United States
Environmental Protection
Agency

Office of Research and
Development
Washington DC 20460

EPA/630/R-96/011
November 1996

EPA

Report on the Benchmark Dose Peer Consultation Workshop

RISK ASSESSMENT FORUM

EPA/630/R-96/011
November 1996

REPORT ON THE

BENCHMARK DOSE PEER CONSULTATION WORKSHOP

Prepared by:

Eastern Research Group, Inc.
110 Hartwell Avenue
Lexington, MA 02173
EPA Contract No. 68-D5-0028

Risk Assessment Forum
U.S. Environmental Protection Agency
Washington, DC

NOTICE

Mention of trade names or commercial products does not constitute endorsement or recommendation for use. Statements are the individual views of each workshop participant; none of the statements in this report 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. (ERG), an EPA contractor, as a general record of discussions during the Benchmark Dose Peer Consultation Workshop. As requested by EPA, this report captures the main points and highlights of discussions and includes brief summaries of discussion topic sessions. The report is not a complete record of all details discussed, nor does it embellish, interpret, or enlarge upon matters that were incomplete or unclear. In particular, each of the five discussion topic summaries was prepared at the workshop by individual discussion topic leaders based on the panel members' discussions during the workshop. Thus, there may be slight differences between the five topic leaders' summaries. ERG did not attempt to harmonize the chairs' comments.

CONTENTS

FOREWORD

This report includes information and materials from a peer consultation workshop organized by the U.S. Environmental Protection Agency's (EPA's) Risk Assessment Forum (RAF). The meeting was held in Bethesda, Maryland, at the Holiday Inn Bethesda on September 10-11, 1996. The subject of the peer consultation was the document entitled *Benchmark Dose Technical Guidance Document* (External Review Draft, EPA/600/P-96/002A). A copy of this report can be obtained through the Office of Research and Development's publications office, Technology Transfer and Support Division, National Risk Management Research Laboratory, U.S. EPA, 26 West Martin Luther King Drive, Cincinnati, Ohio 45268 (telephone: 513-569-7562; fax: 513-569-7566). The expert panel was convened to independently comment on the draft guidance document and make recommendations that will enhance the guidance development process as well as the ultimate product.

Notice of the workshop was published in the *Federal Register* on August 28, 1996 (61 FR 44308). The notice invited members of the public to attend the workshop as observers and provided logistical information to enable observers to preregister. About 40 observers attended the workshop, including representatives from federal government, industry, trade organizations, and consulting firms.

In outlining the scope of the peer consultation, EPA emphasized that the draft guidance document is in a preliminary stage of development and should not be construed as a policy statement. EPA explained that the guidance is intended to be used in conjunction with other Agency risk assessment guidance and to harmonize the methods used to conduct cancer and noncancer quantitative risk assessments. EPA explained further that the draft guidance document is still in a preliminary stage and therefore could benefit greatly from the comments and recommendations of outside experts. EPA asked the expert peer consultants to concentrate their review on technical issues concerning selection of studies and responses for benchmark dose/concentration (BMD/C) analysis; selection of the benchmark response level (BMR); model selection and fitting; use of confidence limits; and selection of the BMD/C to use as the point of departure for cancer and noncancer health effects.

A balanced group of expert panel members were selected from academia, industry, consulting, government, and environmental organizations. Selected panel members provided broad experience and demonstrated scientific expertise in risk assessment. Experts represented the following disciplines: toxicology, biostatistics, risk assessment/risk management policy, and mathematics. Appendix A lists the 18 panel members.

In workshop discussions, EPA sought comments from these scientific experts on the draft guidance document. The draft guidance document presents a procedure that is intended to have reasonable criteria and defaults to assist risk assessors in promoting consistency among analyses of health effects data in the observable range. The procedure is also intended to be useful for determining the point of departure that can be used as the basis for linear low dose extrapolation for cancer, calculation of a margin of error, or application of uncertainty factors for calculating oral reference doses (RfDs), inhalation reference concentrations (RfCs), or other exposure estimates for human health risk assessment. EPA will use the expert panel members' comments and

recommendations drawn from this peer consultation workshop in considering revisions to the draft guidance document.

The workshop report is organized as follows. The report opens with a brief introduction that covers the background of the benchmark dose guidance document, presentations on two ongoing Agency initiatives on the benchmark dose approach, and the purpose of the workshop (section 1). This is followed by the chairperson's summary (section 2) and then the five discussion topic leaders' summaries (section 3). The last section of the report provides observers' comments (section 4). Appendices to the workshop report include a list of panel members, the workshop agenda, the charge to workshop panel members, premeeting comments, and a list of observers.

William Wood, Ph.D.
Executive Director
Risk Assessment Forum

SECTION ONE

INTRODUCTION

This report highlights issues and conclusions from an EPA Risk Assessment Forum-sponsored workshop on the Agency's *Benchmark Dose Technical Guidance Document* (External Review Draft, EPA/600/P-96/002A) published August 9, 1996 (61 FR 44308). The workshop was convened to gather information from scientific experts that will assist EPA in further developing the draft guidance document.

BACKGROUND

EPA has followed distinct practices for evaluating the dose-response relationships of cancer and noncancer-causing agents. The linearized multistage procedure has been applied to extrapolate risk as the 95-percent upper confidence limit for cancer, and the lowest-observed-adverse-effect-level (LOAEL) and the no-observed-adverse-effect-level (NOAEL) approaches have been used to conduct dose-response analyses of noncancer health effects. In 1996, EPA published *Proposed Guidelines for Carcinogen Risk Assessment* (61 FR 17960-18011), which present an approach that will begin to break down the dichotomy between quantitative approaches for cancer and noncancer risks. The proposed cancer risk assessment guidelines emphasize an agent's mode of action in producing tumors and the need to model tumor data as well as other biological responses that might be important in the carcinogenic process. The models can be used to estimate a point of departure for extrapolation below the range of observable effects. The benchmark dose approach is one way of determining the point of departure for linear low dose extrapolation of carcinogens, calculation of a margin of exposure (MOE), or application of uncertainty factors for calculating oral reference doses (RfDs), inhalation reference concentrations (RfCs), or other exposure estimates.

Following a 1990 colloquium recommendation, the Risk Assessment Forum took an active role in promoting research and discussion on benchmark dose issues. A draft report was prepared

that outlined the technique and presented the major questions and decisions involved in applying the benchmark dose method. This draft report was the subject of a 1993 Forum-sponsored colloquium on applications of benchmark dose methods to noncancer risk assessment. Following the colloquium, a Risk Assessment Forum technical panel published a background document on the use of the benchmark dose/concentration (BMD/C) in health risk assessment (EPA/630/R-94/007). In addition, several workshops and symposia have been held to discuss the benchmark dose approach. Subsequent to the development of the background document, a Forum technical panel authored the external review draft guidance document that served as the focus of the August 1996 peer consultation workshop.

In her introductory remarks at the gathering, Carole Kimmel, Ph.D., of EPA's National Center for Environmental Assessment (NCEA), who is a member and chair of the Risk Assessment Forum's technical panel on benchmark dose, explained that because of the Forum's involvement, the draft guidance document is the result of an Agency-wide effort supported by the different EPA offices represented on the technical panel. Dr. Kimmel announced that the Risk Assessment Forum was in the process of developing a framework for health risk assessment that will harmonize approaches for both cancer and noncancer effects. Mode of action data and precursor information are being incorporated into the approach, thereby bringing the issues of the underlying basis for the toxicity of cancer and noncancer health effects closer together.

Dr. Kimmel went on to explain that the benchmark dose document is a working draft that has undergone one round of internal review. Several issues still remain, and EPA felt it was appropriate at this stage in the development of the guidance to solicit comments and input from outside experts on applying the benchmark dose approach to cancer and noncancer risk assessments. Following this workshop and discussions within the Agency, EPA will revise the document, conduct a peer review, and then publish a document under the auspices of the Risk Assessment Forum. The final document will be used in conjunction with other EPA risk assessment guidance.

Rogene Henderson, Ph.D., a senior scientist at the Inhalation Toxicology Research Institute, served as the chairperson of the workshop. In her introductory remarks, Dr. Henderson reviewed the agenda for the workshop (see Appendix B) and the charge to workshop panel members (Appendix C). Dr. Henderson explained that EPA's goals for the guidance document are to have

a procedure that is usable; has reasonable criteria and defaults; can be used for cancer and noncancer assessments when endpoints are relevant to both; and informs our understanding of risk in the range of extrapolation. She then discussed the limitations of using the LOAEL/NOAEL approach, including:

- levels are dependent on study design (e.g., choice of doses, numbers of animals);
- the variability in the data are not taken into account;
- the slope of the dose-response curve is not taken into account; and
- an uncertainty factor is used to connect a LOAEL to a NOAEL.

In contrast, the benchmark dose approach, as an alternative to the LOAEL/NOAEL approach, makes better use of the available data, including taking into account the slope of the dose-response curve and the variability of the data. The BMD/C is defined as the lower confidence limit for the dose that is estimated to produce a given level of change in response (i.e., the benchmark response [BMR]). BMD/C estimates are best when there are doses in the study near the range of the BMD/C; but the BMD/C does not have to be one of the experimental doses.

To help focus the groups' efforts on addressing the charge to workshop panel members, Dr. Henderson reviewed the purpose and goals of the workshop. She reminded panel members that the objective was not to reach consensus on issues, but to identify and elucidate issues relevant to the draft guidance document.

PRESENTATIONS

Prior to discussions by panel members, EPA scientists who were among the co-authors of the draft guidance document presented information to workshop participants on two Agency-sponsored initiatives to support the development of guidance on the benchmark dose approach.

