parameters to be no less than unity should be more strongly worded.

- The power parameter should be restricted to greater than one (one reviewer commented that the power parameter should not be constrained to greater than one).

- The document should provide more explanation and examples about how to use covariates.

- The guidance about using parsimony to drive choice of models is a concern, as is the fact that this approach was emphasized in the examples but not in text.

- How should one deal with control when taking the log?

- More guidance is needed on how to use threshold parameters, particularly since these parameters are not available in the BMDS software.

- The document should provide tables to clarify the use of models and endpoints.

- Parameters should be limited to non-negative values. (This comment was made by just one reviewer.)

A reviewer began the discussion by mentioning that although the BMDS software does not include threshold parameters, other available modeling software does. He pointed out that the term "threshold parameters" can have a variety of interpretations. For example, threshold can refer not only to dose, but also to time. Many effects that do not appear until late in life can be handled in modeling by using what is equivalent to a threshold in time. Regarding covariates, he said that the issue was not so much the use of covariates, but rather what to put in for covariates if one has done modeling and then wants to do a risk extrapolation. He also said he was not enthusiastic about the guidance concerning use of parsimony, since he felt that characterizing dose-response is more important.

A reviewer then commented on the default issue. He said that the guidance essentially tells the reader to use any model as long as it fits. Concerned about the potential for model shopping, he recommended that the document provide a hierarchy of preferred models for certain situations, so that the user would have some sense of which are most appropriate for particular situations. A number of reviewers agreed with this recommendation. Another reviewer was concerned that the examples in the back of the document contained many underlying assumptions concerning default issues that were not clearly discussed. He recommended that the document include a table near the front that clearly states what defaults EPA thinks should be in place for each model. Another reviewer said the text could be strengthened by having examples to illustrate these concepts. She recommended that the text reference the examples in the back of the document, and that the document include a table that provides key information on model choice, including the advantages and disadvantages of each model and guidance on when the model should be used. Although much of this information is in the text, the reader would benefit if EPA could consolidate it into a table.

Reviewers then discussed whether basic mechanistic considerations should be included in the hierarchy for models. Several reviewers felt this would be a good idea. They felt that biology should at least be acknowledged as a matter of principle, since the hope is that the understanding of mechanism of action will increase over time and provide the basis for better risk assessments in the future. Without some acknowledgment now, users may miss out on some opportunities to consider biology when this is possible. Others reviewers were concerned about this idea. The biological underpinning of a model has little to do with whether it is good for describing a particular data set for benchmark, said one reviewer. Also, there currently are so few cases in which enough biological insight is available to describe the nature of the dose-response curve that this should not be a driver at this point.

The central estimates from the models are so close that it would be difficult to distinguish between the models based on model fit, pointed out another reviewer. Biological basis could be used to distinguish between models when the fit is about the same, said another reviewer. The models are simply mathematical forms and are not biologically motivated, responded a reviewer. They are used to estimate where an effect occurs, not where it saturates. It would be inappropriate if putting forth biologically motivated modeling as a goal ends up focusing people on using, for example, the Hill equation because the response saturates at higher doses. It would be better to throw out the high doses than to use the information on the curve for any purpose at all.

A reviewer drew an analogy to EPA's cancer risk guidelines. Here, EPA began by being rigid on model use; as more information became available, the Agency made mechanistic modeling the first priority if it could be well defended. Since that time, mechanistic modeling has been successfully used and defended for cancer risk assessment. For the BMD approach one, however, is simply trying to get a good 10% level, rather than extrapolate. For this reason, it may well be immaterial which model is used. Nevertheless, the BMD guidance could include mechanistic models in the hierarchy as long as their use can be defended, he suggested. Though different models will give similar answers, there will be

differences, responded another reviewer. If these differences are slightly above or below a critical level that triggers action, this will generate conflict. If the document fails to provide guidance in this area, it would create the potential for model shopping and pressure to run every type of model to ensure one did not end up with the low one.

Dr. Haber then summarized her understanding of the discussion about model use and selection as follows. Reviewers would prefer that the guidance provide a hierarchy with some recognition that knowledge about the biology (e.g., mechanism of action or recognition of the distribution) can be used. However, it should be noted that some of the models that apparently have some biological connection, like the Hill model, are really still just curve-fitting exercises in this application.

A reviewer responded that the record should also reflect the feeling, held by many reviewers, that there should be a reluctance to use unconstrained models since lower bounds from these models can be extremely unstable. He felt the guidance provided a mixed message about when to use unconstrained models. Though he generally agreed with the guidance given, he recommended that it be better organized and presented and that concerns about the use of unconstrained models should be strongly stated.

Reviewers then discussed the issue of covariates. In the premeeting comments, several reviewers raised the issue of what the document should say regarding adjustment for covariates. A reviewer began the discussion by saying there are important methodological issues regarding adjustment for covariates. He recommended that the document discuss these issues as well as possible solutions, and indicate that this is an evolving area in which further research is needed. There are currently few references on this topic, he added. Another reviewer mentioned that the premeeting material for a recent NIEHS workshop on low doses of endocrine disruptors included a brief tutorial on issues concerning covariate use. This might serve as a possible model for a discussion of covariates in the guidance, he suggested. A third reviewer agreed that the document does not provide sufficient guidance on covariates. In work on developmental toxicity, she found that use of covariates such as litter size and nonindependence of events improved the assessment. She suggested that the document indicate that it sometimes may be important to consider covariates, especially from a mechanistic standpoint. A fourth reviewer mentioned an example from his work in which age was a significant covariate for the effect of cadmium exposure on beta-2-myoglobulin in the urine. Not only was age associated with longer exposures, it also appeared to change susceptibility. In estimating variability and confidence limits using an overall log probit type model, he found it critical to include age as a covariate.

Some other reviewers expressed concerns about the extent to which the guidance should discuss covariates. One reviewer felt that, given the limited current knowledge about covariate use, the document should briefly refer to the issues and should indicate that significant additional research is needed before additional guidance can be provided. Another reviewer, agreeing that covariate analysis is currently very difficult, recommended that the document focus on straightforward benchmark analysis and simply acknowledge the covariate issue. A third reviewer agreed and suggested removing the one example in the document that currently mentions covariates. Reviewers noted the some commercial BMD software can handle covariates, though the BMDS software cannot.

Dr. Haber summarized the covariate discussion by saying reviewers agreed there are methodological issues concerning covariates, this is an important research area for which there currently are few references, and the text in the document does not adequately highlight the issues. Dr. Haber asked reviewers to submit any relevant references they have.

Reviewers briefly discussed the issue of how to handle the control group when performing a log transformation of the data. A reviewer mentioned an arsenic risk assessment in which a constant was added to the dose measure before taking the log. Another reviewer mentioned that if background is noninteracting, one can control for the background incidence of the effect as is done in conventional log probit analysis.[3]

Next reviewers discussed the issue of threshold parameters. In the premeeting comments, a reviewer suggested that the guidance present more options for dealing with threshold parameters. A reviewer mentioned an assessment in which a threshold parameter was used to improve the fit, not because it represented a biological threshold. This assessment involved a data set in which the responses at several of the dose levels did not differ from controls. She suggested that because there may be instances in which thresholds can be used appropriately when they are not biological thresholds, it would be useful to add threshold parameters to the BMDS software. Two reviewers suggested that, where possible, a benchmark should be done with and without a threshold. If the LCL on the threshold is zero, that indicates there is no threshold.

A reviewer was concerned about routine application of population thresholds. He felt that using a population threshold parameter when the dose is log-transformed could lead to situations in which there seems to be a threshold, when in fact the response is linear. He suggested that a threshold should not be routinely used when there is a plausible mechanistic reason to suggest the response is linear, though one could certainly be used when one is hypothesizing a real threshold response. Another reviewer said he strongly felt that log transformation should not be performed on dose, for both toxicological and statistical reasons. He recommended that the guidelines more strongly indicate that log transformation of dose should not be routinely considered unless there is justification.

Dr. Haber summarized the key points made by reviewers in this part of the discussion: many reviewers opposed log transformation, but there may be times when it is appropriate; threshold parameters can be useful for improving the fit, but they are not input parameters nor do they necessarily represent true biological thresholds; including threshold parameters in the BMDS software would be helpful.

4.3 Discussion of Question 4b

Dr. Haber summarized the key premeeting comments reviewers made in response to charge question 4b: *Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?*

- Several reviewers suggested an additional criterion: biology, especially Michaelis-Menton.
- Several reviewers wanted the guidance to more clearly emphasize graphing and local fit.
- Some reviewers thought the criteria might be overwhelming.
- Some reviewers agreed with using the stricter criterion of GOF p value > 0.1; others thought it could be a problem, especially for nonmonotonic data.

[3]Finney, D.J. 1971. Probit analysis, 3rd edition. Cambridge University Press, London.

- A reviewer requested a plain-English explanation of $p > 0.1$ (i.e., what fails).

- Some reviewers were concerned about using the AIC criteria only if the simplest models fail.

Regarding the issue of the biology criterion, reviewers indicated they had no additional comments beyond what they had previously discussed. Regarding graphing, a reviewer asked how one would know when a graph is working—he himself looks for systematic patterns of departure from the best fit line. Another reviewer responded that he typically looks at how the curve fit relates to the means of the group. He looks for the best fit line to be close to those means and for the means to be equally distributed on either side of the best fit line. The greater the distance of the means from the fit line, the less confidence there is in the fit. He recommended that this approach be more explicitly discussed in the document. A third reviewer added that this relationship of the means to the curve fit should be in the region of the BMR. He also said that balancing of the residuals is important. Another reviewer suggested that the guidance should provide more information on evaluating residuals.

Regarding graphing for continuous data, another reviewer said, it is extremely important to look at the dose-response fit as well as the assumptions behind the typical risk characterization. For example, the variance of the observations needs to be constant over dose and should follow an approximate normal distribution for the residuals. These are both very important for risk characterization involving continuous data. The reviewer recommended that this be discussed in the document. Another reviewer pointed out that this is a different approach than for quantal data. He mentioned that he had observed problems with the way epidemiologists handle this type of problem. When there is hetroscedasticity (the increase in variance with the independent variable), some epidemiologists log-transform the data to get rid of this problem. However, when the x axis is log-transformed, it changes the hypothesized relationship between the variables. If in fact there is good a priori reason to believe that the untransformed data have a direct relationship, but there is an error-of-measurement problem, then the problem should be fixed in various ways (such as weighting the points differently) other than log-transforming.

A reviewer pointed out that an example in the document—in which a one-degree multistage model is used instead of a second-degree multistage model on parsimony grounds, even though both models fit—is in direct contradiction to how a multistage model is used in Global 86, in which all the degrees from one to six are fit and the one with the lowest q1* is selected. Another reviewer responded that the guidance has been overturned by the new cancer guidelines, which simply recommend fitting the data in the region of observation—essentially a benchmark approach. He asked EPA to clarify why there appeared to be a special cancer model coming out. Dr. Setzer responded that the special cancer model under development will be based on the multistage model currently in the BMDS software. The point of the cancer model is to rigorously enforce certain defaults for certain features, including restricting coefficients in the polynomial to be nonnegative. The model will report a slope value at the BMD and its confidence interval, and will provide a linear interpolation down from BMD directly on the graphic. However, it is fundamentally the same model as the multistage model. One purpose of having a specific cancer model is to avoid the potential for model shopping. The reviewer responded that he was concerned there now appeared to be a cancer model, since the cancer guidelines were very clear that there was not going to be a single model. He also pointed out that the BMD guidance uses an example that gives the impression that the use of benchmark in cancer is completely different. EPA's Elaine Frances acknowledged that development of the cancer model goes against the basic premise stated in the BMD document about using the same type of benchmark approach across all types of endpoints. The reviewer responded that the guidance document should avoid appearing to endorse a single cancer model. He recommended that the

document use an example that is consistent with the current cancer guidelines and the model under development, but not mention the specific cancer model in order to avoid giving the impression there is only one way to use the BMD approach for cancer data.

Dr. Haber then asked reviewers to discuss the issue of parsimony versus best fit for model selection. A variety of opinions were expressed on this subject. One reviewer felt that, as a general principle, parsimony is beneficial, since he did not feel the complexity added by using additional degrees of freedom was warranted. He thought the issue of model selection was too complex to automatically rely on the AIC. Another reviewer agreed that, other things being equal, a more parsimonious model is better than a less parsimonious one. A third reviewer pointed out that parsimony can be a concern when it distorts the shape of the dose-response curve. Though the AIC can be useful, he did not want to be regimented by it and recommended that it not be put forward as the only criterion for model selection.

A reviewer pointed out that the number of parameters can affect the calculation of confidence limits. The more parameters one has, especially when using the distribution of the likelihood method for calculating confidence limits, the more one is able to push on the one linear parameter and adjust the others to compensate. So the more parameters one has, the greater can be the gap between the best estimate and the lower bound. The examples on pages 70 and 71 in the guidance document illustrate this point. In these examples, the gap between the lower bound and the best estimate is greater with the two-degree model than the one-degree model. The BMDL for the first-degree model on page 71 is about the same as the BMDL for the second-degree model on page 70, even though the BMD is higher for the second-degree model than for the first-degree model. The reviewer felt it would be ironic if higher degrees of freedom produced lower bounds that are further down than they have to be. For these reason, this reviewer supported the parsimony principle. Another reviewer agreed, and said he supported parsimony because it was better to have a more defined approach.

A reviewer pointed out that the AIC can only be used to compare models within the same family of distributions; it cannot be used when the models are in different families of distributions. The document should make this limitation clear. Also, unlike the chi square, the AIC principle does not indicate the significance of the difference in results between models with different degrees of freedom. The AIC is a useful tool to evaluate models, especially when other criteria do not clearly indicate which model to use, but it is not an infallible guide. Another reviewer agreed that there can be significant problems with using the AIC to compare different model series. A third reviewer said he thought the guidance document did indicate that one should apply AIC to models within the same family, and also that one should decide whether to use more parameters without basing one's decision on a likelihood ratio–type test.

Dr. Haber then asked reviewers to discuss the criterion of GOF p value > 0.1. A reviewer said he was bothered by the fact that, while GOF is judged on the best-fitting model, one often obtains the confidence limits by proposing an alternative model fit. This means that alternative model is being used to describe a response, even though it probably does not fit the data. This has to do with stability of bounds, responded another reviewer. If the best estimate model fits well but the lower bound does not appear to be consistent with the data, this is probably due to an instability or oversensitivity to parameter type.

A reviewer asked whether the GOF value $p > 0.1$ might be too restrictive for some developmental toxicity data sets, in that it might fail to accept any model for some fairly plausible data sets. Another reviewer responded that he was not aware of any data sets in developmental toxicity that would not fit using the models presented. A third reviewer said he had used log-normal theory to describe human data in around

400 cases. He found that the log-normal approach is usually reasonable for large data sets, and generally the variability of the variability measure among chemicals is also reasonably described by log-normal distributions.[4] A fourth reviewer said that out of 11 or 12 developmental toxicity data sets he had evaluated, only about half had a p value > 0.1 when separate endpoints were modeled jointly. For some of these data sets, there was not a monotonic dose-response trend at the lower dose levels. For example, the first dose might have shown a lower malformation rate than controls or a higher rate than the second dose level. This reviewer was concerned that, with a GOF criterion of p > 0.1, data from some experiments will not be usable because they will fail to meet the criterion. Another reviewer referred to published studies using a much larger database of developmental toxicity studies, and she reported that they achieved good fits when separately modeling endpoints such as death or lethality or malformations. A reviewer pointed out that lack of fit for developmental toxicity data can arise if only malformations are modeled and lethalities at high doses are excluded. For risk assessment purposes, it is more important to include both effects, he said. Including the dead embryos often can change an inverted U-shaped dose-response curve for malformations into a dose-response curve that can be easily modeled. A reviewer said the guidance could indicate that one should apply the criterion of p > 0.1 for GOF when looking at endpoints in isolation, not when doing joint modeling. (??)

Another reviewer voiced his support for the p > 0.1 criterion. He was concerned that, without this criterion, too many models would fit, including ones with wide confidence intervals at the bottom. A reviewer pointed out that a GOF test is not always readily available for generalized estimating equations, so for these equations it may not be possible to report a p value. Finally, a reviewer reminded the group that the document does not describe how to incorporate dosimetric adjustments, such as PBPK modeling, deposited dose, etc. He recommended that this be covered in the document.[5]

Dr. Haber summarized the discussion of question 4b as follows: Reviewers recommend that the document provide more guidance on evaluating graphical data, including looking for systematic departures from the best fit line; focusing on the region in the BMR but wanting a balance of residuals on both sides of this region; and a greater emphasis on the fits to the means. Regarding the criterion of p > 0.1, there was initial concern about whether this would work with developmental toxicity models, but there was

---

[4]More information about this can be found in:

- The website www2.clarku.edu/faculty/dhattis .
- Hattis, D., Banati, P., and Goble, R. 1999b. Distributions of individual susceptibility among humans for toxic effects—For what fraction of which kinds of chemicals and effects does the traditional 10-fold factor provide how much protection? Annals of the New York Academy of Sciences 895:286-316.
- Hattis, D., Banati, P., Goble, R., and Burmaster, D. 1999. Human interindividual variability in parameters related to health risks. Risk Analysis 19:705-720.

[5] After the workshop, the reviewer submitted the following references on this topic:

- Clewell H.J. III, Gentry P.R., Gearhart J.M. 1997. Investigation of the potential impact of benchmark dose and pharmacokinetic modeling in noncancer risk assessment. J Toxicol Environ Health 52:475–515.
- Barton, H.A. and Clewell H.J. III. 2000. Evaluating noncancer effects of trichloroethylene: dosimetry, mode of action, and risk assessment. Environ Health Perspect 108(suppl 2):323–334.
- Clewell H.J., Andersen M.E., Barton H.A. 2001. A consistent approach for the application of pharmacokinetic modeling in cancer and noncancer risk assessment. Environ Health Perspect (submitted).

general consensus that $p > 0.1$ is sufficient when endpoints are modeled individually. It was also noted that a GOF value cannot always be calculated for some methods, such as the some applications of generalized estimating equations.

4.4 Discussion of Question 4c

Dr. Haber summarized key comments reviewers had made so far of relevance to charge question 4c: *What are the advantages/strengths of using the methods described to select among "equally" fitting models? What other methods should be considered in making a selection?* She included premeeting comments as well as comments made during discussions of previous questions. Key comments included:

- Some reviewers prefer emphasizing fit over parsimony, while others prefer parsimony because it gives tighter confidence limits.
- The AIC can be a useful tool to inform choice, but it cannot be used to compare across families of models.
- Some premeeting comments concerned the use of the biological basis of models and mode of action when comparing among the models.
- Reviewers wanted additional guidance on, and expressed some concern about, the concept of throwing out outlier BMDLs.
- Reviewers were concerned about the potential for model shopping and the pressure to run all models. They wanted definition of the universe of models that could be used.
- Some reviewers requested additional guidance on dropping the high dose.

A reviewer began the discussion by describing his concern with the Hill model, which is mentioned in an example but not in the text of the guidance document. As far as he knew, only one group was advocating the "fraction of dynamic range" approach; most everyone else appears to be skeptical about its underpinnings. He said he opposed this approach because he did not think that maximal response had anything to do with the onset of chemical effects. He mentioned the example of peroxisome proliferation, in which case the maximal response is a chemical-specific property, rather than a biological property. Though the maximal proliferation response could, for example, be 10-fold greater for one chemical than another, this does not mean that a 10-fold greater increase over baseline would be acceptable for the first chemical. Also, the "fraction of dynamic range" approach is inconsistent with the emphasis in the guidance on the dose region of interest being around the BMR, because it gives too much importance to the high-dose behavior. For these reasons, the reviewer recommended that the example illustrating this approach be deleted. He generally favored dropping the high doses and modeling only the low-dose region if the high doses complicated the dose-response modeling.

A reviewer agreed with the recommendation to eliminate the dynamic range example. Another reviewer agreed with dropping the high doses. He said that the models are very sensitive to the high-dose region, so tweaking something in this region can have a big effect at the low end. This is wrong from a biological point of view, since noise in the high-dose region should not be driving changes in the low-dose region. He suggested that the guidelines be aggressive in recommending caution in the use of high-dose data,

especially when there is a plateau in the high-dose region. In this case, one should use only the first high-dose point. A third reviewer agreed with these comments. A fourth provided some examples of how high doses can distort a model's behavior at low doses. A fifth reviewer agreed with the idea of including the first high-dose point, but excluding the others, in order to see whether there is a decline from the first high dose to the low-dose range.

A reviewer suggested that the Michaelis-Menten interpretation of a plateauing dose-response may be appropriate. Another reviewer agreed. He pointed out that one option when there is a plateauing dose-response is to try to avoid the problem by using biological information or dose transformation of a biological type (such as physiologically based pharmacokinetics) to change the dose scale so that it is no longer hyperbolic. This can be done, for example, with vinyl chloride. The reviewer recommended that the guidelines suggest that, if there is a problem with high-dose data that cannot be solved by dropping the high doses, one might be able to eliminate the problem by identifying and describing a biological basis for the plateau.

A reviewer said he was very concerned about the document's instruction to fit multiple models to multiple endpoints, then throw out any outliers among the BMDLs. He said that he thought he could always make BMDLs vary by at least three-fold and they would still fit, so he felt the document needed a more rigorous criterion than just selecting the lowest BMDL when BMDLs differ by a factor of three. Another reviewer agreed that it makes no sense to do a minimum BMDL across studies when looking at a single endpoint. He said that even if one believes in a 95% confidence limit, the confidence interval is no longer appropriately characterized when one takes the minimum of several BMDLs computed on different bases (a PBPK dose scale, an administered dose scale, etc.). Also, he had performed analyses which showed that taking the minimum over multiple studies with the same endpoint gives an increasingly wrong answer as one continues to repeat the study. For this reason, the guidance to take a minimum BMDL is flawed from a policy standpoint, since it would tend to discourage multiple studies. A third reviewer agreed with this. A fourth reviewer said she thought it important to differentiate between minimizing among BMDLs across studies for a given endpoint and the issue of multiple studies whose purpose is to make sure one has evaluated all the key endpoints.

Dr. Haber summarized the discussion to this point as follows. Reviewers want to eliminate the dynamic range example. The guidance should recommend caution in using high-dose data, especially when there is a plateau, but make sure that the first high dose is included.

A reviewer said he was concerned about how the coverage probability of the confidence limits of a model would be affected when one is using GOF to choose a model from among several models and then doing a bound on the chosen model. He was concerned that this process could generate some undesirable effects on the coverage probabilities of the confidence limits that result from this process. Another reviewer responded that this issue would become less important if EPA established a hierarchy of models and if it is agreed that adversity is outside the scope of the guidance. A third reviewer suggested that there could be some difference of fit among different related models even if there is a hierarchy of models.

Reviewers then discussed the hierarchy of models. One reviewer suggested that the approach used by California EPA's OEHHA might be a good example for ideas about a possible model hierarchy in the guidance. For quantal models, OEHHA uses the log probit model as long as the data fit well and there are not good mechanistic data to suggest that an alternative model would be more appropriate. Another reviewer said that at least two models should be used in order to evaluate model dependence. A third

reviewer agreed that more than one model should be run. He said that, for most "well-behaved" data sets, it would not matter what model was used, since the resulting BMDLs would be similar. However, the only way to tell whether the data are model sensitive would be to run several models. He thought the document already contained some preliminary ideas for a hierarchy, though these ideas were not explicitly set forth as such. For example, the document specifies approximately three models for quantal data and three models for continuous data. One could run those models and then do a limited model comparison on how sensitive the data are to model selection. Based on the results of that comparison, one could proceed to more complex nonlinear models. He recommended that the document provide some flexibility in model selections, since risk assessors will have different model preferences for various reasons. Another reviewer said he found the examples confusing compared to the text in terms of what they implied about a hierarchy. A reviewer mentioned that one advantage for suggesting a hierarchy would be that the information that resulted over time from performing BMD analyses using those models would be more comparable than if a wider variety of models were used.

A reviewer then summarized his sense of the general agreement on model hierarchy as follows. Two to three standard models should be tried first to determine whether there is a model sensitivity. If there is model sensitivity or if none of the standard models work, then more complex models can be used. The guidance should provide some flexibility about what standard models can be used. Another reviewer said he heard general agreement that the collection of models should not be characterized by a minimum, especially when dealing with confidence limits.

A reviewer was concerned that this statement of agreement included no criteria for deciding when a model is an outlier, when the outlier can be tossed out, and how one would choose the model that would be used to continue to the next step in the analysis. Several reviewers responded that the outlier issue for BMDLs is not a concern when only a small number of models are run, in which case the issue becomes which model to select. Another reviewer thought model selection was such a case-specific issue that it would be difficult to provide generic guidance for selection. Another reviewer was concerned that the hierarchy procedure described did not include selection of a point. He supported the idea that one could choose a model and then compare other models to it to see whether there is model sensitivity. If there is none, then one could select that point. Two reviewers indicated their support for this idea. Finally, a reviewer pointed out that there could be situations in which it is unclear whether the models run for comparison purposes agree with the first model. In this case, one could simply present the distribution of the comparison models around the standard model.

4.5 Discussion of Question 5

Reviewers then proceeded to discuss charge question 5: *Please comment on the approaches described to compute confidence limits.* Dr. Haber summarized the key premeeting comments on this issue:

- Several reviewers opposed doing confidence limits.
- Reviewers commented that the document appeared to be too advanced for toxicologists and too simplistic for statisticians.
- Several reviewers opposed log transformation.

- Some reviewers commented that the guidance should identify alternative methods for determining confidence limits, such as the bootstrap method and the likelihood ratio method.

- Concern was expressed about model shopping in relation to the method for computing confidence limits.

Since reviewers who agreed with the use of confidence limits generally did not submit pre-meeting comments to that effect (while those who opposed confidence limits did address that issue in their pre-meeting comments), Dr. Haber invited reviewers who supported confidence limits to speak first. A reviewer responded that he thought confidence limits made sense for some applications but not for others. He thought confidence limits were valuable for comparing potencies among a group of chemicals. Such comparisons should be based on a central tendency estimate, and confidence limits should be calculated on both the upper and lower points of comparison to help determine how robust the comparison is. A lower confidence level is also useful for quasi-regulatory purposes such as calculating RfDs and PODs, for uncertainty factor calculations, for simple linear projections, and to encourage better studies of chemicals, particularly when the only data currently available are from studies conducted decades ago (when it was not possible to envision current uses of the data). This reviewer was not convinced that the cost associated with doing better studies was justified by the value of having an improved toxicological database. Nevertheless, he pointed out, encouraging better studies was one of the original justifications for the BMD approach. Use of confidence limits is consistent with both the past RfD-type approach and to some extent with the linearized multistage approach. A second reviewer agreed that both upper and lower bounds are useful. A third reviewer agreed with the idea of rewarding better study design. He pointed out that data sets from common study designs using the same model can yield different dose-response patterns and uncertainty estimates for the resulting BMDs. In this case, one gets different estimates of precision, which translate into different confidence limits. So LCLs are useful not only to reward better study design but also to give more information about the precision of estimates. Another reviewer wanted the group to be explicit that better study design meant more dose groups and more animals per dose group to decrease the confidence interval. Two reviewers responded that there may be other ways to improve study design, such as more continuous parameters rather than quantal parameters and use of modeling. Another reviewer pointed out that current studies have explicitly prescribed minimum study designs, and it would be rare for a study sponsor to use more animals or a more complex study design to reduce the size of the confidence intervals.

A reviewer said that the data are not a strong tool for deciding what parameterization of a model to choose, and the likelihood ratio used to define the LCL is a reasonable way to reflect that reality. Another reviewer agreed with an earlier statement that confidence limits are useful when one does comparisons. However, he said, when a lower confidence level for a POD and low-dose extrapolation are used, the likelihood ratio criterion for determining a 95% LCL does not give a consistent coverage of probability. When one does a confidence limit on a mean for a normal distribution, there is always a constant 95% coverage, regardless of what the mean is. In a dose-response modeling context, however, the likelihood ratio criterion does not always provide the same level of coverage. In fact, coverage will differ for different dose-response models. It will be close to 95% for a true underlying linear model, higher than 95% for an underlying linear quadratic model, and even greater for something more nonlinear. This is problematic in a regulatory context, because the POD for the regulation will vary depending on the model used. Another reviewer responded by saying that he felt it was important to separate the issue of what constitutes good, adequately studied statistical methodology from the conceptual issue, since these

are very different issues. In fact there are other statistical methods, such as the bootstrap methods, that perform well, so the answer may be to use different tools to do these calculations.

A reviewer said that, in her experience, the confidence limit often is more stable to the different model choices than the central tendency estimate. Another reviewer expressed his support for the use of confidence limits, but wanted to challenge the point about the coverage of probability of the confidence interval. The variation or inaccuracy of the coverage of probability is due to the imperfections of the statistical methods. Use of a central estimate will give consistent coverage of probability, but that coverage is zero because it is for a single value. This leads to a situation where there is no variability and zero confidence. If that is the concern, responded a reviewer, one could compute a bootstrap distribution about the maximum likelihood estimate, which would provide a complete characterization about where the truth is likely to be. The first reviewer replied that several reviewers had suggested in their premeeting comments that the bootstrap method was the most robust way to calculate confidence limits. He mentioned some work he has been involved in that showed that the bootstrap method gave a more consistent estimate of confidence limits than more conventional ways, which resulted in a large degree of fluctuation. He said he agreed with earlier comments that there were drawbacks to the confidence limits not only in terms of methodology but also in terms of data limitations. For example, if a simplistic method is used to calculate a confidence limit, the LCL may be beyond zero. In that case, either the method has failed or the data are insufficient. The use of the bootstrap method for estimating the LCL of the BMD is more complicated than that for estimating a model parameter, because the estimation of BMD is typically an inverse estimation of a covariate. Fine-tuning of the bootstrap method is necessary and helpful. The work of Zeng and Davidian (1997) is particularly relevant. This reviewer provided references for the bootstrap method which are included in the last pages of Appendix G.

Another reviewer responded to a number of earlier comments. She stressed that it was very important to remember that confidence limits are not calculated on a single value, but rather on something that represents a range of values around that point. She agreed that one should distinguish between the method's use and the general concept of confidence limits, and added that one should also distinguish how the method is applied. She emphasized that in proposed methods to improve the approach for generating confidence limits, it was important to distinguish between single point estimates versus estimates over a range of values. A reviewer responded that he supported the use of best estimate, both for the POD and for comparisons. He said that the confidence limits that have been proposed are suitable for the purpose of informing decision-makers such as risk managers about the uncertainty or variability in the statistical procedures used and how those procedures are impacted by the variability of the experimental data. He agreed with the earlier suggestion of characterizing a best estimate and using a bootstrap procedure rather than a likelihood procedure to provide some idea of the variability. He recommended that the bootstrap be used to calculate confidence limits for informative purposes, but he did not recommend that the confidence limit be used as a POD for regulatory purposes. Another reviewer voiced agreement with the suggestions that the guidance should explicitly acknowledge situations in which zero is included in the range.

A reviewer suggested that, while it would be good for the guidelines to recommend the bootstrap procedure, it may not be necessary for the document to recommend whether to use a BMD or BMDL in the risk assessment, since this appears to be outside the document's scope. Another reviewer agreed with this.

A reviewer then showed overheads illustrating a well-behaved data set (see Appendix F). The third overhead was a bar graph showing the results of an evaluation of the data set using the BMDS software. The left-hand set of bars in the overhead show the best estimates (BMDs), the middle set of bars are the

confidence intervals (BMDLs), and the right-hand set of bars represent the p values. The fourth bar from the right shows a very bad p value, which actually represents a linear fit of nonlinear data. This would be rejected. The slide shows how all models give similar results with a well-behaved data set, and that the BMDs reflected by the best fit of the data are also reflected in the lower confidence intervals. Therefore, he said, nothing is gained by using the lower confidence interval, since they essentially provide the same information as the BMDs but in a different range of value. A reviewer reiterated her earlier point that she was uncomfortable focusing on a single BMD without considering the range around the values. Another reviewer said that a distributional characterization would capture the variability across models, but would not capture the performance characteristic of a given model. The performance of a given model would be captured by a bootstrap distribution. A reviewer said the BMDL can provide information about the data—for example, how scattered or tight the data are—but it will not provide information on whether it makes a difference what model one uses.

Dr. Haber then summarized points of consensus in the discussion as follows. Reviewers recommended that the guidance should generally note other approaches to determining confidence limits. The central tendency should be used when comparing values, but this depends on the application. Reviewers disagreed about whether confidence limits should be used for regulatory purposes, but this is outside the scope of the document. It should be noted that there is an interplay between the choice of uncertainty factors and whether the LCL or central tendency is chosen. Confidence limits are of interest, and a reporting of central tendency without confidence limits would be less complete than it should be. Dr. Haber noted that EPA had asked reviewers to provide references on the bootstrap procedure and any other methods for improving the accuracy of confidence limit coverage.

4.6 Discussion of Question 6

Reviewers then discussed the final question in session 2: *What additional concepts, if any, should be illustrated by an example?* Dr. Haber summarized the key premeeting comments regarding examples:

- Many reviewers suggested that the document should provide more basic examples, rather than illustrating relatively obscure things.

- Reviewers suggested that the examples should more closely follow the points they were designed to illustrate.

- Several reviewers asked that the use of BMD for epidemiological data be illustrated, especially using the hybrid model where there is a clear advantage over the NOAEL/LOAEL approach. However, one reviewer commented that the hybrid model should not be included because it is too experimental.

- An example with human risk factor data was requested.

- Reviewers noted that the examples emphasized the hybrid model over other approaches to continuous data, even though the text presents the hybrid model as simply one of the options. This discrepancy should be corrected.

- Reviewers suggested that when constraints are described, the document should clearly describe why something should be avoided.

- Reviewers requested an example illustrating the use of covariates.

- Reviewers suggested that the document should better relate the examples to the roadmap and the decision tree, and should do a better job of walking the reader through the examples.

- A brief example was requested to illustrate when the BMD fails.

- Reviewers felt that the examples should be better documented.

- A recommendation was made to remove the description of dynamic range for continuous data, as this is not a generally accepted approach.

A reviewer said that warning screens in the BMDS software give the impression that the hybrid model is very sensitive. He suggested that the BMDS software and the guidance document should be better aligned. EPA's Dr. Setzer clarified that the hybrid model is the least developed and most unstable of all the models in the software. The warning screens are connected with a beta test and should be there. He also said that the ultimate goal for the software is to drop the hybrid model as a separate model and add it as an option to all the other continuous models.

A reviewer said he had raised the question of how to distinguish between adversity and something that is unusual. Another reviewer responded that he thought that reviewers had generally agreed that implicit dichotomization is better then explicit dichotomization. He supported the hybrid model because it provides an approach that makes sense and is relatable to the approach for quantal data, which makes it easier to be consistent about describing endpoints and BMRs. He said the difficulty in knowing what the 10% represented had given continuous benchmark a bad name. The hybrid approach was helpful in overcoming that problem. Dr. Setzer clarified that, in the guidance, "implicit dichotomization" referred to the hybrid model. A reviewer responded that he supported this usage, but the document should be clearer about what is meant by the term "dichotomizing." It should clarify that the term always means the hybrid model and not explicit dichotomization, and should explain the rationale for this. Another reviewer expressed minor discomfort with the term "hybrid" as the choice for the model's name. He felt it made the model sound too tentative and thought EPA might be wise to reserve this term for other uses.

A reviewer mentioned that the IRIS database contained a few examples of chemical data sets for which benchmark had been used. A second reviewer suggested that these should be included in the document as real-world examples, even if they are not optimal examples. He also suggested the guidance should include an example of the State of California's use of BMD in a regulatory context, which Dr. Alexeeff had mentioned in his premeeting comments. The first reviewer responded that he was concerned about the use of the IRIS examples. He said they generally do not follow the guidelines, since they were created earlier in the development of the BMD approach. He recommended that the document include a recently published arsenic risk assessment by Moralis, which more closely follows the guidelines.

A reviewer asked the group whether the examples should include something that addressed the issue of additive versus extra risk. Another reviewer responded that in this case it may simply be sufficient to indicate there is more than one choice.

Reviewers then discussed the audience for the document. A reviewer asked whether the document should be basic enough to be understandable to someone who was relatively unfamiliar with the BMD

approach prior to reading the document. Dr. Setzer explained that it was never EPA's intention to have the guidance be a basic primer on BMD. The document was targeted for users who had some familiarity with the issues or at least had access to support staff who could provide that experience. He felt the issues were simply too technical to target the document for an uninitiated audience. A reviewer mentioned that the document was targeted to EPA risk assessors who had access to statistical support staff. She wondered whether EPA risk assessors would always have access to this type of support. If not, she said, then basic examples might be helpful for the target audience. Another reviewer added that risk assessors face this type of reality in his organization; he felt it would be useful if the document could provide guidance for this type of reader. A reviewer pointed out that as with so many EPA products, this guidance document would likely be used by risk scientists and others in the private and nonprofit sectors who would have various levels of expertise. While the guidance might not be able to be all-inclusive, it should at least point less experienced readers to other places where they could get help. Another reviewer pointed out that the document would inevitably become a support document for the BMDS software. For this reason, he recommended that the document include a basic example illustrating a step-by-step approach for each of the data types. More refined examples could also be included for advanced users. Another reviewer suggested that different examples could be created for users at different levels. EPA's Jeff Gift mentioned that, eventually, EPA may create some online examples in connection with BMD training the Agency may develop.

Reviewers then discussed ideas for additional examples for the guidance document. Suggestions included:

- An example showing how the bootstrap procedure is used, since this procedure is sufficiently developed to be included in the document.

- An example for quantal and continuous animal data, and for cancer and noncancer epidemiological data.

- An example of application to human data using the risk-factor approach (e.g., cardiovascular data, or use of human sperm count reduction data to project the risks of delays in male infertility or subfertility, which has been done for glycol ethers).

- Examples in which BMD is applied to randomly selected IRIS entries to illustrate the types of real-world issues that analysts are likely to face.

Two reviewers suggested that the examples should reinforce the fact that there are various options and choices to be made when doing a benchmark analysis. The examples should emphasize the process for developing a rationale for the choices to be made.

Reviewers then discussed concerns about how the guidance should handle epidemiological data. Some reviewers were concerned that use of epidemiological data was too complex to include in the document. They did not feel the guidance document should attempt to provide a complete description of the issues associated with use of epidemiological data. Notwithstanding this concern, said a reviewer, the use of benchmark dose modeling in epidemiology may be more important than any other use, because both exposure and response are continuous for epidemiological data. The BMD approach allows evaluation of actual individual exposure levels, whereas a NOAEL can be calculated only by artificially binning continuous exposure into discrete levels, such as 0 to 600 ppm-years, 600 to 1,200 ppm-years, etc., in which case the NOAEL depends on how one arbitrarily splits up exposure. He suggested that, in lieu of abbreviated or simplistic examples, the guidance could reference some acceptable published examples,

such as the arsenic example or the IRIS entry for carbon disulfide, which is based on human data. Another reviewer agreed that benchmark dose modeling for epidemiological data was complex, but said he hoped the document could include an epidemiological example even if it could not be extensive. He mentioned that the Agency is reviewing the ethylene oxide epidemiological analysis, which uses a BMD approach.

A reviewer said that, in lieu of epidemiological studies, one might do human clinical experiments, which are more analogous to animal studies. He pointed out that uncertainty and dosimetry arise in epidemiology and often are not tackled quantitatively. Also, there is the problem that the selection of controls or the comparison group may not be perfectly appropriate, giving rise to the healthy worker effect and the healthy survivor effect. These types of complexities could not be put into the guidance document in detail, but they could at least be mentioned. A reviewer agreed and suggested that even though the document could not go into detail on how to handle these issues, it could mention some of the advantages of using the benchmark approach for epidemiological analysis—for example, that a comparison group is not necessary, removing some of the problems related to the healthy worker effect. A reviewer thought it would be appropriate for the guidance to encourage the use of the BMD approach for epidemiological data, as long as the issues are clearly acknowledged. Another reviewer mentioned that, with epidemiological studies, the high-risk subgroup often is identified by a certain level of covariates. This can provide some very useful guidance, not only with respect to human populations, but also with respect to animal data. This reviewer also said that, given the complex issues regarding application of BMD to neurotoxicological data for example, it may be premature for the document to include guidance on this application.

Finally, a reviewer who was a member of EPA's Interagency Dosimetry Project Working Group commented on the use of dosimetry, including physiologically based pharmacokinetic modeling, with BMD. He said that if the dose-response modeling is being performed on an animal study, then as a general rule the dosimetric adjustments appropriate for the experimental animal should be used, rather than calculating human equivalent doses or concentrations. He said that in the old (1986) cancer guidelines, HEDs were calculated using a scaling factor of (body weight)$^{2/3}$ and the dose-response model was run with the HEDs. In that case it would not matter because the same factor was used for all exposure groups; thus the dose-response model would yield the same result whether the scaling was performed before or after the modeling. This is not necessarily the case, however, when there is a dose-dependent adjustment. An example of the preferred approach for using PBPK in BMD is provided by the EPA risk assessment for vinyl chloride. In this case, PBPK modeling was applied to linearize the dose-response by calculating the amount metabolized per kilogram liver in the animal exposure groups. This internal metric was used in the dose-response model to obtain the benchmarks for the animal bioassays, and the human version of the model was then used to convert the internal benchmarks to the equivalent human exposures. Another example of using dosimetry in BMD is the use of deposited dose instead of inhaled concentration for particulate exposures. In the RfC for nickel, for example, there was a different particle size distribution at each exposure concentration in the animal studies. Therefore, the EPA Regional Deposited Dose Ratio (RDDR) program was used to calculate the HEC for each animal exposure group. The HECs could be used in the dose-response modeling in this case because the dosimetric adjustment in the human was always the same. Thus the HECs would be directly proportional to the deposited dose in the animal, which is the metric that actually should have been used in the dose-response modeling. In general, then, the proper way to perform dosimetry in BMD modeling is to calculate the internal dose metric in each exposure group in the animal study, and use the dose-response model to estimate the BMD in terms of the internal metric in the animal. If a benchmark is needed in terms of administered dose or concentration (e.g., for comparison with NOAELs in other studies), then the appropriate dosimetric adjustment or PBPK model simulation can be performed to convert the internal

benchmark to the equivalent animal exposure. The reviewer said this was the approach recommended by the Interagency Dosimetry Project Working Group.