Discussion of Simulation Studies

Woodrow Setzer, National Health and Environmental Effects Research Laboratory

Dr. Setzer presented the preliminary results of simulation studies that are being conducted to determine the usefulness of the limit of detection (LOD) approach for setting the BMR. Dr. Setzer began his presentation by expressing the opinion that the motivation for adopting the BMD approach was not principally dissatisfaction with existing approaches. Dr. Setzer indicated that the assumption is to use BMDs as plug-in replacements for NOAELs, with little or no change in the structure of uncertainty factors. The recommended BMR of ED_{05} or ED_{10} (effective dose) is likely, however, to result in substantially lower (sometimes higher) RfD/RfCs than the NOAEL approach.

The LOD is a methodology for specifying a BMR in such a way that, hopefully, the overall conservatism of noncancer risk assessments would be similar to what is obtained when using NOAELs (though not necessarily for individual dose-response assessments). The definition of LOD is the magnitude of response just detectable in a two-group design (control and one treatment group) using a one-sided test with a Type I error of 0.05 and a predetermined power. The draft guidance document proposes that, in the absence of determining a "biologically significant" response, the BMR should be set as the LOD of a typical "good" design for the species and endpoint considered. For example, designs recommended in various testing guidelines would be considered good designs. The goal of the LOD methodology is to have a well-designed bioassay where the resulting BMD provides the same level of conservatism as the NOAEL.

Dr. Setzer emphasized that the simulation study is a pilot, and that it is currently incomplete. The results and analyses presented, therefore, must be considered to be preliminary and subject to update. The goals of the simulation study are to:

- determine whether using an approach based on power simplifies the specification of a BMR with respect to maintaining the current level of conservatism in the RfD/C;
- estimate the power to be used in the LOD to maintain the current level of conservatism; and
- explore the behavior of the BMD relative to NOAELs.

The structure of the simulations include assembling a collection of quantal and continuous dose-responses on the dose range 0 to 100. In this study, four distinct quantal shapes and four distinct continuous shapes are used. Each quantal shape is considered in conjunction with background incidences of either 0.05 or 0.15. Each continuous dose-response shape is considered in conjunction with a coefficient variation of either 15 or 30 percent. This makes a total of eight quantal models and eight continuous models (a small sample of possible dose responses). Other components of the simulation structure include:

- Consider experimental designs with either 10 or 20 animals per dose group and either three or four total doses (i.e., four different designs).
- For each of the 32 model x design combinations for each kind of endpoint (quantal and continuous), 100 random data sets using binomial random numbers for quantal endpoints and lognormal numbers for continuous endpoints are generated.
- For each quantal data set, fit log-logistic and Weibull models. For each continuous data set, fit linear, quadratic, and power models. Also, assess the effect of threshold. Reject badly fitting models using the chi-squared goodness of fit test and select among the rest of the models by taking the model with the lowest Akaike Information Criterion (AIC), a measure of the deviance of the model fit adjusted for the degrees of freedom.
- Calculate BMRs given sample size and background (quantal) or coefficient of variation (continuous) for powers of 0.10, 0.25, 0.50, 0.75, and 0.90.
- Calculate the NOAEL using the NOSTASOT (no-statistical-significance-of-trend) approach.

Dr. Setzer reviewed the preliminary results of the pilot simulation study for quantal data only. For each simulated quantal data set, the BMD was compared to the NOAEL. The distribution of the median BMD:NOAEL ratio among the nine dose responses at each power level for each of the designs showed that the ratio increased with increasing power levels. The results indicate that the BMD is more stable than the NOAEL for the dose responses studied in the simulation.

Development of Software for Benchmark Dose/Concentration Analysis

Daniel Guth, National Center for Environmental Assessment

Dr. Guth discussed EPA's work on developing software for BMD/C analysis. Due to the limited choices in commercial software, the inflexibility of available software, and the need for consistent methods and model outputs, the Risk Assessment Forum technical panel identified the need for software to accompany the proposed guidance as a priority. The intended audience of the software is toxicologists, risk assessors, and statisticians.

Dr. Guth described the following software design criteria:

- Freely distributable—Does not require license fees or other software.
- User-friendly—GUI-based and only allows models appropriate to data type.
- Accessible—Windows and Macintosh platforms; able to run on a 486 or better machine; and includes on-line help with explanations of models and parameters.
- Flexible—Standard versus advanced user modes; data entry (direct or import spreadsheet); multiple models selectable; various data types allowed; graphical outputs; and batch operation available.
- Does not set policy—BMR entered by user; parameters are unconstrained; more models available than needed; and exceptions (i.e., method for calculating confidence intervals and exclusion of "threshold" parameter).

The outline for the software capabilities includes:

- Input data—Import ASCII or spreadsheet files; enter data from the screen; modify data structure (add, change, or delete variables); create a new data set as a subset of an existing file; and generate random data for simulation.
- Data file management—Sort data; change or add data records; and transform or compute a variable.
- Data analysis—Select data set (dependent and independent variables); select from available models; save or execute requested analysis; and calculate BMD/C.

- Models available—Dichotomous data (Probit, Weibull, Logistic, Gamma Multi-Hit, Quantal Linear, Quantal Quadratic, Quantal Polynomial [multistage]); nested dichotomous data (Logistic, Rai and van Ryzin, National Center for Toxicological Research [NCTR]); and continuous data (Linear, Polynomial, Power).

- Advanced mode—Specify parameter values; place constraints on parameter values; specify model fitting options; and generate simulated data from specified model, parameters.

- Output—Parameter estimates; statistical report (goodness-of-fit measures and diagnostics); and graphical displays (maximum likelihood estimate [MLE], confidence interval, points).

PEER CONSULTATION WORKSHOP

To involve outside scientific experts in development of the draft guidance document, EPA's Risk Assessment Forum sponsored a two-day workshop, which was held on August 10-11, 1996, at the Holiday Inn in Bethesda, Maryland. The meeting gathered 18 experts (see Appendix A for a list of workshop peer consultants/panel members) with the objectives of describing points of view about issues outlined in the charge to workshop panel members (Appendix C), identifying and elucidating other issues, and highlighting areas for further development.

Prior to the workshop, EPA provided each expert with a copy of the external review draft *Benchmark Dose Technical Guidance Document*. EPA asked workshop participants to review these materials and respond to the following issues:

- the appropriate selection of studies and responses for BMD/C analysis;

- the use of biological significance or limit of detection for selection of the BMR;

- model selection and fitting;

- the use of the lower confidence limit as the BMD/C; and

- selection of the BMD/C to use as the point of departure for cancer and noncancer health effects.

These comments were assembled and sent to all panel members prior to the workshop. See Appendix D for the workshop panel members' premeeting comments.

SECTION TWO

CHAIRPERSON'S SUMMARY OF THE WORKSHOP

Rogene Henderson, Chair
Inhalation Toxicology Research Institute
Albuquerque, NM

The major purpose of the workshop was to solicit the views of experts on the draft of EPA's *Benchmark Dose Technical Guidance Document*. For this preliminary draft, input was sought on key issues concerning the technicalities involved in use of the BMD/C approach so that EPA can appropriately revise the guidance. The meeting was attended by the panel members, several co-authors of the draft document, and public observers (see list of public observers in Appendix E).

The workshop was structured around the premeeting comments solicited from the panel members. As background, however, two of the document co-authors gave informational presentations about topics related to the guidance. Then the discussion leaders presented summaries of the premeeting comments on each of five major issues regarding the calculation of the BMD/C. This was followed by a general discussion of each issue by the panel. Authors of the different sections of the document provided clarification of points as required. Observers were given two opportunities to provide their comments during the meeting.

Informational Presentations

Dr. Woodrow Setzer of EPA presented information about the results of simulation studies that are under way to determine the usefulness of the LOD method for setting the BMR. The panel then discussed issues raised in the presentation. Panelists were in general agreement that biological significance should be the primary factor in setting the BMR and not the LOD. The panel also supported an approach in which biological significance is the first factor to be considered followed by a test of statistical significance.

The simulation studies of Dr. Setzer, which were considered to be well done, indicated that 50-percent power yielded results closest to that of the NOAEL. Based on this information, the panel discussed whether the NOAEL should be considered the "gold standard" for the BMD/C approach or whether the two should be considered separately. Some panel members contended that there is no need to change to the BMD/C method if concurrence with the NOAEL is the validity test for the BMD/C numbers. One might just as well use the NOAEL to start with. Others held that some comparisons of the BMD/C numbers with earlier NOAEL numbers is necessary to determine if the new method is in the "ballpark" of numbers that had previously been considered to be protective of human health. Some panel members strongly disagreed with Dr. Setzer's statement that "The mandate is to use BMDs as plug-in replacements for NOAELs, with little or no change in the structure of uncertainty factors."

The second presentation was given by Dr. Daniel Guth of EPA on a software package that the Agency is developing for calculation of the BMD/C. In the panel discussion that followed, some members expressed concern that the software might restrict some investigators from developing their own software. In this context, the panel discussed the merits of prescriptive versus nonprescriptive approaches to guidance on calculating the BMD/C. The workloads of many people who are doing such calculations on hundreds of new compounds may prevent them from using anything but a standardized, prescriptive approach to making the calculations.

Discussions of Major Issues

Each of the five major issues identified in the charge to the panel were discussed at length. Details of these discussions are summarized in the reports of the individual discussion leaders (see section 3). Regarding these major issues, the panel members reached consensus on only one point: Biological significance should be the basis for the choice of a BMR rather than the LOD approach as proposed in the document. On whether the central estimate of the BMD or the lower 95-percent confidence value should be used for further calculations of risk, the panel engaged in a lengthy discussion. A majority felt that the central estimate should be used, for reasons stated in the premeeting comments. No consensus was reached on this recommendation, however.

In the discussion about the technical points involved in calculating a BMD/C, panel members were generally in agreement that whatever model was chosen, the model and the data should be graphed to help in determining goodness of fit. Also, panelists were in general agreement that background responses should be included in the models; opinions were divided, however, on assuming a threshold for a model. Dichotomizing continuous data was not considered the best approach by many panelists. One panel member described an approach in which continuous data could be used to calculate a BMD/C without dichotomization, and this approach was well received by the panel. Another panelist offered the aid of the American Industry Health Council (AIHC) in helping EPA with some of these difficult technical issues.

General Comments on the Document as a Whole

The following general points summarize the panel's discussion about the document as a whole:

- The panelists generally agreed that the guidance needs to be a "stand-alone" document that can be understood by itself. In the main text, many references were made to other EPA documents or to the appendices. Panelists suggested that the document would be easier to use if excerpts from the cited documents were inserted in the appropriate places and portions of text in the appendices were moved to the main text.

- A panelist suggested that a common nomenclature (rather than ED_x, TD_x, and BMD_x) should be adopted for noncancer and cancer endpoints when the BMD/C approach is used. The panel expressed a general concern that the draft document represents "statistical overkill" because the statistical approaches were much more elegant than the relatively coarse data that one often has to work with. A related concern was that the methods proposed in the document should be transparent. A panelist suggested that the document should be reviewed by a group of risk managers to determine if the method is reasonably transparent to the group that must use the results of the calculations. Also, the value of calculating a range of values rather than a single point estimate was mentioned.

- Panelists discussed use of toxicokinetic data to improve dose estimates and consideration of patterns of responses as well as single endpoints. Authors of the report pointed out that for any calculation of risk, the best scientific information available should be used. The issue of using the best data for dose and for response was not considered unique to the problem of calculating the BMD/C.

- The panel expressed general concern about how the Agency might implement the benchmark analyses in the regulatory arena. How would a benchmark dose be used in a margin of exposure approach? What default options would be used? What uncertainty factors would be used? The panel recommended that EPA explicitly state that the process is an iterative one requiring sound scientific judgment.

- Some panelists found the process described in the draft document to be too prescriptive, while others pointed out that the people conducting risk assessments on multiple compounds often only have time to follow a prescriptive approach.

General Issues Other Than the Five Main Issues

Additional issues arose during the course of the workshop and were addressed during sessions on general considerations. One of these was whether the same uncertainty factors should be used with the BMD/C numbers and with the NOAEL/LOAEL numbers. Because no LOAEL-to-NOAEL conversion is involved in the benchmark approach, the uncertainty factor of 10 that is normally used for this conversion was considered inappropriate for the benchmark approach. The BMD/C, however, is associated with a stated level of response, such as the ED_{10}. Thus, some participants recommended using a factor of 10 to go from the risk associated with ED_{10} to a lower risk. Other uncertainty factors, such as a factor for animal-to-human extrapolation, for human variability, and for differences in study duration would apply to the BMD/C as well as to the NOAEL/LOAEL approach. Some panelists contended that the BMD_{10} should be considered a LOAEL, but no consensus was reached on this point.

The panel briefly discussed the issue of whether both cancer and noncancer adverse health effects, as well as both acute and chronic noncancer effects, could be analyzed by the BMD/C approach. Panelists found no real impediment to using the general approach in all these cases, although some specifics of the analyses might differ for each type of health effect.

Panel members also discussed the objective in calculating a benchmark dose. Is it for comparison across endpoints? to match NOAEL values? to retain the same level of conservatism as a NOAEL? to avoid using NOAELs with high levels of risk?

As had been revealed during Dr. Setzer's presentation, the panel was of two opinions regarding the attempt to match the BMDs to previously determined NOAELs. Some panelists contended that such a course is necessary to determine if the level of conservatism of BMDs is similar to that for the NOAELs, which previously had been thought to protect human health. Others held that such an exercise is not necessary because the BMDs, which were developed to make better use of all available scientific data in completing risk assessments, should be more valid than the NOAELs. The panel did not reach consensus on this issue. The panelists did generally agree, however, that BMDs should be valuable for comparison across endpoints.

The panel also discussed whether EPA should continue to move forward in developing a benchmark approach. The general consensus was that the Agency should continue its work in this area; there was agreement that the analysis of quantal data by this approach is further along than the analysis of continuous data. In continuous data, one must be concerned with the severity of the response. For analysis of continuous data, experts in the field of each type of endpoint measured by such data would need to be gathered to determine what degree of change in an endpoint is considered to be biologically significant as an adverse health effect. Such decisions in the many fields of study concerning noncancer endpoints will involve a considerable investment of time and money by the Agency.

SECTION THREE

DISCUSSION TOPIC SUMMARIES

Selection of Studies and Responses for Benchmark Dose/Concentration Analysis

James Olson, Discussion Leader
Department of Pharmacology and Toxicology
State University of New York at Buffalo
Buffalo, NY

General Comments

The Introduction section of the document clearly presents the limitations of the current risk assessment procedures that utilize LOAELs and NOAELs. While this presents a good background on this issue, it would be helpful if the beginning of the document also presented a clear definition of the BMD/C, the perceived benefits of this approach, and a brief discussion of how the EPA plans to utilize the BMD/C for cancer and noncancer risk assessment. Page 17 may be too far into the document to present a clear definition of BMD/C.

The issue of selection of studies and responses is a critical first step in the process of establishing a BMD/C. The document states that selection of the appropriate studies and endpoints is discussed in Appendix A and in various EPA publications (U.S. EPA, 1991a, 1994c, 1995f, 1996a and b). The panel members suggest that the clarity of the document would be greatly improved if the *Benchmark Dose Technical Guidance Document* could be a stand-alone document. Citing other documents and references is useful for identifying additional information, but whenever possible the present document should contain all necessary key information relevant to developing BMD/C estimates. For example, if only high quality, peer reviewed studies are to be considered for evaluation, this needs to be stated directly in the document. If human studies are given more weight than animal studies, this also needs to be clearly stated. It is understood that the document cannot discuss in detail all aspects of risk assessment that enter into the process of deriving a BMD/C. However, it would be helpful for the document to acknowledge that it is the intention of the Agency

to address issues, such as exposure assessment, that are key to the process of risk assessment. Pharmacokinetic considerations, including physiologically based pharmacokinetic (PBPK) models, tissue dosimetry, body burden, and equivalent human dose, need to be identified in the document as key issues in the process of selecting studies and endpoints for developing BMD/Cs.

Issues Related to Selection of Studies

The document clearly states that the first step in the process is a complete qualitative review of the literature to identify and characterize the hazards related to a particular compound or exposure situation (p. 18, lines 11 and 12). The document goes on to state that "the selection of the appropriate studies is based on the human exposure situation that is being addressed, the quality of the studies, and the relevance and reporting adequacy of the endpoints." Again, it needs to be stated that only high quality, peer reviewed studies will be evaluated. It would also be helpful to include material from Appendix A in the body of the document. The document states on p.18, lines 21-23, that "the process of selecting studies for benchmark analysis is intended to identify those studies for which modeling is feasible, so that BMD/Cs can be calculated and used in risk assessment." Several panelists commented that all studies should be evaluated, without consideration as to suitability for modeling. A number of comments were made regarding the minimum data set for calculating a BMD/C (see 3 bullets at bottom of p. 19):

- The statement on p. 19, line 19, that "at minimum, the number of dose groups and subjects should be sufficient to allow determination of a LOAEL," is not clear; the statement implies that one should not model data sets for which a LOAEL was not actually observed. This criterion also makes a precise definition of LOAEL essential.
- There was some disagreement with the statement that "with only one responding group, there is inadequate information" (p. 19, line 23). This statement needs further justification.
- The existence of only high level response data (criterion on p. 19, line 25) should not necessarily preclude modeling.

- Dose-response modeling should be conducted only when there is evidence of the shape of the dose-response relationship. The presence of evidence suggesting the existence and general shape of dose-response relationships allows for fitting dose-response models to the data sets under consideration. Perhaps consider using a trend test on the doses that exhibit a response as a way to determine if there exists sufficient information about a dose-response relationship for modeling.

There was some concern regarding the requirement that preferred studies should always contain dose-response data in the range of the BMD/C (see Appendix A in the document). To make it a requirement that there be an experimental dose that gives a response about equal to the BMR would be too restrictive. Data used in BMD/C calculations for acute toxicity (noncancer) studies require some flexibility. Data are often comprised of small group size (five or six animals per group) so that observing responses in the range of the BMR will not be possible for many of these studies.

The reader should be cautioned regarding the statement on p. 18, lines 21-23, that for some chemicals, use of a study that provides a NOAEL from a quality study for a relevant, sensitive endpoint is preferable to a BMD (which may be higher) from a study where it can be calculated.

Tree analyses may be a useful tool in organizing, presenting, and communicating BMDs for each of the relevant studies and endpoints. A plausibility distribution could also be used to reflect the relative likelihood of these calculations being relevant to humans.