5. INTERPRETATION AND USE OF THE BENCHMARK DOSE

Discussion Leader: Lorenz Rhomberg

The third discussion session covered "Interpretation and Use of the Benchmark Dose." This included the following charge questions:

- Question 7: The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document's guidance, or would further discussion be useful?

- Question 8: What other comments do you have about the approach to benchmark dose described in the document? What changes to the approach would you like to see, and, very importantly, why?

Section 5.1 provides a summary of this discussion, prepared by the discussion chair, Dr. Rhomberg. Sections 5.2 and 5.3 provide a detailed record of the discussions. Section 5.4 summarizes the observer comments made during this discussion.

5.1 Chair's Discussion Summary

Question 7: The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document's guidance, or would further discussion be useful?

Section II.D of the guidance document lists the reporting requirements. These specify a comprehensive reporting of the data, analyses, and results of a BMD or BMDL determination, including presentation of alternative modeling approaches attempted, their outcomes, and the rationale for choosing among them.

In general, reviewers expressed approval of the guidance on reporting and documentation, both in their premeeting comments and in their discussions at the meeting. The guidance was seen as admirably thorough, and such documentation was seen as necessary in view of the multiplicity of possible analytic pathways allowed. The document appropriately stresses that individual analyses need to be carefully examined with regard to their appropriateness and statistical success in describing the data. The reporting requirements provide the means necessary to conduct such evaluation, to compare alternative models to one another, and to document the basis of conclusions made in choosing among data sets and modeling approaches. The mandated level of documentation is helpful in ensuring that all appropriate examinations have been done, and it is necessary to record the basis of decisions.

No comments or discussion called for dropping any element or for adding further elements to the list of reporting requirements, but there was some discussion aimed at clarifying what each element in fact calls for. Reviewers felt that, as a general principle, the reporting should be sufficiently thorough to enable the analyses to be repeated by other parties. Typically, this would include the provision of the dose-response data used, but reviewers recognized that, for a variety of reasons, it is not always possible to provide all raw data. When the input data are voluminous, it may be preferable to refer to how they can be accessed in electronic form rather than append them to the BMD analysis report itself.

Given the number of alternative approaches allowed under the guidance, reviewers felt that it is important to explain the rationale for choices that are made in the course of analysis. Several reviewers stressed that choice among models should be based on criteria of biological appropriateness as well as strictly statistical criteria. One reviewer noted that, in providing the mandated statements of rationale for choice among data sets or models, analysts should be encouraged to avoid "boilerplate" standard justifications that merely cite precedent and common practice; rather, analysts should provide thoughtful discussions of the criteria mentioned in the guidance.

Reviewers recognized that the documentation asked for in the guidance may be quite voluminous, especially when many data sets must be considered. To the degree that a well-considered decision tree or flow diagram can be established to sort among the possible analyses, as recommended elsewhere by reviewers, this burden can be reduced. Some reviewers recommended that the guidance be clearer as to whether the full set of documentation elements needs to be provided for every endpoint, data set, and model, or only for those that form the analyses eventually chosen as the basis for BMD or BMDL calculations. Reviewers recognized that, while EPA may be able to enforce such thorough documentation for its own analyses, it may be harder to achieve compliance with the full reporting in analyses conducted by other parties based on EPA's methodology.

It was pointed out that, to the degree that analyses focus on central estimates, is important to report confidence limits as well as to provide information about the precision of the estimates. Because BMDs based on quantal data differ in meaning from those based on continuous data, and because there are different ways to define a BMR, reviewers said that, to make such distinctions clear, it may be helpful for the guidance to specify a brief phrase that should be used when referring to a BMD. To the degree that such distinctions are captured in a revised terminology involving subscripts (e.g., BMD_{Q10} or $BMDL_{C05}$), this concern is obviated. Another comment pointed out that it would be useful to note whether the BMD represents an extrapolation up or down from the domain of observation, although it was recognized that this information would be provided by the mandated graphical display of data points, the fitted curve, and the calculated BMD and BMDL.

Question 8: What other comments do you have about the approach to benchmark dose described in the document? What changes to the approach would you like to see, and, very importantly, why?

To a great extent, commenters used this topic in their premeeting comments to reemphasize points they made earlier or to emphasize a particular comment as their primary concern. In particular, several reviewers used question 8 to reiterate their position on whether risk assessment should be based on BMD or BMDL. Since these opinions were discussed in previous sessions, they were not reopened for discussion in the session on question 8. The discussion leader merely mentioned during the session that these topics had reappeared in some reviewers' written comments on question 8. Among the topics treated in this way are those of the application of BMD methodology to epidemiologic data, the comments on percent of dynamic range as a basis for defining a continuous endpoint BMR, the issue of preference for the hybrid model over explicit dichotomization, the call for biological criteria in choosing among models, the suggestion that threshold models be included among the possibilities, and the call for more guidance in choosing among models.

Reviewers discussed whether the guidance should address use of unusually complex data when responses are multivariate in nature, when there are repeated measures on the same individuals, and when there are covariates. Some reviewers noted that the BMD procedure has been oriented primarily toward analyzing

single endpoints at a time, and that analysis of dose-response relationships comprising multiple endpoints assessed simultaneously is outside of the usual application. Other reviewers stated that, although such applications are new and the methodology may be complex, the BMD approach can be applied in principle and there are advantages to doing so. Several reviewers commented on carcinogenesis bioassays in which several tumor types may show elevation in the same set of animals. In these cases, tumors may be statistically independent across individual animals or they may show dependence (positive or negative correlation within animals), perhaps attributable to common individual-specific sensitivity factors (such as metabolic activity) or to immunological suppression of subsequent tumors in animals bearing a first primary tumor. Reviewers briefly discussed methods based on combining incidences versus separate analysis of tumor types and combining risks. They discussed the degree of analogy of this setting with multiple endpoints in a developmental toxicity setting. In general, reviewers felt that dose-response analysis of data sets with multiple endpoints is a developing field, and that application of BMD methodology to such settings, while important, is still too novel to be discussed in the guidance document.

There was discussion regarding the coverage probability when minima among several BMDLs are used, a point raised by several reviewers. It was noted that a minimum is more conservative than the nominal probability. Although some ways of pursuing corrections were briefly discussed, it was felt that the reliance on a tree or flow diagram to identify appropriate models, as reviewers suggested elsewhere, would reduce the incidence of this problem. Some discussion focused on the issue that, particularly in the case of a BMR defined as a percent of the maximum dynamic range, some formulations of continuous endpoint modeling result in a BMR that is not independently defined but becomes in effect a fitted parameter of the model, distorting its coverage vis-à-vis its nominal value. This situation does not arise when BMRs are defined independently of model results; reviewers' suggestion (expressed elsewhere) that the percent of maximum dynamic range not be used at the present time should avoid this difficulty.

Several comments called for more discussion of alternative dose scales, including use of pharmacokinetic modeling to estimate internal doses, and how such metrics should be applied to BMD analysis. The discussion of this topic from an earlier session was briefly reiterated, and it was generally agreed that animal dose-response curves need to be fitted to dose estimates appropriate to those animals, with conversion to human equivalent doses occurring subsequently.

Reviewers noted that two alternative measures of risk over and above background are commonly used—so-called extra risk and additional risk over background. The guidance employs extra risk without discussing its choice. Reviewers generally agreed that, whatever stance is taken, it should be done explicitly and a rationale should be provided. Both measures can characterize risk over background; the distinction only becomes pronounced when background rates are pronounced. Reviewers noted that when extra risk is used it is frequently inappropriately applied to human populations without an allowance for the magnitude of human background, resulting in incorrect projections of human population impact. Reviewers also noted that routine use of additional risk would incorporate the correction into the characterization of risk itself. On the other hand, extra risk calculations would be more readily compared to existing characterizations of risk for other agents, and the projections to human populations can be done correctly if the human background is included in the calculation.

In a brief discussion, reviewers noted that study designs optimized for calculation of BMDs differ from those chosen to best serve other purposes. In specifying how BMDs are to be determined, the guidance implicitly makes recommendations about study design. One commenter mentioned that EPA should gauge these recommendations' consistency with other Agency guidance on toxicological testing.

Finally, there was brief discussion of a comment from a few reviewers that features of a dose-response relationship other than the BMD may be of interest to the risk assessment process—notably, the slope of the fitted curve at the point of the BMD—and that the guidance might encourage the characterization and provision of such information. An EPA employee clarified that steps were already underway to provide for calculation of such a slope.

Reviewers generally agreed that there were no other major additions they wished to suggest for the BMD guidance.

5.2 Discussion of Question 7

Dr. Rhomberg began the discussion by reviewing the key reviewer comments concerning charge question 7: *The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document's guidance, or would further discussion be useful?*

- Many reviewers approved of the way the document handled reporting requirements and had no comments on this question.
- Some reviewers said there should be sufficient information accompanying a BMD analysis to reproduce the analysis.
- Many reviewers commented that documentation of an analysis should include a clear statement of the rationale for the choices made in the analysis.
- Some reviewers noted that the volume of information requested is high and may be too complex. Some reviewers were unclear whether this information was needed for every data set or just the endpoint that becomes the critical endpoint, and for every model considered or just the model that was finally selected.
- Reviewers raised the issue of whether users would comply with the documentation guidelines.
- Several reviewers mentioned that the rationale for choice should be both biological and statistical.
- Some reviewers hoped that a way could be found to encourage users to avoid using boilerplate text in the rationale.
- During earlier discussions, it was mentioned that confidence limits should be reported when central estimates are calculated.
- A reviewer felt it important to document whether an upward or downward extrapolation from the domain of observation was used to calculate the BMD.
- A reviewer also raised the issue of whether some notation indicating the main features of a BMD (such as whether it is quantal or continuous, and at what point the response was taken) should accompany statements of the BMD value.

This latter issue—including information about the main features of a BMD together with statements about the BMD's value—is important, said a reviewer. When this information is not clearly reported, it can require work on the part of the reader to locate the information. Another reviewer mentioned that the old cancer guidelines had asked users to indicate the classification of the carcinogen whenever the potency was stated, but in his experience, this rule was rarely followed.

Reviewers then discussed the decision tree. A reviewer suggested that summary graphics and tables that reviewers had earlier recommended be added to the document could be incorporated into the decision tree. This would consolidate summary information into a single location; the decision tree would be a good place to provide guidance on step-by-step use of the models and the various choices to be made. A reviewer agreed that the decision tree should be expanded, since it currently is minimal. This would serve two functions. It would help users develop a rationale for the choices they had to make in performing a BMD analysis, and it would provide a framework for reporting guidelines. This section could also serve as a map to the document by directing readers to the sections that deal with each of the issues in more depth. In that way, it would help both basic and sophisticated users. It would help make the document more transparent and user-friendly, and it would also be a useful guide for someone tasked with reviewing a BMD analysis. A third reviewer supported the idea of expanding the decision tree to make explicit the decisions or choices that must be made. A fourth reviewer was concerned that the term "decision tree" could be confusing, since it could suggest something that carries probabilities on its branches. Reviewers agreed and decided "decision tree" should be renamed "analysis tree."

Reviewers then discussed concerns about the quantity of data to be reported with an analysis. A reviewer suggested that the document should indicate that an electronic version of the data set, such as an Excel or PDF file, would be acceptable or preferable to a hard copy. Another reviewer agreed with this concern and suggested that the document could simply indicate that access to the raw data should be provided for. A third reviewer agreed with the idea that raw data should be reported, but wanted to make sure this would not be a rigid requirement. He pointed out that the document provided guidance about using data sets that do not even have standard deviations around the means for anything but the control group. Raw data would not be available in these situations. Also, he said, epidemiologists are often resistant to sharing data, in part because of confidentiality concerns or requirements. This is a real-world problem that should not be contradicted by too great an emphasis on data reporting. A reviewer agreed that the guidelines should encourage, but could not require, data sharing. Another reviewer said that even though he would ideally like to see detailed data—for example, statements about what went into the study design, whether dose-range studies were done, what the rationale for dose placement was, and so on—he recognized that this would be hard to do for all bioassays.

A reviewer pointed out that, for some applications, such as using BMDs to develop RfDs, that type of raw data is accessible to the risk assessors who are calculating the RfD. She said that she interpreted "raw data" as meaning the input data rather than the output data. In her experience doing BMD analyses, she often is asked to run every model in the BMDS software. This generates a huge output. So, in her reports, she tends to include only the output for the models that were judged useful. If EPA adopts the hierarchy of models, as reviewers suggested earlier, this would certainly help reduce the volume of output that would need to be reported.

A reviewer pointed out that because of the volume of raw underlying data, the paper version of a report would inevitably have to include some digested version, but digesting this type of information often makes it impossible for anyone else to reproduce or expand the analysis. He therefore suggested that the

document should indicate that an undigested electronic version should be included in the documentation. Another reviewer agreed that reproducibility of analysis is critical. He pointed out that it is also helpful to include significant digits in intermediate results, since this will make it easier for someone else to reproduce the results without introducing compounding of rounding error.

Another reviewer was still concerned that the guidance on reporting could result in voluminous paperwork. He pointed out that the BMDS software for dichotomous data has eight models. After exploring various choices, such as which polynomial to work with and whether to truncate the data at the high end, one can end up with a huge pile of justifications. Another reviewer agreed that it was important to avoid generating unnecessarily voluminous documentation, which would result if the guidance asked users to document every endpoint, data set, and model used. He said he hoped the hierarchy of models would help address this issue. EPA's Dr. Carole Kimmel clarified for reviewers that EPA likely will be calling for calculation of more reference values in the future for chemicals than simply chronic RfDs or RfCs. For example, there likely will be acute and short-term reference values. This may mean that more endpoints will need to be modeled for a given chemical in order to provide the information needed to calculate all these values. A reviewer pointed out that pharmacokinetic analysis can also increase the output, since EPA often wants a NOAEL or BMD analysis run with and without pharmacokinetics.

5.3 Discussion of Question 8

Dr. Rhomberg reviewed reviewers' comments on charge question 8: *What other comments do you have about the approach to benchmark dose described in the document? What changes to the approach would you like to see, and, very importantly, why?* He noted that many, though not all, of these comments had also been made in response to previous charge questions and had already been covered in the discussions. Comments concerned the following issues:

- More complex data. The comments addressed multivariate, repeated measures, and covariates issues. A general theme was that the guidance document should say more on these issues.

- BMDLs and confidence limits. The comments addressed such issues as multiple comparisons; use of joint likelihood in situations where there are several similar BMDLs; and the idea that since the BMR is really a parameter of the continuous model, the uncertainty in its estimate is influenced by the fact that the BMR itself is being fitted.

- Application of BMD methodology to epidemiological data.

- Model choice. Comments included a preference for the hybrid model over dichotomization; use of biological criteria as well as statistical criteria for model selection; the desire for more guidance on model choice to reduce the potential for model shopping and the pressure to run all models in order to produce a result that will be accepted as legitimate; and the suggestion to include consideration of threshold models.

- Dose scale.

- Concerns about using the percent of the dynamic range approach as the basis for defining a continuous endpoint BMR.

- Extra versus additional risk.

- The need to align the BMD guidance with testing guidelines.

- Reporting features of the curve and the slope. This suggestion was made by way of explication of how those features can influence the lower limit and with regard to MOEs and how big they should be.

Dr. Rhomberg suggested that reviewers focus their discussion on topics they had not already discussed. Reviewers began by discussing the issue of extra versus additional risk. A reviewer pointed out that both extra and additional risk are defined mathematically, so the issue is not with their definition but rather how people use extra risk. To get the number of individuals at risk of developing a response to an environmental agent, one should discount the background rate by multiplying the extra risk by the population as a whole *minus* all those individuals who would exhibit the response anyway, without exposure to the agent. Most people, though, simply multiply the entire population by the extra risk. Another reviewer agreed that extra risk can be misused, but pointed out that the background rates with animal studies usually are so low that this error does not make a significant difference. A third reviewer said that when there is a high background rate in the animals, there is less concern about, for example, a 10% absolute increase over that background rate than there would be if the animal had no tumors and there was a 10% tumor incidence. For example, if the animals have an 80% background rate, then a 90% observed tumor rate in exposed animals would mean the extra risk was 50%. This is less of a concern than a 10% tumor rate in exposed animals that have a zero background tumor rate. For this reason, this reviewer was unsure about the value of using the term extra risk. The first reviewer responded that added risk was particularly important and appropriate to use when comparing across chemicals in terms of their risk to human populations.

A reviewer said he preferred to use the term extra risk because it better addressed the issue of whether the tumor incidence happens only in exposed individuals that would otherwise not develop the tumor, or whether the induced tumor actually "co-occurs" in animals that would also have developed the tumors over time without exposure. Extra risk can be helpful in eliminating the double counting that can occur if one assumes that observed tumor incidence is entirely independent of background when if fact it is not. However, he acknowledged that when extra risk is applied to another population, the correction for background rate should be made and often is not. Another reviewer pointed out that traditional probit calculations are usually done with extra risk. He said he tended to use extra risk for calculations, but would not object to a recommendation to use additional risk to report socially relevant consequences. A third reviewer supported the idea that the guidance could specify additional risk as a form for reporting risk. Dr. Rhomberg suggested that reviewers could recommend to EPA that the Agency should consider this issue and provide some explicit guidance on it, rather than leaving it entirely to the user's discretion. At the suggestion of a reviewer, Dr. Rhomberg agreed to write a paragraph summarizing the views on this issue (see Section 5.1).

Reviewers then discussed the issue of multiple comparisons. A reviewer noted taking the minimum of 95% confidence levels does not yield a 95% confidence level for the group. Regarding the comment that BMR is a parameter in a continuous model, he said that when one is computing variability in or evaluating the performance of a procedure, it is important to start at the beginning, since part of the performance includes identifying a cutoff point or a point on which to focus the analysis. He said this type of complete analysis could be done using the bootstrap procedure. He also said that the joint likelihood approach was

reasonable to use for analyses in which data sets may have commonalities such as the slope or the model. The joint likelihood approach could be used to essentially fit all the data sets at the same time under a restriction of that common parameter.

A reviewer expanded on the issue he had raised in his premeeting comments—that the BMR resulting from continous data is in effect a parameter of the model. He said he meant that if the response level is a function of the concurrent data, then it is a statistical quantity or random variable that should be accounted for in the analysis. He thought this was not adequately recognized in the guidance. He said the BMDS software does make it clear where assumptions about how to define the BMR are taken into account as random variables; however, there are options (such as the absolute deviation on the point options) in which the user can use a fixed value to compute the BMR from continuous data. He recommended that the guidance should clearly point out that if the value used in that calculation was derived from the concurrent data, then there will be variability in that parameter that is not explicitly being taken into account. He supported the previous reviewer's comment about the bootstrap method, saying it could automatically take these types of considerations into account if the bootstrap procedure is programmed so that the calculations done to create those values are explicitly stated in the bootstrap process.

Another reviewer commented that the guidance indicates that other BMR values can be reported in addition to the 0.1 value, which should always be reported. He said that when reporting LCLs on several BMR values, it is important to correct for multiple comparisons, because there will not be 95% coverage probability if the confidence limits are considered as a set. There are various statistical ways to do this, he said.

A reviewer returned to the issue of combining effects. He said that in teratology there are a number of conventions for when and how to combine effects. In carcinogenesis, it is important to model for different apparently unrelated tumors separately, but one can then do a combined analysis of cancer slope that takes both sites into account. In doing so, one should add the effects probabilistically rather than adding the upper confidence limits, for example. He said that while counting the animals having any one of several endpoints is statistically acceptable, it is not acceptable from a mechanistic standpoint: it implies a common causative process for both tumor types. Another reviewer countered that tumors that may at first appear to be separate are not always independent, perhaps due to animal susceptibility or some underlying mechanism. Thus it is not always correct to assume that tumors at different sites are independent. The first reviewer agreed there could be dependencies, for example if an animal has a very active liver that creates a more active metabolite that exposes both organs, or if the growth of a tumor in one location makes the animal a poorer host for development of tumors in the other location. A third reviewer agreed that independent modeling was not the best approach and that a combined response may be a reasonable alternative. A fourth reviewer agreed with previous comments about independence and said that rigorous multivariate modeling, where appropriate and possible, has advantages over assuming independence, because one can test for dependence as part of multivariate assumptions.

A reviewer cautioned the group against pushing the envelope too far ahead of current knowledge on combining endpoints. He said the issue of combining endpoints was a good example of where it is critical for statisticians and toxicologists to work together. He recommended that the guidance should not be overly prescriptive about when and how things should be combined, but should indicate the importance of approaching the issue with a biological rationale. Another reviewer agreed and said that developing plausible and mathematically tractable models for combining endpoints is very challenging. A third

reviewer agreed with the previous two comments and added that one must consider severity when combining endpoints.

A reviewer expressed concern that combining endpoints might be beyond the scope of standard benchmark methodology, which he thought was developed primarily for addressing single endpoints. Another reviewer agreed that benchmark was a simple concept, but he pointed out that it allows for much more sophisticated use of the full toxicological database than the NOAEL approach. He therefore felt that the methodology should not be constrained to single endpoints when application to multivariate analysis appears to be appropriate. A third reviewer agreed that benchmark analysis is amenable to more sophisticated applications, such as analysis of multiple endpoints. He said the guidelines should point this out and recommend that this type of analysis be considered when multiple endpoints exist. He said he was personally very interested in hierarchical Bayesian approaches. The first reviewer agreed, but said he just wanted to make sure that the document fully addressed the basic application of the BMD methodology for single endpoints. A fourth reviewer agreed that the document should clearly identify how to apply the methodology to single endpoints, but he disagreed with the earlier point that BMD was a single-endpoint technology. He said it is a general technology that can be used in many other situations, such as multivariate analysis, and there is a growing literature on how the BMD technology can be applied in more complex settings. However, these other applications are not sufficiently advanced for the document to provide detailed guidance on them at this point. Instead, the document should acknowledge that the state of the art for these applications is advancing and should point readers to literature on these topics to help illustrate the flexibility of the BMD approach and encourage use of more complex applications when appropriate. At this point, reviewers ended their discussion of charge question 8 and the meeting was opened for observer comment.

5.4 Observer Comment and Subsequent Discussion

Three observers commented during this session.

Nicole Cardello, Physicians' Committee for Responsible Medicine: Ms. Cardello explained that she was an environmental scientist representing both the Physicians' Committee for Responsible Medicine and People for the Ethical Treatment of Animals. She said members of these organizations were very concerned about using the lower BMDL for regulatory purposes. She noted a number of objections she had heard reviewers make during their discussions regarding use of the BMDL. These included the statistical complications involved, the lack of any added value, and animal welfare concerns. She said the only justification she had heard for using the BMDL was that "it rewards better experiments." As had been pointed out during the discussion and in reviewer premeeting comments, toxicologists may perceive so-called better experiments to mean more dose groups and therefore more animals. In fact, since the width of the confidence intervals is inversely related to the square root of the sample size, toxicologists may feel pressure to experiment on more animals to try to tighten their confidence intervals. Further, she was alarmed at the idea that the use of human epidemiological data would either not be addressed in the document or would be superficially addressed. Little has been said during this workshop about the value and application of human data. Risk assessments involve many assumptions and are plagued with many uncertainties, which can be dramatically reduced if interspecies extrapolation can be eliminated. She said that human data are the gold standard, so it was very troubling to see an exclusive focus on animal data. She said that EPA's bias toward using animal data and encouraging additional use of animals is in stark contrast to the NIEHS implementation guidelines of the 1993 NIH Revitalization Act, which state that agencies with regulatory programs should reduce both the number of animals and reliance on animal tests. She concluded by saying that this use of the lower

BMDL for regulatory purposes was just another example, in her opinion, of EPA moving in completely the wrong direction.

Jeff Gift, EPA National Center for Environmental Assessment, Research Triangle Park: Dr. Gift explained that he currently is the project leader for development of the BMD software. He reviewed some of EPA's plans for the software. Specific changes planned or being considered include:

- Incorporation of the hybrid method into all the continuous models, as mentioned earlier by Dr. Setzer.
- Possible addition of a "run all" button, which would enable the user to run all the models.
- Possible incorporation of a threshold parameter.
- Development of Internet training on the BMD methodology. This would be particularly helpful because the examples in the training could be dynamic and modified, as needed, in response to future developments.
- Modification of the software to correspond to any change EPA makes in the guidance regarding reporting of raw data in the output from BMD analysis.
- Consideration of possible changes in the guidance for a BMD based on neurotoxicity data. Dr. Gift mentioned that he was working on a project with Dr. Zhou regarding data from neurotoxicity studies, in which there often are few animals but repeated measures of continuous parameters. In this type of situation, there can be different BMDs depending on when the measurements were taken.
- Incorporation into the software of more robust methods for calculating confidence limits for BMDLs.
- Addition to the cancer model of a feature that would report the curve slope and confidence around that at the POD.

He acknowledged an earlier request by a reviewer for examples about when it was not appropriate to use a BMD analysis. He also mentioned that categorical regression might be one possibility for severity grade data. He said that another use for BMD calculations that had not yet been mentioned was to provide cost-benefit input to economists involved in risk management.

Dr. Setzer added to this comment that EPA also hopes over the long term to make the BMDS software more amenable to use with epidemiological data. He said that most of the models currently only work with aggregated data, for which one needs a small number of dose groups with many replicates per dose group. This was going to be fixed, he said. He also said that while the Agency would like to include covariates, he shared many of the concerns about modeling of covariates for epidemiological data that had been voiced at the workshop, and he was not sure if, when, and how covariate modeling would be manifested in the software. EPA was also considering adding a time-to-tumor model, he said.

David Gaylor, Sciences International, Inc.: Dr. Gaylor commended the reviewers and thanked them for their input. He said he thought the document would be much improved by their recommendations.

Reviewer Discussion

Following the observer comments, reviewers continued a brief general discussion. In response to Dr. Gift's comments, a reviewer suggested that EPA also consider adding time-dependent dosing and bootstrapping. Because bootstrapping is simply reoptimizing, he said, EPA would have to change the reporting guidance to include reporting a distribution of results, but otherwise bootstrapping would fit within the optimization capabilities already addressed in the software. References on the bootstrap method provided by this reviewer after the workshop can be found at the end of Appendix G.

A reviewer then responded to some of the issues raised by Ms. Cardello. She recalled a remark made in earlier discussions that one might be able to decrease the amount of uncertainty factors used when calculating an RfD or RfC if one could narrow the confidence limits on the value. She said this idea contradicts one of the arguments for going to BMD in the first place, which was that additional studies might not be necessary if the BMD model was accepted as input for regulatory decision-making. A huge benefit of the BMD approach is that it allows movement away from fixation on a single point that has to be tied to experimental design, and this can help to minimize inappropriate use of animal studies. The reviewer who made this earlier remark clarified that he had been referring to an approach used by the California OEHHA. By its nature, the BMD is a more precise estimate of a no-effect level at the 5% level than a NOAEL. Since both the BMD and BMDL are a reproducible estimate of some low level of incidence, whereas the NOAEL might not be, OEHHA risk analysts had decreased the uncertainty factor by a factor of three in general. However, he emphasized, this was not a specific reward for doing studies that decrease the confidence limit. The first reviewer responded that it was nevertheless important from an animal usage standpoint to keep in mind that one advantage of a BMD is the idea that one can calculate a BMD from a study that does not have a NOAEL, which decreases the need to keep performing studies in order to obtain a NOAEL. The second reviewer responded that another way in which the benchmark approach decreases animal usage is that, unlike the NOAEL/LOAEL approach, it gives investigators no incentive to keep performing studies with large groups of animals to identify ever-lower LOAELs. A third reviewer agreed that the use of benchmark studies gets around having to repeat studies for which there is no NOAEL. He felt this was a very important feature of the BMD methodology.

A reviewer said that Ms. Cardello's comment reinforced his earlier point that calculation of a BMD as a best fit is the most robust estimate of the best fit of the data, because it is based on the means of the data and is independent of sample size. He agreed with the first observer that working off the confidence limits or BMDL rewards more dose groups and more animals per dose group.

A reviewer said that he had recently participated in a project run by EPA's Dr. Margosches to reduce animal usage in the up-and-down procedure, which is used in estimating LD_{50}s. He said that a proposal by EPA to add a supplemental procedure to the up-and-down procedure to get a benchmark as well was rejected by committee members, in part because of the additional number of animals that would be required. Also, in his experience in other deliberations with science panels, suggestions about doing more studies have often been rejected because of concerns about animal usage. He thought it likely that, in the future, there will be increasingly greater concern about animal usage and more restrictions on the types of studies performed. That is why he supports the development of mathematical approaches that will enable risk analysts to better synthesize and make use of the data that are already available from existing studies.

Dr. Margosches said that, with regard to confidence limits, she did not think the reviewers had said they should not be looked at, but rather that properly calculated confidence limits can provide valuable information about a study. She said the up-and-down panel that Dr. Alexeeff had been participating in would likely be seeing a proposal for calculating a confidence limit for an LD_{50} from the available data.

A reviewer responded that she wanted to reiterate an earlier comment she had made: that she does not have a strong preference for using the central tendency versus the confidence limits, but she hoped that whatever choice was made, there would be recognition of the interplay between that value and uncertainty factors. For example, if one uses the central tendency, then one is predicting a 10% risk in the sensitive subpopulation in humans unless some additional factor is used to account for the fact that one is using the central tendency. Conversely, if one is using a LCL, there is some additional measure of protection, so one is not predicting an exact risk number, but this makes the result more like the NOAEL. This is an argument for the earlier suggestion of actually predicting a risk and then bringing it down to some societally acceptable value. A reviewer responded that many animal data sets, particularly concave ones, pose a problem for calculating LCLs and there often is valid concern about the uncertainty of where the effect occurs. Since the dose-response is not decreasing at that point, one cannot assume that the NOAEL is within a factor of 10 of the LOAEL. In such cases, the benchmark approach can be helpful because it can help determine what dose is appropriate to use to determine whether this is a health estimate or not. At this point, reviewers ended their discussion and proceeded to summarize the key points made at the workshop (see Section 6).

6. FINAL WORKSHOP SUMMARY

For the final segment of the workshop, the chair, Dr. Park read a list of key points he noted during the discussions. He asked reviewers to let him know if they disagreed with any of these summary statements of if they felt an important point was missing. The key points included:

- Reviewers generally felt this was a good document, and they were supportive of the software.

- Reviewers grappled with the bounds of this document, the scope of the review, and the interaction with, in particular, the use of the process—for example, what happens to a BMD after it is calculated?

- Reviewers discussed whether the document needed to defend its support of the BMD methodology. They generally agreed that the discussion on this topic in the document was appropriate: the BMD approach is a relatively new concept, so an explanation is needed about why EPA is using it.

- A continuing theme was that the document should be more user-friendly. For example, it needs tables of defaults and choices, more examples, and step-by-step process descriptions. It also needs to be made more consistent. The target audience for the document is not primarily statisticians, although they make up a subset of the target audience.

- Reviewers were concerned about model shopping, and generally recommended that EPA establish a hierarchy of models. Some reviewers felt that it should be a strong hierarchy, while others preferred a weaker hierarchy. The California process was cited as an example of a strong hierarchy.

- Reviewers felt there should be some discussion of the operating characteristics of the procedure. EPA has discussed this in the past, and reviewers suggested that perhaps the guidance should include a brief reference to it.

- Looking at all the BMDs within 10 times the NOAEL may not be sufficient, considering that different endpoints may have different cross-species dosimetry and that the use of the BMD may differ depending on the endpoint. So while 10 times is a reasonable rule of thumb, it should not be presented as a hard guideline.

- Related to this, it was recommended that BMD modeling be done on animal doses and then converted at the end, rather than converting to human equivalent doses before doing the BMD modeling.

- Significant concern was expressed about the use of different BMRs. A recommendation was that the 10% response should always be calculated as a common point of comparison. Then, higher or lower responses may be appropriate depending on the robustness of the data. One opinion was that only the 10% response level should be used. Though reviewers had a divergence of opinion on this topic, they generally agreed that for quantal data the 10% level should always be calculated as a minimum and other responses could be calculated as appropriate for the data. Also, the specific response level needs to be considered when the BMR is used for regulatory purposes.

- A reviewer suggested that, since application of the BMD is outside the scope of the discussion, the point about quantal data could be reworded to say that the 10% response could be used for reference within the data set, and that a range of opinion was expressed about what should be used as the POD. Another reviewer added that the group had agreed that 10% could be used for dichotomous data and that a one standard deviation criterion could be used for continuous data as a point of comparison, but these were not necessarily the PODs for regulatory assessment. For the POD, both the percent and the definition of the BMD would change.

Dr. Park then continued with his summary:

- Reviewers were uncomfortable with the selection of the lowest BMDL among a group of BMDLs when they differ by more than three-fold. In this case, selection of the lowest BMDL may not generally be appropriate, and bootstrap confidence intervals may be more appropriate. This relates back to the issue of hierarchy of models and to the issue of model shopping.

- There was a significant divergence of opinion on whether the POD should be from BMD or BMDL. The choice of this decision interacts with the uncertainty factors to be used in the resulting assessment.

- Reviewers recommended that the guidance have more discussion of dropping higher doses when they are not fitting as well as lower doses, much as is done with the multistage model process for cancer.

- The group supports the use of the BMD methodology for epidemiological data, but appropriate application and pitfalls should be more fully discussed.

A reviewer interjected that the group also had questioned whether this document was the place to fully discuss the application of epidemiological data, because of its complexity and state of development.

Dr. Park continued with his summary:

- The fraction of dynamic dose range is a concept that does not currently belong in the document unless more work is done to justify and support it.

- Regarding reporting, concern was expressed that, without a hierarchy of models, presentation of the results of the BMD process would be voluminous.

- Future research topics for application of the BMD process include use of covariates, repeated measures, nested data, multivariate data, Bayesian approaches, and the joint likelihood of numerous BMDLs. These subjects should be acknowledged as future potential uses.

- Reviewers expressed a preference for hybrid models over dichotomization.

- Additional risk and extra risk are both useful, but when extra risk is used the results are sometimes incorrectly applied to the human population. Additional risk is more relevant to the usual method of counting cases in the human population.

- The BMD process helps reduce use of animals compared to the NOAEL/LOAEL approach, because it obviates the need to repeat studies that did not result in the NOAEL.

Dr. Park suggested that latter point should be mentioned in the document as one of the advantages of the BMD approach. A reviewer added that the document should also mention the advantage that the BMD approach has a more consistent definition in noncancer epidemiological studies than the current characterization approaches. He said he hoped the document would encourage the use of the BMD approach for epidemiological data for noncancer endpoints as an alternative to the conventional NOAEL/LOAEL approach, which requires artificially binning the exposure groups. There is a bias inherent in how the data are binned, which is obviated by the use of BMD, he said. Another reviewer responded that he would prefer that the document not advocate this as the only approach for noncancer epidemiological data. He also pointed out that BMD can reduce animal usage if applied properly. In particular, it would be important to design studies properly in the earliest stages of evaluating a chemical. Going back and doing a study designed to use fewer animals after other, less well-designed studies have already been performed on the chemical would use more animals.

Dr. Park asked reviewers if he had missed any important points in his summary. A reviewer said he thought that reviewers felt the document needed more definition of outliers. Another reviewer agreed: without better definition, users may discard an outlier assuming it is a bad data point or a mistake when in fact it just happens to disagree with the others. Dr. Haber reviewed her notes from the discussion in Session 2. She said there was general agreement that the guidance should indicate that a modeler should not automatically choose the lowest BMD/BMDL, but otherwise reviewers did not recommend prescriptive guidance. Reviewers had also noted that the outlier issue will be obviated if EPA establishes a hierarchy of models. Dr. Park said this was an important point that should be captured in the summary of key points from the workshop. At this point, reviewers agreed they had reached the end of their discussions. Representatives of Eastern Research Group and of EPA thanked the reviewers for their input, and the workshop was adjourned.

APPENDIX A

LIST OF REVIEWERS

Peer Review Workshop on the Benchmark Dose Technical Guidance Document

Holiday Inn Capitol
Washington, DC
December 7–8, 2000

List of Reviewers

***George Alexeeff**
Deputy Director for Scientific Affairs
Office of Environmental
Health Hazard Assessment
California Environmental Protection Agency
1515 Clay Street - 16th Floor
Oakland, CA 94612
510-622-3202
Fax: 510-622-3210
E-mail: galexeef@oehha.ca.gov

Kevin Brand
Associate Researcher
Department of Epidemiology
and Community Medicine
University of Ottawa
Ottawa, Ontario K1H-8M5
Canada
613-946-9862
Fax: 613-941-4546
E-mail: kevin_brand@hc-sc.gc.ca

Paul Catalano
Associate Professor
Department of Biostatistical Science
Harvard School of Public Health
Dana-Farber Cancer Institute
44 Binney Street
Mayer Building - Room 222
Boston, MA 02115
617-632-2441
Fax: 617-632-2444
E-mail: pcata@jimmy.harvard.edu

Harvey Clewell
Senior Project Manager
KS Crump Group
ICF Consulting
602 East Georgia Avenue
Ruston, LA 71270
318-242-5017
Fax: 318-255-4960
E-mail: hclewell@icfconsulting.com

George Daston
Research Fellow
Miami Valley Laboratories
Proctor & Gamble
11810 East Miami River Road
Ross, OH 45061
513-327-2886
Fax: 513-627-0323
E-mail: daston.gp@pg.com

Elaine Faustman
Professor
Department of Environmental Health
Institute for Risk Analysis & Risk Communication
School of Public Health & Community Medicine
University of Washington
4225 Roosevelt Way, NE - Suite 100
Seattle, WA 98195-6099
206-685-2269
Fax: 206-685-4696
E-mail: faustman@u.washington.edu

***** Discussion Leader

Clay Frederick
Senior Research Fellow
Toxicology Department
Rohm and Haas Company
727 Norristown Road
Spring House, PA 19477-0904
215-641-7496
Fax: 215-619-1621
E-mail: cfrederick@rohmhaas.com

***Lynne Haber**
VERA Program Manager
Toxicology Excellence for Risk Assessment
1757 Chase Avenue
Cincinnati, OH 45223
513-542-7475, Ext. 17
Fax: 513-542-7487
E-mail: haber@tera.org

Dale Hattis
Research Professor
Center for Technology,
Environment and Development
Clark University
950 Main Street
Worcester, MA 01610
508-751-4603
Fax: 508-751-4600
E-mail: dhattis@aol.com

"Colin Park
Consultant
3352 Bayshore Boulevard, NE
St. Petersburg, FL 33703
727-527-2965
Fax: 727-527-5685
E-mail: crewcolin@aol.com

***Lorenz Rhomberg**
Principal Scientist
Gradient Corporation
238 Main Street
Cambridge, MA 02142
617-395-5000
Fax: 617-395-5001
E-mail: lrhomber@gradientcorp.com

Robert Sielken, Jr.
Vice-President
JSC Sielken
3833 Texas Avenue - Suite 230
Bryan, TX 77802
979-846-5175
Fax: 979-846-2671
E-mail: sielkeninc@aol.com

William Slikker, Jr.
Director, Division of Neurotoxicology
National Center for Toxicology Research
Food and Drug Administration
3900 NCTR Drive
Jefferson, AR 72079
870-543-7203
Fax: 870-543-7745
E-mail: wslikker@nctr.fda.gov

R. Webster West
Associate Professor
Department of Statistics
University of South Carolina
Columbia, SC 29208
803-777-3792
Fax: 803-777-4048
E-mail: west@stat.sc.edu

Yiliang Zhu
Associate Professor
Department of Epidemiology & Biostatistics
College of Public Health
University of South Florida
13201 Bruce B. Downs Boulevard (MDC 56)
Tampa, FL 33612-3805
813-974-6674
Fax: 813-974-4719
E-mail: yzhu@hsc.usf.edu

" Workshop Chair
* Discussion Leader

APPENDIX B

LIST OF OBSERVERS

Peer Review Workshop on the Benchmark Dose Technical Guidance Document

Holiday Inn Capitol
Washington, DC
December 7–8, 2000

Final List of Observers

Leila Barraj
Director
Statistical Services
Novigen Sciences, Inc.
1730 Rhode Island Avenue, NW - Suite 1100
Washington, DC 20036
202-293-5374
Fax: 202-293-5377
E-mail: lbarraj@novigensci.com

Michael Bolger
Risk Assessment
U.S. Food & Drug Administration
200 C Street, SW
Washington, DC 21401
202-205-8705
Fax: 202-260-0498
E-mail: mbolger@cfsan.fda.gov

Marilyn Brower
Risk Assessment Forum
U.S. Environmental Protection Agency
Ariel Rios Building
1200 Pennsylvania Avenue, NW (8601 D)
Washington, DC 20460
202-564-3363
Fax: 202-565-0062
E-mail: brower.marilyn@epa.gov

Richard Canady
Center for Food Safety and Applied Nutrition
U.S. Food & Drug Administration
200 C Street, SW (HFS-308)
Washington, DC 20204
202-205-0136
Fax: 202-260-0498
E-mail: rcanady@cfsan.fda.gov

Nicole Cardello
Research Coordinator
Physicians Committee for Responsible Medicine
5100 Wisconsin Avenue, NW
Washington, DC 20016
202-686-2210
Fax: 202-686-2216
E-mail: ncardello@pcrm.org

Ernest Falke
Senior Scientist
U.S. Environmental Protection Agency
Ariel Rios Building
1200 Pennsylvania Avenue, NW (7403)
Washington, DC 20460
202-260-3433
Fax: 202-401-2863
E-mail: falke.ernest@epa.gov

David Gaylor
Sciences International, Inc.
13815 Abinger Court
Little Rock, AR 72212
501-228-7009
Fax: 501-228-7010
E-mail: dgaylor@sciences.com

Larry Gephart
Group Head, Lubes and Specialties Toxicology
Toxicology Division
Exxon Mobil Biomedical Sciences, Inc.
1545 Route 22, E
P.O. Box 971
Annandale, NJ 08801
908-730-1063
Fax: 908-730-1198
E-mail: lagepha@erenj.com

Jeff Gift
U.S. Environmental Protection Agency
MD-52
Research Triangle Park, NC 27711
919-541-4828
E-mail: gift.jeff@epa.gov

Karen Hogan
U.S. Environmental Protection Agency
Ariel Rios Building
1200 Pennsylvania Avenue, NW (8601D)
Washington, DC 20460
202-564-3403
E-mail: hogan.karen@epa.gov

Jennifer Jinot
U.S. Environmental Protection Agency
Ariel Rios Building
1200 Pennsylvania Avenue, NW (8623D)
Washington, DC 20460
202-564-3281
E-mail: jinot.jennifer@epa.gov

Amy Kim
Graduate Student
Molecular Epidemiology & Dosimetry Group
Laboratory of Computational Biology
and Risk Analysis
National Institutes for
Envrionmental Health Sciences
P.O. Box 12233 (MD - D4-01)
Research Triangle Park, NC 27709
919-541-0001
Fax: 919-541-4704
E-mail: kim3@niehs.nih.gov

Carole Kimmel
U.S. Environmental Protection Agency
Ariel Rios Building
1200 Pennsylvania Avenue, NW (8623 D)
Washington, DC 20460
202-564-3307
E-mail: kimmel.carole@epa.gov

Amal Mahfouz
Toxicologist
Office of Science & Technology
Office of Water
U.S. Environmental Protection Agency
Ariel Rios Building
1200 Pennsylvania Avenue, NW (4304)
Washington, DC 20460
202-260-9568
Fax: 202-260-1036
E-mail: mahfouz.amal@epa.gov

Elizabeth Margosches
Statistician
Office of Pollution Prevention and Toxics
U.S. Environmental Protection Agency
Ariel Rios Building
1200 Pennsylvania Avenue, NW (7403)
Washington, DC 20460
202-260-1511
Fax: 202-260-1279
E-mail: margosches.elizabeth@epa.gov

Mark Nicolich
Statistician
Exxon Biomedical Sciences, Inc.
1545 Route 22, E
P.O. Box 971
Annandale, NJ 08801
908-730-1108
Fax: 908-730-1192
E-mail: mjnicol@erenj.com

Woodrow Setzer
U.S. Environmental Protection Agency
MD-55
Research Triangle Park, NC 27711
919-541-0128
E-mail: setzer.woodrow@epa.gov

Eric Wilson
Researcher
People for the Ethical Treatment of Animals
510 Front Street
Norfolk, VA 23510
757-622-7382
Fax: 757-622-6382
E-mail: eric_w@peta-online.org

Bill Wood
Risk Assessment Forum
U.S. Environmental Protection Agency
Ariel Rios Building
1200 Pennsylvania Avenue, NW (8601 D)
Washington, DC 20460
202-564-3358
Fax: 202-565-0062
E-mail: wood.bill@epa.gov

APPENDIX C

WORKSHOP CHARGE

Peer Review Workshop on the Benchmark Dose Technical Guidance Document

CHARGE TO REVIEWERS

Background

U.S. EPA's Risk Assessment Forum (Forum) has been active in promoting research and discussion on benchmark dose (BMD) issues since 1990. In 1993 the Forum sponsored a colloquium on the applications of BMD methods to noncancer risk assessment. The focus of this colloquium was to review a Forum draft report that outlined the techniques and presented the major questions and decisions involved in applying the benchmark dose method. Following this a Forum technical panel published a background document on the use of BMD in health risk assessment. In the ensuing years the Forum sponsored several workshops and symposia on the BMD approach, including a 1996 external peer review on a previous draft of the *Benchmark Dose Technical Guidance Document*. An Agency review was held on the revised document during the winter of 2000. Comments from this Agency review have resulted in the draft document presently undergoing review. Following the external peer review this December the technical panel will consider comments received and revise the document for final Forum review.