Issues Related to Selection of Responses

Selection of the appropriate endpoints for the BMD/C analysis is the next important consideration. The document (p. 19, line 9) was somewhat vague, stating that the endpoints to model should focus on endpoints that are relevant or assumed relevant to humans and potentially the "critical" effect (i.e., the most sensitive). The document indicates that multiple endpoints can be modeled, but are there sensitive responses (biological vs. toxic) that are not appropriate to model? It might be helpful to discuss the use of specific endpoints, such as an increase in liver weight, increase in hepatic cytochrome P450 protein levels, etc., with regard to their suitability for deriving BMD/C estimates. Perhaps some discussion of biomarkers of exposure/effect would be

helpful in the discussion regarding the selection of appropriate endpoints. The focus on endpoints that are relevant to humans and the most sensitive effect is where extensive toxicological knowledge is required on the part of the risk assessor. Considerable discussion and, hopefully, growing consensus is needed on the identification of relevant endpoints. For example, the BMD approach may not be well suited for neurotoxicity data sets. Further work is needed to address how data sets, which are often unique to a specific endpoint, will be evaluated for modeling.

Endpoint selection should be based on the relevance of the endpoints and quality of the study (good experimental protocol), without regard to the ability to derive BMD/C estimates. The goodness of fit of the data ("smoothly increasing response") should not be a major factor in endpoint selection. It will also be necessary to reduce the number of endpoints that need to be considered in some cases (eliminate redundancy; consider issues such as representativeness and sensitivity).

Have there been any studies or work done to support the claim that having LOAELs differing by a factor of 10 (p. 19, line 13) will ensure that the "critical" BMD/C will not be missed? It appears inappropriate to make the statement that all endpoints whose LOAELs are within an order of magnitude of the lowest LOAEL should be modeled (the critical effect will be selected as simply the lowest BMD).

The BMD/C should be only one of several tools available to the risk assessor. If the data on the endpoint from the best quality study are not conducive to model fitting but provide an adequate point of departure (i.e., an appropriate NOAEL) for extrapolation to derive an acceptable exposure level, then there is no need to use alternative endpoints to determine a BMD/C.

The EPA should consider emphasizing "severity" of impact (in contrast with sensitivity) as the principal consideration. It would be ideal if BMD starting points for different compounds could be selected to be roughly comparable in terms of potential impact on human health. One alternative would involve scoring observable endpoints in terms of their severity and attempting to select the critical endpoint in terms of biological significance.

Issues Related to Selection of Studies and Endpoints for Cancer and Noncancer BMD/C Analysis

BMD/C analyses are intended to be used for a wide range of experimental data sets. Each toxicological discipline has somewhat unique experimental protocols, generating data sets that vary with regard to route of exposure, magnitude of dose (daily and cumulative), duration of the exposure and study, type of data generated during and at completion of the study (dichotomous, continuous, categorical), variability in the data, and potential health significance of the data collected. In general, the panelists were supportive of attempting to model both cancer and noncancer data with the goal of deriving BMD/Cs. However, the document needs to more directly address the inherent differences in these study designs. Little attention was given to the issues of duration of exposure and cumulative versus daily dose for noncancer endpoints. If possible, the document should attempt to address these issues that relate to the use of cancer and noncancer endpoints in developing BMD/Cs. Although separate sections of the document discuss the application of BMD/Cs for cancer and noncancer risk assessment, it might be useful to have a separate section in the first part of the document that addresses the special issues of deriving BMD/Cs from noncancer and cancer studies.

Issues Related to Combining Data Sets

The panel members considered this to be an important issue that requires more clarification. For example, more guidance is needed as to what constitutes biological and statistical compatibility. The suggestion was made to include reference to the peer-reviewed publication by Allen et al.[1] One example of an approach for determining the appropriateness of combining data sets for analysis is given in attachment B of Dr. Fowles premeeting comments (see Appendix E of this report).

[1] Allen, B.C., Strong, P.L., Price, C.J., Hubbard, S.A., and Datton, G.P. 1996. Benchmark Dose Analysis of Developmental Toxicity in Rats Exposed to Boric Acid. Fundam Appl Toxicol 32:194-204.

Possible criteria for determining when studies could be combined are listed with the premeeting comments provided by Dr. Naumann. They are:

- statistical evidence that the study attributes are not different (e.g., population variance, group mean response at similar dose levels);
- similarities in conducting studies (e.g., species, strain, group size, protocol, laboratory);
- similarities in endpoints and data reporting (e.g., individual values vs. summary statistics);
- congruence in modeling results between individual and combined data sets (i.e., does the combined model yield similar values for goodness-of-fit, MLE, and lower bound on dose at the BMR level); and
- ability to clearly state the rationale for combining studies.

The combining and weighing of different endpoints within studies, as with the boron example (#3) in Appendix D of the guidance document, should be approached with caution because, despite the use of expert judgment, it is still subjective. There is a need to avoid any appearance of manipulating the data and to be able to explain the rationale for weighing endpoints. Transparency is very important to avoid the "black box" aspect that is inherent to the BMD approach (and mathematical modeling in general).

Selection of the Benchmark Response Level

Elaine Faustman, Discussion Leader
Department of Environmental Health
School of Public Health and Community Medicine
University of Washington
Seattle, WA

The panelists considered the three approaches (biological significance, limit of detection, and default options) presented in the draft document for benchmark calculation. The panelists also considered the presentation at the meeting by Dr. Setzer on LOD methods and his simulation study results in making the following comments.

The benchmark dose methodology based on biological significance generated the most enthusiastic response. These comments were largely positive and of the three approaches this one generated the most supportive comments. Many panel members expressed the sentiment that a scientific basis for the risk assessment methods was a requirement.

There was also, however, strong support for clarification in the document and a call for more research for all but developmental toxicity endpoints (some panelists even felt more studies were needed for this endpoint as well). In particular, the panel discussed neurotoxicity experiments where patterns of effects in the functional observation battery (FOB) tests were more important than individual responses. In fact, some comments indicated that the biological significance of single responses would be unknown. No additional research evaluations were presented at the workshop to indicate that EPA had applied this methodology to large numbers of neurotoxicity endpoints; however, some limited examples were cited by panel members. For the biological significance approach, there was general agreement that further investigation of the neurotoxicity endpoints would be desirable before wide-scale application for regulatory purposes.

In part, the panel's request for clarification of biological significance could be addressed by incorporating additional examples of guidance information from the specific risk assessment documents on developmental toxicity, reproductive toxicity (draft), neurotoxicity, and general acute

and chronic toxicity. Primarily, the panel sought clarification about the toxicology principles underlying the response that would be pertinent for NOAEL or benchmark methods.

The need for additional information or illustration of how endpoint-specific guidance from the referenced guidance documents would be used was very evident when the continuous endpoints methodology was discussed. The panel spent considerable time discussing the significance of various highly specific endpoints (e.g., what level of change in cholinesterase level is considered biologically significant, what level of fetal body weight is considered adverse). These discussions are not specifically relevant to the benchmark dose discussion but should be discussed for both NOAEL and benchmark approaches. In regards to setting appropriate, biologically defensible levels of change for continuous endpoints, the panel did discuss various approaches. Several panel members discussed specific approaches (either in writing or in verbal comments) that they had used with continuous data. There was general agreement that quantalization of continuous data resulted in loss of information; however, the panelists had differing opinions about how much loss of information would occur. Numerous methods for evaluating the differences in response of treated groups versus control groups were discussed. The need for the document to address in greater detail the body of literature on these approaches was evident from this discussion. One could imagine the usefulness of adding tables listing the various approaches, how the approaches have been applied, and what the authors' comments were concerning the applicability for that endpoint. In general, these approaches centered on using some comparison of the treated response groups with the distribution of the control responses. The panel members noted the lack of evaluations specific for neurotoxicity study designs.

Another issue for benchmark methods and the setting of a response level arose when the panel discussed the ultimate use (goal) of the benchmark dose methodology. The panel discussed whether the benchmark methods were being used to develop a common metric across diverse endpoints of cancer and noncancer effects (holistic view) or were to make comparisons within compounds across endpoints. The panel discussed the need to identify common points of departure (responses) for the benchmark methods that might be defined as the same response level in terms of impact (e.g., equally likely to produce death, equally likely to diminish life quality, equivalent adversity). Some panel members discussed the need to use severity rather than abnormality as the common basis for response comparisons. Although time was spent on this topic, no solutions were

identified. The panel recognized that this topic was better left for subsequent workshops and that such considerations would be pertinent if either NOAEL or BMD approaches were to be used in cross-endpoint comparisons, such as for cost-benefit analysis. The EPA representatives were asked directly about this point, and they responded that the Agency would look to the specific disciplines to define the significance of biological responses within each. The panel raised the possibility of implementing the guidance document in phases, with use of an iterative process.

In regard to the response level issues, the panel introduced the discussion on how uncertainty factors would be used. Panelists expressed the need to discuss uncertainty factors as part of the discussion on response level. Protection of sensitive individuals would still be anticipated to be accounted for in the use of an uncertainty factor approach.

A later discussion of whether the benchmark response should be viewed as a NOAEL or a LOAEL is very pertinent to the response level discussion. The pros and cons of that approach are discussed in that section.

The panelists spent a significant portion of their time at the meeting discussing general issues that are critical for conducting risk assessment based on good science; however, most of these issues were broader than the benchmark methodology and were just as important for calculation of NOAEL values. It was clear from this discussion of science issues for good risk assessment that most of the panel members were frustrated with the default approaches used in general risk assessment but were wary of new methodologies— "rocking the boat." The panel contended that more details are needed in the guidance document on how biological significance would be defined and applied to benchmark does methodologies.

Panel members also discussed other areas of needed clarification, such as whether additional risk or extra risk would be used as the basic response. Additional discussion of potential differences between cancer and noncancer risk and the use of these two reference points is needed in the document, and the definitions of these two descriptions needs to be in the body of the document rather than hidden in an appendix. The panel found it confusing to discuss BMD terminology in the noncancer endpoints and ED terminology in the text of the cancer endpoints. A common

terminology is needed, and one panel member noted that the ED terminology is more intuitive in regards to the use of confidence limits.

Further, the document needs additional illustration of the potential use of individual versus mean responses as well as patterns of response versus individual endpoint responses. Some of this could be accomplished in an expanded section on selection of critical endpoints. The requirement of the benchmark dose methodology to be useful with patterns of exposure versus individual responses seems to be especially important for endpoints like neurotoxicity. This issue also affects NOAEL methods. Two panel members reminded the panel not to overlook human epidemiology data, not only in defining the biological significance of findings in rodent studies but also in defining how we look at population responses.

The second approach that the panel discussed and commented on was the limit of detection methods. Most panelists wanted clarification of this approach. The guidance document referred to the presentation provided by Dr. Setzer at the workshop. Dr. Setzer's presentation illustrated the low power of detection of effects of many study designs used with toxicology testing. A problem that the panelists had with the LOD approach was illustrated in a quote from the presentation of Dr. Setzer: "The draft guidelines propose that, in the absence of the determination of a 'biologically significant' response, the BMR be set as the LOD of a typical 'good' design for the species and endpoint considered. For example, designs recommended by various testing guidelines would be considered good designs." One panelist noted the similarities in this LOD approach as compared with the LOD approach used with environmental monitoring. Abuses of this approach in that application (e.g., remediation being undertaken for no contamination because the LOD for environmental monitoring has been taken as a possible level of contamination and is added with other nondetects) were referenced as an example of why this approach can cause problems. This is specific to the BMD methods because they involve assigning a level of response, versus no response assigned to the NOAEL value. Panel members felt that the LOD approach would lose the advantage of assigning a specific response level—a key advance of the benchmark methodology would be lost.

Other panelists in the group presented an alternative view that the LOD approach allowed the researchers to bound the response. Thus, it would be very useful, not only in showing the

researcher how low the power of currently used biology test procedures are, but also by providing a reference point. Yet other panelists felt that the LOD approach would provide disincentives for improving limitations in experimental design. By accepting a LOD approach, we would not be able to overcome poor experimental design.

Another concern that the panel members raised in regard to the LOD approach is that it was providing the wrong incentive for researchers to identify increasingly sensitive biomarkers for response. The document does not address just how sensitive, yet not clearly adverse, biomarkers of effect would be handled. Would a LOD approach also be used with these types of studies?

The panel also discussed default procedures. It was clear from both the written comments and the verbal comments at the meeting that the panel members were confused about the application of the default procedures for continuous endpoints. Most panelists felt more comfortable with default approaches for quantal approaches than for the proposed LOD approach. Most of the discussion centered on responses of 5 to 10 percent.

The panel members had several key general comments:

- Several panelists stated that the document represented a statistical overkill and was in danger of being too prescriptive.
- Some panel members noted that the document was largely silent on use of biologically based models. There was some support for adding a section to the document on how the BMD methods might fit logically with progression toward these models.
- A number of panel members contended that common nomenclature is needed for use of BMD-like methods across endpoints. Eliminate use of both ED and BMD terms to convey the same concept and use one term consistently for application in cancer and noncancer endpoints.

Model Selection and Fitting

Colin Park, Discussion Leader
The Dow Chemical Company
Midland, MI

Is the Order of Model Application for Continuous and Dichotomous Data Appropriate?

The main comment made here is that there is an apparent inconsistency between continuous and quantal data. For continuous data, it is recommended that a linear model be fit first; whereas in quantal data, more complex models are recommended (e.g., using a polynomial of degree k-1, then reducing the number of terms as appropriate). One reason for this inconsistency is likely due to historical practices of fitting models for these different types of data. There did not appear to be any clear consensus on whether harmonization was important and if so, which way to go, although the subject came up again under the next question.

Apart from this comment, the general consensus appeared to be that the guidelines in this section were appropriate.

Should Other Models Be Considered, or Should the Number of Models Applied Be More Restrictive?

Generally the panelists contended that models should not be more restrictive. In fact, some commenters responded that additional models should be allowed; probit and Michaelis-Menton were specifically recommended. There was, however, a difference in philosophy concerning complexity of modeling.

One general school of thought was that the most simple model that is consistent with the data should be used; that is, start with a linear model (with a threshold if necessary), then check for lack of fit and add more parameters as necessary. It was pointed out, however, that lack of fit tests are quite insensitive. It was recommended by one panelist—with apparent general agreement during the meeting—that a "soft" criterion be used for the goodness of fit test (e.g., p=0.20).

Others felt that complex models could be fit, then parameters eliminated as appropriate (backwards elimination) (e.g., start with a polynomial of degree k-1, then look for the most parsimonious model). (Personal note: This is the same discussion that went on 20 to 30 years ago as to whether stepwise forward or stepwise backward regression was the most appropriate). One concern mentioned in the workshop is that there is something to be said for simplicity, given that the output will be used by non-statisticians.

Another school of thought approached the question from the point of view that results should be calculated from a number of models, then the range (or distribution) of the results displayed or represented as a summary statistic (e.g., the mean be calculated and professional judgment be used).

Are the Parameters Proposed as Defaults for Model Structure Appropriate?

a. What should be the default approach for selecting the degree of the polynomial to use?

A number of commenters responded that the models should be as parsimonious as possible (see above on alpha levels for goodness of fit). A few thought that the best fit should be the criterion, although it is not clear if they had considered the impact on confidence limits.

b. Is the default of not including a background parameter appropriate unless there is some indication of a background response level?

The apparent consensus on this issue was no.

c. Is the use of extra risk as a default for quantal data appropriate?

Most commenters had no opinion or said that they had no strong reason for answering one way or the other. One commenter did say, however, that the use of extra risk for estimation is inconsistent with how program offices use risk estimates to calculate population cancer burden and that added risk should be used instead.

One panelist suggested extra risk be used but said that the recommendation was based on policy considerations of public health protection (i.e., using extra risk as the estimation procedure results in higher potency/risk estimates).

d. Is the default of not including a threshold parameter appropriate?

No consensus was reached on this question, although the discussion at the workshop appeared to provide some support *for* inclusion of a threshold parameter, particularly in the case of linear models for continuous data. The comment was made that the estimated (threshold) parameter had minimal biological interpretation relative to the existence of true thresholds. It was also pointed out, however, that most of the parameters in the statistical models had little biological relevance. It was suggested that the correct interpretation of estimates of the threshold parameter is that it represents an *apparent* threshold in the *observed* data, and might be more appropriately referred to as an intercept.

e. Is the default of modeling continuous data as such appropriate?

A large majority agreed that continuous data should not be dichotomized, although it was pointed out in this session, and in others, that there has been inadequate research into the operating characteristics of different approaches to calculating benchmark doses from continuous data. For example, how much sensitivity is lost by dichotomizing the data?

The issue of the need to first determine the biological significance of changes in many continuous endpoints was raised in this session and continually through the workshop.

Is the Approach for Determining the Fit of the Model Appropriate? Are There Additional or Alternate Criteria That Should Be Used?

- See above on alpha levels.
- It was mentioned, even by the statisticians, that more description and/or familiarity with the AIC criterion was necessary.

- It was mentioned by a number of panelists that software that included a graphical presentation of results was a good idea.

- There was discussion in this session, and later, on the issue of evaluating results from different models. The general feeling was that an arbitrary requirement that results from different models be within a factor of 3 was probably not very supportable. One suggested alternative was to carry forth all outputs in the form of a range or distribution. Another suggestion was that if different models give widely different results, this indicates that utilizing the BMD as a point of departure may not be a good idea. Instead the traditional NOAEL/LOAEL approach should be used. (Personal note: This would be okay for noncancer risk assessment, but what about cancer? It appears that EDx's will be more consistent from model to model than LEDx's, which is another argument in favor of using central estimates rather than upper bounds.)

Additional Comments

A case was made by some participants for keeping the process more simple than is currently being proposed. It was held that calculating the limit of detection, fitting numerous and somewhat complex models, and calculating confidence intervals is unnecessary. The idea of a BMD is to calculate a point of departure that is more data-driven than NOAEL's and LOAEL's (noncancer) and that more accurately reflects the limitations of the data than the linearized multistage (LMS) model (cancer). The complexity being proposed, however, was inconsistent with some of the objectives of the risk assessment process (e.g., transparency).

It was held that confidence intervals raise the following problems:

- From the point of view of non-statisticians, confidence intervals introduce a "black box" component into the process.

- The interpretation of the regulatory limits after the incorporation of uncertainty factors is not clear. For example, 95 percent of the population is protected? We are 95 percent sure that all the population is protected? Neither of these interpretations is correct nor are they implied in the methodology, but it is possible (likely) that these kinds of interpretations could be made.

- The additional complexity of a lower bound rather than a best estimate has a very small effect relative to the magnitude of uncertainty factors added on in the next step.

- There is more than one method for calculating the limits resulting in different answers, thereby adding confusion to the process.

On the other hand, the use of confidence intervals does reward good experimentation, although simulation results show the rewards to be small. There was consensus, however, that LED's should at least be calculated in conjunction with ED estimate's. The question concerned which one to use if a single value was used as a point of departure. It appeared to be generally felt that reporting a range or distribution of the Ed_x would solve this problem. A very few appeared to favor the LED_x as a single reported value.

The suggestion was again made in a later session that the point of departure be calculated as a range or distribution, reflecting the statistical variability of the ED_x. There was minimal discussion of this suggestion, with no apparent strong objections.

There was discussion throughout the workshop on the need to further validate the methodology, particularly for continuous data.