In your review, please address the following issues and questions on the *Draft Benchmark Dose Technical Guidance Document*.

Charge Questions

I. Preparation for Computing a Benchmark Dose: Selecting Data and An Appropriate Benchmark Response Level — *Discussion Leader - George Alexeeff*

1. What concepts and terms related to the benchmark dose (BMD), if any, do we need to define more clearly?
2. The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?
3. What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?

II. Modeling to Compute a Benchmark Dose: Model Selection, Fitting, and Confidence Limits — *Discussion Leader - Lynne Haber*

4. Model selection and fitting
   a. What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?
   b. Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?
   c. What are the advantages/strengths of using the methods described to select among "equally" fitting models? What other methods should be considered in making a selection?

5. Use of confidence limits
   a. Please comment on the approaches described to compute confidence limits.
   b. Please comment on the soundness of the criteria used to deviate from using a 95% one-sided confidence interval for the BMDL.

6. Examples: What additional concepts, if any, should be illustrated by an example?

III. Interpretation and Using the Benchmark Dose — *Discussion Leader - Lorenz Rhomberg*

7. The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document's guidance, or would further discussion be useful?

8. What other comments do you have about the approach to benchmark dose described in the document? What changes to the approach would you like to see, and, very importantly, why?

APPENDIX D

REVIEWER PREMEETING COMMENTS

Workshop for Peer Review of the Benchmark Dose Technical Guidance Document

Premeeting Comments

Washington, DC
December 7**B**8, 2000

Notice

The U.S. Environmental Protection Agency (EPA) strives to provide accurate, complete, and useful information. Neither EPA nor any person contributing to the preparation of this document, however, makes any warranty, expressed or implied, with respect to the usefulness or effectiveness of any information, method, or process disclosed in this material. Nor does EPA assume any liability for the use of, or for damages arising from the use of, any information, methods, or process disclosed in this document.

Any mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Table of Contents

Charge to Reviewers D-7

Peer Reviewer Comments

George Alexeef..... D-9

Kevin Brand D-19

Paul Catalano D-41

Harvey Clewell D-51

George Daston..... D-59

Elaine Faustman D-67

Clay Frederick..... D-77

Lynne Haber..... D-83

Dale Hattis..... D-93

Colin Park D-109

Lorenz Rhomberg.....D-113

Robert Sielken, Jr.....D-123

William Slikker, Jr. D-147

R. Webster West D-155

Yiliang Zhu D-161

Note: Premeeting comment materials have been reproduced as received.

Peer Review Workshop on the Benchmark Dose Technical Guidance Document

CHARGE TO REVIEWERS

Background

U.S. EPA=s Risk Assessment Forum (Forum) has been active in promoting research and discussion on benchmark dose (BMD) issues since 1990. In 1993 the Forum sponsored a colloquium on the applications of BMD methods to noncancer risk assessment. The focus of this colloquium was to review a Forum draft report that outlined the techniques and presented the major questions and decisions involved in applying the benchmark dose method. Following this a Forum technical panel published a background document on the use of BMD in health risk assessment. In the ensuing years the Forum sponsored several workshops and symposia on the BMD approach, including a 1996 external peer review on a previous draft of the *Benchmark Dose Technical Guidance Document*. An Agency review was held on the revised document during the winter of 2000. Comments from this Agency review have resulted in the draft document presently undergoing review. Following the external peer review this December the technical panel will consider comments received and revise the document for final Forum review.

In your review, please address the following issues and questions on the *Draft Benchmark Dose Technical Guidance Document*.

Charge Questions

I. Preparation for Computing a Benchmark Dose: Selecting Data and An Appropriate Benchmark Response Level **C** *Discussion Leader - George Alexeeff*

1) What concepts and terms related to the benchmark dose (BMD), if any, do we need to define more clearly?
2) The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?
3) What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?

II. Modeling to Compute a Benchmark Dose: Model Selection, Fitting, and Confidence Limits **C** *Discussion Leader - Lynne Haber*

4) Model selection and fitting
    1) What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?
    2) Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?
    3) What are the advantages/strengths of using the methods described to select among Aequally@ fitting models? What other methods should be considered in making a selection?

5) Use of confidence limits
   1) Please comment on the approaches described to compute confidence limits.
   2) Please comment on the soundness of the criteria used to deviate from using a 95% one-sided confidence interval for the BMDL.

6) Examples: What additional concepts, if any, should be illustrated by an example?

III. Interpretation and Using the Benchmark Dose **C** *Discussion Leader - Lorenz Rhomberg*

7) The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document=s guidance, or would further discussion be useful?

8) What other comments do you have about the approach to benchmark dose described in the document? What changes to the approach would you like to see, and, very importantly, why?

George Alexeeff

GEORGE V. ALEXEEFF, Ph.D., DABT

OEHHA, Cal/EPA
301 Capitol Mall, Room 205
Sacramento, CA 95814-4327
(916) 322-2067

Oakland: (510) 622-3202
(510) 622-3210 (fax)

e-mail: galexeef@oehha.ca.gov

EDUCATION: Ph.D. in Pharmacology and Toxicology, University of California at Davis, September 1982.
B.A. in Chemistry, Swarthmore College, Swarthmore, PA, May 1976.

CERTIFICATION: Diplomat of the American Board of Toxicology, Inc., (DABT) 1986-2001.

EMPLOYMENT HISTORY: Deputy Director for Scientific Affairs, Office of Environmental Health Hazard Assessment, CAL/EPA, February 1998-present.

Chief, Air Toxicology and Epidemiology Section, California Office o Environmental Health Hazard Assessment, CAL/EPA, October 1990-February 1998.

Joined California Department of Health Services 12/86.

MAJOR RESPONSIBILITIES:
Oversee a staff of approximately 80 scientists (including physicians, toxicologists and epidemiologists) in multidisciplinary evaluations of the complex health impact of pollutants and toxicants. Areas include ambient air quality standards, toxic air contaminants, air toxics hot spots, public health goals for water, Proposition 65 evaluations for carcinogens, developmental/reproductive toxins, fish advisories, hazardous waste site risk assessment, ecological risk assessment, and multi-media risk assessment.

Prior to becoming Deputy Director for Scientific Affairs, was Chief, Air Toxicology and Epidemiology Section in OEHHA.

PUBLICATIONS:
Over 50 publications in toxicology and risk assessment.

Peer Review Workshop on the Benchmark Dose Technical Guidance Document

Comments from George Alexeeff on "Benchmark Dose Technical Guidance Document, October, 2000"

I. Preparation for Computing a Benchmark Dose: Selecting Data and an Appropriate Benchmark Response Level.

1. What concepts and terms related to the benchmark dose (BMD), if any, do we need to define more clearly?

The term 'point of departure' (POD) needs to be further defined on page one, elsewhere in the text. The term should be defined in a manner that clearly explains its relationship to the NOAEL and LOAEL. Specifically, it should be clear what type of uncertainty factors would typically be applied to the POD. The term also needs to be further defined vis-a-vis the BMR. Without a clear understanding of the meaning of the POD, the value of the BMR is unclear. Currently, it is too flexible, allowing risk assessors to come to completely different conclusions and interpretations without having the groundwork available to clearly lay out and clarify the differences.

The document claims (page 2, line 1-2) that it is beyond the scope of the document to provide guidance for RfC/RfD calculation. While that may be the case, the purpose, use and application of the BMD needs to be placed in proper context and perspective. If it isn't, it is not possible for the reviewer to determine if sufficient information has been provided in the BMD documentation.

Specifically, it should be clearly stated if the BMD is to be treated as a NOAEL or LOAEL in the computation. There should be some discussion about the inherent uncertainty in the POD or BMDL in contrast to the NOAEL or LOAEL. That is, if you had two PODs, one determined by the NOAEL approach and one using a BMDL, if all other factors are equal should the level of uncertainty for the BMDL be considered less than the NOAEL? If the uncertainty isn't less, then what is the purpose of calculating a BMDL? If it is less, how does it impact a total uncertainty of 10 or 100 in a data set? This type of perspective should be laid out in a concise manner.

The reference to "practical threshold" (page 3, line 22) is unclear and unnecessary. It appears that the document may be referring to the NOAEL. In any case, the issues raised in the paragraph appear applicable to concerns regarding the NOAEL, so that is what should be referred to. Further, the reference to "practical threshold" differs from the points made on page 4 (lines 18-22) that refer to the NOAEL.

On page 4, (line 10), clarify the term "response levels."

I do not understand the bullet on page 4 (line 8-9). Clearly the suggestion is that the BMD approach does account for the uncertainty in the estimate. However, how this applies to the NOAEL/LOAEL approach is not clear to me. Possibly the use of the term "uncertainty" requires clarification.

On page 4 (line 15) the term "derive" is used. Many programs considered LOAELs with an uncertainty factor to be an "adjusted NOAEL." Thus, it is unclear why one cannot derive an estimated NOAEL from a LOAEL. I suggest the bullet be rewritten. Note that on page 5 (line 26) it states that the NOAEL or LOAEL is used as the POD, so it is not clear how that is a limitation.

On page 10 (line 14) the term "multinomial modeling" should be explained.

2. *The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?*

The issue of dose selection is raised on page 4 (line 2-3), but no specific citation is given. The statement indicates that the NOAEL/LOAEL is highly dependent on dose selection, basically because it was one of the doses selected. The implication is that the BMDL is not dependent on dose selection. However, I think it would be important to indicate if its dependence on dose selection has been tested. As indicated on page 14 (line 8-10), "it is be preferable to have studies with one or more doses near the level of the BMR to give a better estimate of the BMD." On page 5 (lines 14-15) the document states "response data are modeled in the range of empirical observation;" this factor would seem to make the method dependent on dose selection. Further, the document states (page 18, line 23-24) "the major aim of benchmark dose modeling is to model the dose-response data for an adverse effect in the observable range." So, it appears to this reader that the BMD is also dependent on dose selection. Thus, I wonder if this is much of a limitation with regard to the value of the BMD approach.

The bullet regarding sample size (page 4, lines 4-7) is also without citation. While I agree with the inherent logic of the bullet, I wonder if it has been tested. Clearly the ability to determine an effect, especially a continuous effect, is related to the number of subjects tested. Do we have any evidence that the "NOAEL for a compound (and thus the POD) will tend to be higher in studies with smaller numbers of animals per dose group." Possibly other experimental factors have compensated for the sample size for the identification of a NOAEL in a study? Do we know that there is a specific tendency or is there simply a sample size that is very unreliable for calculating the effect? I ask these questions for several reasons. First, the up-and-down procedure for acute toxicity uses very few animals (as few as one per dose group) but is very reliable for calculating an LD50. Second, we have to be careful about encouraging the increased use of animals for experimentation. Unnecessarily large group sizes should be discouraged. Third, as indicated on page 10 (line 1-2) the simulation study by Kavlock et al. (1996) found "virtually no advantage in increasing the sample size" when two dose levels had response rates above background; although effects were seen when on dose level responded above background. Finally, in preliminary work we have conducted looking at the ratios of the NOAELs to LOAELs for mild acute inhalation effects, the ratios appear to be independent of sample size.

Two documents that should be included were developed by OEHHA, Cal/EPA. One document entitled "Air Toxics Hot Spots Program Risk Assessment Guidelines, Part I, The Determination of Acute Reference Exposure Levels for Airborne Toxicants," (Office of Environmental Health Hazard Assessment, Cal/EPA, March 1999), includes 14 examples of benchmark dose calculations. This document can be found at www.oehha.ca.gov/air/acute_rels/acuterel.html. Another document entitled "Air Toxics Hot Spots Program Risk Assessment Guidelines, Part III, The Determination of Noncancer Chronic Reference Exposure Levels for Airborne Toxicants," (Office of Environmental Health Hazard Assessment, Cal/EPA, February, 2000, and May, 2000), includes four examples of benchmark dose calculations. This document can be found at www.oehha.ca.gov/air/chronic_rels/chronicrel.html.

3. *What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?*

Graphical depictions, i.e., examples, are needed for many of the examples discussed in the document.

- On page 9-10 (lines 25-8) the information would be strengthened if the description were shown graphically to determine the meaning of "were benefits" or "poorest results." It is not clear how much variation is involved and if it would have much bearing on a risk assessment.
- It would be useful to provide an example of the work of Gaylor and Slikker (1990) regarding the implicit dichotomization.
- It would be helpful to show an example how explicitly dichotomizing the data (Gaylor 1996, West and Kodell 1999) results in a loss of precision when compared to implicitly dichotomizing the data (page 11, line 1-2).

Under data evaluation and endpoint selection (page 13, line 23) states that the method "allows for the possibility that NOAELs will continue to be used for some endpoint." I would like this statement to be fleshed out further. Based on the generally poor quality of the toxicological database, I see the NOAEL approach dominating 90% or more of the assessments, once we move on from the standard 30 or 40 chemicals commonly assessed. I have concerns that there will be a requirement that risk assessments should not be performed without the higher quality BMDL data. This would mean that most chemicals could not be assessed. Consequently, I suggest that the statement be rewritten in a more realistic manner that may state a preference, and yet still accepting the NOAEL approach when the data require it to be used.

Several sections of the report (e.g., page 11, 9-15, page 15, lines 17-28) discuss the categorical analysis approach in the context of the BMD approach. It is not clear if the report is suggesting that the categorical analysis approach is a type of BMD approach, can be modified to be a BMD approach, or incorporates data that may be useful in a BMD approach. This needs to be clarified and explained further, possibly using examples. Especially line 27-28; "these regression models can be used to derive a BMD." This needs to be described more completely. It is unclear if the phrase "by estimating the probability of effects of different levels of severity" is referring to modeling the effects of differing severity together or separately.

Under selection of endpoints (page 16, line 20-22) the text refers to choosing endpoints from within a study and modeling those within 10-fold of one another. It is unlikely that one study will have evaluated so many endpoints, with such an extensive dose-response range, except possibly a chronic bioassay or a developmental study. Some guidance should be given with regard to evaluating endpoints from different studies. Does the same 10-fold rule apply?

In general, the criteria for selecting studies for BMD modeling are very inclusive. The document suggests modeling every endpoint within a 10-fold range in the study. Does this imply that every study within 10-fold of the lowest value should also be modeled (assuming the above rule applies)? This would likely result in modeling studies within a greater than 10-fold range. As indicated later, for each analysis a number of models are used. This could result in quite a few calculations. Studies unsuitable for BMD analysis, with NOAELs or LOAELs will also have to be considered. This suggests a potentially extensive data management issue.

Regarding minimum data sets, page 17 (line 12-13) states that "the BMD may be … orders of magnitude lower" than the lowest dose tested. It is not clear to me how this can be the case. It would be useful to provide an example of a useful result that meets this criterion.

The suggestion to combine data sets is a good one (page 18, line 1-16). We conducted such an analysis for the endpoint of eye and respiratory irritation in humans following acute ammonia exposure (Alexeeff et al., (1999) Determination of Acute Reference Exposure Levels for Airborne Toxicants, Office of Environmental Health Hazard Assessment, Cal/EPA, Oakland, CA, pp. C13-C22). In the analysis we combine data from four studies and calculate a benchmark dose. This is shown in the following table taken from the document:
Table 2. Ammonia, Human Irritation, 60 Minute Exposures (adjusted), ppm

Study Conc.	32	30	50	50	72	50	80	134	110	140	500
Exposure Time (min.)	5	10	5	10	5	120	120	5	60	60	30
60 min. adjusted Conc.	19	20	29	34	42	43	69	78	95	120	430
Response	0/10	0/5	0/10	4/6	3/10	7/16	9/16	8/10	12/16	15/16	7/7
Study	2	3	2	3	2	1	1	2	1	1	4

Table adapted from: (1) Verberk, 1977; (2) Industrial Biotest Laboratories, 1973; (3) MacEwen et al., 1970; (4) and Silverman et al., 1949. The two lowest concentrations were combined for the log-probit analysis since this improved the fit of the data.

The benchmark concentration approach used a log-normal probit analysis (Crump and Howe, 1983; Crump. 1984). The 95% lower confidence limit of the concentration expected to produce a response rate of 5% is defined as the BC_{05}. The BMD for a 5% response was 20.1 ppm and the $BMDL_{05}$ for ammonia from this analysis was 13.6 ppm.

BMR	BMD (ppm)	BMDL (ppm)
1%	13.4	7.8
5%	20.1	13.6

In fact, considering the paucity of data for many compounds and the cost and difficulty of conducting animal studies, it is likely that sufficient data for BMD calculation may require combining studies. For this reason, it would be helpful to clarify the statement "which will require more elaborate modeling to include properly" means (page 18, line 11).

Regarding the criteria for selecting BMR, the proper selection depends on how the value will be interpreted (e.g., as a NOAEL or LOAEL) and used (i.e., what uncertainty factors are likely to be applied). Without a clear understanding of its meaning and use, there will be no really basis for choosing one. In OEHHA, Cal/EPA we have considered the BMDL a NOAEL, but with greater certainty than the traditional NOAEL. For this reason we chose a BMR of 5% based on our analyses that indicate a BMR of 10% might be a LOAEL. We also use a 3-fold smaller uncertainty factor. However, while the justification for doing so, is a higher quality NOAEL, it can be difficult to justify in the traditional uncertainty factor approach. If the data are human data, it is fairly straightforward to apply an uncertainty factor of 3 for intraspecies variability since BMDL accounts for some of the variability. However, if the data are animal data, the justification to implement 30-fold factor can be unclear. Because in this case the analysis is addressing experimental variability and response but not interspecies or intraspecies uncertainties.

The document states (page 18, line 23-24) "the major aim of benchmark dose modeling is to model the dose-response data for an adverse effect in the observable range." This suggests that the purpose is to identify a LOAEL since we are identifying an effect in the dose-response range.

On page 19 (line 6-8) it states that "it is not critical that a common response level be used for all chemicals or endpoints, … it may be desirable to use response levels below 10%, if possible, in order to minimize the degree of low-dose extrapolation required." This is an important statement regarding the need for some flexibility to derive a BMR different than 10%. The current statement could be interpreted that if one dose not want to use a LOAEL to NOAEL uncertainty factor, a smaller BMR could be used. That is clearly one reason. However, I think it is important to discuss reasons why the BMR may be less than or greater than 10%. If these reasons were laid out it may give greater guidance regarding the purpose of the BMDL. I am able to identify the following reasons for using a BMR smaller than 10%:

- to minimize the degree of low-dose extrapolation required;
- to evaluate a relatively rare effect;
- to take advantage of the increased sensitivity of a very large study or combination of studies;
- to insure the identification of a NOAEL for a severe effect such as lethality;
- when applying the result to steep dose-response slope;
- the percent change or incidence level of the effect considered to be biologically significant is less than 10%.

I am able to identify the following reasons for using a BMR larger than 10%:

- the study variability is much greater than 10%;
- the study doses are much greater than 10%;
- the effect measured may be a NOEL rather than a NOAEL;
- the percent change or incidence level of the effect considered to be biologically significant is greater than 10%.

It may be helpful to develop more comprehensive lists to help clarify the ultimate purpose of the BMR.

On page 19 (line 26-28) it states that "the 10% response rate is at or near the limit of sensitivity in most cancer bioassays." While this may be a justification for the LED10 approach, I don't see it as a justification for the BMDL and thus the BMR, a noncancer risk assessment.

4. *Other points regarding the Background Section.*

Page 4 (lines 12-13) discusses the issue of dose-response. It appears to me that dose selection is dependent on dose-response. Generally, doses are selected to encompass the dose response curve. If it is steep, the doses are close together, if it is shallow, the doses are spaced apart. While this may be an intuitive decision on the part of the investigator, often a range-finding study is conducted prior to dose selection. Clearly, if the study is poorly designed then dose-response is not taken into account for the doses selected, and consequently, the NOAEL and the LOAEL. I question whether the BMD approach could provide a good BMDL if the study is inherently poorly designed.

On page 5 (lines 1-2) the text indicates that the BMD approach does not make any assumptions about the nature of the toxicological responses. However, as the response frequency increases for many responses, the response severity also increases. Thus, the BMD approach appears to require the assumption that the response severity is not increasing to any great degree. On page 15 (lines 17-19) it states "after combining data in severity categories," which may be interpreted as keeping data in severity categories separated or combining data from multiple severity categories. This should be clarified.

Page 5 (line 6-7) states that the BMD approach is "more independent of study design." It is not clear to me if this has been shown. One would have to vary key study design factors and show that the results were not affected. My understanding of the BMD approach was that it more fully incorporated the study design features in the analysis. For example, the number of subjects responding at each dose is incorporated. If possible, all the dose levels are used in the analysis and not just the NOAEL or LOAEL. The variability in the responses at different doses is incorporated into the analysis. Consequently, the text up to page 5 (lines 4-7) has not laid out the reasons for choosing the BMD approach over the NOAEL approach.

The clearest statement for choosing the BMD approach over the NOAEL approach appears on page 6 (lines 1-3). For me, the statement conveys the point that we can have more confidence in the BMDL than in a NOAEL or a LOAEL as a POD. Part of the reason is the rigor and systematic manner in which the BMDL is calculated. Another reason is that the general quality of studies from which BMDL can be calculated is better than the general quality of studies used for the NOAEL approach.

II. Modeling to compute a Benchmark Dose: Model Selection, Fitting, and Confidence Limits

5. Model selection and fitting

a. *What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?*

Some additional examples in the text would be helpful, particularly regarding the use of continuous data. However, it would be useful to display some basic dichotomous data as well.

b. *Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?*

If there were some information on the biological nature of the response, it would be helpful to incorporate that into the criteria for model selection.

c. What are the advantages/strengths of using the methods described to select among "equally" fitting models? What other methods should be considered in making a selection?

Currently the procedure suggests modeling with multiple models and then eliminating some results based on stated statistical criteria. There is little discussion of the biological or empirical basis of the respective models. It would be useful to discuss the biological implications of the different models if there are any. Most toxicological texts describe the dose-response curve as a log-normal distribution. This is based on extensive toxicological data fitting. It is not clear to me how the plotting of one, two or three points can provide a better biological basis for dose-response modeling. How to differentiate the variability from a log-normal model from a different model is not discussed. In the OEHHA, Cal/EPA approach for acute data, the log-normal model is the primary default choice. The log-normal model is among the most widespread models used for toxicity testing and has traditionally been used extensively for determination of acute lethality and other dichotomous responses (Finney, 1971; Rees and Hattis, 1994). Furthermore, the log-normal distribution aspect of the model is biologically plausible and accounts for some degree of inter-individual variability (Rees and Hattis, 1994). Hattis (1996) showed that the log-normal relationship effectively modeled data from 126 human studies.

6. Use of confidence limits

a. *Please comment on the approaches described to compute confidence limits.*

I will leave this response to others more qualified.

b. *Please comment on the soundness of the criteria used to deviate from using a 95% one-sided confidence interval for the BMDL.*

I agree with the choice of the 95% confidence interval. We use the same criterion in our department.

7. Examples: what additional concepts, if any, should be illustrated by an example?

For a particular chemical it would be useful to see how an RfD value would be calculated by the two approaches.

III. **Interpretation and Using the Benchmark Dose**

8. *The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document's guidance, or would further discussion be useful?*

The rationale for study and endpoint selection should be covered in the text in the context of reviewing the study. I am not sure what value it is to say that a trained toxicologist selected the data (page 75, line 2). Without any biological information regarding the appropriateness of various models the "rationale" for choosing a dose-response model (page 35, line 24) seems meaningless. The statement on page 75 (line 5) stating "fits a wide variety of dose-response shapes for nested data," does not seem helpful to me. Maybe it has some statistical meaning to others. For the other criteria, it would seem reasonable that the model prints out the relevant information in a summary table. If the model doesn't print it out, I don't think it should be included. The procedure is complicated enough.

9. *What other comments do you have about the approach to benchmark dose described in the document? What changes to the approach would you like to see, and, very importantly, why?*

The major changes I would like to see are some examples of how the BMDL data are incorporated into a risk assessment. As I indicated above, without a clear understanding of the context of its use, it is difficult to determine the appropriate BMR. Also, once a clear description of the context was provided, the appropriateness of many of the criteria could be more definitively evaluated.

It would appear that some of the criteria would have little impact on the BMDL. Some examples that indicate the sensitivity of the BMDL to changes in the BMR, LCL, steepness of the dose-response curve and model choice would be helpful.

Kevin Brand

Dr. Kevin P. Brand
Research Associate
Dept. of Epi. & Comm. Med., University of Ottawa

Dr. Brand has broad expertise in environmental health risk assessment, and a specific interest in uncertainty analysis and its interplay with policy formulation. His doctoral thesis (1999) developed and demonstrated frameworks for improved, policy oriented, interpretations of bioassay evidence. A significant part of this work studied the influence of experimental design upon BMD (and NOAEL) estimates. Although the work focused on ratios (of either type of estimate) and their use in exploring extrapolation relationships, broader implications for BMD estimation were found. He is currently collaborating on additional BMD simulation studies, and in separate work has applied a decision analysis framework to high to low dose extrapolation (and to the EPA's 1996 guidelines on that topic). He has served as a peer reviewer for the Journal of Risk Analysis for 5 years and as an instructor for the recent workshop on "Advanced Methods for Dose Response Assessment: Bayesian Methods." Kevin has published four papers. He received a BASci in civil engineering (U. of Toronto), and a MS in environmental engineering (Carnegie Mellon U.). He then worked (1 year) as research engineer at Lawrence Livermore National Laboratory (working on fate and transport modeling), after which he obtained a SM and a Sc.D in environmental health science (Harvard School of Public Health). Kevin is currently a Research Associate at the Institute of Population Health, U. of Ottawa.

Topic I: Preparation for computing a BMD: Selecting data and an appropriate BMDR

Question 1: What concepts and terms, do we need to define more clearly.

Response: I begin this response with a list of terms, which need to be defined. I then make a stylistic comment that may help with understanding, and then move onto some more substantive concerns, firstly regarding the definition of dose, and secondly regarding the challenge of defining an appropriate BMR. Finally, I suggest the use of schematics for clarifying some issues that recur in the text.

Several terms could be clarified. In no particular order, these include, graded monotonic, dynamic range, sensitivity, limit of detection (what kind of statistical power are you targeting), minimum dataset, and risk assessment principles.

The notation, RfD/C, presumably used as short form for "RfD or RfC," is too easily misunderstood (especially to those new to this subject) as "RfD divided by C." The same applies to NOAEL/LOAEL, or BMD/BMDL. Why not take the opportunity at the beginning of the document to define the terms as encompassing the variants. For example, stating `that the RfD should be construed of as encompassing an RfC (where necessary).' And similarly, for the other cases (e.g., NOAEL/LOAEL).

The units of dose need to be defined. Are they expected to be scaled by body-weight, or is some other scaling rule implicit? Notably, this has ramifications for the subsequent application of Adjustment Factors (aka, Uncertainty Factors). Depending on the presumed scaling rule, additional adjustment may be necessary to account for inter-species extrapolation. Although such adjustments often amount to a reduction in dose, they should not be confused with adding a margin of safety.

It is suggested that a consistent choice of a BMR, is less important from the standpoint of simply estimating the point of departure (POD), and more important when the BMD is to be directly

used in toxicity ranking exercises. This is an important distinction. However, we can not assess the importance of a consistent choice of BMR for definitions of PODs. It is difficult to evaluate this statement without having more detail on how toxicity evaluations and risk assessments are to proceed after the determination of the POD. It is quite conceivable that a margin of exposure (MOE) approach would work better with a consistent BMR; how can one be specific about what might be an acceptable MOE without knowing what the BMR was that helped establish the POD? Perhaps, the guidance's insistence upon reporting the ED10 or LED10 in addition to whichever EDq (LEDq) is chosen will obviate this concern.

A figure showing three or four archetypal quantal datasets similar to (Fig R1; attached), could be helpful for illustrating several recurrent themes throughout the document. See for example, pages, vi, 14, and 17, where, one or more of these archetypes are described (in words) in order to clarify a concept (such as the notion of a LOAEL, the types of outcomes causing undefined maximum likelihood, and nonsignificant outcomes).

The document appears to pay much less attention to categorical outcomes. This may be appropriate. However, it would be helpful to state in the introduction that categorical outcomes will not be covered by the document in any great detail (perhaps adding that categorical outcomes could be treated analogously to either quantal or continuous outcomes). This would save the reader from wondering, what should be done for categorical data at every caveat (as happened to me).

Topic I: Preparation for computing a bmd: Selecting data and an appropriate bmd level

Question 2: The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?

Response: The references are appropriate and the text is generally appropriately referenced. I have some suggestions for additional references and some comments on the way the text has

summarized the literature on BMD estimation.

I have chosen to list all recommended references, and not just those specifically dealing with the development of the BMD approach. I divide these by topic.

General background on noncancer risk assessment:

Baird, S., J.C. Cohen, J.D. Graham, A.I. Shlyakhter, and J.S. Evans.
Noncancer risk assessment: A probabilistic alternative to current practices.
HERA, 2(1):79--102, 1996.

Alternative approaches or perspectives on BMD calculation

Bosch-RJ; Wypij-D; Ryan-LM A semiparametric approach to risk assessment for quantitative outcomes. Risk-Anal. 1996 Oct; 16(5): 657-65

Vermeire, T., H. Stevenson, M.N. Pieters, M. Rennen, W. Slob, and B.C.
Hakkert. Assessment factors for human health risk assessment: A discussion
paper. *Crit. Rev. Toxicol.*, 29(5):439--490, 1999.

Slob W., and M. N. Pieters, A probabilistic approach for deriving acceptable human intake limits and human health risks from toxicological studies: general framework, Risk Analysis,v18:787-798, 1998.

More information on experimental design

Brand, K.P., L. Rhomberg, and J.S. Evans. Estimating noncancer uncertainty factors:Are ratios [of] NOAELs informative ? *Risk Analysis*, 19(2):295--308, 1999.

Brand KP, P Catalano, JK Hammitt, L. Rhomberg and JS Evans,
Limitations to empirical extrapolation studies: The case of BMD ratios,
accepted by Risk Analysis, November, 2000.

Literature pertaining to the combining of toxicological evidence.

DuMouchel WH and JE Harris, Bayes methods for combining the results of cancer studies in humans and other species, *JASA*, v78(382):293-315, 1983.

DuMouchel WH and Groer PG, A Bayesian methodology for scaling radiation studies from animal to human, *Health Physics*, v57(Suppl 1):411-18, 1989.

Gelman A., J. Carlin, H. Stern, and D. Rubin. Bayesian Data Analysis.
Chapman & Hall, 1995.

NRC, Combining information: Statistical issues and opportunities for research (National Acadamy Press, Washington, D.C., 1992).

The literature review dealing with the BMD approach is impressive in its breadth. My only complaint is that the review does not place enough emphasis on what is not yet known about the probable operating characteristics of the BMD approach (labeled "Properties of the BMD" in the document) in actual practice. The review tells us what is known, and what studies have established, but does not do as good a job of telling us what has yet to be studied, and what issues are insufficiently understood. A relatively impressive volume of papers have studied the operating characteristics of various BMD approaches, and are cited by the document. One may mistakenly conclude from this volume, that the operating characteristics are well understood. But the studies to-date are unlikely to have examined the full spectrum of possible outcomes. Studies of (relatively) newly developed techniques, for example those modeling multiple outcomes, and accommodating intra-litter correlation, are commonly tested-out on a restricted number of datasets (sometimes one, and not uncommonly, a `re-cycled' dataset). This is a sensible start, but it limits generalize-ability. The database of Faustman and co-workers (1994), the one exception, a large collection of bioassay outcomes, only serves to exemplify the point. This study found that roughly 40 percent of bioassay outcomes preclude the identification of a LOAEL. This important characteristic would not be discovered by studies on single datasets. So although, the cited studies have advanced our understanding

tremendously, the limits of that understanding should be recognized.

Even simulation studies, which increase the spectrum of outcomes encountered in an analysis, have typically examined a restricted set of outcomes. They typically generate synthetic datasets using an assumed underlying dose response curve. The spectrum of synthetic datasets remains restricted, because the set of explored truths is typically small. Colleagues and I, have done some analysis which suggests that important characteristics may be overlooked by focusing in on only a restricted set of conditions (Brand *et. al*, 2000).

On a more specific, but related, point, the Fowles acute study [P9-17], has limited relevance to chronic or sub-chronic noncancer outcomes. The summary of it's findings should be tempered accordingly.

Topic I:
Question 3: What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?

Response:
In a slight twist of the question, I would like to suggest that some discussion currently in the document be removed. I do, however, believe that additional discussion is required to clarify how data on historical controls can be used. These two points are discussed in turn.

The document puts too much effort into defending the BMD approach (over the NOAEL approach) and the concept of harmonizing noncancer and cancer risk assessment. Will the readers who pick up this document for guidance really be in need of convincing? I suggest that the sections enumerating the advantages of the BMD approach (and harmonization) be shortened.

The document suggests ways in which data on historical controls can be used (see for example,

viii-2 and p20-22). It suggests that such data can be used to better characterize the Standard Deviation (SD) expected in controls, but that the mean response (in the controls) must always be estimated from the outcome data alone. I have two concerns. First, what if the control group's response is atypical and there is reason to believe it is not representative of true background conditions? Would it not make better sense to consider the historical control's mean response in such a case? Second, the document may be interpreted as sanctioning the use of the historical SD, even for characterizing the SD in dosed groups. I believe the SD observed in the dosed groups should be preferred over the historical SD, which is incapable of reflecting any chemical specific impact on heterogeneities. More generally, it may be worth rethinking the permission to use the control group's SD as a surrogate for that in dosed groups; particularly, because as the document points out certain biological endpoints are expected to show heteroscedastic patterns with dose.

What is the origin of the 10-fold rule? (p vii-line 9) You may want to elaborate on the intuition behind 10-fold rule.

Topic II: Modeling to compute a bmd: Model selection, fitting, and confidence limits:

Question 4a: What additional discussion is needed to ensure adequate presentation of the proposed defaults for parameters for various models?

Response: I have a general comment about organization, and then some specific points of detail.

With two primary outcome types (quantal and continuous), the various possible complications (intra-litter correlation, or joint outcomes), and the various possible approaches to modeling (in particular for continuous outcomes), the array of choices can be hard to process. A clear structure, perhaps making use of tables, is going to be critical for helping the reader to process all the choices, and ideally quickly focus in on the one that interests him or her. This latter facility, would save the reader a lot of confusion by allowing them to focus on only those

caveats/details applicable to their circumstance.

It will also be helpful to specify the order in which alternative approaches are preferred. For example, when modeling continuous data, what is the rank ordering of preference across the various approaches (explicit dichotomiziation, the hybrid approach, etc.). The order in which the document currently lists the continuous approaches, does not seem to reflect the proper preference. Implicit dichotomization is listed second. But, it should arguably be first. A rationale should be provided alongside the ordering of preference.

The document should explicitly state what is to be done with datasets showing no significant trend with dose (p35-10). Such datasets may well have no LOAEL. If so, what should be used to determine the POD or RfD ? Will the highest test dose be accepted as a provisional NOAEL (as it is in the standard NOAEL approach)?

On a more specific point, the document suggests that Weibull models can have an upper horizontal asymptote, other than the expected 100 percent line. This would not apply to the Weibull models I'm familiar with. To avoid such misunderstandings, it may be worthwhile presenting the algebraic forms of the models referenced in the document.

Question 4b: Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend alternative criteria?

Response:

The proposed goodness of fit (GOF) test for evaluating fit, seems appropriate, but a more careful explanation would help. I also have some reservations about the suggestion of opting for a more strict criteria for GOF.

The goodness of fit test may require some explanation, especially for the uninitiated. Readers may have a natural inclination to associate a lower alpha-level with a more restrictive test, and

yet the opposite holds for a GOF test. When explaining, it may help to associate the p-value with an index of deviance (or sum of squared deviations), and to remind the reader that the greater the deviance the worse the fit (and the lower the `p-value'). In anticipation of this point of confusion, it will also be helpful to explicitly state the decision rule associated with the GOF test (e.g., a fit having a p-value less than 0.10, will be rejected, and deemed an inadequate fit).

The proposed stricter standard (alpha-level = 0.10) for the GOF test, may warrant reconsideration. Based on my discussions with biostatisticians at Health Canada, I suspect that some datasets, in particular those showing some non-monotonic behavior, may preclude attaining such a standard for virtually any monotonic model. It could be a particular concern for developmental studies. In fact some preliminary simulation analysis (L. Marro, Health Canada) suggests, that such a change in criteria could change the fraction of non-significant fits from 10 percent to 30 percent (a substantial impact). Some consideration should be given, to providing different GOF criteria for different types of studies, perhaps recommending a more lenient one for developmental studies.

At [p29-1ine7] the document argues that, on occasion, a goodness of fit statistic will suggest large deviance, but the deviance will be predominantly due to data non proximate to the anticipated BMD. In this case the text suggests that it may be appropriate to drop the non-proximate data. I have no disagreement here, but suggest that the complementary case should also be of concern. Namely, a model goodness of fit may have small aggregate deviance primarily due to the non-proximate points, and may indeed have poor (large) aggregate deviance when only the proximate points are included. Caution is required here as well.

Question 4c: What are the advantages of using the methods described to select among "equally" fitting models. What other methods should be considered?

Response:

The transparency of the proposed rationale for choosing among alternative dose response

models is appealing. Akaike's Information Criterion (AIC) would prove useful in supporting that transparency. I have one major concern about the document's current explanation of AIC, and a couple of minor reservations about the reliance on such an index.

My major concern is easily addressed. It relates to the computation of the AIC, which the document describes in enough detail that a user may be tempted to compute his or her own. Users should be strongly cautioned to confirm that the maximum likelihood values they determine for each model are indeed determined in the exact same manner. It is conceivable that users would fit competing models with different optimization packages, and further that the different packages would deal with likelihood differently. Specifically, some packages may drop proportionality constants, while others may not. AIC are only valid, if the likelihood values are calculated in a consistent manner. The document may already allude to this issue (P30-10) when they say "… using similar fitting methods." If so, they should be more explicit.