Use of Confidence Limits

Lorenz Rhomberg, Discussion Leader
Harvard Center for Risk Analysis
Harvard School of Public Health
Boston, MA

The topic of use of confidence limits was the focus of a number of written premeeting comments as well as of lively discussion during the workshop. The charge to panel members included three questions:

- Should the lower confidence limit on dose be the definition of the BMD/C?
- Are the defaults for the method of confidence limit calculation appropriate?
- Is the default of the 95-percent confidence limit appropriate?

These are best considered in reverse order, since the later ones presume answers to the earlier.

Is the Default of the 95-Percent Confidence Limit Appropriate?

That is, should we consider other percentiles? Note that the question is about the default; specifically, it applies to the BMD/C definition (presuming it is defined as the estimated lower limit on the dose producing the BMR).

Two participants had noteworthy written comments on this issue, which they further discussed at the meeting. One suggested "soft" limits (i.e., less than 95 percent) as a way to avoid the "linearization" of the confidence interval at lower dose levels. Such soft limits were discussed elsewhere in regard to the goodness-of-fit determination. Here, however, they were aimed at preserving the ability to track curvature in the data's dose-response pattern. The value of this consideration was debated, it not being clear to some why linearity of the lower limit with dose in the lower dose range was to be considered problematic. The lower bound on dose is required for

one dose only—that producing the BMR—and its behavior at other doses is not really at issue, in this view.

Another panel member reminded us that the choice of 95 percent represents a tradeoff between the costs of making the interval's coverage wider than necessary to include the target value and the cost of missing the target value. The choice of coverage probability therefore has implicit policy aspects. He also noted that confidence interval construction methods need to be made to achieve their nominal coverage for all possible sets of parameter values; for particular data sets and curve shapes, this nominal coverage may be achievable with narrower intervals, an advantage more easily realized by bootstrapping methods than by other methods of confidence interval construction. (It should be noted that neither panel member favored use of the confidence interval in definition of the BMD/C.)

Are the Defaults for the Method of Confidence Limit Calculation Appropriate?

This was the second question in the charge to workshop panel members, and it should be noted that the direct question is again about defaults. As a default, the present document suggests reliance on asymptotic methods. The discussion presumed that confidence limits would be calculated and presented even if the BMD/C definition did not rely on them.

There are several methods for calculating confidence intervals. These are based on different approaches and invoke different assumptions. Moreover, intervals can be placed on various aspects—estimated parameters, slopes of lines, the population of instances or the mean—that have very different meaning and interpretation, distinctions that are sometimes lost in risk communication. Complex models may have complex methods for confidence interval calculation. Certain models may have parameters difficult to estimate (or to estimate independently), affecting the size of the confidence interval.

There were some written comments on this question, but it received little discussion at the workshop. One panel member preferred likelihood-based methods for confidence intervals, since

maximum likelihood is the preferred curve-fitting method. Another questioned use of asymptotic methods for the typically small sample sizes of most toxicological studies.

Two panel members asked why bootstrap methods could not be considered. This introduces fewer problems with model complexity, multiplicity of parameters, and small sample sizes. A potential difficulty may be that risk managers using such assessments might be disturbed by the lack of exactly repeatable interval calculations.

One panel member wrote that confidence intervals do not replace a good uncertainty analysis, which is what is needed to characterize uncertainty. Another noted, however, that confidence intervals are a natural way to express the degree of uncertainty in the calculation of the BMR from a set of experimental data.

Is the Default of the 95-Percent Confidence Limit Appropriate?

This third question received the bulk of written comment and of discussion during the workshop. The thrust of the question is whether the BMD/C *definition* should be based on the statistical lower (95 percent) bound on the dose/concentration that is estimated to produce the BMR (as proposed) or on a central estimate of that dose/concentration. Again, we presumed that confidence intervals would be calculated and presented in any case, even if the BMD/C were to be based on a central estimate of the dose producing the BMR.

The balance of opinion was for central estimates, although a significant group argued for a definition of BMD/C based on the lower bound. Those favoring a lower bound-based BMD/C definition cited reasons that were few and basic, while those arguing for central estimates gave reasons that were many and varied.

The main argument of those defending the use of the lower bound in BMD/C definition is that there is uncertainty in estimation of the BMD owing to experimental error in the particular study used to estimate it. Given the aim of the risk assessment process to identify doses unlikely to produce adverse effects, we should allow for this potential error to avoid underestimating the

intended BMD. At the workshop, one panel member noted that *any* choice of a point in the distribution of estimates—be it an upper or lower bound or a central estimate—implies a particular choice of weights given to some kinds of errors versus others. No stance is free of such values, and so the particular stance adopted should give weights appropriate to the aim of the exercise—in this case, not to underestimate risk. Others argued that central estimates are more appropriate for making comparative choices and for conveying the most plausible interpretation of the data; any desire to gauge the probability of underestimation should, in this view, be addressed by a separate examination of confidence limits or uncertainty analysis, not in the definition of the BMD/C.

A second argument offered in favor of a BMD definition based on lower limits was that use of a lower bound encourages good experimental design, since good design leads to tighter limits and thus higher reliable estimates of the BMD. Several participants, in written comments and at the workshop, raised doubts as to whether the amount of incentive for more powerful experiments was large enough to be of practical value; they cited their simulations that suggested that practically foreseeable changes in experimental design had but minor effect on narrowing the confidence interval. They questioned whether this benefit was worth the shortcomings of a lower limit-based BMD definition. Another panel member pointed out that most testing is done according to approved protocols that are evidently felt to be sufficiently powerful given practical constraints, so that real design flexibility may be limited in any case.

Several commenters noted that using the lower bound would produce BMD/Cs that are protective of public health. One panel member argued that such use of the lower bound is similar to the upper bound used in cancer risk assessment, and would be appropriate from the point of view of the goal of harmonizing cancer and noncancer methods. Workshop discussion noted that harmonizing on central estimates might be a better alternative, in the view of some.

There were several main arguments (plus a number of ancillary ones) presented in favor of defining BMD/Cs in terms of central estimates of the dose/concentration producing the BMR. These were discussed in written comments and during the workshop. It was noted by several participants that at a 1994 workshop on BMD procedures, the use of lower bounds had already been debated and the use of central estimates endorsed. Some questioned why this issue kept returning.

The argument most often cited in one form or another was that a central estimate constitutes the single "best" interpretation of the data at hand. It provides "an unbiased (not intentionally biased) starting point for risk assessment;" it constitutes "more precise use of experimental data;" it provides "our best understanding of the response and accurately portray[s] this to the risk manager:" lower bounds, on the other hand, indicate "where we think the BMDs might be...instead of where we think they are." At the workshop, some commenters felt that central estimates could also mislead by failing to emphasize the existence of a range of plausible lower dose estimates as causes of the BMR.

A second argument is that the risk assessment process is "already conservative enough" and needs no special accounting for experimental variability. "There are significant conservative assumptions to make up for the animal variability;" "many other health-conservative steps are already built into the risk assessment process." Some commenters countered that uncertainty should be dealt with wherever it is found. Several discussants pointed out that the amount of uncertainty accounted for by the use of the lower bound is trivial compared to that acknowledged in the application of the several ten-fold uncertainty factors used in determining RfD/Cs. In this view, the slight numerical adjustment afforded by the lower bound implies an unwarranted degree of precision in the risk assessment process. Some workshop discussion questioned whether the amount of conservatism contributed by the lower bound on an effective dose was worth the "baggage" of accusations of overblown conservatism that would likely accompany its use.

Third, it was argued that use of lower bound-based BMDs will hamper comparison among experiments and endpoints that differ in sample size and hence in the width of the confidence interval on dose producing the BMR. A lower bound "confounds the evaluation of relative locations;" a central estimate, on the other hand, "facilitate[s] the comparison of critical effects for different endpoints." One panel member provided a hypothetical example showing how endpoints determined with poor dose-response resolution would often be chosen as critical effects if lower bound-based BMDs were used for comparison, even if these endpoints did not appear critical in terms of central estimates. The comparability issue received considerable discussion at the workshop. It was pointed out that the desired consistency of central estimates would only be achieved if a single, risk-based definition of BMR were adhered to, and not if the proposed definition of BMR on limit of detection were employed. It was widely agreed that comparisons

among experiments and endpoints, including the choice of critical effect, might be better made on the basis of central estimates of BMD. Those favoring use of lower bounds, however, suggested that, once critical effects were chosen, a BMD definition based on a lower bound would still be possible and appropriate. In addition, a panel member cautioned that confidence intervals should also be examined during comparison among studies so as to help gauge the probability that rankings and comparisons are robust to the uncertain values of the central estimates.

A fourth argument cited in favor of central estimates is that confidence limits are widely misinterpreted by users of risk assessments. Their introduction, therefore, hampers risk communication. One panel member argued that central estimates were much simpler to grasp and allowed unsophisticated users to conduct and interpret assessments without delving into statistical arcana. Several commenters related accounts of misapprehensions among users regarding the nature and meaning of confidence limits. A 95-percent bound may be believed to refer to 95 percent of the population being free of the effect, or the confidence limit may be interpreted as addressing all of the uncertainties in the risk assessment, not just the experimental error in the single critical study. The connection of the lower bound on dose to produce a risk with the upper bound on risk at a given dose has confused many users. Some commenters argued, however, that the potential for misinterpretation is a risk communication challenge to be faced, not grounds for omitting valuable analysis. The development of EPA software to conduct BMD/C analysis, if well documented and explained, may obviate some of these concerns.

A fifth issue that was raised is that of the stability and robustness of the confidence interval calculation in the face of variation in methods, models, and data sets. Several commenters said that choice of mathematical model to fit to data had more influence on the estimated value of the lower bound on dose than on the central estimate. There are several alternative statistical approaches to calculating confidence limits (raising the issue named in the second question of the charge to workshop panel members regarding preference among them); confidence limits are somewhat dependent on which is chosen. One panelist noted that in the context of a given model, the instability of maximum likelihood estimates, an issue for low dose extrapolation, was not a serious concern for estimation in the range of the BMR. There was some concern expressed that reliance on lower bounds might constrain the choice of models and parameterizations to those with well-behaved, relatively narrow confidence limits. For example, the statistical difficulty of reliably

estimating a threshold parameter may discourage consideration of modeling that might produce useful insight. (A panel member noted that the difficulty in estimating a value for a threshold parameter may be largely obviated by constraining the degree of high-dose extreme nonlinearity allowed.) Another panelist noted that confidence bounds are less sensitive to trend than central estimates. A panel member expressed concern that models with many parameters would, owing to their few degrees of freedom, have particularly wide confidence intervals. In response to the above issues, proponents of confidence intervals in BMD definition defended the value of addressing experimental uncertainty; many of the issues could be addressed by appropriate specification of default procedures.

A panelist noted that the NOAEL has no confidence interval or explicit allowance for uncertainty in its determination. In a sense, its uncertainty has been addressed in the form of the uncertainty factors, an element that now is being made more explicit analytically.

Another panel member noted that central estimates of the BMD are usually in the experimental range, while lower limits on these estimates may not be.

As discussed earlier under the choice of the 95-percent limit, one panel member argued that, since confidence limits become linear with low doses, use of a lower bound in the BMD definition amounts to adoption of a linear extrapolation assumption essentially the same as that in the LMS procedure of cancer low dose extrapolation.

One panelist noted that central estimates are appropriate for use in cost-benefit analysis. Others noted that current noncancer risk assessment methods do not make explicit risk estimates for different dose levels (as cost-benefit analysis would presumably demand) and that the uncertainty factors do not produce central estimates, even when a BMD is based on a central estimate.

In the workshop discussion, it was suggested that the issue of experimental uncertainty in the BMD determination could be addressed in an explicit uncertainty factor. This would preserve the advantages of a central estimate, while allowing consideration of uncertainty in its proper realm—risk management choices in the face of uncertainty. A panel member pointed out that ED_{10}s were somewhat like LOAELs and ED_{05}s somewhat like NOAELs (at least in dose magnitude)—the use

of an uncertainty factor instead of a lower bound-based BMD is thus similar to the traditional LOAEL-to-NOAEL uncertainty factor. Several participants questioned the wisdom of using a crude and approximate uncertainty factor in place of a well-defined and statistically justified lower bound that addresses the specific uncertainty of the given experimental design and response levels.

There was considerable discussion of the idea that experimental uncertainty could be considered as an element separate from the BMD definition. One method for doing this is to make the uncertainty an explicit uncertainty factor, as suggested by a panel member. There was discussion of how big such a factor might be and how to make it address specifics of the data in particular instances. Several panel members stressed using central estimates for comparisons, choices, and estimations, and then examining confidence limits as a separate step to gain perspective. Another panelist (in his written comments) said that a full quantitative uncertainty analysis is possible and preferable to any attempt to fold particular experimental error concerns into any estimation procedure such as BMD definition. The workshop participants discussed the possibilities of expressing the BMD as a range or distribution, reflecting not just a central estimate or a single lower bound, but the full spectrum of tenable possibilities weighed by their relative support. This could enter into a distributional approach to other elements of the assessment, including distributions on exposure and on the values of the uncertainty factors used in extrapolating animal results to levels deemed safe for human exposure. Some participants, however, doubted whether risk managers would welcome diffuse answers to safety questions.

Summary

After considerable discussion of the issues, there continued to be disagreement among the workshop participants regarding whether BMD/Cs should be defined as the lower bound or the best estimate of the level associated with the BMR. Many workshop participants argued for use of central estimates. Their arguments noted several properties of lower limit-based BMD definitions they deemed undesirable, but they focused on the notion that best estimates were most useful for comparison among endpoints and studies, while considerations of error in BMD estimation should be considered separately in the risk assessment process (so that risk management choices could take account of it as decision-makers see fit). There was no consensus on this point, however, there being

a substantial minority of participants arguing that a lower bound-based BMD definition was appropriate given the purpose of its estimation: to help define a dose unlikely to cause human toxicity. All agreed that some consideration of the uncertainty in estimation of BMDs owing to experimental error had to enter into the risk assessment process in some way, and that central estimates and lower bounds should always both be reported.

If estimation error is to be considered separately, there was disagreement as to whether it is best to do so in a separate uncertainty factor, with a full quantitative uncertainty analysis including a distribution on the estimated BMD, or simply as a reported lower bound. The question is hard to separate from that of how other sources of uncertainty in the risk analysis are to be handled. The central issue is whether uncertainties associated with each element or step of the analysis are to be somehow incorporated into the results reported for that step (as with a lower bound-based BMD definition), or whether a separate exercise examines the uncertainties of all steps comprehensively.

Selection of Benchmark Dose/Concentration to Use as the Point of Departure

William Pease, Discussion Leader
Environmental Defense Fund
Oakland, CA

The panel members supported EPA's general approach of establishing a series of decision points with defaults as a way of selecting a BMD/C from various model predictions to serve as the basis for deriving regulatory standards. Panel members criticized several technical aspects of the Agency's default approach, however, and raised several general issues regarding the attributes that a "point of departure" for low dose risk assessment might exhibit. The following sections summarize the technical concerns of the panel and then present the range of opinions expressed about various attributes of a BMD/C that might be useful in a regulatory risk assessment context.

Comments Requested in Charge to Workshop Panel Members

a. The determination of equivalence of methods

EPA proposes to assess the "equivalence" of different BMD model predictions using statistical procedures (a goodness-of-fit test), expert judgment (based on visual examination of model fits to observed data), and an arbitrary default definition of equivalence (if model estimates of the BMD/C are within a factor of 3).

The panel was in general agreement that this approach required revision and further explanation. It was noted that goodness-of-fit tests need to be designed so that they evaluate models in the low dose region of interest (i.e., fit in the area of the ED_{10}) rather than across the entire observed data range. There was general support for the use of a visual assessment of model fit. Most concerns raised regarded selection of a factor of 3 as the default definition of equivalence: Opinions ranged from a statement that even this size factor could have substantial regulatory impact, to a request for at least some rationale to support what must admittedly be an arbitrary criterion.

One panel member noted that the Agency should consider adopting different definitions of equivalence based on whether the selection choice involved different BMD model estimates for the same endpoint based on a single study's data or different BMD/Cs generated for different endpoints from multiple studies.

b. Use of the Akaike Information Criteria for comparing the fit of models

For "equivalent" models, EPA proposes to select the final BMD/C based on application of the Akaike Information Criteria (AIC).

Virtually all panel members requested additional description and references for this proposed procedure. Statisticians on the panel raised several concerns about the AIC: It does not focus its evaluation of model fit on the low dose region of interest, and it has generally been applied to select among models in very data-rich situations (e.g., time-series data). Considerable concern was raised about the ability of the method to usefully discriminate between model results based on typically sparse dose-response data sets. Several alternatives to such heavy reliance on statistical techniques to guide the decision process at this point included taking the geometric mean of equivalent model results or selecting the lowest of the equivalent BMDs as a health protective default.

c. Is the default approach for selecting the BMD/C to use as the point of departure for cancer and noncancer dose-response analysis appropriate?

For non-equivalent models that have passed statistical and visual goodness-of-fit testing, EPA proposes to select the lowest estimated BMD/C as a health protective default.

Panel members acknowledged that some default procedure is required to select among results when BMD-estimates are clearly model dependent. There was considerable discussion about whether the need to rely on such defaults could be reduced by altering the Agency's current definition of BMD/C so that it is redefined as a model's maximum likelihood estimate rather than a lower 95th-percentile confidence bound. Some members noted that central estimates of BMD/Cs from various models are much less variable than lower bounds, so that an altered definition of the BMD could produce fewer instances of non-equivalent results.

In clear situations of model-dependent BMD/C estimates, most members of the panel generally agreed that a default approach of selecting the lowest estimate as a point of departure could be justified as a public health policy choice. It was noted that there would generally be no biological or statistical rationale available for selecting one model's result over another at this point. A panel member opposed to this approach recommended that whenever there is clear model dependence in BMD/C estimates, the BMD approach should be dropped and the Agency should shift back to using a NOAEL as a point of departure. Another panel member raised the possibility of carrying multiple model-dependent BMD estimates forward into the risk management process, rather than excluding some possible BMD values through the application of defaults. Different model results (with their associated uncertainties and some estimate of their overall plausibility) would be presented as part of a decision tree to risk managers.

Comments on General Issues Regarding Selecting a Point of Departure

The panel was unanimous in expressing concern that EPA had not clearly expressed its goals in adopting the BMD approach, and that this had prevented a thorough evaluation of the potential attributes that should be exhibited by BMD/Cs. Through several presentations at the meeting, it became apparent that EPA primarily conceived of BMD/Cs as an improved replacement (a "plug-in") for NOAELs in noncancer risk assessment that should generally not affect the conventional application of uncertainty factors to derive reference doses. Panel members raised a number of concern about this narrow conception of the BMD approach and identified three other desirable attributes for BMD/Cs that the Agency should consider as it proceeds with defining and implementing the BMD approach. The following sections summarize the discussions of the four potential attributes that BMD/Cs could be designed to support.

a. BMD/Cs should be a "plug-in" for NOAELs with minimum impact on uncertainty factors or the RfD process

EPA has been motivated to a considerable degree in its BMD development effort to develop a new approach to noncancer risk assessment that addresses the widely acknowledged problems of NOAELs as a starting point, but that preserves the current "level of conservatism" associated with

the conventional process. While this motivation is understandable from a political perspective (since it will result in minimal revisions to current standards and does not require altering standard uncertainty factors), panel members expressed considerable skepticism about this goal. Members could reach no consensus on what the current "level of conservatism" provided by NOAEL-based reference doses was: Some argued it was clearly adequate, others that we have little or no empirical evidence about the degree of safety provided by most noncancer health standards.