If AICs are computed appropriately, they may be an appropriate way to identify the best model. I understand that their use in distinguishing among non-nested models is still controversial, but then there may not be any better criterion. I don't feel qualified to say much more on the topic of AIC, except, to recognize that barring incorrect computation, its use should not be too controversial, given that it will only be applied among models that already agree within a factor of three. So the consequences of making the wrong choice are not that large.

Some observers will disagree with overriding the AIC rank ordering, when models differ by more than a factor of three. (In this case the guidelines recommend, erring on the side of safety.) Given that the BMD (at least in the case of quantal data) should be pretty close to the data, I 'm not sure how often one can expect inter-model disagreements of larger than a factor of 3. Is it possible to say anything about this?

Topic II:
Question 5a: Please comment on the approaches described to compute confidence limits.

Response: The approaches appear to be appropriate. Although the document does state the circumstances under which the various alternative approaches apply, or are preferred, it could be more explicitly stated.

Question 5a: Comment on the soundness of the criteria used to deviate from using a 95% one-sided confidence interval for BMDL?

Response: No Comment.

Topic II:
Question 6: What additional concepts should be illustrated in the examples?

Response: The examples are useful in illustrating a number of important concepts. The example related to cancer risk assessment, and that related to epidemiological outcomes, are clearly works in progress; as they stand, it would be best to omit them. The example on continuous outcomes, demonstrates some important concepts, but could do a bit more. Specifically, it is a good platform for demonstrating the different approaches to modeling continuous response (the old mean response, explicit dichotomization, the hybrid, and … are there others?). I think the example currently uses the hybrid approach, but could not confirm. It would be helpful, to apply the other approaches, and compare and contrast the results.

Topic III: Interpretation and using the BMD
Question 7: Reporting requirements: Are they reasonable, and should there be more discussion of this issue?

Response:

Raw data should be a reporting requirement (if it is not already).

It would be helpful to have some statement about what evidence was used in choosing the dose placement used in the experiment. Was it based on the results of dose ranging studies (if so how powerful were those studies), or were previous studies consulted in making the

choice of doses (what was the effective power of those studies). This type of information would help determine how confident we can be that the doses were well placed.

To the extent that the guidance provides choices in the analysis of the data (e.g., how to deal with nonsignificant datasets), the users should be required to report what choices were made.

Topic III:

Question 8: What other comments do you have about the approach to BMD. What changes to the approach would you like to see, and, very importantly, why?

Response:

The definition of a BMR for continuous outcomes is challenging. The concept is more natural to quantal data, where its choice largely depends on the issue of a "being close to the data." Specifically, it is chosen to be "close" enough that the BMD is not overly sensitive to model choice. The discussion of a BMR for continuous data does not appear to be motivated by the same concern. Rather than focusing on `being close to the observed data' the concern is instead, defining a cut-point that can be reasonably thought of as separating normal from abnormal (continuous) responses. This is most clearly an issue in the (old) mean response approach where the cut-point essentially amounts to a z-score applied to the control group. It will be less of an issue in the explicit dichotomization approach, but appears to be an outstanding issue in the hybrid approach. Given the inconsistency in purposes, is it reasonable to strive towards comparability of BMDs obtained from continuous and quantal data? The document seems to suggest, yes. I have my doubts. But given my lack of experience in dealing with continuous (dose response) data, these doubts may more reflective of my ignorance. I do, however, think it is fair to say that the documentation is not sufficiently clear on this issue. It may help if the document made a stronger statement about which continuous approach is preferred, and if it discussed the limitations of the various approaches in facilitating comparability across quantal and continuous outcomes.

The document briefly mentions the results of epidemiological studies as appropriate for

BMD applications. This discussion (see for example, p20 or p26) is brief, and as a result, superficial. Although BMD applications to epidemiological data, are likely to share much in common with applications to bioassays, the commonality is not sufficient, in my opinion, to warrant, including epidemiological applications in this guidance. In particular, much more discussion would have to be dedicated to the specifics of epidemiological studies, to do justice to epidemiological applications. More thought should be given to this special application, before guidance is given.

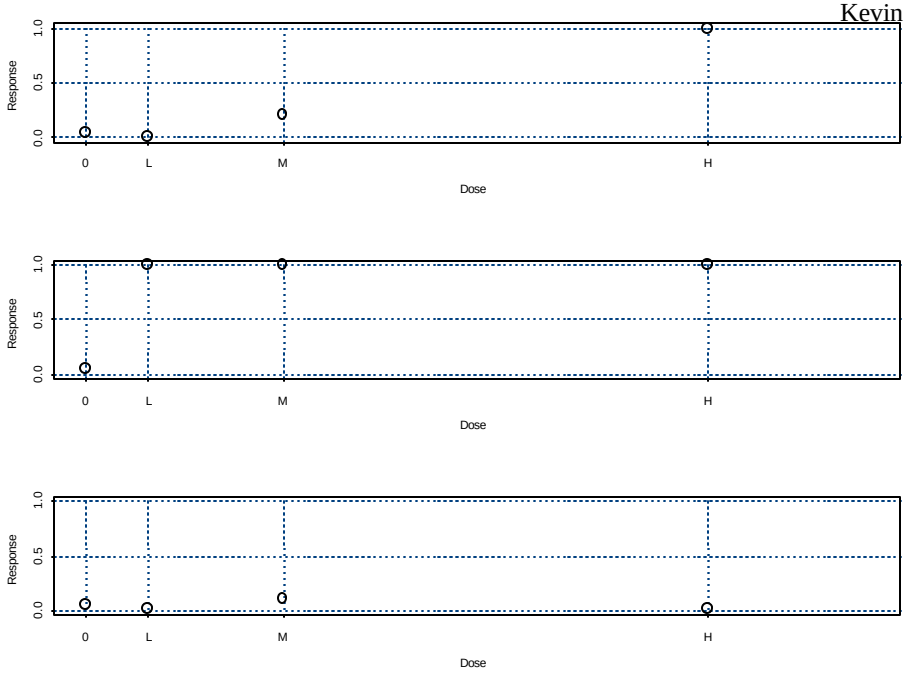


Fig R1: Archetype datasets for use in illustrating concepts in the document. First panel illustrates the type dataset causing unbounded MLE estimates of shape (e.g., in case of 3 parameter Weibull function). Second panel illustrates a similar concept, and also shows an example where there is not NOAEL. The third panel illustrates a dataset showing no significant trend with dose, and having no LOAEL.

The following are line-item comments, made using the notation:

P1-2]

Where the first # (1) is the page # and the second (2) is the line number

---

P3-15] change "while the NOAEL is the highest dose at which no …" to "while the NOAEL is the highest dose *at or below which* no adverse effects have been detected" because the latter is more strictly correct.

P5-2] Sentence beginning "Benchmark dose …" is not strictly correct. Any model makes fairly strong parametric assumptions far beyond the simple assumption of monotonicity. Also, why "toxicological dose-response" and not simply "dose-response." (On the same sentence you may note that for some endpoints the dose response curve will actually decrease with dose)

P5-23] Adjustments to the POD are described as uncertainty factors but a component of these adjustments could be required to account for what are thought to be systematic interspecies, inter-regiment, and etc differences in toxicity. Perhaps it is assumed that the doses have already been scaled to adjust for any anticipated systematic differences, if so this should be stated.

P6-1] "… accounts for uncertainty …" please be specific. It only accounts for sampling uncertainty. The latter part of the sentence is too strong.

P6-8] "… such as when all individuals respond." Clearly this statement only applies to quantal data.

P6-11] "not informative" this depends on what you mean by informative … in this case we can be fairly confident in stating that environmental exposures in excess of the lowest test dose, are unsafe! (Since, the test system showed 100% response at comparable exposures). This is

clearly more informative than no information whatsoever. In short, bounding type statements can amount to all the information you need depending on the question you are asking.

P6-11] Sentence beginning "Another …" should be modified. The sentence appears to advise not adhering too exclusively to statistical criteria for significance. It suggests that outcomes failing to show statistical significant trend may still be judged significant, based on expert judgment. This is appropriate, however, the concern should not be confined to cases involving a rare outcome (which the text seems to do). Clearly statistical power should be considered. Studies with low power may warrant placing less importance upon statistical tests and more attention to judgment. Importantly, high background undermines power, and therefore over-ruling statistical significance criteria need not be reserved for rare events alone.

P9-28] suggests that shortness of confidence interval is a sufficient measure of a good design. Clearly, optimal design is a more complicated endeavor than simply minimizing the span of a confidence interval.

P10-14] If the issue is more general than multinomial modeling, please modify this sentence to say joint modeling rather than multinomial modeling.

P10-19] Define explicit versus implicit dichotomization. How do they differ from the hybrid approach? My understanding is that the implicit and hybrid approaches are one and the same, however this may not be clear to the reader.

P14-6] This sentence suggests that the a BMD can be computed even with only one dose group showing an elevated response. Although, this may be occasionally appropriate, caution should be strongly advised.

P14-23] Not clear what the sentence "Occasionally …" is getting at. Is it referring to studies

that might report only the proportion responding, but not the total number of animals at risk?

P16-7] What do you mean by feasible?

P16-12] `Select a subset as representative.' Please illustrate via example.

P18-23] Parse this sentence. Too many concepts for one sentence.

P20-25] The sentence beginning "This gives an …10% … 2nd percentile …" is not self evident. Please elaborate.

P21-5] This introduction can be better worded.

P21-17] Again poorly worded. Suggest something like,
A model's parameters are chosen so as to maximize the accord between the model predictions and the observed data (or minimize the descrepancy between the observed data and the model predictions).

P21-21] Discussion of a taxonomy of model types is misplaced and confusing. Its not clear how this discussion benefits the reader. This section could be better motivated, and explained, or it could be deleted.

P22-19] Sentence beginning "The initial …." What do you mean by the nature of the measurement? If this refers to the distinction among quantal, categorical or continuous outcomes, then say so. You refer to this as "type of endpoint" in the subsequent enumeration. The inconsistency is confusing.

P23-2] Poor wording.

P23-15] Incorrectly refers to models as probability density models. Although the models can be thought of as cumulative density functions, and therefore related to probability density functions, they are not, themselves, probability density functions.

P23-26] Instead of "… level of the response that expert opinion holds is …" say "… level of the response judged to be adverse." Otherwise, you need to specify what kind of expert you are talking about, and further why an expert is qualified to make such a judgment, so imbued with societal values.

P23-22] Seem to be describing the difference between implicit and explicit dichotomization. If so why not say as much.

P24-3] References are provided for the hybrid approach, but are not provided for the implicit and explicit approaches.

P24-9] The justification of the lognormal distribution is an un-necessary digression.

P24-19] Define "coefficient of variation."

P24-19] Delete sentence "Again this is a property … ." It is an unnecessary distraction.

P24-22] Delete sentence "This model includes … ." Again an unnecessary distraction.

P25-3] You may want to cite Chernoff, to support the assertion that the number of parameters should not exceed the number of dose groups. Is this out of interest for identifiability, or for optimal estimation properties or what?

P25-4] Can you clarify what you mean by "those parameters affecting overall shape." What other types of parameters are there? Can you clarify by an example.

P26-1] "… outcome is impossible …" is unclear. How about "… that zero background is logically implied … ." Or perhaps it would be best to say that it is never appropriate to fix any of the parameters, although it may make sense to put lower- or upper- bound constraints.

P26-4] Slob (1999) makes a compelling argument that thresholds have no practical meaning and are only of theoretical interest. His work could be cited here, and it should be noted, more emphatically that so-called precipitous rising response with dose, can be handled without needing to invoke a threshold, i.e., with the use of shape parameters.

Slob W. Thresholds in toxicology and risk assessment, *Internat. J. of Tox.*, v18:259-268, 1999.

P26-20] Paragraph on model fitting could be improved. To illustrate a possible re-write.

For any specified model, there are many possible values to choose for the parameters. Some choices fit the data better than others. The purpose of parameter estimation is to choose that set of parameters which maximizes the agreement between the model and the observed data.

P27] The wording on this page could be improved.

P27-27] `… often normal (guassian) or log-normal.' Best to say have a normal or lognormal distribution.

P28-28] I find this statement hard to believe. Can you give a reference or an example.

p20-13] Is this paragraph referring to the process of cross-validation? This should be stated

and references can be cited.

P30-24] Define “localize.”

P31-28] How is this section related to the section on the next page (entitled Quantal Data). There seems to be some overlap between these to sections, and as a result some unnecessary redundancy. (Well written though.)

P33-10] You neglect to mention the Rao-Scott approach to dealing with overdispersion. This straight-forward approach, performs quite well relative to the other approaches mentioned. See Krewski and Zhu (1994).

P33-18] This sentence is probably only referring to the parenthetical statement in the previous sentence. Please clarify.

P34-12] Does the EPA software accommodate all types of data, including those subject to interlitter correlation, multiple endpoints, and continuous endpoints. If not, it may be worth pointing out what types of conditions it can accommodate and what plans there are for increasing the set of conditions it can accommodate.

P54-3] Seem to use the rate of response observed in the control group as the estimate of background parameter. Would it not be more proper use the estimate returned by the MLE fit?

Paul Catalano

Paul J. Catalano, D.Sc.

Associate Professor of Biostatistics

Dana-Farber Cancer Institute and Harvard School of Public Health

Dr. Catalano's research interests are in the area of statistical modeling of dose-response data and particularly in the problem of methods for quantitative risk estimation and techniques for analyzing multiple outcomes. His recent work has investigated multiple outcomes models for the analysis of dose-response data from developmental toxicity and neurotoxicty experiments, applications where multiplicity plays a key role in the data analysis. Methods for computing benchmark doses from continuous data, from multiple continuous and discrete outcomes and from clustered data are the focus of his current research as well as small sample techniques for analyzing multiple categorical outcomes. In addition to working on methodological problems, Dr. Catalano has served as collaborative statistician for a number of studies investigating environmental exposures including large studies of the non-cancer effects of ozone and the cardiopulmonary effects of concentrated air particles. Dr. Catalano has published papers in the *Journal of the American Statistical Association, Biometrics*, the *Journal of Agricultural, Biological and Environmental Statistics, Statistics in Medicine, Risk Analysis* and a number of other journals in environmental science and toxicology.

Peer Review Workshop on the

Benchmark Dose Technical Guidance Document

Premeeting Comments

Paul J. Catalano
Dana-Farber Cancer Institute and Harvard School of Public Health
November 17, 2000

Topic I, Question 1. (Concepts and terms)

The term LED used on page vii line 22 is not defined in the glossary.

The document uses the terms EDx / ECx / ED10 but the glossary terminology is ED_x / EC_x, a minor inconsistency.

The document uses the term "categorical data" (page 15, line 10) to mean essentially ordinal data and the terms "ordinal" (glossary) and "ordered categorical" (page 23, line 6) to mean ordinal data identified with arbitrary integer codings. There seems to be no technical distinction between these concepts yet the terminology differs. Also, the terminology leaves undefined data that are nominal, multinomial which some readers might think of as "categorical data." A suggestion might be to use the term ordered categorical throughout to help clarify meaning.

The terms α and "critical value" (page 28, lines 20 and 21 respectively) are not defined in the glossary.

The term "bootstrap" is mentioned only parenthetically on page 36, line 7 (and is also defined in the glossary). Was it intended to provide no further discussion about using this technique for lower limit calculation? It is commonly used and seems worth mentioning. Many readers unfamiliar with this technique may gloss over its small mention in the text. Should the document elaborate on this

method?

The term “extra risk” is used throughout the document. Perhaps “additional risk” (P(d) – P(0)) should also be defined and listed in the glossary. If the guidance document prefers the extra risk calculation over additional risk, perhaps this should be explicitly addressed.

Topic I, Question 2. (Literature review)

In addition to the simulation study by Kavlock *et al*. (1996) cited on page 9, line 25 there is also a simulation study from our group (Weller, *et al.*, 1995, *Risk Analysis*) that examined study design in the clustered data, developmental toxicity setting with particular emphasis on model fitting and benchmark dose estimation. Perhaps this paper should be cited.

Page 10, around line 10 mentions work on multiple outcomes but the literature review does not cite some recent, very relevant work in this area. In particular, papers by Regan and Catalano (1999, *Biometrics*; 1999, *Journal of Agricultural Biological and Environmental Statistics*; 2000, *Risk Analysis*) that investigate multivariate modeling and BMD estimation of joint continuous and binary endpoints might be referenced. Additionally, a recent paper by Molenberghs and Ryan (1999, *Environmetrics*) on multiple outcomes in developmental toxicity might be referenced. These references should also be listed on page 34 around line 5 in the Multiple Outcomes section.

Around line 21 also on page 10 and also on page 24, line 4 are some important references to novel approaches for modeling continuous outcomes. Perhaps the RACO method of Bosch *et al.* (1996, *Risk Analysis*) should also be cited and discussed. It has direct relevance and offers an alternative to the methods that are covered in detail in the guidance document.

Topic I, Question 3. (Selection of studies and endpoints)

The executive summary (page vii around line 5) should also address the problem that for clustered

data (which is given substantial attention in the document) more than just a measure of response variability per dose group may be needed in order to perform an appropriate analysis. In particular, data at least at the level of each cluster (litter) are typically needed. Individual level data or, minimally, intra-litter correlations for each litter may also be required for some models. The document does not address these potential requirements. Not only does this have implication for people collecting new data but it underscores the problem of trying to fit models and compute BMDs from summary data in publications and online databases that do not contain individual level data.

The executive summary (page vii line 27) and other places throughout the document (e.g., page 20, line 16) mention the concept of "dichotomizing" the data in a way that is somewhat vague and can be subject to multiple interpretations. For example, does the document mean that the data are dichotomized before modeling (i.e., turning measured outcomes into binary ones), dichotomized after a continuous model fit (using a cutpoint after a continuous data fit) or dichotomized in a "hybrid" model sense (see Bosch, Wypij, and Ryan 1996, *Risk Analysis* where binary data result from pairwise comparisons between measured outcomes from control and dosed animals). For a technical guidance document a bit more attention should be given to exactly what is meant by "dichotomize" in the various places that the term is used.

Page 14, line 26 mentions that number of responses and number of subjects per group are needed for modeling but for clustered data these summary statistics are needed *per cluster (litter)*, not just per dose group. Perhaps this should be addressed. The same argument can be made for the discussion on page 15, around line 3, for continuous data. The document mentions that number, mean and variability estimates are needed for each dose group but in fact for clustered data a finer level of reporting is required, specifically reporting at least at the cluster level (including a measure of intra-cluster correlation for each cluster). Perhaps more attention could be given to this issue.

The issue of studies with multiple endpoints is taken up briefly on page 16, line 10. Perhaps it should be mentioned that individual level data (and specifically *not* simply dose level summaries of each outcome tabulated separately) are required for any rigorous multivariate analysis of the data.

Topic II, Question 4a. (Proposed defaults for parameters)

The issue of adding covariates to dose-response models is mentioned briefly on page 26, line 18 and in the developmental toxicity example (page 73, line 24 and page 78, lines 68-69) but no guidance is given on how to compute or interpret BMDs computed from non-linear models that include covariates. This is a difficult problem and there are no standard techniques to solve it. The reader is left wondering what to do (or, worse, not fully appreciating the problem and thinking it is trivial). Since the answer to the question is not at all apparent, it is suggested that, at least, the problem should be discussed and issues raised.

It is suggested that to clarify issues a bit the phrase "For most non-linear models," be added at the beginning of page 26, line 27. (That is, to clarify that iterative fitting is often, but not always, required when fitting dose-response models.)

On page 32 line 15 and elsewhere in the document the technique of taking logarithms of dose is discussed. Perhaps it should be mentioned exactly what is meant here, specifically regarding the issue of what to do with the control dose (0) when taking logs. This is often a point of confusion to many people and there are several approaches to handling the problem. Perhaps the document could list some common ways to handle this and provide guidance on which are considered preferable.

Topic II, Question 4b. (Model fit criteria)

No comments.

Topic II, Question 4c. (Selecting among models)

The example on quantal data (page 55 line 2) uses the log-logistic model but does not provide

parameter estimates for the fitted model (nor does it actually define the log-logistic model). Perhaps a bit more detail should be given here to help readers understand the fitting. Showing the fitted model (mathematically) would help.

In the continuous data example (page 65-66) several models ("A1", "A2", "A3" and "R") are fit to the example data and tests related to these models are shown on page 66 but no description of these models is given in the document making it impossible for readers to understand exactly what is being fit and what tests are being conducted. It is strongly suggested that more material be added here to make the document self-contained.
It would be helpful if the document provided the fitted power parameter value for the model described on page 67, line 14. It would provide useful information when comparing the results to the preferred linear (called polynomial in the document but really linear) model on page 67, line 13.

On page 75, line 9 the BMDL is listed as 410 but in the output on page 79 line 5 it is 409.

On page 76, the standard error estimates for phi4 and phi5 (lines 51 and 52) are zero in the output. This seems anomalous. Perhaps it should be discussed in the text of the example. Also, how is it that the "Default Initial Parameter Values" (lines 29-38) are identical to the (presumably) final "Parameter Estimates" (lines 45-52) obtained after fitting the model?

Topic II, Question 5a. (Confidence limits)

The method of using the LR (likelihood ratio) to obtain the BMDL is discussed in some detail on page 31 around line 24. The example given is for finding the confidence interval for a direct parameter of the model. On the next page (32, line 3) it is mentioned that the calculations are more complicated when the value of interest cannot be expressed as a model parameter. Since most BMD estimation problems fall into this class perhaps more discussion of how to compute LR intervals should be given. It is suggested that at least a reference on using the method for this case be given. A good one is Chen and Kodell, 1989, *Journal of the American Statistical Association*.

It is curious that the LR confidence interval method is not listed as an option under the "Quantal Data" section starting at line 5 on page 32. Perhaps it should be mentioned as an alternative to the delta method techniques that are discussed in this section.

On page 33, lines 6-7 the sentence "A simple example is the moment estimates." is a bit vague. It is unclear if this refers simply to the dispersion parameters or to all parameters in the GEE model. Presumably it is the latter but perhaps the sentence could be clarified.

Topic II, Question 5b. (Deviating from 95% one-sided CI)

No comments.

Topic II, Question 6. (Additional examples)

Related to the response above regarding use of covariates in dose-response modeling (Topic II, question 4a) a new example (or part of one of the existing examples) that provides detail on how to approach the problem of defining the BMD in the presence of covariates should be included in the guidance document. The examples given do not address this issue directly but the guidance document does mention adjusting for covariates and in practice it is an important issue.

Topic III, Question 7. (Reporting requirements)

No comments.

Topic III, Question 8. (Other comments and suggested changes)

In general the topic of multivariate/multiple outcomes modeling does not receive much attention in the guidance document but since many non-cancer assays collect several (sometimes *many*) endpoints that could be candidates for BMD estimation and multivariate modeling, perhaps the document should stress more the data requirements for analysis (usually individual level data) and provide more references to recent papers in the growing literature on this subject.

In the guidance document, several techniques are described for defining and finding benchmark doses for continuous data. Perhaps it should be mentioned that any technique involving estimation of the BMR or response cutpoint to define the BMR from the concurrent data are subject to increased variability in the final BMD estimates exactly from the estimation of the BMR or cutpoint. That is, when the concurrent data are used to define it, the BMR or cutpoint is actually a *parameter* of the BMD estimation process yet the methods described in the guidance document treat these quantities as fixed constants. Because they are actually estimated from the data, and themselves subject to variability, treating these quantities as fixed constants will ultimately underestimate the true variability of the BMD estimated in this fashion. This is in contrast to modeling binary data where the BMR is, typically, a fixed constant (e.g, 10%). The literature on this subject does not seem to address this issue but nevertheless the problem remains. An example of how this technique can introduce variability in the final BMD estimate is given on page 67 of the guidance document in the use of the dynamic range of the concurrent data to obtain a 5% quantile of the dynamic range as a basis for finding the BMD. The 5% value (82.9, line 9) is estimated from the data yet no accounting of its variability is used in obtaining the variability of the final BMD (or BMDL). The document considers this “creativity” (line 28) and although it is quite a novel idea, perhaps the discussion should also address variability issues in using this method (and any other method like it, such as defining the BMR through the tail area of the control distribution or through a certain multiple of control standard deviations shift from the control mean).

References mentioned in the above premeeting comments.

Bosch RJ, Wypij D and Ryan LM (1996) A Semiparametric Approach to Risk Assessment for Quantitative Outcomes. *Risk Analysis.* **16**: 657-665.

Chen JJ and Kodell RL (1989) Quantitative risk assessment for teratologic effects. *Journal of the American Statistical Association.* **84**: 966-971.

Molenberghs G and Ryan LM (1999) An exponential family model for clustered multivariate binary data. *Environmetrics.* **10**: 279-300.

Regan MM and Catalano PJ (1999) Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics.* **55**: 760-768.

Regan MM and Catalano PJ (1999) Bivariate dose-response modeling and risk estimation in developmental toxicology. *Journal of Biological, Agricultural and Environmental Statistics.* **4**: 217-237.

Regan MM and Catalano PJ (2000) Regression models and risk estimation for mixed discrete and continuous outcomes in developmental toxicology. *Risk Analysis.* **20**:363-376.

Weller EA, Catalano PJ and Williams PL (1995) Implications of developmental toxicity study design for quantitative risk assessment. *Risk Analysis.* **15**, 567-574.

Harvey Clewell

Harvey J. Clewell III is a professional research manager with over twenty years of experience in environmental quality research, toxicology research, and hazardous materials management. He has gained an international reputation for his work in the application of physiologically based pharmacokinetic (PBPK) modeling to chemical risk assessment. He has played a major role in the first uses of PBPK modeling in cancer and non-cancer risk assessments by EPA, ATSDR, OSHA, and FDA, for such chemicals as methylene chloride, trichloroethylene, vinyl chloride, and retinoic acid. He has also been heavily involved in initial efforts to implement the new EPA cancer guidelines regarding the use of mode of action considerations in cancer risk assessment. Mr. Clewell has authored numerous scientific publications, has provided testimony both in civil tort cases and congressional hearings, and frequently provides invited lectures and computer workshops in the areas of pharmacokinetics and risk assessment. Prior to joining ICF Consulting, Mr. Clewell served for 20 years as an officer in the U.S. Air Force Biomedical Science Corps; his duties included Deputy Director of the Air Force Toxic Hazards Research Unit, Director of Hazardous Materials Safety for the Air Force Aeronautical Systems Center, and consultant to the Air Force Surgeon General on Chemical Risk Assessment.

Comments on Benchmark Dose Technical Guidance Document

I. Preparation for BMD: Selection of Data and BMR

1. Definition of concepts and terms: The various concepts and terms used in the document are very clearly described. I strongly agree with the standardized use of the abbreviations BMD and BMDL for the central estimate and lower bound, respectively. I would suggest recommending the use of a subscript for the BMR (e.g., $BMDL_{10}$ for the lower bound on the benchmark at a BMR of 10%).

2. Citations: The literature review provides adequate citations to the most important published work on the BMD approach, although I would suggest including some of the work by Wout Slob and colleagues, e.g.: Vermeire T, Stevenson H, Peiters MN, Rennen M, Slob W, Hakkert BC. 1999. Assessment factors for human health risk assessment: a discussion paper. Crit Rev Toxicol 29(5):439-90.

3. Selection of studies and endpoints: The discussion of study and endpoint selection is excellent. I am very much in favor of the recommended use of a standard BMR for comparison purposes. I also agree with the suggestion that the comparison be made on the basis of the BMD and BMDL corresponding to a 10% BMR for quantal data, and on the basis of a change in the mean equal to one control standard deviation from the control mean for continuous data.

I think it should be clarified (on p. vi and p. 17) that while a positive trend test is a normal prerequisite for performing BMD analysis on a data set, there may be cases in which dose-response modeling of an endpoint in a well reported but Anegative@ epidemiological study may be justifiable based on positive associations obtained in other studies of the same endpoint that are not amenable for BMD analysis. A published example is the BMD analysis of the Anegative@ Seychelles study of methylmercury developmental neurotoxicity cited in the last example in the BMD Technical Guidance Document (Crump et al. 2000). A similar approach has also been published for BMD analysis of neurological endpoints in a Anegative@ study of occupational manganese exposures: Gibbs, JP, Crump, KS, Houck, DP, Warren, PA, and Mosley, WS. 1999. Focused Medical Surveillance: A Search for Subclinical Movement Disorders in a Cohort of U.S. Workers Exposed to Low Levels of

Manganese Dust," NeuroToxicology 20, 299-313.

In the executive summary (p. vi), it is implied that having a dose near the BMR will reduce the confidence interval. I don=t think this is necessarily true.

II. Computing the BMD: Model Selection, Fitting and Confidence Limits.

4. Model selection and fitting:

a. Model parameter defaults: I am reasonably satisfied with the discussion of model selection and bounding of parameters. However, I think the discussion of the need for constraining power parameters to be no less than unity (p. 25) should be more strongly worded, and should be expanded to provide the information that was intended to be included by the reference to AExample 1". I can=t think of a single case where I have felt that there was a justification for considering the unconstrained versions of such models.

b. Model fit criteria: The criteria for evaluating model fit are comprehensive, to say the least, but they may be overwhelming for a typical practitioner. I agree with the use of 0.1 as a cutoff for the goodness-of-fit statistic, and with the importance of visual evaluation of fit (plots). I also agree with most of the suggestions for improving model fit, particularly the elimination of high doses. However, I am very strongly opposed to recommending log-transformation of the dose, which severely distorts the dose-response. For example, the appropriateness of log-transformation of dose in BMD modeling for the Faroes study (Budtz-Jorgensen et al. 2000, cited in the last example in the BMD Technical Guidance Document) was considered by the NAS during their recent review of the RfD for methylmercury, because the lowest BMDLs in that study were obtained from modeling of log-transformed doses. The NAS determined that the log-transformation of doses was inappropriate, and recommended the use of the untransformed dose-response modeling.

Note: Despite its appearance, the log-logistic model, as described on p. 21 of the document, does not constitute a log-transformation of dose. It can be rewritten as $p = P0 + (1-P0)/(1+K/dose^b)$ where $K=e^{-a}$.

I agree with the usefulness of Aikake=s Information Criterion, but I think its importance is overemphasized in the first example. The AIC only needs to be used if the simplest models fail; it would then be useful for evaluating the value of increasing the complexity (parameterization) of the models considered.

c. Selection among Aequally@ fitting models: I think the discussion of how to deal with multiple model estimates is satisfactory.

5. Use of confidence limits:

a. Description of computational approaches: The discussion of confidence intervals is acceptable, but is highly technical compared to the rest of the document. It is probably not very useful to the typical practitioner, who will undoubtedly be using a software package (such as the EPA=s BMDS program) that handles it transparently to the user. It might be better to eliminate some of the detail or relegate it to an appendix.

b. Soundness of criteria to depart from 95% one-sided CI: I couldn=t find any discussion of this question in the document, and don=t understand why such a departure would ever be necessary anyway.

6. Examples:

The numbering of the examples should be fixed. Note that the example referred to in the text on p. 25 as AExample 1" is actually in chapter 2 of the Examples section, and is not referred to there as Example 1.

The examples for both quantal and continuous data provide excellent illustrations of the process of model selection, but it should be pointed out that they have been designed to illustrate a number of points about BMD modeling and do not necessarily represent a typical situation for BMD modeling. In the first example, the unconstrained versions of several models are included in the comparison. However, as mentioned in the text of the document on p. 25, the unconstrained versions of models with estimated power parameters should not even be considered in a typical BMD analysis. I assume that they are only included in the example to demonstrate why they are not recommended, but I=m not sure a typical reader will understand that they are only included for didactic reasons. Similarly, the Hill model should only be considered if there is a compelling reason

for using it (and I have yet to come across a case where there was a compelling reason for using it). The second example should be rewritten to put more emphasis on the calculation of a conventional BMR (increase from control) rather than a non-conventional one (fraction of maximal response).

Again, I think the points made in both of these examples are good; I just feel that the visibility they give to the less plausible and more complex models could actually lead some users (who may tend to look at the examples more than the main text) to give them more attention, rather than less. All of the problems just described result from trying to use the examples both to demonstrate issues in BMD modeling and to provide illustrations of typical procedures. It might be useful to include additional examples of simpler applications. Alternatively, the current examples could be made simpler by focusing on illustrating a typical analysis rather than demonstrating advanced issues. It's really not necessary to demonstrate the problem with unconstrained power models in an example; it's enough to just say they shouldn't be used. Similarly, its not necessary to introduce the Hill model to have an example where dropping the high doses improves the ability of the standard models to estimate a conventional (increase over control) BMR.

Finally, there should definitely be an example of the use of BMD modeling for human epidemiological data for a continuous endpoint using the hybrid approach. This is probably the single most important use of BMD modeling, since the NOAELs determined by conventional approaches (e.g., categorization, summarization) with such data are artificial. None of the studies cited as examples in chapter 6 of the Examples section actually comes close to conforming to the recommendations in the BMD Technical Guidance Document and could be considered to be an appropriate example.

III. Interpretation and Use of BMD

7. Reporting requirements: The listing of reporting requirements seems adequate, and no additional discussion seems to be needed.

8. Other comments: The guidance document should be more directive in terms of the preferred approach for conducting BMD modeling. Allowing flexibility is good, but the most appropriate models and modeling approaches should be clearly specified. The first example gives the impression that a large number of different models, including over-parameterized and implausible ones, should always be run for every data set. This would not only be a waste of time, it would also be extremely confusing, particularly when a number of data sets must be considered. In practice, it seems like it would be more efficient to recommend starting with the simple models (polynomial, constrained Weibull, power, etc.) and only bring in more complex models if the simple ones fail to adequately describe the dose-response. This approach appears to be implied by the text on p. 29, but I think the point should be made more clearly both there and in the examples.

In particular, it should be clearly stated (on p. 24 and in the continuous example) that the hybrid modeling approach is the preferred approach for modeling of continuous data. It should also be made clear that the hybrid models can be used with any of the BMR definitions listed on p. 20 of the BMR Technical Guidance Document, since the cutoff for abnormal can be defined in terms of either a percentile of the control distribution or a continuous response value. As pointed out in the guidance, using the hybrid model avoids the loss of information inherent in discretization, and therefore the hybrid approach should definitely be preferred to dichotomization and use of a quantal model.

George Daston

BIOGRAPHICAL SKETCH

George P. Daston
Research Fellow
Procter &Gamble

EDUCATION:

INSTITUTION	DEGREE	YEAR CONFERRED	FIELD
Univ. of Miami	Ph.D.	1981	Biology (Teratology)
US EPA, RTP, NC	Post-doctoral	1983	Developmental Toxicology

EMPLOYMENT:

1985-present Procter & Gamble, Miami Valley Labs
1997 on: Research Fellow
1993-97 : Principal Research Scientist, developmental and reproductive toxicology
1990-93: Group Leader, Developmental, Reproductive and Neurotoxicology
1985-90: Staff Scientist, developmental toxicology

1983-85 Univ. of Wisconsin-Milwaukee, Dept. of Biological Sciences, Asst. Professor

BIOGRAPHICAL SUMMARY AND RESEARCH INTERESTS

Dr. Daston has spent his entire career in research to understand the effects of exogenous chemicals on the developing embryo, fetus and child. His research interests include teratogenic mechanisms, in vitro methodologies, and risk assessment. He has published over 70 peer-reviewed articles, reviews and book chapters, and has edited three books. His most recent research includes 1) toxicant-nutrient (especially zinc) and maternal-embryonal interactions in developmental toxicity; 2) the role of pattern formation genes in abnormal development; 3) genomic approaches to endocrine disrupter screening; and 4) improvements in risk assessment methodology for non-cancer endpoints. Dr. Daston's activities in professional societies include serving as Chair of the Reproductive and Developmental Effects Subcommittee of the American Industrial Health Council (1990-99); Chair of the Developmental and Reproductive Toxicology Technical Committee of ILSI-Health Effects Sciences Institute; President (1994-95) of the Society of Toxicology's Reproductive and Developmental Toxicology Specialty Section; President (1999-2000) of the Teratology Society; member of the National Academy of Sciences Board on Environmental Studies and Toxicology (1995-98); and member of EPA's Endocrine Disrupter Screening and Testing Advisory Committee (EDSTAC). Dr. Daston has recently served on the organizing committees for: ILSI/EPA/AIHC workshops on benchmark dose methodology and human variability in toxic response; an EPA workshop on endocrine-mediated toxicity; and as co-chair of an AIHC/EPA workshop on Leydig cell tumors, an ILSI/EPA workshop on interpreting reproductive toxicity endpoints and a NIEHS workshop on the state of validation of the FETAX assay for teratogen screening. Dr. Daston is an Associate Editor of *Toxicological Sciences*, Field Editor for *Teratogenesis, Carcinogenesis and Mutagenesis*, on the Editorial Board of *Human and Ecological Risk Assessment* and *Reproductive Toxicology*, and an ad hoc reviewer for *Teratology*, *Journal of Nutrition* and other journals. Dr. Daston is an adjunct professor in the Department of Pediatrics and Developmental Biology Program at the University of Cincinnati and Children's Hospital Research Foundation, and lectures in courses on teratology, developmental biology, toxicology, and risk assessment. Dr. Daston was a Visiting Scientist at the Salk Institute, Molecular Neurobiology Laboratory, 1993-94. Dr. Daston was elected a Fellow of AAAS in 1999.

George P. Daston
November 22, 2000

Comments on EPA's Benchmark Dose Technical Guidance Document

General Comments: The draft provides a very good description of how data should be analyzed to determine a BMD as the point of departure for risk assessment. I found it to be in good shape; I had few substantive comments. EPA should be encouraged to move ahead in implementing the guidance provided here, and in making more routine use of BMD methodology as a cornerstone of its harmonization of risk assessment approaches.

There are, however, two implementation issues that concern me: the first is the use of a lower confidence limit of the BMD, instead of the central estimate, as the starting point for risk assessment; the second is the decision that the benchmark level of response can vary depending on endpoint. Both of these issues have engendered considerable debate in previous discussions of benchmark dose implementation. To my knowledge there has not been a resolution to those debates, at least not a consensus resolution. I have presented my opinions on these issues in my comments to the specific questions we were asked (below).

Regardless of the decisions that EPA ultimately makes on these issues, it will be important to readers of the guidance document (many of whom will have the responsibility for implementing its recommendations) to understand the pros and cons of each view. I urge EPA to add these when it revises the document.

Responses to Questions:

I.1. All of the terms pertaining to benchmark dose are defined clearly, and the concepts are well described. Though not directly pertinent to the definition and usage of benchmark dose, I believe that the term "practical threshold" (p. 3, line 22) is being incorrectly used in the draft document. The draft describes practical threshold as the dose level at which no effects are observed in a particular study; however, in the literature the term has been used to describe a dose level for which the theoretical response is so small that it is unlikely to be measurable by any practical means (Gaylor et al., 1988, Teratology 38: 389-391; Daston, 1993, Iss. Rev. Teratol. 6: 169-197). (NB: Gaylor et al. termed this a "pragmatic threshold", but the intended meaning is unchanged.) This is a minor point.

I.2. The literature review is good. The only addition I would make is to cite Allen et al. (1996) in the context of the discussion on p. 15, lines 10-28. (The paper is already cited elsewhere in the text to illustrate a different point.) The discussion deals with how to model responses that are characterized in terms of severity of effect. Allen et al. (1996) analyze a study on the developmental toxicity of boric acid. One of the most notable effects in this study was a change in the rate of a particular skeletal variation (lumbar rib) with increasing dosage, with the highest dosage producing frank malformations of the13th thoracic rib and vertebra. If one assumes that the variation is a less severe manifestation of the same mechanism of action that caused the malformation, then one can assign a particular weighting factor that indicates the severity of the effect that each fetus bears. This provides a single, weighted variable that can be plotted on the same dose-response curve. Allen et al. present models in which one weights the variation as ½ as severe as the malformation, 5/6 as severe (an assumption that the variation is tantamount to a malformation in severity), or 1/6 as severe (an assumption that the severity of the variation is trivial).

I.3. On p. 14, lines 1-2, the document states that studies with more dose groups and a graded dose response will be optimal for BMD analysis. Unfortunately, EPA testing guidelines are very clear about the number of dose groups and animals per dose group that should be used, and these guidelines are not optimized for determining shapes of dose-response curves. While this document is not the place to revise test guidelines, it would be useful for it to suggest that added flexibility in those guidelines might improve the quality of dose-response analysis, and by extension of risk assessment. One of the downsides of current study design is the case presented on p. 17, lines 7-28; it would be nice, to revise lines 23-24 to indicate that the ideal situation is not to collect more data but to allow flexible study design so that the dose-response curve is better defined after the first study.

On p. 15, lines 5-8: I disagree that it is reasonable to assume that the variance of treated groups is comparable to that of the control group. A study that does not report measures of variance for treated groups should be judged as insufficient to support BMD modeling (or NOAEL determination) and should not be used for risk assessment.

On p. 19, lines 6-7, I don't understand why EPA thinks that it is not critical to use a common response level as the POD for all endpoints or chemicals. Isn't this injecting subjective value

judgments about relative levels of concern for particular endpoints? This same problem occurs again in the default procedure presented on p. 20, lines 1-6. In this latter instance it is stated that the reason for selecting a lower BMR in some instances is because the study designs permit it. In other words, the BMR will be endpoint-specific and selected on the basis of the sensitivity of the study designs for each endpoint. The NOAEL's dependence on sample size is one of the principal reasons for contemplating a BMD approach (p. 4); the decision to not use common response levels appears to be a failure in correcting this problem and thereby limits the attractiveness of the method.

I'm not sure that we know yet what will be the best procedure(s) for handling continuous variables, but I think that EPA makes a good case for the procedures outlined on p. 24. It would be very worthwhile to revisit these recommendations after we have had a few years experience in analyzing continuous data in this way.

II.4.a. For the most part this is a good discussion. On p. 26, lines 3-10, it would be useful if the document provided some guidance on how to select a threshold term for modeling purposes. Some ideas that come to mind are 1) use the x-intercept of a straight-line fit of the experimental data, or 2) use a value that provides the best fit.

II.4.b. This section is satisfactory. I would like to see a greater emphasis placed on graphing the data, though. While this doesn't obviate the need for quantitative measures of fit, it does make an excellent reality check. It also expands the pool of experts evaluating the model fitting exercise to include the biologists, most of whom (myself included) tend to become glassy-eyed when faced with too many tables of residuals.

II.4.c. In my opinion, having a number of models that fit the data equally well is a nice problem to have, in that it suggests a greater level of confidence that the data have been fairly represented in the analysis. Since the purpose of the whole exercise is to have a POD that EPA has a great deal of confidence in, I would suggest as a default procedure that the model with the tightest confidence intervals at the BMD be chosen from among models with comparable fit.

II.5.a. No comments, other than to say that the potential problems listed for computing confidence limits appear to be the result of having low sample sizes or choosing a BMR rate that is too far below

the experimental dose region. The former suggests that the study being used as the basis for risk assessment is inadequate to satisfy regulatory requirements for statistical power and shouldn't be used in the first place. The latter suggests that the study being used is inadequate in its dose selection and shouldn't be used, either. In other words, if a good job is done in selecting the studies used for risk assessment, then potential problems in computing confidence intervals will be obviated.

II.5.b. No comments

II.6. The examples are adequate. It would be nice to have references added to the Examples section citing instances in which BMD has been used in setting RfD/RfC values. I would like to review those to see how the guidance provided here is applied, or if different guidance has been used whether it will be informative in revising this document.

III.7. These seem reasonable. I don't know whether the omission was intentional, but I was pleased to see that there wasn't a requirement to report the NOAEL for the critical effect along with the BMD. I think we need to resist the temptation to constantly check our results against the NOAEL, which is a flawed standard. There is an ample number of objective reality checks built into the BMD reporting requirements.

III.8. My biggest concern remains the use of the lower confidence limit instead of the central estimate for BMD as the starting point for risk assessment. The principal reason for doing so is to "reward better experimental designs and procedures" (p.31). It needs to be pointed out that the aspects of experimental design that contribute to low variability and high resolving power have already been painstakingly considered in the development of most modern testing guidelines. These testing guidelines represent the collective efforts of the best minds in toxicology, within and outside EPA and are regarded in the toxicology community as being satisfactory for detecting and characterizing toxic responses. The guidelines are highly standardized in protocol and conduct. The toxicity studies that are submitted to EPA for risk assessment are, for the most part, compliant with those highly standardized guidelines. Therefore, the issue of "rewarding better studies" is, in large part, moot, as all guideline studies fall into the category of "better studies".

For guideline studies, I recommend that EPA use a central estimate of BMD as the POD for risk

assessment. The use of a BMDL should be reserved for those studies that aren't compliant with testing guidelines. In those cases, use of the lower confidence limit will fulfill the purpose of ensuring that the BMR is not exceeded.

Elaine Faustman

Elaine M. Faustman, Ph.D., D.A.B.T., received her A.B. in Chemistry and Zoology from Hope College (1976) and her doctorate in Pharmacology/Toxicology from Michigan State University (1980). She took her postdoctoral training in Toxicology and Environmental Pathology in the School of Medicine at the University of Washington. She is currently Professor and Director, Institute for Risk Analysis and Risk Communication in the School of Public Health and Community Medicine at the University of Washington, where she has received the Outstanding Teaching Award. Her research interests include quantitative risk assessment for non-cancer
endpoints, reproductive and developmental toxicology of metals, and “in vitro” and molecular biological methodologies. She is the Director of a EPA-NIEHS funded Child Health Center which is evaluating key mechanisms defining the children's susceptibility to pesticides. She is an elected fellow of the American Association for the Advancement of Science and has recently served as chair for the National Academy of Sciences Committee on Developmental Toxicology. She is a member for the NIEHS-NTP Committee on Alternative Toxicology Methods. She has served on the NIEHS-NTP Board of Scientific Counselors and the National Academy of Sciences Committee in Toxicology. She has served as Associate Editor of *Fundamental and Applied Toxicology* and editorial boards of *Reproductive Toxicology* and *Toxicology Methods*.

Benchmark Dose Technical Guidance Document

Introductory Comments

The document preparers deserve a great deal of praise for putting together this much needed technical guidance document. I felt that most of my comments, responses were of a nature to fine tune and focus rather than reinvent, however having said that I feel that significant more work is needed on the document. A particular issue that I, as a reviewer kept returning to, was that of who is the intended audience. Lack of clarity on this point results in a document that is somewhat uneven with a mixture of both very basic information and some rather technical material with the middle-road users left somewhat unaddressed. I have tried in my comments to add questions, and points that need clarification specific for that user. I think that this document has the potential and need to address this type of user and to bridge the interface between the toxicologist ant the statistician to be useful for risk assessment.

A more methodical approach is needed or at a minimum rationale given for each example presented. Specify what points are to be illustrated by each example. This would ensure more completeness in our guidance. Give a table as a user's "troubleshooting" guide so a user who is advanced can rapidly go to specifics. I have made numerous suggestions where figures or tables which summarize text should be added to clarify text and be helpful to both the basic and advanced user of this guidance document.

Mention is given in the technical document to related documents and programs either developed or under development. More information and specifics about how all these materials will connect is probably needed in the technical document. This would assist the reviewer in understanding what needs to be repeated and yet avoid complete duplication or re-interpretation of existing guidance. Obviously similar examples should be available for both the software packages and in this technical document. Has that been coordinated? Please explain the larger picture for benchmark development within the agency.

REVIEWERS RESPONSES TO CHARGE QUESTIONS:

Question 1: Preparation for Computing a Benchmark Dose…

The document seems clear in the concepts and terms regarding benchmark dose (BMD). See my comments on each section of the technical document for comments on redesigning Figure 1 for clarity in this section of the document. I do feel however that the abbreviations used for BMD and BMDL should always have the response level specified as part of the BMD or BMDL abbreviation. For example us a subscript to designate 10 percent response or just immediately following the response—BMD_{10} or BMDL05, etc.

Question 2: References

Although the document has a robust reference list, there are some additional references that should be mentioned.

Bogdanffy M., Daston G., Faustman E.M., Kimmel C.A., Kimmel G.L., Seed J., and Vu V. 2000. Harmonization of Cancer and Non-Cancer Risk Assessment: Proceedings of a Consensus-Building Workshop. (This is a manuscript submitted for publication that summarizes the larger cancer/noncancer harmonization meeting that took place in DC in 2000. There are many relevant discussions in this paper.)

Faustman E.M., and Bartell S.M. 1997. Review of Noncancer Risk Assessment: Applications of Benchmark Dose Methods. *Human and Ecological Risk Assessment* : Vol. **3**. No. 5, pp. 893-920.

Foster P.M., and Auton T.R. 1995. Application of benchmark dose risk assessment methodology to developmental toxicity: and industrial view. *Toxicol Lett* **83,** 555-559.

Haag-Gronlund, M., Fransson Steen, R., and Victorin, K. 1995. Application of the benchmark method to risk assessment of trichlorethene. *Regul Toxicol Pharmacol* **21**(2), 261-69.

Kimmel, C.A., Kavlock, R.J., Allen, B.C., and Faustman, E.M. 1995. The application of benchmark dose methodology to data from prenatal developmental toxicity studies. *Toxicol Lett* **82/83**: 549-554.

National Research Council (NRC) 2000. Methods for Developing Spacecraft Water Exposure Guidelines. Chapter 4, and Appendix B. National Academy Press.

Question 3: Additional clarification and discussion:

Please see the following comments on Section I introduction and Section II Benchmark Dose Guidance, Section A Data Evaluation and Endpoint Section.

Page 1--Introduction, Section A. Purpose:
Purpose of guidance document starts with a detailed discussion of application of benchmark dose yet fails to start with an introduction paragraph that defines what is the benchmark dose. Please add.

Pages 2-3--Introduction, Section B. Background:
Add reference here to other documents evaluating cancer versus noncancer issues. See especially workshop report for recent harmonization workshop jointly sponsored by Society of Toxicology (SOT) and the U. S. Environmental Protection Agency (EPA) (Bogdanaffy et. al.,

2000).

Page 4 lines 13-14:
How is the slope (when very steep or shallow) considered with NOAEL or LOAEL? Please give a few more clarifying details or give references.

Page 4 footnote:
It was unclear to this reviewer what "too high" meant in the context of this sentence, please clarify.

Page 7 figure 1:
Axis are confusing since below 0% fraction affected appears possible as are responses below a zero dose. Is there another approach that could be used to illustrate these points? Label BMR. Label BMD and BMDL using arrow so they are not confused as part of graph axis.

Pages 8-12: Introduction. Section C. Review….

Pages 8-9:
As results are presented from a variety of papers in this section to justify response levels at which study NOAELs are ascertained, please specify total number of studies and size of studies evaluated for each type of endpoint. Obviously, size and number of studies evaluated can make a difference in confidence in the NOAEL level responses. This is important as not all of these endpoints have equally robust studies to support these statements. Also, as the review specifies later, specific responses such as lethality (see page 9 lines 17-24) were evaluated. Include reference to Allen et. al. as examples of studies that evaluated viability endpoints as well as specific malformation endpoints.

Page 9:
This section discusses comparisons between model- see lines 23-24, however does not discuss the extensive comparisons of models done in Allen et. al. 1994b.

Pages 13-20: Benchmark Dose Guidance

Pages 13-15: Section 1-Data Evaluation and Endpoint Selection:
Although this section is correctly focussed on the use of benchmark calculations, there appears to be a need for a more continuous effort to add comments on what to do when the identified problems/issues arise for NOAEL/LOAEL usage. Although some comments/suggestions are present there appears to be a need for a more systematic approach to each of these points.

Page 16: Section 2-Selection of Studies to be Modeled:
Should statement be added that at a minimum all studies selected for calculation of NOAELs would also be reviewed for calculation of BMDs? As it is stated now, there would be very different studies chosen.

Page 16 paragraph 2 lines 6-14:
Contrast recommended BMD practice with current NOAEL/LOAEL practices?

Page 16 line 11:
Change the word large to larger.

Page 16: Section 3-Selection of Endpoints to be Modeled:

Page 16 paragraph 1:
Excellent discussion on why endpoints surrounding but not limited to lowest LOAEL should be evaluated. These points are very important.

Page 17 lines 23:
Add the following phrase after "about dose response…" "relationship for calculation of either NOAEL or BMR."

Page 18: Section 5-Combining data for a BMD Calculation.
Ensure that this also discusses option for NOAEL when such a section occurs.

Pages 18-19: Section B-Criteria for Selecting the BMR.
Page 18 lines 20-22:
Note that the statement in lines 20-22 regarding other agency activities on selection of the BMR probably need to be repeated clearly in the executive summary and introduction if it is not already there. This reviewer missed this comment in earlier sections, did others?

Page 19 lines 28-29 and 19 line 1:
This sentence needs to acknowledge both study design and endpoint evaluation as contributing to varying levels of detection.

Page 19 lines 14-19:
This reviewer felt it was a good idea to request the ED10 as a comparison for all evaluations.

Page 20 lines 1-6:
Is there a need to recognize that some study designs and some endpoint evaluations will result in less sensitive studies where 30-40% response level would be present? For example, for neurotoxicity studies would we also be allowing for a calculation of a higher BMR? Do both sides of this response need to be considered?

Page 20 lines 8-12 Guidance on Continuous Data:
Paragraph should specify guidance documents a user would utilize for determining "if there is a minimal level of change in the endpoint that is generally considered to be biologically significant or adverse." Specify by adding references, tables or more text.

Question II. Modeling to Compute a Benchmark Dose

Question 4. Model Selection and Filtering.
Question 4a: Additional discussion.
Please see example of points for clarification listed as follows by page location. In particular please note that the discussion on pages 23-25 needs a table that lists in order, options for matching models and endpoints.

Pages 22-24: Section 2-Background for Model Selection, Point A. Selecting the Model:
This reviewer would suggest preparing a table that compares model choices and advantages and disadvantages as well as suggested usage to accompany this text.

Page 25 Part ii. Experimental Design:
Starting on line 6 and extending through line 16, label this section as "Non independence of Observations".

Pages 25-26 Part iii. Constraints and Covarients

Page 26 lines 3-9:
This section needs an example to clarify how the term "threshold" is used and to illustrate what occurs when the response is "precipitous" and a threshold parameter is fit.

Page 26 lines 10-12:
Cite Allen et. al., 1994b.

Question 4b. Model Fit Criteria

Pages 26-28:
I found this section to be uneven in the level of detail that it provided to the reader. I would suggest that this section needs additional and consistent level of detail. Again this would be dictated by the anticipated audience which I have already suggested needs clarification. One approach to help illustrates concepts in this section would be to use a table or figure entitled something like "Examples of Model Fit Methods" which lists each method, gives basis for fit assessment, lists assumptions and gives examples of suggested applications. This would make the text more useful and easier to scan.

Other examples of where this section needs additional clarification are seen on:

Page 28 lines 4-11:
Define GEE. Explain the term "correlation structure" using an example. Please keep in mind that a potentially broad (cross disciplinary) audience may be using this guidance document.

Pages 28-29:
This section provides a start at describing methods to evaluate model fit however, it should provide more details. For example, 6 lines of text mention the utility of plotting the residuals, yet the text dose not show a plot nor show how these would be interpreted for assessing model fit. In fact, the examples attached to the document do give some more details (I would still argue somewhat inadequately) but no attempt to integrate or cross reference the examples is done with the document text or to "walk through" concepts in a systematic manner. The examples appear to be an after thought written by someone who was different than or separately engaged from those who were writing the text. Let's coordinate these sections and the critical concepts that we need to illustrate.

Page 29 lines 7-9:
Authors of the document are to be applauded. Yes, it is very important to do graphical plots of the data.

Question 4c: Advantages/Strengths of Model Comparisons

Page 29 lines 10-22:
The technical document discusses the potential option of "dropping results from the highest dose group" to improve model fit. As a reviewer of this document I am concerned that this option is casually mentioned in this section without any reference to a biological review of the data to ascertain if this is appropriate and in the absence of any clear criteria this option could be abused? Please add criteria for use.

Page 29-30 Section D. Comparing Models:
This section would benefit from a table listing approaches and applications. Table should give specific reference to examples where this approach is demonstrated in back of technical document.

Page 30 line 13:
Refers to "an external consideration " please clarify for readers.

Question 5a,b Use of Confidence Limits

Page 30 lines 24-25:
Editorial changes needed in this sentence.

Page 30 lines 26-29:
Very unclear and Page 31 lines 1-10 what "convenience" is served by this explanation of confidence intervals. Seems to be a statement of fact, not discussed at it relates to risk assessment discussions. Make it real for users by giving a specific example of BMD and illustrating choice of 95% CL as detailed.

Page 31 lines 1-10:
Document should clarify, not muddle the meaning of confidence limits. Document should state 95% CL on BMR represents with 95% confidence that the BMR will fail within this interval.

Page 31 lines 11-12:
Add the statement "for example" to this sentence.

Page 31 lines 21-23:
Clarify the implication of such statements.

Question 6: Examples

Page 31 lines 24-29 and Page 32 lines 1-3:
This paragraph is an example of the uneven treatment and statistical methods received in this document. Just enough details are provided to be confusing for general use and uneven for statistician use. This could be corrected with examples list concepts that need to be illustrated (perhaps we can engage review groups to do this methodically) address each point systematically.

Page 32 Quantal Data:
This reviewer believes the addition of a table for clarification of the classification system would be useful.

Page 32 line 13:
Shouldn't document provide a few more details within for the delta method? Don't force the users to go to references to obtain basic concepts/approaches.

Page 32 lines 22-23:
Either make sentence terminology consistent with human situations (pregnant mothers, children) or animals (pregnant dam, litter mates). As it is currently written it is mixed and confusing.

Page 33 lines 3-4:
Finally, a definition for GEE. See comments on page 28 where it was first used.

Page 33 lines 6-7:
Document should explain "moment estimates".

Page 33 lines 9-10:
General statements such as "There is still incentive to correctly specify the variance function since it improves statistical efficiency" should be modified to explain implication for BMD modeling.

Question III. Intrepretation and Using the Benchmark Dose

Question 7. Reporting Requirements:

Page 36-37. Section E. Decision Tree:
This section looks less like a decision tree and more like an outline. Please fix with a graphic.

Page 36 line 24:
At a minimum, define "reasonable candidate" in footnote to decision tree.

Page 37 lines 5-6:
Define or reference criteria in decision tree for determining when an "alternate model" is superior.

Page 37 lines 8-9:
Reference specific section in document that discusses human data and the "case-specific" methods

Page 37 Point 3, lines 10-15:
Decision tree should provide more details and at a minimum reference context or examples that clarify each statement. Readers are left with a feeling that the BMR model can do "Model Shopping" to see best outcome. This needs clarification.

Page 37 lines 22-23:
Text on these lines need clarification. "In which case, further analysis may be appropriate", without references, explanations such comments are useless.

Question 8-Additional Needs:
Please see my initial comments and reviews throughout this critique.

Clay Frederick

Dr. Clay B. Frederick
Sr. Research Fellow and Research Section Manager
Rohm and Haas Company

Dr. Frederick has expertise in mechanistic toxicology research and in biologically-based risk assessment. He and his coworkers and collaborators have extensively investigated the mechanisms underlying a variety of toxic responses observed in descriptive toxicity studies. The information from these mechanistic studies has been used to construct physiologically-based pharmacokinetic and pharmacodynamic (PBPK/PD) models to describe animal and human physiology, metabolism, and the biological responses associated with xenobiotic exposure. These models have been extensively evaluated relative to risk assessment methods based primarily on default methodologies. He is currently a member of the Society of Toxicology, American Association for Cancer Research, Environmental Mutagen Society, and the Society for Risk Analysis. He has served the Society of Toxicology as President of the Risk Assessment Specialty Section and as a member of the Risk Assessment Task Force. He is an Associate Editor for *Toxicology and Applied Pharmacology*. He holds an appointment as an adjunct associate professor in the Pathobiology Department at the University of Pennsylvania, and he has participated in numerous peer review and advisory committees for government and industry. He has been a diplomate of the American Board of Toxicology since 1983.

Peer Review Workshop on the Benchmark Dose Technical Guidance Document
December 7-8, 2000
Response to Charge Questions --- Clay Frederick

Charge Questions:

I. Preparation for computing a benchmark dose: selecting data and an appropriate benchmark response level

I. 1. *What concepts and terms related to the BMD, if any, do we need to define more clearly?*

Response: The BMD Technical Guidance Document does an excellent job of describing BMD methodology and defining the terms associated with its use.

I. 2. *The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?*

Response: The following references were noted on RiskAnal recently:
Bailer, A.J., Stayner, L.T., Smith, R.J., Kuempel, E.D. and Prince, M.M. (1997) Estimating benchmark concentrations and other non-cancer endpoints in epidemiology studies Risk Analysis 17: 771-780.
For aquatic/environmental tox, other references might include:
Bailer, A.J., and Oris, J.T. (1997) Estimating inhibition concentrations for different response scales using generalized linear models. Environmental Toxicology and Chemistry 16: 1554-1559.
and
Bailer, A.J. and Oris, J.T. (2000) Defining the baseline for inhibition concentration calculations for hormetic hazards. Journal of Applied Toxicology 20: 121-125.

I. 3. *What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?*

Response: This issue is covered on page 16 of the Technical Guidance Document. The text on this issue ends with the conclusion that "The selection of the most appropriate BMDs for use for determining the POD must be made by the risk assessor using scientific judgement and principles of risk assessment, as well as the results of the modeling process." However, the text above this conclusion appears to encourage 'endpoint shopping', which is to say, the calculation of BMDs on virtually every endpoint measured and then choosing the one with the lowest BMDL for use in the risk assessment. A clearer statement regarding the decision criteria to be used would be useful to the user.

II. Modeling to compute a BMD: model selection, fitting, and confidence limits.

II. 4. Model selection and fitting

II. 4. A. *What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?*

Response: The discussion in the Technical Guidance Document appears to be adequate.

II. 4. B. *Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?*

Response: The discussion in the Technical Guidance Document on pages 34-35 on this issue appears to be adequate.

II. 4. C. *What are the advantages/strengths of using the methods described to select among "equally" fitting models? What other methods should be considered in making a selection?*

Response: I am generally comfortable with the text on this issue on pages 34-37. However, when reviewing the use of these guidelines in the examples, I became uncomfortable. Example 4 is particularly problematic, since it suggests using a first-degree multistage model with a relatively poor curve fit (based on inspection) relative to second degree multistage model with a much better curve fit (also based on inspection). The fallacy lies in the overdependence on the use of the confidence intervals and the desire to use the most 'parsimonious' first degree model. This is really just a bias toward the use of more linear models that are typically encompassed by relatively wide confidence intervals rather that simply focusing on the best fit to the means of the dose groups. It was not clear to me from the reading of the text that 'parsimony' was a critical determinant in model selection (despite an 'offhand' comment on page 30), nor am I convinced that it should be.

II. 5. Use of confidence limits

II. 5. A. *Please comment on the approaches described to compute confidence limits.*

Response: The approaches described/used appear to be adequate. However, I am not a strong supporter or the use of confidence intervals for BMD calculations based on the typical datasets that are used. The underlying problem is noted on page 31 in the discussion of various methods for calculating the confidence intervals. Many guideline noncancer toxicity evaluations use 5-10 animals per dose group, and there is pressure to minimize the number of animals used for toxicity evaluations. As noted on page 31, the underlying assumptions associated with the calculation of confidence intervals tend to be problematic with small sample sizes, i.e., both the calculation and use of confidence intervals becomes increasingly problematic with smaller dose groups. The underlying issue is whether additional animals above the recommendations of standard testing guidelines will need to be sacrificed to meet the needs of calculating acceptable

confidence intervals for BMDs. Generally, I am not convinced that the justifications provided for the use of confidence intervals are adequate and that more consideration should be given to minimizing the use of animals for toxicity evaluations.

II. 5. B. *Please comment on the soundness of the criteria used to deviate from using a 95% one-sided confidence interval for the BMDL.*

Response: The criteria seem a little fuzzy to me, but I'm not sure how they can be refined. Note that changing the focus of the discussion away from the confidence interval and simply focusing on the best curve fit to the means of the data simplifies this issue considerably.

II. 6. *Examples: What additional concepts, if any, should be illustrated by an example?*

Response: An example in which the use of BMD methodology fails and a discussion of the rationale for the failure would be instructive.

III. Interpretation and using BMD

III. 7. *The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document's guidance, or would further discussion be useful?*

Response: The criteria listed seem reasonable.

III. 8. *What other comments do you have about the approach to benchmark dose described in the document? What changes to the approach would you like to see, and very importantly, why?*

Responses:
1. I am highly supportive of the EPA's development of the BMD software and the widespread dissemination of the software at no charge via the internet.
2. I have reservations about the bias toward the selection of linear models that do not fit the means of the data very well (or at least as well as a more nonlinear model) as illustrated in one of the examples.
3. The development of the software and the guidelines for its use appear to come from statisticians. Little or no consideration appears to have been given to the relatively fixed structure of many regulatory guidelines for the conduct of tox studies or to the ongoing efforts to minimize the use of animals in toxicity evaluations. I strongly suggest a better alignment of the guidelines for the use of BMD methodology with the EPA's guidance to minimize the unnecessary use of animals in toxicity evaluations. Changing the focus of the POD to the BMD rather than the BMDL would be an appropriate response to this issue.

Lynne Haber

Dr. Lynne Haber
Research Program Manager
Toxicology Excellence for Risk Assessment (*TERA*)

Dr. Haber has extensive experience in the development of noncancer and cancer risk values, and in research related to risk assessment methods. She has developed more than 15 noncancer and cancer assessments for EPA's Integrated Risk Information System (IRIS) or for EPA program offices, as well as numerous individual RfDs and RfCs. Her current interests are in the application of mechanistic information in risk assessment and in methods for extending the dose-response curve to low doses. She was the lead author for a series of case studies on the application of benchmark dose modeling, conducted for EPA/NCEA, in which issues related to the application of the modeling and use of the results were identified and discussed. The collected case studies are available as an NTIS document, and one of these case studies has been published in a peer-reviewed journal. Dr. Haber has also conducted several case studies on the application of categorical regression for the development of toxicity values for inhalation and oral databases, and made a presentation to the Risk Assessment Forum on categorical regression. Her published work includes lead authorship of the chapter on noncancer risk assessment (including dose-response modeling methods) for Patty's Toxicology. She has been a peer reviewer for a number of manuscripts, and has served on EPA peer review panels for the noncancer benzene IRIS assessment and a mixtures assessment, and on a DOD panel on an assessment for a nerve agent.

I.

1.

Response: No comment. The current general text is clear.

2.

Response: It would be useful to include some of the epidemiology applications of benchmark dose (BMD) modeling in the main text, rather than only addressing them in the last appendix. A reference on the use of BMD for epidemiology data to supplement those listed in the appendix is:
Bailer, A.J., Stayner, L.T., Smith, R.J., Kuempel, E.D. and Prince, M.M. (1997) Estimating benchmark concentrations and other non-cancer endpoints in epidemiology studies. Risk Analysis 17: 771-780.

3.

Response: The statement (p. 16, line 8) that "in most cases the selection process will identify a single study or very few studies for which calculations are relevant" is a bit overly optimistic. In my experience in using BMD modeling for doing chemical assessments, even a moderate database can yield several studies with data appropriate for modeling. In addition, even one study may yield a number of different endpoints that are appropriate for modeling, if one is using the approach of modeling all biologically relevant effects with a lowest observed adverse effect level (LOAEL) up to 10-fold the lowest LOAEL. The suggestion to use a subset of endpoints as representative of effects in the target organ or study sounds good in principle, but it is unclear how to do that while recognizing that the critical effect can often not be identified by inspection.

The statement about differences in the slope of the dose-response curve affecting the relative values of BMDLs is clear to readers experienced in BMD modeling, but is not necessarily clear to the naïve reader. The text should refer to earlier guidances where this concept is discussed in greater detail and with figures, or such a discussion should be inserted here.

The text correctly notes a statistically or biologically significant dose-related trend (rather than a LOAEL) as part of the minimal data set. It would be useful, however to explicitly note, with examples, that useful modeling can be done if there is a positive trend to the data, even if there is no

LOAEL (e.g., the carbon disulfide and methylmercury examples cited in the appendix).

Haber et al. (1998) present an example where the response at noncontrol doses jumps between no response and near-maximal response, and useful modeling results are obtained, and could be cited to illustrate this concept.

II

4. Model Selection and fitting

a.

Response: No comment. The existing discussion is clear.

b.

Response: No comment. The presented criteria are clear and adequate.

c.

Response: The approach described for selecting among "equally" fitting models has the advantage of being public-health protective after models that clearly fit poorly are eliminated. Additional rationale would be useful for the approach used when the BMDL estimates are not within a factor of 3. Since such a situation does reflect some model dependence, an argument could be made to use the better fitting-model(s), based on the Akaike Information Criterion (AIC), rather than the one(s) with the lowest BMDL. The guidance addresses this issue to some degree by saying that further consideration is appropriate if the lowest BMDL is an outlier, but it does not address the situation in which the model results fall into families, so that there is a group of "outliers" with lower BMDLs, but somewhat higher AIC values. Some perspective on what are meaningful and what are small differences in the AIC (or reference to such a discussion) would also be useful.

5.

a.

Response: No comment.

b.

Response: No comment.

6.

Response: Carbon disulfide is a nice example illustrating both the use of human data and the calculation of a BMDL when a LOAEL is not identified in the study. Both concepts are useful ones to illustrate. It would also be nice to refer the reader to a complete assessment (if one exists) that meets the criteria specified in the guidance. This would allow the reader to have a model for how to address the issues related to documentation and rationale for choice of endpoints and models when the complete database for a chemical is being considered.

Appendix example 2: It would be useful to provide the BMD/BMDL calculated initially, for comparison with the improved results. The example illustrates some useful points about local maxima. However, additional discussion about the relative importance of the larger chi-squared residual at a dose of 3 in the second analysis (P. 65) than in the first (P. 64) vs. the better estimate of the standard deviations at low doses in the second analysis. For example, it would be useful to remind the reader how the standard deviation estimate enters into the determination of the BMR, and hence the BMD. Finally, it would be interesting to know if the chi-squared residual at a dose of 3 was improved when the top doses were dropped.

III.

7.

Response: The guidance presents clear rationale as to why the required reporting elements are desirable. However, the definition of "case" is unclear and has significant implications for how burdensome these reporting requirements are. Is EPA asking for each of these elements to be reported for every model used for every endpoint? Or is this just for the chosen model for every endpoint? I agree that it is important to document the model results (BMD and BMDL) and goodness-of-fit test statistics for all of the models used for an endpoint, and to provide the rationale for the choice of model for a given endpoint. The guidance also presents a clear rationale for the utility of presenting additional model information for at least the model chosen. However, in a fairly typical case when five endpoints are modeled, using a variety of different models (and sometimes different options for some models in order to address fit issues), presenting all of the requested data for every model and endpoint can become quite burdensome. Attaching the output simplifies the issue for the risk assessor, but the total output for an assessment could run to a substantial number of pages, and would be difficult for the reader to digest. Summary tables help to organize the data,

but organizing summary tables for all of the requested data for multiple endpoints and models can also be difficult. Therefore, I propose that the BMD, BMDL, and goodness-of-fit test statistics be presented for all models run, and then that the estimates of model parameters with standard errors, and the standardized residuals only be presented for the model(s) of choice for each endpoint. This fully documents the modeling without requiring substantial data reporting from models that are excluded.

If the second major bullet in the reporting requirements (dose response model(s) chosen for each case) refers the models chosen for conducting the modeling, a bullet needs to be added for documenting the final choice of model and BMDL based on the modeling results. If this bullet refers to the final choice of model and associated BMDL, it would make more sense to list this near the end, after the calculation of the BMDL.

8.

Response:

One of the stated advantages to the BMD approach is the use of a consistent point of departure. Given that goal, it is not clear why the ability of a study to estimate a lower response level would mean that a lower BMR should be used. Unless the lower BMR is accounted for in the use of a smaller uncertainty factor (or by a modifying factor), this would appear to go back to the pattern of the NOAEL/LOAEL approach, in which more sensitive studies can be "penalized" by resulting in lower NOAELs/LOAELs. It appears that the text on p. 19 could result in a similar situation for BMDs. The exception in which it would be appropriate to use a lower BMR for a more sensitive study would be when the additional information on the shape of the dose-response curve at low doses is also used to decrease the composite uncertainty factor. The other situation when it might be appropriate to use a lower BMR would be when human data are used, the endpoint is a clear toxic effect (rather than an early precursor effect or a very mild effect), and the data are such that using a lower BMR does not involve extrapolating well below the data. In this case, use of a lower BMR would address the issue of the use of the lower bound on a 10% response for an overt toxic effect with a small uncertainty factor being possibly nonconservative, and being associated with some nonzero predicted response rate.

More information needs to be presented on data array analysis after the best model (and associated

BMDL) is chosen for each endpoint. This would include arraying all endpoints and the associated BMDLs (if available) and the NOAEL/LOAEL (if no BMDL could be calculated for that endpoint). As described in the guidance, biological relevance (and relevance to humans) should be the first basis for choosing the critical effect. However, the guidance should also address situations where biological relevance can not be used to differentiate among endpoints. It may be the case in such situations that a study that is not amenable to modeling defines a lower NOAEL/LOAEL boundary than the lowest BMDL.

Please provide references for the statements that (1) epidemiology studies often have greater sensitivity and (2) a BMR of 1% has typically been used for quantal human data.

The guidance includes some useful discussion of defining the BMR in terms of a defined adverse response. It would be useful to address the implications of whether the definition of adversity is based on an individual change or a change in mean. Doses that cause a specified response (e.g., a response 10% below the control mean, if that is defined as adverse) in some individuals will be lower than those that cause the same response in the mean.

Other/Minor Comments

The document is generally well written, and does a nice job of presenting some complex concepts. The introduction is quite nice, especially the text regarding the need to consult with statisticians.

P. 3, lines 9-10: The use of the lower bound for cancer modeling also addresses variability in human response – not just the variability in the data.

P. 6, line 24: The text should state "BMD and BMC *(not BMD and BMDL)* generically to refer to oral and inhalation values."

P. 16, line 7: Modeling should be done for studies for which the modeling is feasible *and useful*. There may be a number of studies or endpoints for which modeling is feasible, but the effect is clearly not the critical one, and so no useful information would be obtained.

P. 20, lines 14-20: Note that such dichotomization of the data results in the loss of a lot of information, and is generally not the preferred approach.

P. 31, lines 1-2: Rephrase to "but *these confidence limits* do not account for…" As written, the text could imply that the *values* do not account for….

Appendix example 1: Before going into detail about issues related to constrained/unconstrained models, it would be useful to walk through the decision tree regarding the choice of BMDL for the model results presented in the table on p. 54.

P. 61: Please explain "dynamic range of response" the first time it is used, rather than later in the example. (This term is not really defined until page 66.)

P. 62, line 10: The dose response is not "*clearly* sigmoidal." It is apparently sigmoidal, but, as noted later in the text, a concave-down model would also fit the data within the data confidence limits.

P. 63: Please explain how the value of 1600 was estimated for *V*, given that the maximal increase over background is approximately 1800.

P. 64, line 23: Please be clearer about the source of the new estimates. Do these need to be estimated manually, or does can software such as BMDS be used to conduct such estimates?

P. 66, lines 27-29: Please reduce the numbers of parentheses in the sentence.

P. 67, line 4: Please provide and refer to the appropriate portions of the output, so the reader can track the source of these numbers.

P. 67: Need some more text before the table, stating that the 5% dynamic range represents modeling in which the BMD was defined based on a specified degree of change. Again, it would be useful to refer explicitly to the decision tree shown earlier in the document – that the model with the lowest AIC is chosen because the BMDs are all within a factor of 3.

P. 68: To avoid confusion, it would be useful to state that, because the data from the 1988 DBCM assessment were used, the body weight to the 2/3 conversion was used, rather than body weight to the ¾.

P 73, lines 18-19, and p. 75, line 1: Should be Tables A-5.1 and A-5.2.

P. 73, line 23: It would help the reader in following and applying the example to show in the output the coefficients that gauge the influence of litter size. It would also be useful to explain whether combining of litters with adjacent levels of the litter-specific covariates was done manually, by the program, or some combination thereof.

Dale Hattis

Dale Hattis
Professor
Clark University

Dale Hattis, is Research Professor with the Center for Technology, Environment and Development (CENTED) of the George Perkins Marsh Institute at Clark University. For the past twenty-five years he has been engaged in the development and application of methodology to assess the health, ecological, and economic impacts of regulatory actions. His work has focused on the development of methodology to incorporate interindividual variability data and quantitative mechanistic information into risk assessments for both cancer and non-cancer endpoints. Specific studies have included quantitative risk assessments for hearing disability in relation to noise exposure, renal effects of cadmium, reproductive effects of ethoxyethanol, neurological effects of methyl mercury and acrylamide, and chronic lung function impairment from coal dust, four pharmacokinetic-based risk assessments for carcinogens (for perchloroethylene, ethylene oxide, butadiene, and diesel particulates), an analysis of uncertainties in pharmacokinetic modeling for perchloroethylene and an analysis of differences among species in processes related to carcinogenesis. He is a councilor and has recently been named Fellow of the Society for Risk Analysis, and serves on the editorial board of its journal, Risk Analysis. He holds a Ph.D. in Genetics from Stanford University and a B.A. in biochemistry from the University of California at
Berkeley.

Dale Hattis
Clark University

Peer Review Workshop on the Benchmark Dose Technical Guidance Document

General Comments

Overall, this is a good and useful guide to the resolution of statistical issues involved in calculating benchmark doses. Unfortunately there are important toxicological/mechanistic issues and policy/risk management questions that are not addressed with the same sophistication. I hope and trust that the final draft will reflect a better interdisciplinary balance of considerations.

Question #1 What concepts and terms related to the benchmark dose (BMD), if any, do we need to define more clearly?

In general, the Glossary at the end of the document could benefit from a few more equations—e.g. for the Michaelis Menten model, the gamma distribution, the Hill equation, . Also, I use the term "convex" applied to a dose response relationship to mean one where the slope becomes less steep with increasing dose. I don't know whether this is the most common usage or whether the opposite sense given in the glossary is more common.

Additionally I think there is a need for better explanation of the contrasts and contrasting implications of two different uses of bench mark dose calculations. The first of these is the use of the BMDL as a surrogate for the NOAEL in conventional calculations of RfD's (or, similarly RFC's). The second is the use of the BMDL as a Point of Departure for linear projections of potential low dose risks. The emphasis of the current guidance seems to be on the latter, but defining benchmark responses and the incidence of those responses for BMD calculations seems more critical to for the NOAEL-surrogate context.

An important example of where this makes a big difference is on pp. 18-19. After an introductory paragraph saying that the Agency is currently developing guidance for selection of an appropriate response level/percentage for bench mark dose calculations, the document describes how this is approached in the context of defining a Point of Departure for low dose risk projection. After explaining that a 10% response level is generally used for the purpose of making potency comparisons among chemicals, the document says,

> "For the POD, on the other hand, it is not critical that a common response level be used for all chemicals or endpoints, and for purposes of deriving quantitative estimates at doses below the observable range, it may be desirable to use response levels below 10%, if possible, in order to minimize the degree of low-dose extrapolation required. Thus, while it is important to always report ED10s and LED10s for comparison purposes, the actual "benchmark dose" used as a POD may correspond to response levels below (or sometimes above) 10%, although for convenience standard levels of 1%, 5%, or 10% have typically been used rather than a floating level dependent on the actual limit of detection of the relevant study.

The authors of this passage seem to have lost sight of the fact that one of the main intended benefits of the bench mark dose procedure for defining BMDL's as NOAEL surrogates was to encourage better-designed toxicological studies with lower detection limits. Industrial sponsors of such superior studies would be rewarded with higher BMDL's (and, by extension, higher RfC's and ADI's) in recognition of the decreased uncertainty of the toxicological database. If, instead (as implied by the quoted passage above) more sensitive studies are penalized rather than rewarded by lower BMDL's then we are back to the same kinds of perverse incentives that were intrinsic in the system where NOAELs are the basis of RfCs and ADIs. (A similar emphasis on the POD/low dose risk projection use is also apparent on p. 20, lines 1-2 where a lower BMR is suggested based on greater than usual sensitivity). Clearly the balance of the document between low dose projection and NOEL surrogate uses must be restored before publication, lest a major part of the purpose behind replacing NOAELs with BMDLs is defeated.

Question #2 The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?

In general, even though BMD model fitting is conceived of as primarily an exercise in statistical analysis, I think there must be some room for mechanistic considerations (e.g., Hattis, 1996). To start with, there is an extensive literature on PBPK modeling (one starting reference is Hattis, 1991), though PBPK analysis should be used ideally with some consideration for plausible definition of peak vs AUC delivered internal dose surrogates in the cases of specific endpoints. Three papers of mine discuss some options for modeling dynamic effects (dose-time-response relationships) in the context of both carcinogenesis and presumed threshold non-cancer (neurological) effects (Hattis, 1990; Hattis and Shapiro, 1990; Hattis and Crofton, 1995). As a general matter I would hope that when

analysts encounter an apparently saturating dose response relationships, as in the example on page --, then one of the principal models considered would involve a Michaelis-Menten transformation of the administered dose. I illustrate this in more detailed comments a little later. Additionally, I think it would be relevant to cite recent attempts to analyze human data on interindividual variability and use these in quantitative projections of likely risks (Hattis, 1998; Hattis et al., 1999). This is perhaps more of an alternative to the BMD approach than an example of it, but it may be considered a relevant and helpful option for some chemicals and data sets.

Question #3 What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?

I think the generic advice to do modeling for essentially all studies and endpoints where this is feasible is very good. There is only one potential source of ambiguity in the guidance given on p. 16, line 22 (that studies be considered relevant if their LOAEL is up to 10-fold above the lowest LOAEL). This is, what units are to be used to express the LOAEL for comparison in this case? The usual choice would be mg/kg, but I think it should be mg/body weight$^{2/4}$ because this rule seems most generally accurate for interspecies comparisons (Watanabe, 1992; Travis and White, 1988).

Question #4: Model selection and fitting

a. What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?

There is a mechanistic consideration that I think deserves some mention here. That is the fact that there may often be some knowledge about the mode of action of related chemicals (e.g., cholinesterase inhibitors) and that quantitative modeling approaches that have proven helpful for the action of other members of a common-mechanism group should be among the first to be considered for a new group member. Sometimes, where there are only limited data for the member of the group under study, but some comparative data allow an inference about the relative potency of the chemical under study in comparison to a reference chemical, then dose response data for the reference chemical could be used to infer benchmark doses for the chemical in question (after adjustment for the estimate of relative potencies).

Another topic that is not mentioned in the "combining data" discussion on p. 18 is the possibility that one can combine observations from different levels of severity. I do this in my analyses of human data when I have, for example, the numbers of an exposed group who report different levels of severity of irritation when presented with a variety of exposure levels. In this case I use a probit model where the slope has a direct interpretation in terms of the interindividual variability. By fitting a common probit slope but different intercepts I am able to get the benefit of the data for multiple levels of severity of response in estimating the slope—which is the primary parameter of interest in my interindividual variability analyses (Hattis et al., 1999). I usually apply a similar technique for the analysis of continuous response data for humans—defining multiple levels of response and estimating a common slope from the proportions of subjects who are pushed beyond various levels of the continuous response variable as a function of dose.

Finally, I have serious concerns about the prospect of adopting the default given on lines 21-27 of page 20—that for continuous response data "a change in the mean equal to one control standard deviation from the control mean (see Section II C2e) can be used." This level of response does indeed have the effect of shifting the population such that there are about 11.7% of the affected group at or beyond the level of the variable (2.1 standard deviations from the mean) that, in controls, represents the about the 1.7nd percentile—for an "extra risk" of being 2.1 control standard deviations out from the mean of . $\frac{0.118 - 0.018}{1 - 0.018} \cong 0.1$. However despite this superficial relationship to a quantal ED10 (superficial because there is nothing whatever inherently definitively "abnormal" about the 1.7nd percentile) the authors do not seem to have thought through either the biological implications of such a standard for important risk-related continuous variables or the issue of what is likely to often represent a statistically detectable level of effect:

- The statistical delectability issue is easy to examine from first principles. Let us imagine that we have as few as 25 animals per group—a minimal group size for a respectable toxicological experiment. From the usual relationship between standard deviations and standard errors of a normally distributed variable, a one standard deviation shift in such a group would represent $1/(25^{.5}) = 5$ standard errors. Surely this is well beyond any reasonable standard for what constitutes a minimally statistically detectable effect. A more usual choice might be 2 standard errors, meaning that a shift of 0.4 standard deviations should routinely suffice to define a minimally detectable effect in a group of

25 animals or human subjects. And it should be noted that this 0.4 standard deviation definition of a usual minimally detectable effect should be considered to correspond to a LOAEL, not a NOAEL. A NOAEL level on statistical grounds might generally be 2-3 fold lower in typical cases.

- Biologically, at least for variables that are strongly linked to adverse health outcomes such as human birth weights (possibly analogous to fetal weight reductions commonly observed in animal toxicology tests), a one standard deviation change seems very large. In the case of human birth weights, a one standard deviation change would correspond to about a 500 g shift—or about 15% of approximate average birth weights of about 3400 g. Such a shift is two and a half times as large as the approximately 200 g shift typically associated with cigarette smoking—which is in turn associated with quite serious increases in risks of infant mortality and other adverse outcomes. Calculations I did some time ago (Hattis, 1998; Hattis and Ballew, 1988) indicate that the strong relationship between birth weight and infant mortality (see Figure 1) implies, if it reflects causal processes, that even a 1% shift in mean birth weights would be expected to be associated with an excess risk of infant mortality of about 0.5/1000 babies; similarly a 15% shift would be associated with an excess mortality risk of about 10/1000 (Figures 2-3, Table 1). These are definitely in a range that should be considered to be of public health concern. Certainly it would be just flat wrong to regard a 500g shift in humans to be a NOAEL or even a LOAEL. The 1% (34 g) shift level, which approximates a level that would definitely be of public health concern, but barely detectable in good epidemiological studies is perhaps a better approximation of a LOAEL—with a value 10 fold or more less than this perhaps approximating a NOAEL. I would suggest that the authors could gain a fuller perspective on how seriously to regard quantitative changes in key parameters by applying this same kind of analysis to other variables known to be of public health significance for the human population—e.g., blood pressures, LDL cholesterol, fibrinogen, FEV1, etc.

a. Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?

I think general mechanistic plausibility should always be considered in choosing models. For example, for this reason I will nearly always use a log probit model as a first choice for application to quantal data for putative individual threshold responses because there is a plausible mechanistic interpretation—many multiplicative factors determining individual thresholds. So even in the very common case where there is no meaningful difference in the fit of, say, a logistic vs a probit model to specific data sets, I would nearly always prefer the probit because there is no mechanistic interpretation I know of for a logistic distribution of individual thresholds.

In the case of the first example given on pp. 53-59 with the saturating-type quantal response, I would first think to use a Michaelis-Menten element as an important feature. Saturation of activating enzyme systems is a frequent observation in carcinogenesis, and very likely non-cancer effects resulting from internal exposure to activated intermediates. Similar Michaelis Menten kinetics can also result from the saturation of binding sites in the kidney that are an important step in the process of urinary excretion of some chemicals. Moreover, it is also quite likely that there are cases where binding of toxicants to cellular receptors is also saturable in the same way that binding of substrates to the active sites of enzymes is saturable. It is by no means necessarily so that all dose response relationships with decreasing slope at

Figure 1
Relationship Between Weight at Birth (in 500 Gram Increments) and Infant Mortality (Data of Hogue et al., 1987)

Dale Hattis
Clark University

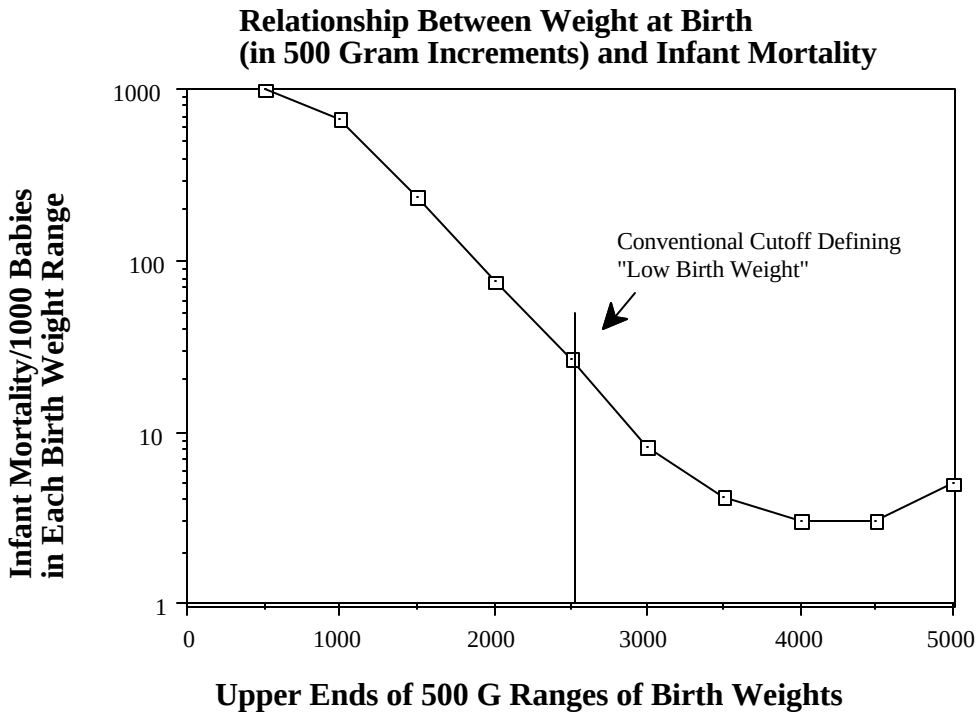
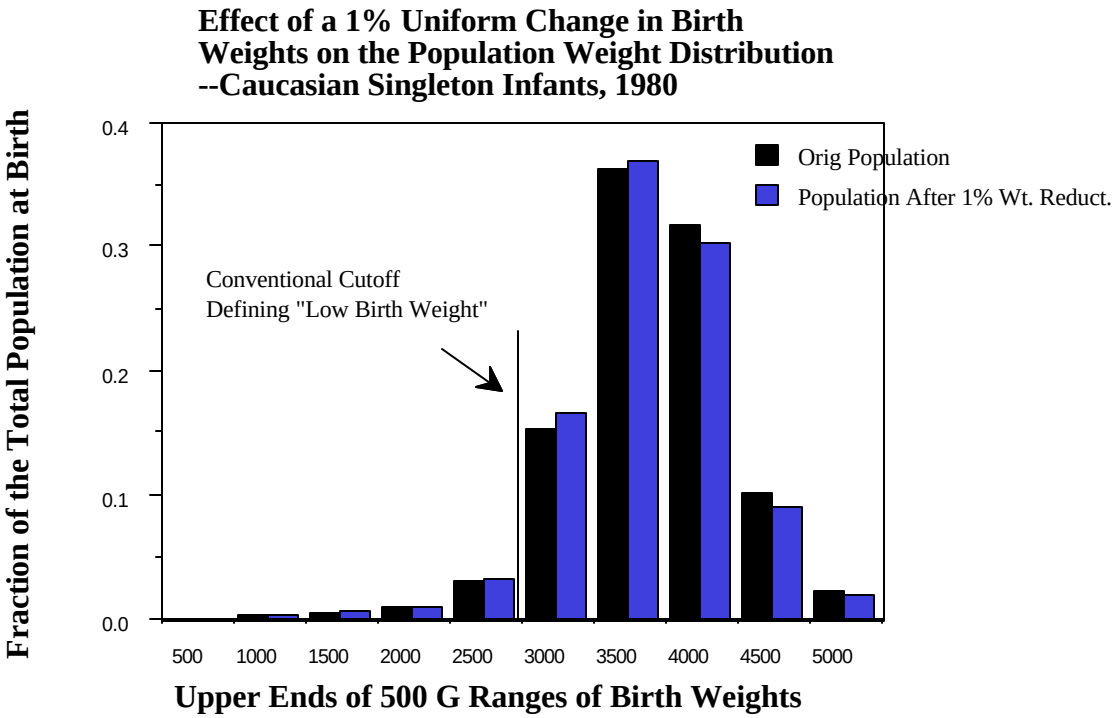


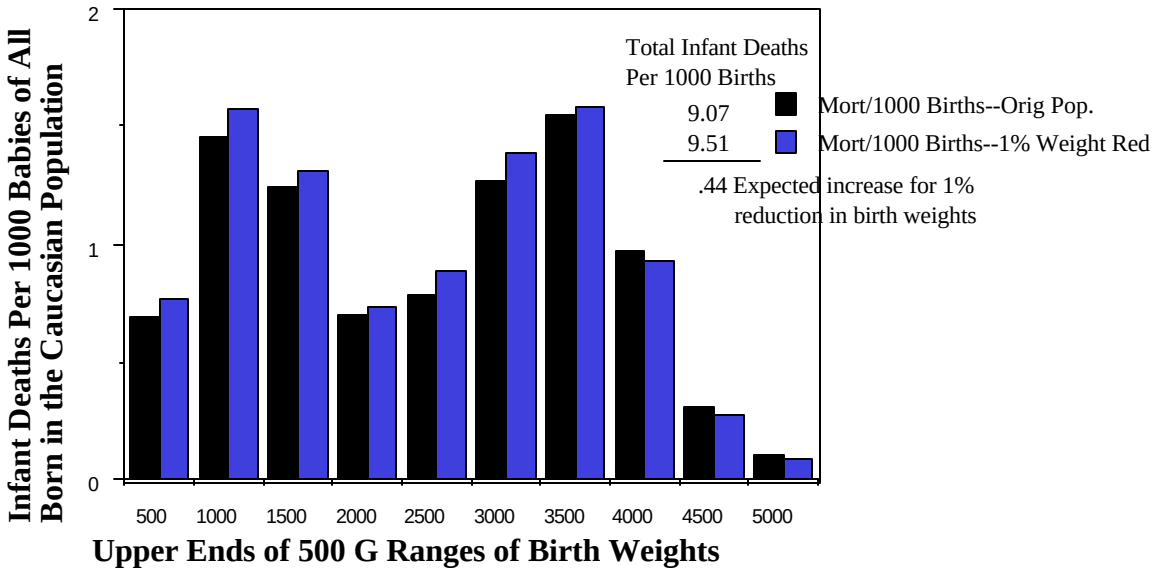
Figure 2
Effect of a Uniform Change in Birth Weights on the Population Weight Distributions
(Source: Rees and Hattis, 1994)



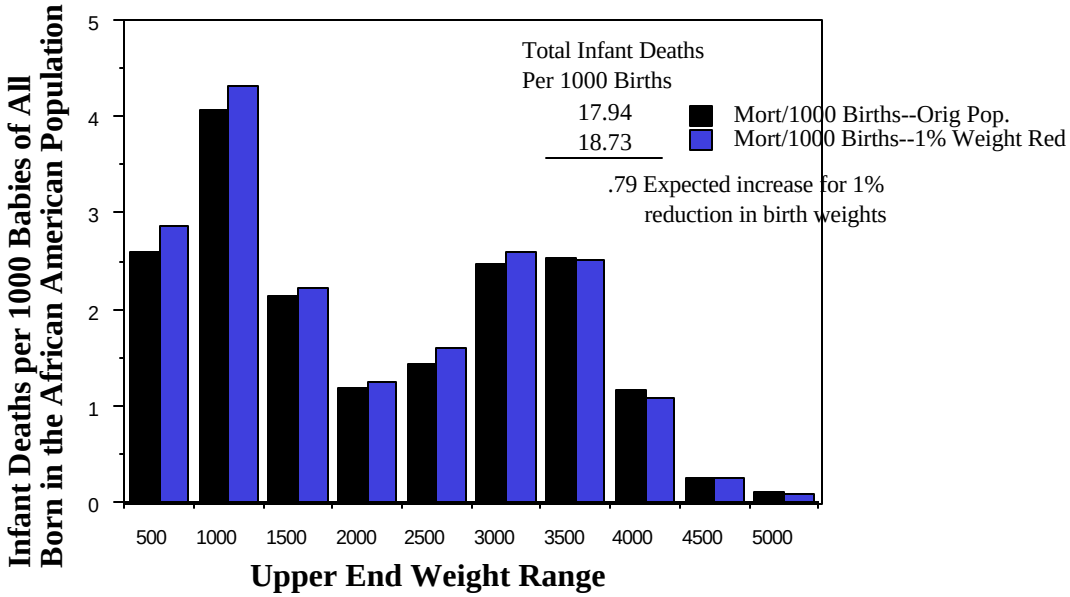
Source: Hattis, 1998

Dale Hattis
Clark University

Figure 3
Expected Effect of a 1% Reduction in Birth Weights on the Distribution of Overall Infant Deaths Per 1000 Babies Born--Caucasian Population



Expected Effect of a 1% Reduction in Birth Weights on the Distribution of Overall Infant Deaths Per 1000 Babies Born--African American Population



Source: Hattis, 1998

Table 1

Summary Implications for Infant Mortality of Reducing Birth Weights by Equal Amounts in Grams at All Percentiles of the Black and White Birthweight Distributions

% Birthwt reduction	White excess mort./1000	Black excess mort./1000	Total excess mort/1000
0.01	0.00426	0.00768	0.00484
0.1	0.0427	0.0770	0.0485
1	0.440	0.790	0.499
10	6.00	10.2	6.70

Source: Hattis and Ballew, 1989.

higher doses operate by this kind of process. But these examples indicate that there are many ways that Michaelis-Menten type dose response shapes can arise, and they should in general be considered in preference to mathematical forms for which there is no plausible connection to specific mechanisms. For a response where it is plausible to combine a Poisson one-hit underlying response function with one of the Michaelis-Menten modifiers mentioned above the dose response function would be:

$$P(D) = Bkgd + (1 - Bkgd)*\{1 - e^{[-V_{max} D/(K_m + D)]}\}$$

where Bkgd is the background incidence, Vmax is the maximum of Poisson "hits" per dose, and Km is the dose at which half the maximal Poisson "hits" per dose are produced. Applied to the data given in example 1, this would lead to a maximum likelihood estimate of the BMD (the dose producing an incidence of 11.8% with a background of 2%) of about 4.4 dose units. In the time I devoted to the process, I could not get my implementation of this model to converge in calculations of upper and lower 95% confidence limits.

I can imagine that some will object to this notion of generally preferring mechanistically-connected models to those which fit as well (or perhaps a little better) but have no mechanistic association. I believe models with mechanistic connections are to be preferred because (1) adding the mechanistic considerations adds at least a bit of information to the raw numbers reported by the experimenters, and (2) incorporating a mechanistically interpretable component can allow that hypothesis to be evaluated and modified if indicated either with the data at hand or in subsequent experiments. In other words, I would argue that these models make better use of the totality of available potentially relevant technical information and can contribute to the iterative processes of information gathering, assessment and interpretation that we think of as science.

b. What are the advantages/strengths of using the methods described to select among "equally" fitting models? What other methods should be considered in making a selection?

See my previous comment. I would want to include a model's inherent connection to plausible mechanisms among the characteristics used for selection. Exactly how much weight this consideration is given, of course, depends on how plausible the mechanistic ideas are. Ideally there might be some Bayesian/decision analytic evaluation in cases where there is some tradeoff between the virtues of having a mechanistic connection and short term goodness of fit considerations for the existing data.

Question #5 Use of confidence limits

a. Please comment on the approaches described to compute confidence limits.

In general the reasoning in the document seems sound, although it would be desirable to include some more equations and explanation to supplement the references to primary literature provided. I did have one reservation reading p. 32 lines 15-16. Here the document seems to suggest that in cases where a dose response model does not have a positive lower 95% confidence limit that can be used as the BMD, the dose should be log transformed to ensure a positive value. I would caution that this is a material change in the mathematical form of the model, and therefore should trigger a complete review of the available choices of model form for both mechanistic criteria and goodness-of-fit

criteria. In general if mechanistic considerations suggest use of an arithmetic representation of dose, the mechanistic connection will be lost if the dose is log transformed.

I also think it would be good for the authors to discuss the mechanistic rationale, if any, for the suggestion on p; 32, lines 23-25 for the use of a beta-binomial distribution for the number of affected individuals in a liter. Why not assume a lognormal distribution of individual susceptibilities within litters and then express that in binomial terms?

Finally, as already indicated, I have very strong reservations about the proposal to use a shift of one standard deviation as the benchmark response for continuous variables that might have important implications for human health.

b. Please comment on the soundness of the criteria used to deviate from using a 95% one-sided confidence interval for the BMDL.

I must have missed these criteria in my reading of the document. I don't know what section is being referred to.

Question #6 Examples: What additional concepts, if any, should be illustrated by an example?

I would like to see an illustration for a compound that affects an established cardiovascular or reproductive risk factor. For example, what if a chemical increases blood pressures or fibrinogen levels. For reproduction, one for example take the effects of a glycol ether such as ethylene glycol ethyl ether on male sperm counts (e.g., Hattis and Ballew, 1989).

Question #7 The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document's guidance, or would further discussion be useful?

In general the reporting requirements seem fine to me. It is important that enough documentation be provided to allow each of the choices in BMD modeling and estimation to be understood and reproduced.

Question #8 What other comments do you have about the approach to benchmark dose

described in this document? What changes to the approach would you like to see, and why? As already discussed, I think the document would benefit from a broader disciplinary orientation. What is now an almost exclusively statistical orientation should be supplemented with regulatory risk management/incentives considerations (e.g., in the guidance for fixing % response levels for purposes of calculating benchmark doses that are to be used as replacements for NOAELs in current RfD calculations), and considerations of the quantitative implications of different mechanistic hypotheses with respect to both pharmacokinetics and pharmacodynamics.

In general I think the advice on selecting models for use in POD comparison on p. 35 lines 1-9 is vague and likely to lead to abuse. This is particularly true for the prescription to throw out results from a model that is an "outlier" when compared to the BMDLs estimated from other models. Either one should just always use the most conservative BMDL from the array of models with roughly similar goodness-of-fit statistics (particularly at the low end of the dose range) or a much clearer rationale should be available than a subjective judgment just that one model seems to depart materially from the results obtained from other putatively equally plausible models. In this context, I also don't quite understand the reason for averaging BMDL's from different models specified on the previous page. It seems an odd thing to do to average conservative percentile estimates. Probably a more appropriate procedure, if indeed one wishes to treat the estimates of all eligible models on an even handed basis would be to calculate the required percentile of a combined likelihood distribution in which each of the component models contributes an equal share (or, perhaps, share derived from their relative overall likelihood statistics in "predicting" the observed data).

References

Baird, S. J. S., J.T. Cohen, J.D. Graham, A.I. Shlyakhter, & J.S. Evans. 1996. Noncancer risk assessment: A probabilistic alternative to current practice. Hum. Ecol. Risk Assess. **2**: 78-99.

Hattis, D., 1990. "Pharmacokinetic Principles for Dose Rate Extrapolation of Carcinogenic Risk from Genetically Active Agents," Risk Analysis, Vol. 10, pp. 303-316.

Hattis, D., 1991. "Use of Biological Markers and Pharmacokinetics in Human Health Risk Assessment," Environmental Health Perspectives, Vol. 89, pp. 230-238.

Hattis D. 1996. "The Challenge Of Mechanism-Based Modeling In Risk Assessment For Neurobehavioral Endpoints." Environmental Health Perspectives, Vol. 104, Suppl. 2, pp. 318-390.

Hattis, D. 1998. "Strategies For Assessing Human Variability In Susceptibility, And Using Variability To Infer Human Risks" In Human Variability in Response to Chemical Exposure: Measures, Modeling, and Risk Assessment, D. A. Neumann and C. A. Kimmel, eds., CRC Press, Boca Raton, FL, pp. 27-57.

Hattis, D., Banati, P., and Goble, R. 1999b. "Distributions of Individual Susceptibility Among Humans for Toxic Effects--For What Fraction of Which Kinds of Chemicals and Effects Does the Traditional 10-Fold Factor Provide How Much Protection?" Annals of the New York Academy of Sciences, Volume 895, pp. 286-316.

Ballew, M., and Hattis, D., 1989. Reproductive Effects of Glycol Ethers in Females--A Quantitative Analysis, M.I.T. Center for Technology, Policy, and Industrial Development, CTPID 89-7, July.

Hattis, D., and Shapiro, K. (1990) "Analysis of Dose/Time/Response Relationships for Chronic Toxic Effects--The Case of Acrylamide," NeuroToxicology, Vol. 11, pp. 219-236.

Travis, C. C. and White, R. K. 1988. Interspecific scaling of toxicity data. *Risk Anal.8*:119-125.

Watanabe, K., Bois, F. Y., and Zeise, L. 1992. Interspecies extrapolation: A reexamination of acute toxicity data. *Risk Analysis 12:*301-310.

Colin Park

DR. COLIN N. PARK

BIOGRAPHICAL INFORMATION

Dr. Park worked for The Dow Chemical Company for 29 years and retired in 1998 as a senior associate environmental consultant and issues manager in the environmental and health area. His main interest was in the use of risk assessment in regulation, legislation and communication.

He received a B.S. degree in mathematics in 1965 from the University of British Columbia and earned a M.S. and Ph.D. in applied statistics from Purdue University in 1970. He joined Dow in 1970 as a statistical consultant and held a number of statistics, computer, and risk assessment positions in Dow.

Dr. Park has served on the Science Advisory Board of the National Center for Toxicology Research, and on the Risk Assessment Committee of the Council on Environmental Quality for the State of Michigan, as well as on numerous EPA review committees.

Topic # I
Question 1
Concepts and terms seem to be well defined although I am not sure that BMD, used as the procedure and also used as the central estimate (as opposed to BMDL) is the best terminology.

Question 2
No other suggestions

Question 3
I think there should be a little more discussion on, or at least reference to, the concept of adverse effects vs. effects that are not adverse, or are adaptive, or are not biologically plausible.

Topic # II
Question 4
EPA has done a good job of presenting a general discussion on model selection and model discrimination, but I believe additional guidance needs to be included if this tool is to be used in a regulatory context. I think the process needs to be more prescriptive for use as a regulatory tool. My concern is that the current guidance says, in brief, to fit a number of models, and if the BMDLs are not within a factor of three-fold, then select the lowest BMDL subject to some qualification on the lowest being an outlier. It would seem to me that there could be a number of data sets for which the BMDLs will vary by more than three–fold if a number of different models are fit. In particular, this could occur when models are over parameterized. This encourages an additional layer of conservativeness in the process.

I would prefer to see an approach which uses default models depending upon the type of data, and an ordering of the default models in the case of lack of fit. I think this is necessary to avoid model shopping.

A somewhat related concern is that I would like to see a discussion of parameter minimization. It seems to me that there is not much guidance against over parameterizing a model. Guidance could include suggesting an approximate test of significance for each parameter, or deleting parameters and comparing the Akaike Information Criterion.

Question 5.
I suppose this horse has already left the barn, but I have to comment on confidence intervals.

Previous workshops have recommended that the concept of a BMD is a good idea, and could be very simple in application if confidence limits were not used. The use of confidence intervals turns the process from a simple, easy to apply, biological concept into a complex modeling issue, which requires a statistician and is much more model dependent.
The argument FOR confidence intervals is that they reward better experimentation, both from the point of view of number of animals and better experimental procedures. Width of confidence intervals are roughly inversely proportional to the square root of “n”, so the gain in changing from 10 to 20 animals per dose group is to reduce the width of the confidence interval by approximately 1/1.4 = 30%, a minimal gain for a lot more work.
Given the uncertainty in extrapolating from inbred rodents to the human population, I see very little point in being so statistically exact in one small piece of the process, except for job security for statisticians.

Question 6:
No comment

Topic III
Question 7:
No comment

Question 8:
See Question 5

Lorenz Rhomberg

Lorenz R. Rhomberg, Ph.D.
Principal Scientist
Gradient Corporation
Cambridge, MA

Dr. Rhomberg is a Principal Scientist at Gradient Corporation, a Cambridge MA environmental consulting firm. Before joining Gradient he was an Assistant Professor at the Harvard School of Public Health, where he maintains an adjunct appointment. From 1984-1994 he was a risk assessor at the U.S. Environmental Protection Agency in Washington. Dr. Rhomberg earned his Ph.D. in population biology from the State University of New York at Stony Brook and his B.Sc. in biology from Queen's University in Ontario. His interests lie in methodology and science policy for quantitative risk analysis, including pharmacokinetic modeling and probabilistic methods with special emphasis on cross-species extrapolation, chlorinated solvents and endocrine active agents. Dr. Rhomberg is a member of the Office of Pesticide Programs' FQPA Science Review Board, has served on several FIFRA Scientific Advisory Panels, on NAS Committees, and other panels. He is a past President of the New England Chapter of the Society for Risk Analysis. He has published two books and over 50 articles and book chapters on risk analysis topics.

Gradient Corp.; 238 Main St.; Cambridge, MA 02142 617-395-5000
lrhomber@gradientcorp.com

Comments from Lorenz Rhomberg

Topic I – Selecting Data and Benchmark Response
Question 1 – *Definitions of concepts and terms.*

The definition in the text of terms particular to the BMD context is quite clear. The special clarification of the distinction between BMD and BMDL is necessary and helpful (and the chosen solution is a good one).

The BMD Guidance would be a good place to advocate differentiating between a Dose-*Response* relationship (different proportions of a population responding with discrete outcomes as a function of dose) and a Dose-*Effect* relationship (differing magnitude or severity of an individual's outcome as a function of dose). This terminology is advocated by some but is by no means standard. Since making the distinction is useful for BMD analysis, the Guidance could usefully call for making the distinction.

The inclusion of a Glossary is a good idea, and the set of terms defined therein is about right. Many of the definitions given need some reconsideration, however. In doing so, thought should be given not only to the term-by-term definitions, but also to the overall role of the Glossary, the technical level at which it is pitched, and the utility of the set of definitions for the user.

Particularly for statistical terms used by (but not exclusive to) BMD analysis, the Glossary is inconsistent, sometimes giving the general statistical definition and sometimes simply naming the context in which the term is used in BMD analysis. I suggest that most terms should be addressed in *both* ways; a technically sound but non-technically expressed *definition* should explain the concept followed by an explanation of the term's meaning and use in the BMD context. The technical level should not be high, but it should nonetheless be sufficient to convey some understanding of the concept and its applicability to the BMD context.

In contrast to the body of the document, many of the Glossary definitions are not well written. A few of the more salient definitions in need of attention are:

Cancer Potency – estimates *incremental* risk, in practice it is usually an upper bound instead of an estimate per se, should explain how it has units of inverse dose and is applied by multiplying by dose;
Chi-square Test – an example of a sloppy technical definition that gives no context for why it is relevant to BMD analysis;
Clustered Data – another unhelpful sloppy technical definition, it should say that individual subjects belong to defined groups, and that features common to the group could affect outcomes in its members;
Confidence Interval – "a specified confidence (percent of *instances*)", since the answer doesn't fluctuate from one moment to another.
Coverage – the definitions referred to do not define the term;
Degrees of Freedom – an example of confusing a particular application with the general definition;
EC_x and ED_x – x's should be subscripts, the sentence "Dose may be expressed…." should be omitted as unnecessary;
Estimate – a very bad definition;

Extra Risk – should be contrasted with Additional Risk (which should also be defined) and both should be described as incremental or "excess risks" above background.
Goodness-of-Fit – drop "…in order to provide a test for rejection…", define it as a description of the probability that the observed lack of fit of data to a model would result if the fitted relationship were true, and then explain how it is used to judge model adequacy (*i.e.*, don't define it in terms of its use, but separate the definition from the description of its application).
Hazard Identification – a bad definition that is at odds with the NAS definition and EPA's policies;
Hybrid Model – refer to Crump and Kodell *et al.*;
Likelihood Ratio – add explanation of applicability to BMD and reference to Likelihood Ratio Test and its proposed use;
Linear Dose-Response Model – the definition is incorrect—a cubic equation would be called linear under it since response is proportional to the cube (which is "some function") of the dose. Moreover, in application, models are considered "linear" if they are linear at low doses, even if they curve at high doses (as does the linear one-hit model, for instance). The reference to "change in response" being proportional to dose is not clear—is this an attempt to refer to the underlying hazard function?
Linearized Multistage Model – an example of a definition that is not precise enough to be useful. Many models that are not the multistage would fit this definition. A non-unique property of a thing cannot serve as a definition of that thing;
Local Maximum – "in a region *of parameter space*…" (*i.e.*, this is not an example of risk assessment differences between EPA Region III and Region VII). Also, one needs to describe the relevance to BMD analysis;
Likelihood Function – this doesn't seem technically correct—the likelihood is the probability that a given set of parameter values would result in the observed data, and the likelihood function is the variation in that probability as a function of the parameter values;
Logistic Model – it is a *particular* sigmoid function—all sigmoid functions are not logistic;
Log Transformation – explain "transformation" and say why it is used;
Maximum Likelihood Estimate – add "under a specified model of experimental error" and explain that it is one of several methods for fitting a model to data;
Michaelis-Menton Equation – is *not* a dose-response curve;
Parameter – a poor definition that is neither technically sound nor helpful to someone who needs to look it up in a Glossary;
Profile Likelihood – a plot of the *value* of the likelihood function versus *alternative* values of the parameter;
Quantal Data – not just dichotomous—dichotomous data are a subset of quantal data;
Quantile – not a percentile—a percentile is a particular kind of quantile, as are deciles, quartiles, *etc.*;
Rectangular Hyperbola – perhaps technically correct (although I'm not sure) but totally unhelpful to someone looking the term up in a Glossary;
Risk Characterization – not the full NAS definition and out of accord with EPA's policies on risk characterization;
S-Plus; SAS – these are *particular* software packages, not synonyms for "statistical software".

These are the ones I felt compelled to comment on explicitly, but all the entire Glossary should be scrutinized carefully.

Topic I – Selecting Data and Benchmark Response
Question 2 – *Suggestions for inclusion in literature review.*

No specific suggestions.

Topic I – Selecting Data and Benchmark Response
Question 3 – *Selection of studies and endpoints for the BMD.*

The selection of studies seems well discussed. The procedures should ensure that the critical effect is not excluded, although the role of the Uncertainty Factors subsequently to be applied to any calculated BMD is not very explicitly discussed.

More of an issue is that there is little discussion about how to regard distinctions between frankly adverse endpoints and physiological or biochemical changes that are to be regarded as markers or precursors. Particularly for continuous endpoints, there is the technical capability of defining BMRs and associated BMDs for endpoints that should not become the basis of an RfD. This issue is not unique to the BMD approach, but the focus on the dose-response, and the emphasis on data sets that can be modeled, increases the difficulties. (I would contend that the Dioxin reassessment is an example of the problem.) As we improve in elucidating the underlying physiological basis of toxicity (and as we gather data to do so) the issue of distinguishing the detectability of the operation of such processes from the manifestation of the adverse consequences of those processes becomes more important yet more problematic. I think the Guidance has to address this issue more directly.

The issue of defining the BMR has points needing further discussion. In the case of quantal endpoints, the Guidance implies in some places that the BMR should be chosen as a lower response than 10% if the data support estimating such a level. This is fine for use of the BMD/BMR as a Point of Departure for a low-dose linear extrapolation, but it causes problems for Margin of Exposure approaches, since the margin is assessed for a different point on the curve in different cases (with consequent differences in interpretation and acceptability of a given MOE). Elsewhere, the Guidance suggests that a 10% BMR always be included as a standard, but it is not clear if this is just as a secondary result to be used in comparisons or if consistency in use of BMRs for RfDs is intended.

Moving the BMR around depending on what can be estimated leads to a lack of meaning of the BMR/BMD concept. It is not a "benchmark" if it doesn't mean the same thing from case to case. It confuses detectability of response with response—the very problem that one seeks to avoid in moving from NOAELs to BMDs.

On the side of continuous endpoints, the issue of defining a BMR is problematic, and the Guidance understandably makes little headway in sorting this out. Several points ought to be taken up and discussed, however:

The guidance on p.20 names several methods for defining a BMR for continuous endpoints, but this section fails to name a method (5% of the maximum dynamic range) that is used later on in the

examples (p.60) and (as 1%) in EPA's Dioxin reassessment. The use of a fraction of maximum dynamic range can be questioned on several grounds: it may be hard to estimate, since the maximum may not be clear, and its toxicological relevance is not clear (since the eventual behavior at high doses says nothing about the adversity of a response at lower doses). In any case, its interpretation would be very different from that for other BMR definitions, and it raises the possibility of "shopping" for a BMR definition by comparing a changed value to controls on the one hand or to some hypothetical high-dose population on the other. My advice is to omit the method of fraction of maximum dynamic range from the Guidance.

Even with the remaining methods for defining a continuous BMR, the issue of adversity remains, as mentioned above.

When continuous endpoints are modeled, the continuous variable responds differently from one individual to another, and what is modeled is actually the mean value in the population as a function of dose. If a BMR is defined as an amount of change that is considered adverse, then the population's reaching this degree of change represents something close to a 50% prevalence of bearing an adverse value (assuming a symmetric distribution). The discussion of BMRs should consider more thoroughly the implicit connections of continuous and associated implicit quantal definitions of endpoints and how they interact with the way the continuous BMR is chosen. This is done explicitly in the hybrid approach, but similar considerations are implicit in other BMR definitions.

Topic II – Model Selection, Fitting, and Confidence Limits

Question 4a – ***Model Selection and Fitting: Presentation of proposed defaults for the parameters for various models.***

If the "defaults" mentioned in this question refer to standard definitions of the BMR, then the discussion elsewhere (Question 3) applies. If they refer to models, the question is not clear, because presumably the parameters are found by fitting, not specified as defaults. What is at issue is guidance about alternative forms of a family of models, *e.g.*, the degree of a multistage model, the use or disuse of covariates, restrictions on ranges of shape parameters, and the like. This matter could receive more discussion. Is use of covariates required when data are available? Is the choice of using them or not using them to be decided strictly on statistical grounds, or are other judgment bases legitimate? (and if so, what are they?)

If covariates are used, what values of the covariate are to be used in the calculation of a BMD? Strictly speaking, covariates are supposed to be independent of dose. What if they are not? (*e.g.*, lower litter sizes at high doses).

It is noteworthy that the guidance on parsimony in choosing the degree of the multistage model (*i.e.*, to use the model with fewer fitted parameters if the larger model does not significantly improve fit using a likelihood test) contradicts current EPA policy and practice. If this change is intended, it should be noted as a departure from current practice.

Topic II – Model Selection, Fitting, and Confidence Limits

Question 4b – ***Model Selection and Fitting: Criteria to evaluate the fit of the model.***

The goodness-of-fit is assessed on the best-fitting set of parameters for each model, but the eventual use of the BMDL (rather than the BMD) means that in practice, an alternative set of parameters is being used, namely (if based on the distribution of the likelihood ratio) the set that just fails to *deteriorate* the fit significantly at the 5% level. In practice, it will often happen that the curve implied by these alternative parameters fails the goodness-of-fit test. There are differences of opinion as to whether this constitutes a problem, but it ought to be discussed explicitly. (The same phenomenon happens with traditional cancer risk assessment, and no one has seemed to care, but perhaps they should.)

Topic II – Model Selection, Fitting, and Confidence Limits

Question 4c – *Model Selection and Fitting: Methods to select among "equally" fitting models.*

The operational rules (best AIC if BMDs within 3-fold, otherwise lowest BMD) seem appropriate and practical. The exception that the lowest BMD should not be used if it is an "outlier" needs better specification. What constitutes valid evidence of being an "outlier" and what rationale needs to be given? Since the situation only arises when different BMDs are already quite different, this seems a recipe for controversy without further guidance.

It would help to make a clearer distinction between choosing among alternative models as descriptors of a single data set and choosing among alternative data sets as a basis for the critical BMD. (I presume the former is intended in the present question and in the guidance given on pp.34-35.) In choosing models, there should be a more prominent role for assessing relevance to the toxicologic endpoint and its presumed mode of action. The best model is the one that best embodies the actual underlying reasons why different levels of dose lead to different response levels and that has a shape that is to be expected from the operation of the presumed process. At the most basic level, tolerance distribution models should apply to some modes of action and stochastic process models to others as a matter of principle, and such appropriateness should not be overruled by (nonsignificant) differences in goodness of fit.

I am somewhat disquieted by some aspects of the approach of attempting several models and then choosing which to use after the fact. It conceivably has some of the same pitfalls as the failure to distinguish between *a priori* and *a posteriori* statistical tests, although whether this is a real practical concern is not clear.

First, is there any danger in not specifying the universe of available models beforehand? The result will always be conditional on the models attempted, and an infinite number of models are not attempted in every case. We presume (probably with good reason) that no additional model that fits the data well and has a "reasonable" rationale would give an answer that is much different. Might ad hoc models be defined that are designed to push the answer one way or another? What grounds are to be used to defend against such manipulation?

Second, given that a set of "usual" models exist, would there be pressure to attempt all of them in every case so as to avoid criticism that one was avoiding some possible answers? Would this encourage blindly using models without regard to their biological appropriateness? Would results be deemed illegitimate if they did not start out with trying certain models?

Third, one can easily imagine a situation where the results emerging from different models would be enough to engender regulatory actions sufficiently different to be worth fighting about. If we have marginally meaningful statistical criteria for choice somewhat at odds with marginally

supported biological arguments about which models are appropriate, it could lead to problems.

Fourth, one wonders about the effect on coverage probabilities of confidence limits when the model around which such limits are being set was chosen among several alternatives on the basis of fit to the data.

On p.67, in the discussion of one of the examples, the Guidance offers several observations about the kinds of judgments necessary in choosing models (even on statistical grounds alone, setting aside biological ones). Some of the pitfalls of failing to exercise such judgment are illustrated. It is not clear how the guidance on model selection in the body of the text is to be accompanied or modified by the principles illustrated in the example.

Topic II – Model Selection, Fitting, and Confidence Limits
Question 5 – *Computation of confidence limits.*

The Guidance presumes that the method for computing confidence limits will emerge naturally from the choice of models and the nature of the data. It may be worth discussing whether this is necessarily so, and if a problem of shopping for a method with desired results might be created. Perhaps a hierarchy of preferences can be defined.

The explanations of the statistical methods for defining confidence limits attempt to convey the essence to a nontechnical audience. This is laudable, but the explanations fall somewhat short of what would seem necessary to engender real understanding. (A statistician would understand the explanations, but wouldn't need them.) If this is deemed important, perhaps the explanations could be expanded somewhat.

Topic II – Model Selection, Fitting, and Confidence Limits
Question 5b – *Criteria to deviate from using a 95% one-sided confidence interval.*

I didn't find this discussed in the Guidance. If there are to be such criteria, clearly they must be more prominently discussed.

Topic II – Model Selection, Fitting, and Confidence Limits
Question 6 – *Examples: What additional concepts should be illustrated by an example?*

The examples are generally good, but they do illustrate how much room for maneuvering there is in the procedure. It is hard to specify criteria for soundly applying judgment, so practice will have to reveal how much controversy is generated by how different practitioners exercise the judgment and flexibility offered.

As argued above, the example of 5% change in dynamic range should in my view not be used in the document.

Topic III – Interpretation
Question 7 – *Reporting requirements.*

An admirably thorough set of reporting requirements are presented, aimed at making all the

bases for analyses clear. It will be a challenge to adhere to them because of the volume of documentation they represent. Although the agency may do well itself, it is unlikely that other practitioners undertaking BMD analyses not subject to EPA's approval (but nonetheless intended to be done according to EPA's methodology) can be counted on to be as thorough.

Given the choices available and judgments that are necessary, it is important to document and explain the basis of all of the decisions. An assessment that lacks this documentation could be seen as being somewhat arbitrary.

A key part of the documentation and reporting is the explication of the rationale for study selection for dose-response model choice, for estimation procedure and confidence limit calculation, and for definition of the BMR. Given that there will be limited case-specific arguments to be made, these can quickly devolve into boilerplate descriptions based chiefly on the rationale that similar choices were made in the past, or that it is the usual practice. This has to be guarded against in implementation of the Guidance.

The inclusion of the specified "standard" ways of defining BMRs (10% extra risk for quantal endpoints and a 1 control-SD change in the mean response for continuous ones) is important, because the very idea of a "benchmark" is that it is an index that has similar meaning and interpretation from case to case. (Indeed, under Question 3, I question whether case-specific definitions of BMRs is a good idea, for this very reason.)

Topic III – Interpretation

Question 8 – *Other comments and proposed changes.*

The Guidance should say something about the use of alternative dose metrics and how this might affect issues of model fit and hence model choice. This especially includes the use of pharmacokinetically based dose measures.

The Guidance should be more explicit about the issue of assessing adversity of endpoints. While this is not necessarily part of the BMD calculation, it is important to establish that BMDs based on merely detectable physiological change are not necessarily indicative of a compound's toxicity (even if the changes detected are in some sense part of an eventual toxic response).

Characterization of shape of the dose-response curve, the strength of evidence in favor of linear *versus* nonlinear shapes, and the slope at the lower end of the observable range are all aspects that one would want to pass on to further parts of the risk assessment process. While these aspects are not directly involved in defining a BMD, the BMD so defined should be interpreted in light of them, and the Guidance should emphasize this.

More explicit guidance for what kinds of models are considered legitimate for what kinds of data would be helpful. Similarly, preferences for means of calculating confidence intervals should be indicated. These should not be prescriptions, but there should be some basis for judging what kinds of choices are sound and which are merely manipulations.

The issue of covariates needs more discussion—about when they can be and must be used, when it is alright to forgo them, what values to use when defining BMDs, and so on.

In the end, all guidance must be considered interim, and practical experience will help suggest what parts of the current Guidance need bolstering, clarification, or modification. Good guidance establishes the general rules and principles to be applied; it encourages appropriate flexibility while guarding against manipulation. It is also important, however, for guidance to provide the means for avoiding or at least resolving conflicting interpretations. To some degree, one needs

dependable rules to settle close calls.

Robert Sielken, Jr.

Robert L. Sielken, Jr. is a biostatistician, president of Sielken, Inc., and Vice President of Jellinek, Schwartz, & Connolly, Inc. He received his Ph.D. degree in Probability and Statistics from Florida State University in 1971. His research in dose-response modeling was supported in the 1970's by the Food and Drug Administration, the National Center for Toxicological Research, and the National Institutes of Health. Since serving on the Society of Toxicology ED_{01} Task Force in 1981 and helping to develop some of the earliest probabilistic risk assessments in the 1980's for Superfund (e.g., the Rocky Mountain Arsenal), Dr. Sielken has published more than 50 papers on cancer, noncancer, and ecological risk assessment (using animal bioassays, human epidemiological studies, and ecological data) and participated in more than 100 workshops, short courses, and conferences on exposure and risk assessment. He also co-authored a book entitled Quantitative Cancer Modeling and Risk Assessment. In 1985, after 15 years of teaching in the Department of Statistics at Texas A&M University, he exchanged his duties as Professor for those of an Adjunct Professor and formed Sielken, Inc., which does statistical research and consulting primarily in the area of quantitative health risk assessment. Dr. Sielken is also an Adjunct Professor in the Department of Carcinogenesis at the University of Texas. In 1996, Sielken, Inc., became a subsidiary of Jellinek, Schwartz & Connolly, Inc., which provides integrated science-based advocacy and management solutions to business and industry. Dr. Sielken serves as a consultant to professional societies, industry, and state and federal governments and provides litigation and advocacy support. A major emphasis in his current research is probabilistic, Monte Carlo, and weight-of-evidence based approaches to quantitative health risk assessment using distributional characterizations of aggregate and cumulative exposure and risk.

MEMBERS OF THE IT GROUP OF COMPANIES:
Sielken, Inc. Bryan, Texas 77802 USA (409) 846-5175 Fax (409) 846-2671
Jellinek, Schwartz & Connolly, Inc. Arlington, Virginia 22209 USA (703) 527-1670 Fax (703) 527-5477
JSC International Ltd. Harrogate, North Yorkshire HG1 5YQ UK (1423) 520245 Fax (1423) 520297

Comments Prepared by Robert L. Sielken Jr., Ph.D.

Sielken Inc., 3833 Texas Avenue, Suite 230, Bryan, TX 77802

Tel: 979-846-5175, Fax: 979-846-2671; Email: SielkenINC@aol.com

November 27, 2000

I. Preparation for Computing a Benchmark Dose: Selection Data and An Appropriate Benchmark Response Level — *Discussion Leader – George Alexeeff*

1. What concepts and terms related to the benchmark dose (BMD), if any, do we need to define more clearly?

Comment 1.1

The differentiation between BMD and BMDL is appropriate.

2. The literature review cites works that have helped to develop the BMD approach. Do you have suggestions for inclusion of other work?

Comment 2.1

The literature review on the development of the BMD approach should include the work on the distributional characterization of the BMD. Some examples of such work that I have been involved in are as follows. The literature review should include these works as well as the works by other developers.

Sielken, R.L. Jr. (1989) "Useful Tools for Evaluating and Presenting More Science in Quantitative Cancer Risk Assessments," Toxic Substances Journal, Vol. 9, 353-404. Hemisphere Publishing Corporation.

Holland, Charles D. and Robert L. Sielken Jr. (1993) Quantitative Cancer Modeling and Risk Assessment, Prentice Hall, Englewood Cliffs, New Jersey.

John S. Evans, John D. Graham, George M. Gray, and Robert L. Sielken Jr. (1994) "A Distributional Approach to Characterizing Low-Dose Cancer Risk," Risk Analysis, Vol. 14, No. 1, pp. 25-34.

John S. Evans, George M. Gray, Robert L. Sielken Jr., Andrew E. Smith, Ciriaco Valdez-Flores, John D. Graham (1994) "Use of Probabilistic Expert Judgment in Distributional Analysis of Carcinogenic Potency," Regulatory Toxicology and Pharmacology, Vol. 20, Number 1, 15-36.

Evans, John S., John D. Graham, George M. Gray, and Robert L. Sielken Jr. (1995) A Distributional Approach to Characterizing Low-Dose Cancer Risk, Low-Dose Extrapolation of Cancer Risks, pp. 253-274, ed. by Stephen Olin, William Farland, Colin Park, Lorenz Rhomberg, Robert Scheuplein, Thomas Starr, and James Wilson, ILSI Press, Washington, D.C.

Sielken, Robert L. Jr., and Ciriaco Valdez-Flores (1996) Comprehensive Realism s Weight-of-Evidence Based Distributional Dose-Response Characterization, Special Issue of the Journal of Human and Ecological Risk Assessment on: Theoretical, Toxicological and Biostatistical Foundations for Deriving Probability Distribution Functions for Reference Doses and Benchmark Doses with Application to Carcinogens and Noncarcinogens, Vol. 2, No. 1, pp. 175-193.

Sielken, Robert L. Jr., and Ciriaco Valdez-Flores (1999) Probabilistic Risk Assessment's Use of Trees and Distributions to Reflect Uncertainty and Variability and to Overcome the Limitations of Default Assumptions, Environment International, Vol. 25, No. 6/7, pp. 755-772.

3. What additional discussion, if any, is needed to clarify the description of the selection of studies and endpoints for the BMD?

Comment 3.1

A distributional characterization of the BMD provides an opportunity to explicitly reflect the BMD corresponding to each study and each endpoint.

Several explanations of distributional characterizations of the BMD and examples have been published. Some of this work is explicitly referenced in my Comment 2.1.

In general, distributional characterizations of the BMD call for the calculation and explicit presentation of the BMD corresponding to each combination of alternatives that are scientifically defensible. If there were two defensible studies and two candidate endpoints evaluated in each study, then there would be four BMDs calculated. (Alternatives for other factors in the BMD calculation, such as the dose-response model, would increase the number of calculations.) The calculated BMDs would be explicitly presented and also summarized in the form of a distribution. The distribution can be as

simple as a frequency distribution of the calculated BMDs. This frequency distribution would treat each of the calculations as equally likely to be relevant. More sophisticated distributions can be developed if the current state of knowledge about the relevance of each specific calculation is incorporated. One method of incorporating the current state of knowledge about the relevance of different combinations of alternatives is through the use of expert judgments and subjective probabilities.

Distributional characterizations of the BMD can more accurately reflect the current state of knowledge about a specific chemical than a default characterization based on a default alternative for each factor in the BMD calculation. Distributional characterizations bring out the range of possible BMD values, the relative likelihood of the different BMD values in that range, and combination of alternatives leading to each calculated BMD value.

Distributional characterizations of the BMD can reflect the uncertainty and variability in the BMD calculation better than a single value (either a single BMD or a single BMDL). Because distributional characterizations of the BMD do a better job of reflecting the uncertainty and variability in the BMD calculation than a single value, distributional characterizations of the BMD are the best thing to combine with distributional characterizations of the dose from exposure to characterize risk. Distributional characterizations of the BMD are also the best thing to combine with distributional characterizations of uncertainty factors to characterize Reference Doses or Concentrations. Furthermore, distributional characterizations of the BMD are the most complete basis for comparing the BMDs of two chemicals.

Comment 3.2

There is another approach to dose-response modeling and BMD calculation that combines the dose-response information from multiple experiments/studies rather than selecting one specific study. This approach calculates the BMD from the combined dose-response information in multiple studies rather than calculating the BMD from a single dataset.

The approach assumes that there is one underlying true dose-response relationship for excess risks. However, the approach allows the background response rates in different studies to be different. Also, the approach allows the response

frequencies to be proportionately higher or lower than average in different studies.

The method of combining dose-response information from multiple studies on the same endpoint is illustrated using the multistage-Weibull time-to-response model. In the current multistage-Weibull model, the probability that the time T of a response (the occurrence of the endpoint under consideration) at dose d occurs by time t is

$$P(T \leq t; d) = 1 - \exp[-g(d) \times H(t)]$$

where

$$H(t) = t^{power},$$

$$g(d) = a0 + a1 \times d + a2 \times d^2 + \ldots$$

This model is expanded for multiple experiments to

$$P(T \leq t; d) = 1 - \exp[-g^*(d) \times H(t)]$$

where

$$H(t) = t^{power},$$

$$g^*(d) = \{ a0 \times [1 + alpha(i)] + a1 \times d + a2 \times d^2 + \ldots \} \times [1 + beta(i)] \}$$

and

alpha(i) = departure from the average background rate in the i-th study,

alpha(1) + alpha(2) + ... = 0,

beta(i) = departure from the underlying dose-response relationship in the i-th study, and

beta(1) + beta(2) + ... = 0.

The alpha's allow the background response rates to differ from study to study with the average departure from a0 being zero. The beta's allow the overall response rates in different studies to differ. The study is allowed to have a proportional impact on the response rates at each dose but not to change the relative impact of the different dose levels.

The expanded multistage-Weibull time-to-response model is fit to the combined dose-response data for a given endpoint from all studies. If maximum likelihood estimation is used, the likelihood for all studies combined is the product of the likelihoods for the individual studies, and this product of the likelihoods is maximized. The excess risks and BMDs are calculated from the expanded multistage-Weibull time-to-response model with the alpha's in the fitted model set equal to zero and the beta's in the fitted model set equal to zero.

Comment 3.3

The endpoint for the BMD should be an ADVERSE health effect. The emphasis needs to be on an adverse health effect and not just an effect. Effects that are unusual (e.g., below a 5-th percentile, greater than a 95-th percentile, and more than one control standard deviation from the control mean) are not necessarily adverse.

II. Modeling to compute a Benchmark Dose: Model Selection, Fitting, and Confidence Limits — *Discussion Leader – Lynne Haber*

4. Model selection and fitting

a. What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?

Comment 4.a.1

Non-negativity restrictions on the parameters in the function of dose in the dose-response model may not always be appropriate. Sometimes the elimination of some non-negativity restrictions can result in substantially improved fits to the observed data without resulting in logical inconsistencies in that dose range. Furthermore, not all dose-relationships are monotone in the dose. Removing some non-negativity restrictions is often critical to characterizing non-monotone dose-response relationships.

b. Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?

Comment 4.b.1

The Akaike Information Criterion (AIC) should not be over-emphasized. A small difference in AIC values should not be used to eliminate an otherwise more appropriate model.

c. What are the advantages/strengths of using the methods described to select

among "equally" fitting models? What other methods should be considered in making a selection?

Comment 4.c.1

Distributional characterizations of the BMD can bring out the range of possible BMD values resulting from different models, the relative likelihood of the different BMD values in that range, and combination of modeling alternatives leading to each calculated BMD value.

5. Use of confidence limits

a. Please comment on the approaches described to compute confidence limits.

Comment 5.a.1

The discussion on the computation of a one-sided confidence limit for a model parameter based on the distribution of the likelihood ratio is generally okay, but the target reduction in the log-likelihood is $(\chi^2_{1,1-2\alpha})/2$ not $(\chi^2_{1,1-\alpha})/2$ for a one-sided $100(1-\alpha)\%$ confidence limit. This mistake appears on page 31, line 29. A slightly different but related correction needs to be made on page 58, line 2. Here, the target reduction in the likelihood is $(\chi^2_{1,1-2\alpha})/2$ not $(\chi^2_{1,\,2\alpha})/2$ for a one-sided $100(1-\alpha)\%$ confidence limit. For example, for a 95% upper confidence limit for a model parameter based on the distribution of the likelihood ratio, the target reduction in the log-likelihood is $(\chi^2_{1,\,0.90})/2$ where $\chi^2_{1,\,0.90}$ is the 90-th percentile of a chi-square distribution with one degree of freedom. Specifically, $\chi^2_{1,\,0.90}$ is approximately 2.71.

The following discussion explains why a 90-th percentile in the chi-square distribution is used to generate one-sided 95% confidence limit. Roughly speaking, the reason is that the likelihood ratio acts like the square of a sum of normal random variables; so that, extreme values of the likelihood ratio correspond to the case where the observed values are too small as well as the case where the observed values are too large. Thus, of the 10% of the times that the 90-th percentile in the chi-square distribution is exceeded, 5% of the times are because the observed values are too small, and 5% of the times are because the observed values are too large. Only, one set of

these 5% occurrences needs to be guarded against in the one-sided 95% confidence limit.

A mathematical example is as follows. Let $X_1, \ldots, X_n$ be independent normal random variables with unknown mean μ and known variance σ^2. Then, the likelihood is

$$L(\mu,\sigma^2) = [(2\pi\sigma^2)^{1/2} \exp(- (X_1 - \mu)^2 /2\sigma^2)] \times \ldots \times [(2\pi\sigma^2)^{1/2} \exp(- (X_n - \mu)^2 /2\sigma^2)].$$

Here, the maximum likelihood estimate of μ is the sample mean, and the logarithm of the likelihood ratio is (after some algebra)

$$\ln [L(\mu,\sigma^2) / L(xbar,\sigma^2)] = - 0.5 \times (\mu - xbar)^2 / (\sigma^2/n)$$

where xbar is the sample mean. Thus, the likelihood ratio statistic is

$$- 2 \ln [L(\mu,\sigma^2) / L(xbar,\sigma^2)] = (\mu - xbar)^2 / (\sigma^2/n)$$

which is the square of a standard normal random variable. Hence, the likelihood ratio statistic has a chi-square distribution with one degree of freedom in this case. A one-sided 95% upper confidence limit on the mean μ is the largest μ^* such that

$$(\mu^* - xbar)^2 / (\sigma^2/n) \le \chi^2_{1, 0.90}.$$

That is,

$$\mu^* = xbar + [(\sigma^2/n) \times \chi^2_{1, 0.90}]^{1/2}.$$

This is a one-sided 95% upper confidence limit on the mean μ because

$$\begin{aligned} P(\mu \le \mu^*) &= P\{ \mu \le xbar + [(\sigma^2/n) \times \chi^2_{1, 0.90}]^{1/2} \} \\ &= P\{ -[\chi^2_{1, 0.90}]^{1/2} \le (xbar - \mu)/ (\sigma^2/n)^{1/2} \} \\ &= P\{ -[2.71]^{1/2} \le (xbar - \mu)/ (\sigma^2/n)^{1/2} \} \\ &= P\{ -[(1.645)^2]^{1/2} \le (xbar - \mu)/ (\sigma^2/n)^{1/2} \} \\ &= P\{ -1.645 \le (xbar - \mu)/ (\sigma^2/n)^{1/2} \} \\ &= .95. \end{aligned}$$

This follows because the 90-th percentile of a chi-square distribution is the square of the 95-th percentile of a standard normal distribution (i.e., $2.71 = 1.645^2$), -1.645 is the 5-th percentile of a standard normal distribution, and $(xbar - \mu)/ (\sigma^2/n)^{1/2}$ has a standard normal distribution.

In this example, the 95% upper confidence limit μ^* based on the likelihood ratio statistic is xbar plus a positive increment. The increment is not needed when xbar exceeds μ which happens 50% of the time. Specifically, the increment is not needed in the 5% of the times when xbar exceeds its 95-th percentile. Even those these high

values of xbar contribute to the frequency of high values of the likelihood ratio statistic, they do not contribute to the frequency of the upper confidence limit failing. A larger increment would only be needed when xbar is below its 5-th percentile. Thus, only half of the 10% of the time that the likelihood ratio statistic exceeds its 90-th percentile is the upper confidence limit going to fail. Thus, a 95% upper confidence limit is based on the likelihood ratio statistic's 90-th percentile.

Comment 5.a.2

The guidance document proposes that, in general, the smallest BMDL be chosen as the BMDL. Choosing the minimum BMDL as the BMDL has considerable impact on the properties of the BMDL. For example, if there is only one BMDL to choose from and that BMDL is a 95% lower confidence limit on the true BMD, then the probability that the BMDL is less than or equal to the true BMD is 0.95. However, if there are three 95% lower confidence limits to choose from, then the probability that the BMDL is less than or equal to the true BMD is $1-(1-0.95)^3 = 0.99875$. Similarly, if there are five 95% lower confidence limits to choose from, then the probability that the BMDL is less than or equal to the true BMD is $1-(1-0.95)^5 = 0.999999675$. Thus, the minimum of several 95% lower confidence limits is no longer a 95% lower confidence limit. The minimum of several 95% lower confidence limits is much more conservative than a single 95% lower confidence limit.

If the minimum of M lower confidence limits is chosen as the BMDL, then the confidence levels for the M lower confidence limits should be increased. For example, if the minimum of two 77.6% lower confidence limits is chosen as the BMDL, then the probability that the BMDL is less than or equal to the true BMD is $1-(1-0.776)^2 = 0.95$. Similarly, if the minimum of three 63% lower confidence limits is chosen as the BMDL, then the probability that the BMDL is less than or equal to the true BMD is $1-(1-0.63)^3 = 0.95$. If the minimum of four 53% lower confidence limits is chosen as the BMDL, then the probability that the BMDL is less than or equal to the true BMD is $1-(1-0.53)^4 = 0.95$. If the minimum of five 45% lower confidence limits is chosen as the BMDL, then the probability that the BMDL is less than or equal to the true BMD is $1-(1-0.45)^5 = 0.95$.

Comment 5.a.3

Choosing the BMDL to be the minimum of the BMDLs from individual studies does not encourage multiple studies. Furthermore, by choosing the BMDL to be the minimum of the BMDLs from individual studies, the BMDL can only get worse with additional research; so that, additional research is not encouraged.

Comment 5.a.4

Choosing the BMDL to be the minimum of 95% lower confidence limits from individual studies and/or the minimum of 95% lower confidence limits from individual models and/or the minimum of 95% lower confidence limits from individual endpoints does not facilitate comparisons across chemicals. If the BMDL is chosen as a minimum of 95% lower confidence limits, then the confidence levels for different chemicals are generally going to be different and not readily comparable.

Comment 5.a.5

Choosing the BMDL to be the minimum of 95% lower confidence limits from individual studies causes the distribution of the BMDL to move further away from the true BMD. Thus, choosing the BMDL to be the minimum of 95% lower confidence limits from individual studies causes the BMDL to be farther away from the true BMD.

For example, suppose that the response model is

$$y = beta0 + beta1 \times dose + epsilon$$

where

$beta0 = 0$

$beta1 = 1$

epsilon = normal random variable with mean 0 and standard deviation 2.

Suppose there are 3 dose levels (0, 5, and 10) with 50 independent observations at each dose. Suppose that the BMR is 10%. Then the true BMD is 1.

The estimation of the dose-response model is a standard linear regression problem. Here, a 95% lower confidence limit on the BMD corresponding to an added risk of 10% is

$$BMDL = BMR / (b1 + t_{n-2,\ 0.95} \times sb1)$$

where

b1 = standard linear regression estimate of beta1

sb1 = standard estimated standard deviation of beta1

$t_{n-2,\ 0.95}$ = 95-th percentile of student t distribution with n-2 degrees of freedom, n = 150.

The distributions of BMDL in 10,000 Monte Carlo simulations of this example with BMDL being the minimum 95% lower confidence limit from 1, 3, and 5 studies are as indicated in the table below. The distributions move further and further away from the true BMD (1.0) as the number of studies increases from 1 to 3 to 5.

Distribution of BMDL when BMDL is the minimum of 95% lower confidence limits True BMD = 1.0			
Percentage	Minimum of 1 95% lower confidence limit	Minimum of 3 95% lower confidence limits	Minimum of 5 95% lower confidence limits
5%	0.43	040	0.38
10%	0.46	042	0.40
25%	0.51	0.45	0.43
50%	0.60	0.50	0.47
75%	0.71	0.55	0.51
90%	0.86	0.61	0.55
95%	0.98	0.65	0.58
99%	1.35	0.74	0.64

Comment 5.a.6

The way in which 95% lower confidence limits on BMDs are often computed using dose-response models usually results in a lower confidence limit (LCL) that is below the BMD much more often than 95% of the time.

It is a very common misconception to believe that the probability that a 95% LCL based on an unknown dose-response model is below its target approximately 95% of the time. In fact, it is usually below its target much more than 95% of the time.

This seems to contradict "elementary statistics." However, a closer look at the statistical definition of the nominal 95% coverage level will explain the seeming contradiction. The definition of the coverage level is as follows. Let S be the set of all possible parameter values. For example, if the dose-response model was a multistage model with the probability of a response at dose d equal to

$$P(d) = 1 - \exp(-[\text{beta0} + \text{beta1} \times d + \text{beta2} \times d^2]),$$

then S would be the set of all possible values of (beta0, beta1, beta2) or the set of all possible values with the betas restricted to be non-negative. Then the coverage for a lower confidence limit is defined to be the minimum over all possible values of the parameters (beta0*, beta1*, beta2*) in S of the probability that the lower confidence

limit is less than or equal to its target when beta0 = beta*, beta1 = beta1*, beta2 = beta2*. Thus, a 95% lower confidence limit BMDL on the BMD satisfies

$$0.95 = \text{minimum of } \{ P(\text{BMDL} \leq \text{BMD given that (beta0, beta1, beta2)} = (\text{beta0*, beta1*, beta2*}) \}$$

where the minimum is over all (beta0*, beta1*, beta2*) in S.

This means that the 95% coverage level is actually the minimum coverage level where the minimum is taken over all possible parameter values. That is, in general, the probability that a 95% lower confidence limit on BMD is less than or equal to BMD depends on the true value of the parameter (here, (beta0, beta1, beta2)). For some parameter values, the probability that a 95% lower confidence limit on BMD is less than or equal to BMD can be greater than 95%. The minimum coverage probability is 95%.

In practice, for a dose-response model like the multistage model, the coverage probability is almost always greater than 95% and sometimes much greater than 95%. The coverage probability happens to be 95% only when beta2 = 0; that is, when the dose term is a linear function of dose. For multistage models, the parameters corresponding to linear multistage models are the parameters with the minimum coverage probability. Hence, it is the presence of linear models in the parameter space that drives the form of the BMDL and its coverage probability.

An illustrative example may be helpful. Suppose that an animal bioassay has 50 animals at each of four doses. The four doses are 0, MTD/4, MTD/2, and MTD where MTD is the maximum tolerated dose. A multistage dose-response model of the form

$$P(d) = 1 - \exp(-[\text{beta0} + \text{beta1} \times d + \text{beta2} \times d^2])$$

is fit to the observed response frequencies with the parameters restricted to be non-negative. A nominal 95% lower confidence limit (BMDL) on the BMD for a BMR corresponding to an extra risk of 1/100,000 is computed using the standard linearized multistage model and GLOBAL software. Thus, the BMDL is based on the likelihood ratio. The nominal coverage being 95% means that the minimum coverage over all possible non-negative values of (beta0, beta1, beta2) is 95%. What is the coverage at specific values of (beta0, beta1, beta2)? Three values of (beta0, beta1, beta2) are explicitly evaluated. Each value corresponds to P(0) = 0.0 and P(MTD) = 0.60. The three values of (beta0, beta1, beta2) are

Case 1. (beta0, beta1, beta2) = (0, 0.92, 0)

Case 2. (beta0, beta1, beta2) = (0, 0.08, 0.84)

Case 3. (beta0, beta1, beta2) = (0, 0, 0.92)

The three cases correspond to increasingly nonlinear models; Case 1 is a linear model, Case 2 is linear-quadratic model, and Case 3 is a quadratic model.

The coverage in 100 Monte Carlo simulation trials is 94% in Case 1, and 100% in Cases 2 and 3. Thus, the coverages are substantially different than 95% in Cases 2 and 3. That is, the BMDL will be below the BMD approximately 95% of the time in those cases when the function of dose in the model is linear but substantially more than 95% of the time when the function of dose in the model is nonlinear. Thus, the coverage depends on the specific underlying dose-response model. This means that when a BMDL is computed using the standard linearized multistage model its coverage is unknown and dependent on the true dose-response relationship.

The following table indicates where the BMDLs were in relation to the true BMD in the above example.

The Conservative Bias in the Linearized Multistage Model BMDL Intensifies as Sublinearity in the Underlying Dose-Response Model Increases			
Specified Interval: Ratio = True BMD / BMDL	Percentage of Time the Ratio is in the Specified Interval in 100 Monte Carlo Simulations		
	P(d) = $1 - \exp[-0.92d]$	P(d) = $1 - \exp[-0.08d - 0.84d^2]$	P(d) = $1 - \exp[-0.92d^2]$
0.5 to 1	**6%**	0%	0%
1 to 2	**94%**	0%	0%
2 to 5	0%	**20%**	0%
5 to 10	0%	**72%**	0%
10 to 20	0%	**8%**	0%
20 to 50	0%	0%	**1%**
50 to 100	0%	0%	**14%**
100 to 1000	0%	0%	**85%**

Thus, in these examples, the BMDL is within a factor of 2 of the true BMD when the true underlying dose-response model is linear. However, if the underlying dose-response model is linear-quadratic, then the BMDL is 2 to 20 fold below the BMD. Furthermore, if the underlying dose-response model is quadratic, then the BMDL is 20 to 1,000 fold below the BMD.

This example clearly illustrates that the behavior of a BMDL is heavily dependent on the underlying dose-response model. Specifically, the degree to which the BMDL is below the BMD is heavily dependent on the underlying dose-response model.

This example is a good illustration of why comparisons between chemicals should be based on estimated BMDs rather than BMDLs.

Returning briefly to "elementary statistics": Although the coverage of 95% confidence limits in the context of dose-response models is not usually 95%, the coverage in simpler situations is often always 95%. For example, a 95% confidence interval on the mean μ of a normal random variable is equal to 95% regardless of the value of μ. In fact, most non-dose-response-modeling situations have confidence limit

procedures which have the same level of coverage for all parameter values in the parameter space. However, this is not true in most dose-response-modeling situations.

Comment 5.a.7

Basing BMDL on a bootstrap procedure is a better than basing the BMDL on the usual non-bootstrap confidence limit procedures.

The usual non-bootstrap confidence limit procedures base the BMDL on the asymptotic properties of procedures; that is, the properties as the sample size tends to infinity. These properties do not apply for finite sample sizes. Furthermore, their small sample properties are very case dependent (as illustrated in Comment 5.a.6). Loosely speaking, you don't really know what you have when you compute the BMDL using the usual non-bootstrap confidence limit procedures such as those based on the asymptotic distribution of the likelihood ratio or the asymptotic distributions of the parameters.

Bootstrap procedures focus on the case at hand. Bootstrap procedures reflect the given sample sizes and the underlying dose-response model rather than all possible dose-response models.

My professional preference is to use a non-parametric bootstrap procedure. For a BMD, a non-parametric bootstrap procedure would estimate the BMD using the original study data. Then, the procedure resamples (with replacement) the original study data to create a simulated experimental outcome (same number of doses, same doses, same sample size per dose, etc.). The estimated BMD is computed for the simulated data set. This procedure is repeated as often as desired. The result is one original estimate of the BMD and several simulated estimates of the BMD. The combined distribution of estimated and simulated BMDs is used to characterize the true BMD. For example, a 95% lower bound (say, BMDL) on the BMD would be the largest estimated or simulated BMD value such that at least 95% of the estimated and simulated BMDs are greater than that value.

(If there were 100 estimated or simulated BMDs, the BMD would be the 6-th smallest BMD. Here, the 6-th, 7-th, ..., 100-th smallest values are 95 out of 100 values, and these 95% of the values are greater than or equal to the 6-th smallest value. If there were 1,000 estimated or simulated BMDs, the BMD would be the 51-th smallest BMD. Here, the 51-th, 52-th, ..., 1000-th smallest values are 950 out of

1000 values, and these 95% of the values are greater than or equal to the 51-th smallest value.)

The advantage to a bootstrap procedure is its ability to reflect the case at hand rather than a set of cases containing one relevant case (namely, the case at hand – the existing data set and the true underlying dose-response model) and a host of irrelevant cases (e.g., parameter values other than the true parameter value).

The improvement in the behavior of a BMDL based on a 100 bootstraps versus the BMDL based on the linearized multistage model (LMS) is illustrated in the following table. The three dose-response models are the same as in Comment 5.a.6. For each of the three dose-response models, there are 20 original data sets generated. The BMR is 10% here. For the linear model, the LMS procedure's average BMDL is 91% of the true BMD, and the bootstrap procedure's average BMDL is 98% of the true BMD. (The average is over the 20 original data sets.) For the linear-quadratic model, the LMS procedure's average BMDL is 71% of the true BMD, and the bootstrap procedure's average BMDL is 77% of the true BMD. For the quadratic model, the LMS procedure's average BMDL is 68% of the true BMD, and the bootstrap procedure's average BMDL is 73% of the true BMD. The two procedures are much more similar for BMR = 10% than for they are for smaller BMRs. Nevertheless, on average, the bootstrap procedure's BMDL is closer to the true BMD than the LMS procedure's BMDL.

This example also illustrates the reasonableness of the bootstrap procedure for those who are seeking information about its empirical behavior.

The Improvement in the Behavior of a BMDL based on a Bootstrap Procedure versus the BMDL based on the Linearized Multistage Model (LMS) Increases as the Sublinearity in the Underlying Dose-Response Model Increases						
Specified Interval: Ratio = True BMD / BMDL	Percentage of Time the Ratio is in the Specified Interval in 100 Monte Carlo Simulations					
	P(d) = $1 - \exp[-0.92d]$		P(d) = $1 - \exp[-0.08d - 0.84d^2]$		P(d) = $1 - \exp[-0.92d^2]$	
	LMS BDML	Bootstrap BDML	LMS BDML	Bootstrap BDML	LMS BDML	Bootstrap BDML
0.60 to 0.70	5%	**5%**				
0.70 to 0.80						
0.80 to 0.90		**15%**				
0.90 to 1.00	15%	**20%**		**5%**		**10%**
1.00 to 1.10	25%	**10%**	10%	**10%**	10%	**5%**
1.10 to 1.20	20%	**30%**	10%	**10%**	10%	**15%**
1.20 to 1.30	25%	**20%**	5%	**5%**	10%	**20%**
1.30 to 1.40	10%		10%	**10%**	15%	
1.40 to 1.50			15%	**15%**	20%	**35%**
1.50 to 1.60			10%	**10%**	15%	
1.60 to 1.70			10%	**20%**		**10%**
1.70 to 1.80			10%	**5%**	10%	
1.80 to 1.90			10%	**5%**	5%	
1.90 to 2.00			5%	**5%**		
2.00 to 2.10						
2.10 to 2.20						**5%**
2.20 to 2.30			5%			
2.30 to 2.40						
2.40 to 2.50					5%	

b. Please comment on the soundness of the criteria used to deviate from using a 95% one-sided confidence interval for the BMDL.

Comment 5.b.1

A BMDL provides a limited characterization of the impact of experimental variability on the estimate of the BMD. A BMDL is a single number that provides a lower bound on the true BMD but does not explicitly indicate the relative likelihood that the true BMD takes on specific values. The BMDL does not characterize the relative likelihood of any specific number in the range between the estimated BMD and the BMDL (or below) being the true BMD. For example, it does not characterize the relative likelihood that the true BMD is halfway between the estimated BMD and the BMDL, nor the relative likelihood that the true BMD is one-fourth of the way between the estimated BMD and the BMDL, etc. However, if the BMDL is computed using a bootstrap procedure, then there is a distribution of estimated BMDs available. We have discussed in Comment 5.a.7 how this distribution can be used to determine an improved BMDL. Another use of the bootstrap distribution is to characterize the relative likelihood that different values are the true BMD. Thus, the BMDL can be replaced or supplemented with a distribution indicating the relative likelihood of different values for the true BMD.

6. Examples: What additional concepts, if any, should be illustrated by an example?

No comments at this time.

III. Interpretation and Using the Benchmark Dose — *Discussion Leader - Lorenz Rhomberg*

7. The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document's guidance, or would further discussion be useful.

No comments at this time.

8. What other comments do you have about the approach to benchmark dose described in the document? What changes to the approach would you like to see, and, very importantly, why?

Comment 8.1

The benchmark dose approach should emphasize the estimated BMD and not the BMDL as the target value for risk assessment. Estimated BMDs have an estimated toxicological risk whereas the BMDL does not. Specifically, the estimated excess risk at the estimated BMD is BMR. The estimated excess risk at the BMDL is less than BMR, but how much less is unknown; thus, the estimated excess risk at the BMDL is unknown.

Comparisons among chemicals should be based on the estimated BMDs and not the BMDLs. Comparisons based on estimated BMDs are comparisons based on the same excess risks. Comparisons based on BMDLs are comparisons based on unknown and generally different excess risk levels.

The example in Comment 5.a.6 shows that how often a BMDL is below the true BMD and how far the BMDL is below the true BMD is very case dependent. Thus, comparisons across chemicals should be based on estimated BMDs and not based on BMDLs.

Comment 8.2

Human epidemiological data are important sources of information about the human dose-response relationship. The dose-response models used for these data are quite varied and usually quite different from the models used for animal experimental data. Appropriately, the guidance document does not restrict the models used for

human epidemiological data. On the other hand, BMRs of 10% are usually not appropriate for BMDs based on human studies. Rarely, is a 10% excess risk observed in human studies. Quite often a BMD for a BMR of 0.10% (one hundred fold smaller than a BMR of 10%) or an even smaller percentage is more in the heart of the human data. Extrapolating upwards from observed dose levels to dose levels corresponding to a BMR of 10% is not very reasonable or reliable. Furthermore, BMDs for BMRs of 0.10% or smaller are generally less dependent on the choice of the dose-response model than BMDs for a BMR of 10% when the observed data is human epidemiological data. Thus, BMRs considerably less than 10% are more appropriate for human epidemiological data than a BMR of 10%.

Comment 8.3

Choosing a BMR assumes that the risk assessment based on the BMD for that BMR will adequately reflect the choice of the BMR. How is an acceptable margin of exposure determined differently for a BMD based on a BMR of 10%, 5%, 1%, 0.1%, etc.?

Comment 8.4

BMDs should be computed on the animal dose scale when the true human equivalent dose scale is unknown. Unfortunately, the BMD is often computed on an assumed human equivalent dose (HED) scale (usually based on a default body-weight scaling factor) when the true human equivalent dose scale is unknown, and then the risk assessment based on this BMD incorporates additional conservative adjustments because the BMD is based on animal data. In order to avoid redundant conservative adjustments, BMDs should be computed on the animal dose scale when the true human equivalent dose scale is unknown.

Comment 8.5

BMDs should be calculated using <u>added</u> risks instead of <u>extra</u> risks. Benchmark doses based on added risk refer to the whole population whereas benchmark doses based on extra risk refer only to the portion of the population that would not have an adverse response when the dose was zero. Because added risks refer to the whole

population and because risks for a population are usually calculated and communicated in terms of added risks rather than extra risks, added risks are the more appropriate basis for benchmark doses.

Also, BMDs based on added risks are more comparable across chemicals. This follows because added risks are on an absolute scale which is the same for all chemicals whereas extra risks are on a relative scale that varies as the background response rate varies.

Comment 8.6

The best estimate of the BMD should be emphasized. If a lower bound (e.g., a BMDL) is presented in order to reflect the uncertainty associated with the estimated BMD, then an upper bound (based on the same approach that was used to generate the BMDL) should also be presented. This will provide additional useful information to the risk manager.

William Slikker, Jr.

William Slikker, Jr., Ph.D.

Dr. Slikker is currently Director, Division of Neurotoxicology, National Center for Toxicological Research, FDA, and Adjunct Professor in the Departments of Pediatrics and Pharmacology and Toxicology at the University of Arkansas for Medical Sciences and in the Department of Medicinal Chemistry, University of Tennessee.

Education: University of California at Santa Barbara, 1972 BA, Biology;
University of California at Santa Barbara, 1974, M.A., Biological Sciences;
University of California at Davis, 1974, Ph.D., Pharmacology and Toxicology;
Postdoctoral Staff Fellowship, 1978-1979, Perinatal Research Program, NCTR/FDA;
Sabbatical Study, 1989, Freie Universitat, Berlin.

His professional activities include the Teratology Society: Councilor, 1995-1998, Publication Committee Chairperson, 1998-2000 and Vice-President Elect, 2000-2001; Society of Toxicology: South Central Chapter, President, 1988-1989, Neurotoxicology Specialty Section, President, 1995-1996; American Society of Pharmacology and Experimental Therapeutics: Executive Committee Chairperson for the Developmental Pharmacology Section, 1993-1997 and Program Committee member, 2000-2002.

Editorial Board member of *Fundamental and Applied Toxicology*; 1989-1995; *Reproductive Toxicology*, 1997-present; *Teratology* Section Editor,1994-1998; Associate Editor of *NeuroToxicology*, 1994-present and *Toxicological Sciences*, 1999-present. Reviewer for NIH, HEI, EPA, ATSDR and NIEHS developmental and neurotoxicology grants and manuscripts.

Research interests include developmental pharmacology, transplacental pharmacokinetics, neurotoxicology and risk assessment; mechanisms of toxicity and approaches to neuroprotection for substituted amphetamines, excitotoxicants and metabolic inhibitors. He has authored or co-authored over 235 peer-reviewed research articles and book chapters and co-edited the first comprehensive text on Developmental Neurotoxicology.

Peer Review Workshop on the Benchmark Dose

I. Preparation for Computing a Benchmark Dose: Selecting Data and An Appropriate Benchmark Response Level

Questions #1: What concepts and terms related to the benchmark dose (BMD), if any, do we need to define more clearly?

Response: No comment.

Question #2 The literature review cites works that has helped to develop the BMD approach. Do you have suggestions for inclusion of other work?

Response: On page 9, lines 9-16: Other examples of the use of continuous developmental toxicology data in the risk assessment process have been published. Slikker and Gaylor, 1998.

Quantitative models of risk assessment for developmental neurotoxicants. In: Handbook of Developmental Neurotoxicology, Eds W. Slikker and L. Chang, Academic Press, San Diego, pp. 727-732.

On page 9, line 16, page 11, line 8: An additional reference presenting an example of how continuous data can be used to derive BMDs using the hybrid approach may be added.

Slikker, W., Jr., Crump, K.S., Andersen, M.E., and Bellinger, D. Biologically-based, quantitative risk assessment of neurotoxicants. Fund. Appl. Toxicol., 29:18-30, 1996.

On page 15, line 9: The importance of the variance to the BMD outcome in both control and experimental groups needs to be emphasized and discussed.

Question # 3: What additional discussion, if any, is needed to clarify the description of selection of studies and endpoints for the BMD?

Response: On page vi, line 5 and page 13, lines 22-25: Clear criteria should be stated for when the BMD is not calculated but rather a NOAEL/LOAEL approach is appropriate.

On page vii, line 25 – viii, line 5: It should be stated that if continuous data are available, they should be used to calculate a BMD because the process of "quantalization" of the continuous data may result in loss of precision (Gaylor, 1996).

On page 16, lines 6-9: What criteria are used to select studies for benchmark dose analysis?

II. Modeling to Compute a Benchmark Dose: Model Selection, Fitting, and Confidence Limits

4. Model selection and fitting

a. What additional discussion is needed to ensure adequate presentation of the proposed defaults for the parameters for various models?

Response: On page 22, line 6: It is true that nonlinear models do not necessarily have biological interpretation but then neither do linear models. At

least the nonlinear model (especially those of the Michaelis-Menten-type) are consistent, shape wise, with the saturation of enzyme systems, uptake systems and receptor binding that occur in biological systems.

Slikker, W., Jr. and Gaylor, D.W. Biologically-based dose-response model for neurotoxicity risk assessment. Korean J. Toxicol., 6(2):205-213, 1990.

Slikker, W., Jr. and Gaylor, D. Concepts on quantitative risk assessment of neurotoxicants. In Neurotoxicology: Approaches and Methods, Chang and Slikker, Eds., Academic Press, Chapter 51, pp. 771-776, 1995.

On page 24, lines 22-24: An example from the literature can be used here to show the importance of variance on the calculation of the BMD from continuous data.

West, W.H., and Kodell, R.L. (1993). Statistical methods of risk assessment for continuous variables. Communications Statistics-Theory Methods 22, 3363-3376.

b. Please comment on the adequacy of the criteria to evaluate the fit of the model. Would you recommend additional criteria?

Response: On page 29, line 19-22: The point that higher doses may be dropped if they reflect the up region of a “U” shaped dose response curve is very important. It should be emphasized that these higher dose points definitely be dropped if there is any mechanistic data to suggest an explanation for the “U” shaped curve.

On page 33, line 20: Another reference on how to deal with change in mean response relative to the standard deviation is provided by West and Kodell, 1993

(see 4a).

c. What are the advantages/strengths of using the methods described to select among "equally" fitted models? What other methods should be considered in marking a selection?

Response: See II 4b above

5. Use of confidence limits

a. Please comment on the approaches described to compute confidence limits.

Response: See II 4b above

b. Please comment on the soundness of the criteria used to deviate from using a 95% one-sided confidence interval for the BMDL.

Response: No comment.

6. Examples: What additional concepts, if any, should be illustrated by an example?

Response: A simple case of the use of continuous data should be presented. The current single example is complicated with some issues that pertain to other types of data sets.

In addition, an example of the "hybrid" type of BMD calculation should be provided (page 24, line 3). Examples could include Gaylor and Slikker, 1992 and Slikker et al, 1998.

Gaylor, D. and Slikker, W., Jr. Risk assessment of neurotoxicants. Neurotoxicology, Eds. Hugh Tilson and Clifford Mitchell, Raven Press, Ltd., New York, 1992, pp. 331-343.

Slikker, W. Scallet, A. and Gaylor, D.W. Biologically-based dose-response model for

neurotoxicity risk assessment. Toxicology Letters, 102-103:429-433, 1998.

III. Interpretation and Using the Benchmark Dose

7. The guidance document summarizes the elements of the reporting requirements needed to document BMD computations. Is the inclusion of each of these elements reasonable in light of the document's guidance, or would further discussion be useful?

Response: Because many of the new data sets that will be reported to EPA in the future will be continuous in nature, it may be important to set this category of data on an equal footing with the quantal data approach and arrange this document and reporting requirements accordingly. As neurotoxicology and immunotoxicology and most of the genomic and proteomic data become more the rule rather than the exception, the need to deal with continuous data sets will become much greater. Therefore, EPA should prepare now to encourage the submission of these continuous data sets and not give the impression that these data are an unusual or extra problem. By providing clear guidance and several clear examples of how to deal with continuous data, EPA will be properly preparing for "modern toxicology" which is already upon us.

8. What other comments do you have bout the approach to benchmark dose described in the document? What changes to the approach would your like to see, and, very importantly, why?

Response: See #7.

R. Webster West

Dr. West is an Associate Professor in the Department of Statistics at the University of South Carolina. He joined USC after completing his Ph.D. at Rice University in 1994. While in graduate school, he began doing research on risk assessment for continuous responses with members of the Biometry Department at the National Center for Toxicological Research (NCTR). During his tenure at USC, he has continued to do research in this area with the NCTR group, and he has also began developing methods for applying simultaneous confidence bands in risk assessment with his colleague, Walter Piegorsch. This research project is now in its fourth year of funding from NIH. Dr. West has published several articles on risk assessment in journals such as Computational Statistics and Data Analysis and Risk Analysis. He also serves on a study section for NIH.

Topic 1

Question1

The document ignores any formal definition or discussion of some key terms in risk assessment such as risk, additional risk and excess risk (a formula for this quantity is given on line 1 of page 54). Risk is typically defined as the probability of an adverse effect at a certain dosage level. Additional risk at that dose level is then defined as the risk at that level minus the risk for controls (background risk). Excess risk is additional risk divided by one minus the background risk. The BMD as defined in this document is the dose corresponding to an excess risk of 0.1. For dichotomous responses, the observed data is often times given on a risk scale as the fraction of adverse effects are reported at each of the dose levels. In this case, it is easy to estimate and understand excess risk in terms of the original data scale. For continuous responses, the document suggests that a dose level corresponding to a mean response that is one standard deviation above/below the control mean will correspond to an excess risk of 0.1 for normally distributed data. (As a side note, I get an excess risk of about 13% in this case.) This is only true if the background risk is 0.02. The document correctly identifies the problem of defining a cutoff for adverse effects with continuous data, but the document does not associate the chosen cutoff with background risk. If a risk assessor chooses a cutoff corresponding to a background risk that is significantly different from 0.02, the excess risk at the recommended BMD level may be far above or far below 0.1. Without a better description of risk and its connection to the cutoff, the guidance provided by the document may be severely flawed in terms of interpretation.

Question 2

The document overlooks some of the original applications of the BMD approach to continuous responses. Some examples are given below. Since I found the examples section to be lacking in this area, I feel that these references should definitely be included.

1. West, R. W. and Kodell, R. L. (1993). ``Statistical methods of risk assessment for continuous variables," *Communications in Statistics*, 22:3363-3376.

2. Kodell, R. L. and West, R. W. (1993). ``Upper confidence limits on excess risk for quantitative responses,'' *Risk Analysis*, 13:177-182.

Question 3

No comments.

Topic 2

Question 4a

I don't think I understand this question completely. I am unaware of any default parameter values that should be used for any of the models discussed in the document. I also don't believe the suggestion made in refernce to the first example that the slope for the log-logistic model should be constrained to be greater than one. The data for this example is somewhat poor, and I think extrpolating based on this example is a big mistake.

Question 4b

The discussion of model fitting seems to be somewhat biased towards quantal responses in terms of the outlined approach. The suggestion that all model fitting is iterative is misleading because this is not the case for continuous responses. In terms of criteria for evaluating the fit of a continuous dose response model, one should also look at a histogram of the residuals to see if the assumed distribution for the responses appears to be met. Any subsequent inferences depend greatly on this distribution. This could also be done more formally with a goodness-of-fit test.

Question 4c

No comments.

Topic 3

Question 5a

I found the discussion of confidence intervals at the bottom of page 30 somewhat scary. The relationships described between one-sided and two-sided confidence limits assumes a symmetric distribution. I realize that this may be very frequent in practice, but I think this discussion should make some note of this assumption. The document does identify the profile likelihood procedure as the prefereed method for obtaining the BMDL, but this name is not given to the procedure discussed at the bottom of page 31 until the examples section. I feel this should be added so that the document provides a reference term for this approach.

Question 5b

I found very little mention (lines 7-10 on page 31) of any criterion for deviating from the 95% lower confidence limit for the BMDL. However, I am not sure that more discussion is required on this

topic.

Question 6

I found the examples section to be somewhat lacking. There is very little discussion of the data used in these examples, and I think the choice of these data sets is unfortunate (specifically for the first two examples). It is a bad practice to fit a model with two or more parameters to three data values as is done in the first example. A more standard set of data would better outline the general approach one should take in this case. The dose-response shown in Figure A-3.1 is as messy as I have ever seen. Several clean data sets could be used as better examples in this case. Also, there is no plot of the fitted model for this example as is suggested by the document.

Topic 3

Question 7

I think the overall guidance provided by the document is quite reasonable, but I think more attention needs to be given to determining a BMR for continuous responses. The document lacks a clean description of how this quantity affects the resulting BMD analysis.

Question 8

The document does identify several motivating reasons why the BMD approach is preferable to the NOAEL approach. The bulk of these reasons, however, center on the fact that the NOAEL is a bad estimate of an unknown dose threshold. It may be a bad idea to throw out threshold models completely due to the problems with one approach. When deemed biologically appropriate, a reasonable way to approach risk assessment would be to first fit threshold models (changepoint models) which incorporate information from the dose-response to estimate a dose threshold. Lower confidence limits for this dose threshold will not suffer from the same problems exhibited by the NOAEL. If the lower confidence limit on the threshold is 0, then one may move directly to BMD appraoch. This would in many cases represent a much more unified appraoch to doing risk assessment.

The document sometimes suggests that one should always return the BMDL for the dose corresponding to an exces risk of 0.1 along with other BMDLs for other BMRs. Caution should be emphasized in this situation because the BMDLs are not simultaneous and to interpret them properly requires a correction for multiple comparisons.

Yiliang Zhu

Yiliang Zhu, Ph.D.
Associate Professor
Department of Epidemiology and Biostatistics
College of Public Health, University of South Florida

Dr. Yiliang Zhu has extensive research experience in quantitative risk assessment of health effects due to environmental exposure, and has expertise in the benchmark dose methodology as applied to risk assessment of carcinogenicity, developmental toxicity, mutagenicity, and neurotoxicity. He is currently the principal investigator of a National Science Foundation grant that supports statistical researches to explore effective study designs, dose-response modeling, and benchmark dose estimation for neurotoxicity screening. His current work also includes developing biologically-based dose-response models to test the validity of linear low-dose extrapolation in carcinogenicity risk assessment. He has published numerous articles related to environmental health risk assessment, and also served regularly as a peer reviewer for numerous statistical and environmental science journals such as the Journal of American Statistical Association, Risk Analysis, Biometrics, Fundamental and Applied Toxicology, Teratology, Environmental Health Perspectives.

Topic I

Question 1
Response: None

Question 2
Response: None

Question 3
Response:

Endpoint selection: In situations where multiple endpoints are available for dose-response modeling and BMD calculation, all endpoints with LOAEL up to 10-fold of the lowest LOAEL should be analyzed to ensure the detection of the lowest BMDL (P.16, L.20-24). Some references or explanation on the choice of 10-fold cap would be helpful.

Whereas some guidance is given on combining data sets for dose-response modeling, relatively little is given on combining endpoints within a single study (II.A.3. p.16). In addition to dose-response modeling of multiple endpoints (I.C.3), another alternative approach is dose-response modeling of combined endpoints when combining endpoints is judged as relevant such as in the case of neurobehavioral screening test (Moser at al., 1997).

Topic II

Question 4a
Response: None

Question 4b
Response:

BMR Selection for continuous data: The default amount of change used to define BMR is based on a mean shifting (P.19, L.14-19). Although conventional and easy to use, this approach has limitations. For example, adverse effects may be characterized by some individual's deviation from the population average to either end, resulting in the same mean, but increased variation in the distribution of the exposed population. It seems possible to unify the various options of BMR for continuous data by first defining a region A of values for the endpoint, such that any response y belonging to A is a risk indicator. Given the distribution of the response, the region A can be defined in principle by, for example, shift in mean, a fraction of the range of the control population responses, or a threshold of abnormality. The probability, Prob(y belongs to A), measures that risk in relation to exposure level. A special case of this risk is the proportion of affected subjects. This probability is computed from a fitted dose-response model. Based on the risk Prob(y belongs to A), computation of BMD would be essentially the same as that of quantal data.

Question 4c
Response:
Model selection criteria: The default acceptance criterion for goodness-of-fit is a p-value >= 0.10, instead of a conventional p-value >=0.05, for example. Suppose we have a data set that reveal a slightly higher response rate at the first dose group than at the second, and an otherwise monotone dose-response trend. It is probable that a dose-response model fit to the data set will have a p-value greater than 0.05 but less than 0.10. It is not clear to what degree the estimated BMD/BMDL will differ from the true one, if this fitted dose-response model is accepted as a basis for BMD/BMDL computation.

Question 5a
Response:
Confidence Limits: Most conventional methods such as the delta-method and likelihood ratio require strong assumptions for correct confidence limits or intervals, especially when the size of the data set is small and the distribution of the data is skewed. As an alternative, yet robust method, the bootstrap method should be explored and encouraged to use. As we gain more experience and user-friendly software become available, bootstrap-based confidence limits (BMDL) may be considered default.

Question 5b.
Response: None

Question 6
Response:
Other examples or concept: Future version of the guideline may consider including an example of repeated measure data set, or an example of covariates. In either case, BMD will be a function of the covariate (or time in the repeated measure case).

Topic III

Question 7
Response: None

Question 8
Response: See comments related to question 6. When covariates are included in a dose-response model the resulting BMD will be a function of the covariates (e.g. time), guidance will be necessary as to at which covariate point(s), BMR, BMD, and BMDL should be defined. Expert opinion on relevant testing time is critical, epidemiological and statistical considerations on issues such as higher risk sub-population, first peak effect time would be important.

Reference

Moser, V.C., Tilson, H.A., MacPhail, R.C., Becking, G.C., Cuomo, V., Frantik, E., Kulig, B.M. and Winneke, G. (1997) The IPCS collaborative study on neurobehavioral screening methods: II. protocol design and testing procedures. *Neurotoxicology* **18(4),** 929-938.

APPENDIX E

WORKSHOP AGENDA

United States
Environmental Protection Agency
National Center for Environmental Assessment

Peer Review Workshop on the Benchmark Dose Technical Guidance Document

Holiday Inn Capitol
Washington, DC
December 7-8, 2000

Agenda

Workshop Chair: Colin Park

T H U R S D A Y , D E C E M B E R 7 , 2 0 0 0

8:00AM	**Registration**
8:30AM	**Welcome and Announcements** *Kate Schalk, Peer Review Manager, Eastern Research Group, Inc. (ERG)*
8:35AM	**Opening Remarks** *Workshop Chair: Colin Park*
8:45AM	**Reviewer Introductions/Conflict of Interest Disclosure** *Facilitated by Colin Park*
9:00AM	**Background on the Benchmark Dose Technical Guidance Document** *Woodrow Setzer, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency* (EPA)
9:20AM	**Charge to the Reviewers and Groundrules for Discussions** *Workshop Chair: Colin Park*
9:30AM	**Summary of Premeeting Comments and General Discussion** *Workshop Chair: Colin Park*
10:00AM	B R E A K
10:15AM	**Plenary Discussion of Charge Question 1** *Discussion Leader: George Alexeeff*
12Noon	**Observer Comments**
12:30PM	L U N C H

THURSDAY, DECEMBER 7, 2000 (continued)

1:30PM **Plenary Discussion of Charge Question 1: Recommendations Summary**
Discussion Leader: Georg Alexeeff

2:00PM **Plenary Discussion of Charge Question 2**
Discussion Leader: Lynne Haber

3:00PM B R E A K

3:15PM **Plenary Discussion of Charge Question 2**
Discussion Leader: Lynne Haber

4:15PM **Plenary Discussion of Charge Question 2**: **Recommendations Summary**

4:45PM **Wrap Up and Charge for Day Two**

5:00PM A D J O U R N

FRIDAY, DECEMBER 8, 2000

8:30AM **Opening Remarks**
Colin Park, Workshop Chair

8:35AM **Plenary Discussion of Charge Question 3**
Discussion Leader: Lorenz Rhomberg

10:00AM B R E A K

10:15AM **Plenary Discussion of Charge Question 3: Recommendations Summary**
Discussion Leader: Lorenz Rhomberg

11:15AM **Observer Comments**

11:45AM **Synthesis and Development of Recommendations/Wrap-Up**
Workshop Chair: Colin Park

12:30PM A D J O U R N

APPENDIX F

OVERHEADS

Overheads from Dr. Setzer's Presentation

Benchmark Dose Technical Guidance

Background for Peer Review Workshop

Technical Panel Members

- David Gaylor
- Karen Hogan
- Jennifer Jinot
- Carole Kimmel
- Woodrow Setzer - chair

Audience and Purpose of BMD Technical Guidance

- Target audience is Agency risk assessors and their statistical support
- Provide basic technical instruction in the application of BMD methodology
- Provide background in the methodology to facilitate flexible decision making, assistance in selecting software, and reviewing other applications of BMD
- Establish a set of consistent defaults

Guidance Context

- BMD Technical Guidance is one of several documents that inform dose-response assessment processes that use BMD, for example:
  - Endpoint-specific guidelines, Framework for Harmonizing Approaches to Cancer and Noncancer Risk Assessment, review of the RfD process (in development), and the revision of the cancer guidelines.
- In particular, other activities are in progress that will provide guidance on BMR and uncertainty factor selection in the RfD process, including situations where the BMD is to be used.

Risk Assessment Context

- The general problem which BMD is proposed to address:
- How do we quantify dose, both for comparison across studies and endpoints, and for extrapolating?

Risk Assessment Context (cont.)

- Original Approach: NOAEL - Highest dose in a study for which there is insufficient evidence to infer that the response differs from the control response.
- NOTE: NOAEL stands for No Observed Adverse Effect Level, not No Adverse Effect Level.
- Thus, comparisons between studies or endpoints may not be based on the same currency.

Risk Assessment Context (cont.)

- Crump (1984) Proposal: BMD - Use the dose corresponding to a prespecified response, derived by *interpolating* dose-response data using a fitted curve.
  - BMD does not equal NOAEL *or* LOAEL, though
  - we may *calibrate* the BMD (through choice of BMR) so that the BMD falls in the same general range as the NOAEL.
  - The goal is, given a response level, to get the best available estimate of the dose that should yield that response level, on average, and quantify its uncertainty.

Data Context

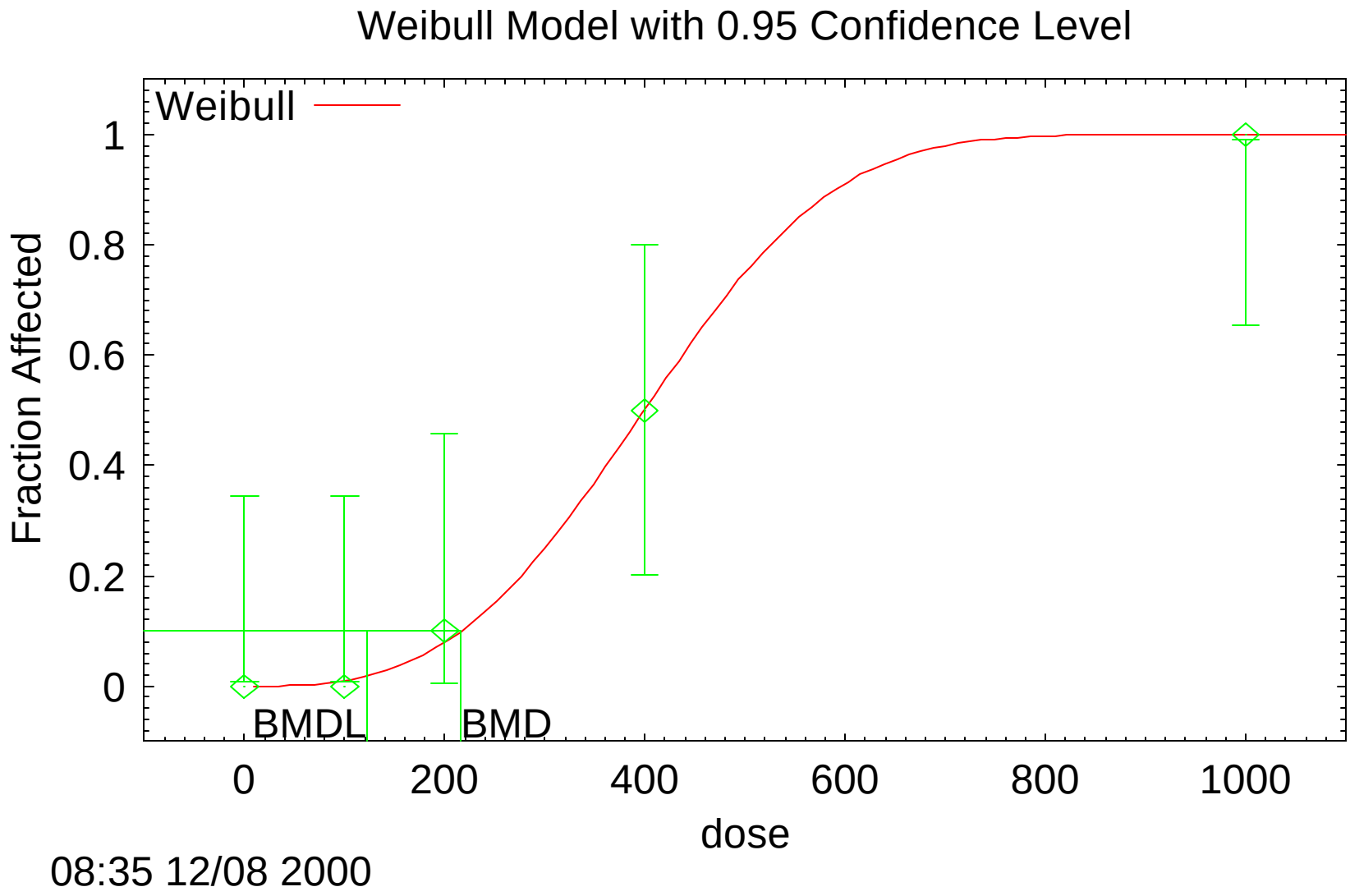
- Conformity with EPA testing guidelines (though guidelines were not necessarily written from the standpoint of estimating dose-response)
- Level of detail ranges from full access to individual data to only summary data (standard deviations reported to 1 significant digit, if at all).
- Ability to require further studies ranges from no to yes (within restrictions).
- Epidemiology data to bioassays involving invertebrates.

Overheads Used by Dr. Frederick

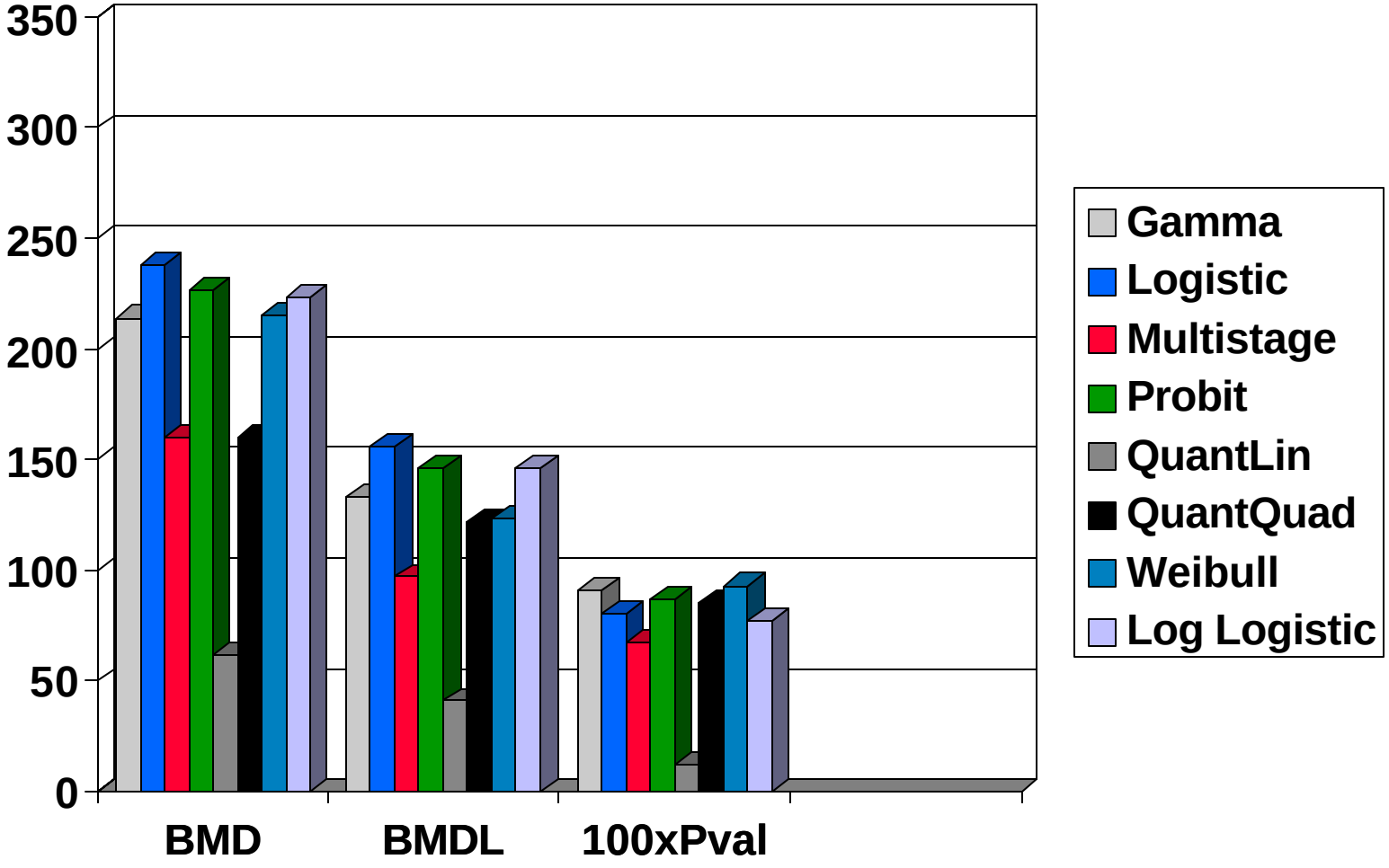
Test 1 Dataset

- Doses = 0, 100, 200, 400, 1000 mg/kg
- 10 animals/dose group
- Responses of 0, 0, 10, 50, 100% incidence of effect
- This is an example of a monotonic dataset for evaluating the BMD methodology. Varying the experimental design parameters (e.g. mid and high doses, number of animals/dose group, etc.), the data resulting from a study (e.g. incidences of effect at each dose level), and the curve fit methodology is instructive relative to the factors affecting the calculations of the BMD and BMDL.

Example of One of the BMDS Curvefits for Dataset 1



Results of BMDS Evaluation of the Test 1 Dataset



APPENDIX G

ADDITIONAL REFERENCES SUGGESTED BY REVIEWERS

ADDITIONAL REFERENCES SUGGESTED BY REVIEWERS

Note: This appendix contains excerpts from reviewer premeeting comments in which additional references were suggested, as well as references supplied by reviewers during and after the workshop. At the workshop, reviewers agreed to compile this reference list and submit it as part of the workshop report.

It is suggested that, if possible, the specific limitations of the NOAEL/LOAEL (page 4) be referenced, instead of having some of them referenced specifically and some of them referenced generally.

Two documents that should be included were developed by OEHHA, Cal/EPA. One document entitled "Air Toxics Hot Spots Program Risk Assessment Guidelines, Part I, The Determination of Acute Reference Exposure Levels for Airborne Toxicants," (Office of Environmental Health Hazard Assessment, Cal/EPA, March 1999), includes 14 examples of BMD calculations. This document can be found at www.oehha.ca.gov/air/acute_rels/acuterel.html. Another document entitled "Air Toxics Hot Spots Program Risk Assessment Guidelines, Part III, The Determination of Noncancer Chronic Reference Exposure Levels for Airborne Toxicants," (Office of Environmental Health Hazard Assessment, Cal/EPA, February, 2000, and May, 2000), includes four examples of BMD calculations. This document can be found at www.oehha.ca.gov/air/chronic_rels/chronicrel.html.

General background on noncancer risk assessment: Baird, S., J.C. Cohen, J.D. Graham, A.I. Shlyakhter, and J.S. Evans. Noncancer risk assessment: A probabilistic alternative to current practices. *HERA*, 2(1):79--102, 1996.

Alternative approaches or perspectives on BMD calculation: Bosch-RJ; Wypij-D; Ryan-LM A semiparametric approach to risk assessment for quantitative outcomes. Risk-Anal. 1996 Oct; 16(5): 657-65.

Vermeire, T., H. Stevenson, M.N. Pieters, M. Rennen, W. Slob, and B.C. Hakkert. Assessment factors for human health risk assessment: A discussion paper. *Crit. Rev. Toxicol.*, 29(5):439--490, 1999.

Slob W., and M. N. Pieters, A probabilistic approach for deriving acceptable human intake limits and human health risks from toxicological studies: general framework, Risk Analysis,v18:787-798, 1998.

More information on experimental design: Brand, K.P., L. Rhomberg, and J.S. Evans. Estimating noncancer uncertainty factors:Are ratios [of] NOAELs informative ? *Risk Analysis*, 19(2):295--308, 1999.

Brand KP, P Catalano, JK Hammitt, L. Rhomberg and JS Evans, Limitations to empirical extrapolation studies: The case of BMD ratios, accepted by Risk Analysis, November, 2000.

Literature pertaining to the combining of toxicological evidence: DuMouchel WH and JE Harris, Bayes methods for combining the results of cancer studies in humans and other species, *JASA*, v78(382):293-315, 1983.

DuMouchel WH and Groer PG, A Bayesian methodology for scaling radiation studies from animal to human, *Health Physics*, v57(Suppl 1):411-18, 1989.

Gelman A., J. Carlin, H. Stern, and D. Rubin. Bayesian Data Analysis. Chapman & Hall, 1995.

NRC, Combining information: Statistical issues and opportunities for research (National Acadamy Press, Washington, D.C., 1992).

Important characteristics may be overlooked by focusing in on only a restricted set of conditions (Brand *et. al*, 2000).

The Fowles acute study [P9-17], has limited relevance to chronic or sub-chronic noncancer outcomes and the summary of it's findings should be tempered accordingly.

In addition to the simulation study by Kavlock *et al.* (1996) cited on page 9, line 25 there is another simulation study (Weller, *et al.*, 1995, *Risk Analysis*) that examined study design in the clustered data, developmental toxicity setting with particular emphasis on model fitting and BMD estimation.

Page 10, around line 10 mentions work on multiple outcomes but the literature review does not cite some recent, very relevant work in this area. In particular, papers by Regan and Catalano (1999, *Biometrics*; 1999, *Journal of Agricultural Biological and Environmental Statistics*; 2000, *Risk Analysis*) that investigate multivariate modeling and BMD estimation of joint continuous and binary endpoints might be referenced.

A recent paper by Molenberghs and Ryan (1999, *Environmetrics*) on multiple outcomes in developmental toxicity might be referenced. These references should also be listed on page 34 around line 5 in the Multiple Outcomes section.

Around line 21 also on page 10 and also on page 24, line 4 are some important references to novel approaches for modeling continuous outcomes. Perhaps the RACO method of Bosch *et al.* (1996, *Risk Analysis*) should also be cited and discussed. It has direct relevance and offers an alternative to the methods that are covered in detail in the guidance document.

Vermeire T, Stevenson H, Peiters MN, Rennen M, Slob W, Hakkert BC. 1999. Assessment factors for human health risk assessment: a discussion paper. Crit Rev Toxicol 29(5):439-90.

Cite Allen et al. (1996) in the context of the discussion on p. 15, lines 10-28. (The paper is already cited elsewhere in the text to illustrate a different point.) The discussion deals with how to model responses that are characterized in terms of severity of effect. Allen et al. (1996) analyze a study on the developmental toxicity of boric acid. One of the most notable effects in this study was a change in the rate of a particular skeletal variation (lumbar rib) with increasing dosage, with the highest dosage producing frank malformations of the13th thoracic rib and vertebra. If one assumes that the variation is a less severe manifestation of the same mechanism of action that caused the malformation, then one can assign a particular weighting factor that indicates the severity of the effect that each fetus bears. This provides a single, weighted variable that can be plotted on the same dose-response curve. Allen et al. present models in which one weights the variation as ½ as severe as the malformation, 5/6 as severe (an assumption that the variation is tantamount to a malformation in severity), or 1/6 as severe (an assumption that the severity of the variation is trivial).

Bogdanffy M., Daston G., Faustman E.M., Kimmel C.A., Kimmel G.L., Seed J., and Vu V. 2000. Harmonization of Cancer and Non-Cancer Risk Assessment: Proceedings of a

Consensus-Building Workshop. (This is a manuscript submitted for publication that summarizes the larger cancer/noncancer harmonization meeting that took place in DC in 2000. There are many relevant discussions in this paper.)

Faustman E.M., and Bartell S.M. 1997. Review of Noncancer Risk Assessment: Applications of Benchmark Dose Methods. *Human and Ecological Risk Assessment*: Vol. **3**. No. 5, pp. 893-920.

Foster P.M., and Auton T.R. 1995. Application of benchmark dose risk assessment methodology to developmental toxicity: and industrial view. *Toxicol Lett* **83,** 555-559.

Haag-Gronlund, M., Fransson Steen, R., and Victorin, K. 1995. Application of the benchmark method to risk assessment of trichlorethene. *Regul Toxicol Pharmacol* **21**(2), 261-69.

Kimmel, C.A., Kavlock, R.J., Allen, B.C., and Faustman, E.M. 1995. The application of benchmark dose methodology to data from prenatal developmental toxicity studies. *Toxicol Lett* **82/83**: 549-554.

National Research Council (NRC) 2000. Methods for Developing Spacecraft Water Exposure Guidelines. Chapter 4, and Appendix B. National Academy Press.

Bailer, A.J., Stayner, L.T., Smith, R.J., Kuempel, E.D. and Prince, M.M. (1997). Estimating benchmark concentrations and other non-cancer endpoints in epidemiology studies Risk Analysis 17: 771-780. F

For aquatic/environmental tox, other references might include: Bailer, A.J., and Oris, J.T. (1997) Estimating inhibition concentrations for different response scales using generalized linear models. Environmental Toxicology and Chemistry 16: 1554-1559. And Bailer, A.J. and Oris, J.T. (2000) Defining the baseline for inhibition concentration calculations for hormetic hazards. Journal of Applied Toxicology 20: 121-125.

It would be useful to include some of the epidemiology applications of BMD modeling in the main text, rather than only addressing them in the last appendix. A reference on the use of BMD for epidemiology data to supplement those listed in the appendix is: Bailer, A.J., Stayner, L.T., Smith, R.J., Kuempel, E.D. and Prince, M.M. (1997) Estimating benchmark concentrations and other non-cancer endpoints in epidemiology studies. Risk Analysis 17: 771-780.

The literature review on the development of the BMD approach should include the work on the distributional characterization of the BMD. Some examples include:

— Sielken, R.L. Jr. (1989) "Useful Tools for Evaluating and Presenting More Science in Quantitative Cancer Risk Assessments," Toxic Substances Journal, Vol. 9, 353-404. Hemisphere Publishing Corporation.

— Holland, Charles D. and Robert L. Sielken Jr. (1993) Quantitative Cancer Modeling and Risk Assessment, Prentice Hall, Englewood Cliffs, New Jersey.

— John S. Evans, John D. Graham, George M. Gray, and Robert L. Sielken Jr. (1994) "A Distributional Approach to Characterizing Low-Dose Cancer Risk," Risk Analysis, Vol. 14, No. 1, pp. 25-34.

— John S. Evans, George M. Gray, Robert L. Sielken Jr., Andrew E. Smith, Ciriaco Valdez-Flores, John D. Graham (1994) "Use of Probabilistic Expert Judgment in Distributional Analysis of Carcinogenic Potency," Regulatory Toxicology and Pharmacology, Vol. 20, Number 1, 15-36.

— Evans, John S., John D. Graham, George M. Gray, and Robert L. Sielken Jr. (1995) "A Distributional Approach to Characterizing Low-Dose Cancer Risk," Low-Dose Extrapolation of Cancer Risks, pp. 253-274, ed. by Stephen Olin,William Farland, Colin Park, Lorenz Rhomberg, Robert Scheuplein, Thomas Starr, and James Wilson, ILSI Press, Washington, D.C.

— Sielken, Robert L. Jr., and Ciriaco Valdez-Flores (1996) "Comprehensive Realism's Weight-of-Evidence Based Distributional Dose-Response Characterization," Special Issue of the Journal of Human and Ecological Risk Assessment on: Theoretical, Toxicological and Biostatistical Foundations for Deriving Probability Distribution Functions for Reference Doses and Benchmark Doses with Application to Carcinogens and Noncarcinogens, Vol. 2, No. 1, pp. 175-193.

— Sielken, Robert L. Jr., and Ciriaco Valdez-Flores (1999) "Probabilistic Risk Assessment's Use of Trees and Distributions to Reflect Uncertainty and Variability and to Overcome the Limitations of Default Assumptions," Environment International, Vol. 25, No. 6/7, pp. 755-772.

On page 9, lines 9-16: Other examples of the use of continuous developmental toxicology data in the risk assessment process have been published. Slikker and Gaylor, 1998. Quantitative models of risk assessment for developmental neurotoxicants. *In: Handbook of Developmental Neurotoxicology,* Eds W. Slikker and L. Chang, Academic Press, San Diego, pp. 727-732.

On page 9, line 16, page 11, line 8: An additional reference presenting an example of how continuous data can be used to derive BMDs using the hybrid approach may be added. Slikker, W., Jr., Crump, K.S., Andersen, M.E., and Bellinger, D. Biologically-based, quantitative risk assessment of neurotoxicants. Fund. Appl. Toxicol., 29:18-30, 1996.

The document overlooks some of the original applications of the BMD approach to continuous responses. Some examples are given below. Since I found the examples section to be lacking in this area, I feel that these references should definitely be included.

— West, R. W. and Kodell, R. L. (1993). "Statistical methods of risk assessment for continuous variables," *Communications in Statistics*, 22:3363-3376.

— Kodell, R. L. and West, R. W. (1993). "Upper confidence limits on excess risk for quantitative responses," *Risk Analysis*, 13:177-182.

— In addition to the simulation study by Kavlock *et al.* (1996) cited on page 9, line 25 there is also a simulation study from our group (Weller, *et al.*, 1995, *Risk Analysis*) that examined study design in the clustered data, developmental toxicity setting with particular emphasis on model fitting and BMD estimation. Perhaps this paper should be cited.

Page 10, around line 10 mentions work on multiple outcomes but the literature review does not cite some recent, very relevant work in this area. In particular, papers by Regan and

Catalano (1999, *Biometrics*; 1999, *Journal of Agricultural Biological and Environmental Statistics*; 2000, *Risk Analysis*) that investigate multivariate modeling and BMD estimation of joint continuous and binary endpoints might be referenced. Additionally, a recent paper by Molenberghs and Ryan (1999, *Environmetrics*) on multiple outcomes in developmental toxicity might be referenced. These references should also be listed on page 34 around line 5 in the Multiple Outcomes section.

Around line 21 also on page 10 and also on page 24, line 4 are some important references to novel approaches for modeling continuous outcomes. Perhaps the RACO method of Bosch *et al.* (1996, *Risk Analysis*) should also be cited and discussed. It has direct relevance and offers an alternative to the methods that are covered in detail in the guidance document.

A published example is the BMD analysis of the "negative" Seychelles study of methylmercury developmental neurotoxicity cited in the last example in the BMD Technical Guidance Document (Crump et al. 2000). A similar approach has also been published for BMD analysis of neurological endpoints in a "negative" study of occupational manganese exposures: Gibbs, JP, Crump, KS, Houck, DP, Warren, PA, and Mosley, WS. 1999. Focused Medical Surveillance: A Search for Subclinical Movement Disorders in a Cohort of U.S. Workers Exposed to Low Levels of Manganese Dust," NeuroToxicology 20, 299-313.

In Background add reference here to other documents evaluating cancer versus noncancer issues. See especially workshop report for recent harmonization workshop jointly sponsored by Society of Toxicology (SOT) and the U. S. Environmental Protection Agency (EPA) (Bogdanaffy et. al., 2000).

Clewell HJ III, Gentry PR, Gearhart JM. 1997. Investigation of the potential impact of benchmark dose and pharmacokinetic modeling in noncancer risk assessment. J Toxicol Environ Health 52:475-515.

Barton, H. A. and Clewell, H. J., III. 2000. Evaluating Noncancer Effects of Trichloroethylene: Dosimetry, Mode of Action, and Risk Assessment. Environmental Health Perspectives, 108(suppl 2):323-334.

Clewell, H.J., Andersen, M.E., and Barton, H.A. 2001. A consistent approach for the application of pharmacokinetic modeling in cancer and noncancer risk assessment. Environmental Health Perspectives (submitted).

Andersen ME, Clewell HJ, Krishnan K. 1995. Tissue dosimetry, pharmacokinetic modeling, and interspecies scaling factors. *Risk Anal* 15:533-537.

Clewell HJ, Andersen ME. 1985. Risk assessment extrapolations and physiological modeling. *Toxicol Ind Health* 1:111-131.

H. V. Huikuri, T. Makikallio, J. Airaksinen, R. Mitrani, A. Castellanos, and R. J. Myerburg. 1999. Measurement of Heart Rate Variability: A Clinical Tool or a Research Toy? J Am Coll Cardiol 34, 1878-1883.

H. V. Huikuri, T. H. Makikallio, K. E. J. Airaksinen, T. Seppanen, P. Puukka, I. J. Haiha, and L. B. Sourander. 1998. Power-Law Relationship of Heart Rate Variability as a Predictor of Mortality in the Elderly. Circulation 97, 2031-2036.

Finney, DJ. 1971. Probit analysis, 3rd edition. Cambridge University Press, London.

Hattis, D., Banati, P., and Goble, R. 1999b. Distributions of Individual Susceptibility Among Humans for Toxic Effects—For What Fraction of Which Kinds of Chemicals and Effects Does the Traditional 10-Fold Factor Provide How Much Protection? Annals of the New York Academy of Sciences, Volume 895, pp. 286-316.

Hattis, D., Banati, P., Goble, R., and Burmaster, D. 1999. Human Interindividual Variability in Parameters Related to Health Risks. Risk Analysis, Vol. 19, pp. 705-720.

REFERENCES RELATED TO BOOTSTRAP METHODS

BASIC TEXT

Efron, Bradley and Robert J. Tibshirani (1993) An Introduction to the Bootstrap, Chapman & Hall, New York. (Chapters 12 to 14 discuss Confidence Intervals.)

LITERATURE REFERENCES

BMD-Specific

Zeng, Q. and Davidian, M. (1997) Bootstrap-adjusted calibration confidence intervals for immunoassay. Janual of American Statistical Association, Vol 92, 278-290.

General

Efron, Bradley (1985) "Bootstrap confidence intervals for a class of parametric problems," Biometrika 72, 45-58.

Efron, Bradley (1987) "Better Bootstrap Confidence Intervals," Journal of the American Statistical Association, Vol. 82, No. 397, 171-185.

DiCiccio, Thomas and Robert Tibshirani (1987) "Bootstrap Confidence Intervals and Bootstrap Approximations," Journal of the American Statistical Association, Vol. 82, No. 397, p. 163

Schenker, Nathaniel (1985) "Qualms About Bootstrap Confidence Intervals," Journal of the American Statistical Association, Vol. 80, No. 390

Tibshirani, Robert (1988) "Variance stabilization and the bootstrap," Biometrika, 75, 3, 433-44

Hall, Peter (1987) "On the bootstrap and likelihood-based confidence regions," Biometrika, 74, 3, 481-93

Hall, Peter (1992) "On bootstrap confidence intervals in nonparametric regression," The Annals of Statistics, 20, 2, 695-711

Hall, Peter (1988) "Theoretical comparison of bootstrap confidence intervals," The Annals of Statistics, 16, 3, 927-953

Leger, Christian and Robert Cleroux (1992) "Nonparametric Age Replacement: Boostrap Confidence Intervals for the Optimal Cost," Operations Research, 40, 6, 1062

Comparisons of Likelihood-Based and Bootstrap Confidence Regions

Owen, Art B. (1988) "Empirical likelihood ratio confidence intervals for a single functional," Biometrika 75, 2, 237-49

Monte Carlo-Based Confidence Intervals

Buckland, S. T. (1984) "Monte Carlo Confidence Intervals," Biometrics, 40, 811-817.