In the absence of a way of assessing health protectiveness, it appears that this approach of "aiming" the BMD to be as close to existing NOAELs as possible (i.e., designing the BMD approach to generate a BMD:NOAEL ratio of one over the complete set of compounds assessed) could actually lead the Agency to duplicate some of the problems associated with the NOAEL approach. For example, EPA proposes to use the LOD method to establish BMD/Cs rather than a fixed incidence level. This approach was developed to ensure that the BMD/C estimated from insensitive studies (e.g., some neurotoxicity assays) would generally be equivalent to NOAELs that could be estimated from such studies. Panel members generally rejected this approach because it rewards the existing detection limits of conventional testing protocols (e.g., by providing a higher percentage BMD/C as a starting point for neurotoxicity than developmental toxicity) and provides no incentive to improve detection limits for inadequately assessed endpoints.

The panel generally agreed that it would be more appropriate for the agency to conceive of BMD/Cs as aiming for a low effect level rather than a conventional NOAEL. Particularly if the Agency redefines the BMD/C to be a central estimate of a dose associated with a 10-percent incidence of biologically significant adverse effects (ED_{10}), it will simply not be plausible to equate this with the "no observed adverse effect level" of conventional noncancer risk assessment. The BMD/C will be, by definition, an effect level where something biologically significant is occurring. The panel emphasized, however, that this does not make the BMD/C equivalent to a LOAEL. (In conventional noncancer risk assessment, LOAELs may serve as a point of departure for deriving a reference dose when a poor study has failed to ascertain a NOAEL and requires application of an additional 10-fold LOAEL->NOAEL uncertainty factor.) In contrast to LOAELs, the ED_{10} will be associated with a defined level of adverse incidence, often lower than that of LOAELs from well-designed studies. Because it is an "effect level," however, several members of the panel recommended that EPA should reexamine its current uncertainty factor practice. The Agency

should consider whether a new uncertainty factor (to extrapolate to a no risk dose from a low effective dose to obtain a point of departure) is required and should also examine other impacts that the BMD methodology could have on conventional uncertainty factors.

The panel noted that the guidance document needs to explicitly address under what conditions the Agency anticipates continuing to conduct noncancer risk assessment with NOAELs rather than BMD/Cs. Concerns were raised about the confusion that may arise if the two approaches are mixed (e.g., used for quantal endpoints, but delayed application for continuous or neurotoxic endpoints). The document would benefit from a clear statement of the limited conditions under which NOAELs will continue to serve as points of departure.

b. BMD/Cs should provide a consistent and comparable point of departure for calculating reference doses or margins of exposure

As proposed, EPA's BMD/Cs can represent different incidence proportions (from a quantal default of 10 percent to as high as 40 to 50 percent for some continuous data with high levels of detection) of adverse impacts of widely varying severity. Risk characterizations for different compounds (based on the margin of exposure between current exposure and the BMD) will be more comparable if they employ a common level of adverse impact as their point of departure. This is a feature that is (at least rhetorically) possessed by the current NOAEL/uncertainty factor (UF) approach: Standard setting begins for all compounds from a level observed to have no adverse impacts. Uncertainty factors are then applied to derive an RfD that is likely to be below the population's threshold for any adverse impact. The hazard indices that are conventionally used to conduct noncancer risk assessment (the ratio of current exposure to RfD) are therefore interpreted as an exposed population's distance from a "safe" level of exposure. EPA's current BMD approach complicates this interpretation of hazard indices because it replaces the NOAEL with a starting point that represents doses associated with different percentage increases in responses of differing severity for different compounds. To maintain the integrity of an exposure/RfD ratio as a risk characterization tool, it would be ideal if BMD starting points for different compounds could be selected to be roughly comparable in terms of the potential impact on human health.

Several panel members emphasized that the BMR (percentage incidence) should be the same for most applications to allow for consistent interpretation of derived reference doses or margins of exposure.

There was general agreement that EPA must address "severity" of impact in defining points of departure for noncancer risk assessment. Panel members noted that it would be advisable to derive BMD/Cs based on incidence of adverse endpoints of comparable severity (either all BMDs could refer to some consistent minimal level of severity, or each BMD could be accompanied by a categorical indicator of the severity of the critical effect it is based on). Either option would involve using the concept of biological significance as an organizing concept for defining BMR. Expert judgment would be required to classify commonly observed adverse endpoints by severity (using different endpoint-specific measures of incidence and adversity to define a common severity scale). The panel noted that EPA will need to invest substantial effort in developing consensus definitions for biological significance (for many continuous endpoints) and that this effort could be extended to develop a common severity scale. Further effort should be devoted to issues raised by attempting to define a "common level of adverse impact" as a point of departure for standard setting.

One additional alternative approach to this problem involves addressing the variations in the seriousness of impacts observed at the point of departure with an additional uncertainty factor based on severity (the guidance does acknowledge that the nature of the response should be a consideration when evaluating the adequacy of calculated margins of exposure or hazard indices).

c. BMD/Cs should provide a suitable basis for low dose extrapolation

The use of BMD/Cs as a point of departure for low dose extrapolation elicited a wide range of expert opinions among panel members. Several supported the guidance document's general dismissal of this potential use, arguing that extrapolation beyond the range of observation using curve-fitting models is not credible or appropriate for noncancer endpoints. Other panel members noted that limiting the BMD approach to replacing NOAELs in a conventional uncertainty factor-based margin or exposure assessment would forgo exploring a much needed improvement in noncancer risk assessment: the ability to generate quantitative estimates of the incidence of

noncancer effects at various exposure levels. Such quantitative estimates of low dose risks for noncancer endpoints are required if these effects are to be considered in cost-benefit analyses.

Several panel members supported an intermediate position, noting that risk estimation within the experimental data range (of applied doses, not only of observed effects) was appropriate if it was constrained on an endpoint-specific basis (e.g., estimation down to an ED_{01} might be appropriate for cancer endpoints, given the power of cancer bioassays). Low dose estimation outside this range would only be appropriate if analysts could provide a plausible theoretical or biological basis for use of specific models (as in the case of low dose cancer risk estimation).

There was general panel agreement that the issues surrounding use of BMD/Cs as points of departure for low dose risk estimation required further discussion in the guidance document: Further guidance on risk estimation for exposure situations within the experimentally observed range is needed (noting, for example, how such estimation is currently being applied to the premature morbidity and mortality associated with ozone and particulate matter); discussion of other peer-reviewed, low dose risk estimation applications is warranted; and a clear policy statement on potential uses of BMD approaches for risk estimation is needed.

d. BMD/Cs should be derived to provide information useful for modifying uncertainty factors or evaluating margins of exposure

Several panel members noted that the process of estimating a BMD/C provides information about the shape of a compound's dose-response curve in the low dose region that is very valuable to risk managers. Indications of steep or shallow slopes is helpful in evaluating whether margins of exposure equate to margins of protection for public health (e.g., a steep slope in the low dose region indicates that risks will decrease quickly in the low dose region, and that the linear margin of exposure provided by standard uncertainty factor applications is likely to be protective). Information about the slope of the dose-response curve is also useful for evaluating the potential severity of impacts in the low dose region, which can be used to classify endpoints to establish consistent starting points or to establish the appropriate magnitude for any new uncertainty factor aimed at extrapolating from low effective doses to no risk doses. Information provided by the confidence limits on a BMD/C could be used to help establish the appropriate magnitude of the conventional

modifying factor for data quality (if the lower confidence bound is omitted from the definition of the BMD/C).

Another panel member noted that considerable important information is compiled during the benchmark dose process that warrants being presented to decision-makers. BMD/Cs could be conceived as ranges instead of single points of departure (to reflect the variety of model options as well as choices between best estimates and confidence limits). While the risk management process has a limited history with this type of detailed risk assessment data, adopting a distributional approach to BMD/C development could combine with a distributional approach to uncertainty factors to support more probabilistic risk assessment for noncancer endpoints.

SECTION FOUR

OBSERVERS' COMMENTS

Observers were given two formal opportunities to provide information and make public statements during the workshop. Observers were asked to sign up if they intended to make a statement. The following comments were made by observers at the workshop.

Arnold Kuzmack of EPA's Office of Water provided two comments. First, on the issue of including the threshold parameter in models, Dr. Kusmack expressed the opinion that it will be extremely difficult to communicate to the regulatory and legal communities that the BMD is not a true estimate of the biological threshold. Second, Dr. Kusmack concurred with a comment from one of the panel members about describing sets of comparable endpoints. Dr. Kusmack suggested that ways to describe this information need to be developed over time.

Lynne Haber of ICF/Clement, Inc., had one comment regarding the choice of the BMR. Ms. Haber supported the use of biological information for selecting the BMR and expressed the opinion that the draft guidance needs more information on how to select and use biological information (e.g., 10 percent change in body weight).

Joseph Siglin of Exxon Biomedical Sciences, Inc., offered an opinion concerning experimental design. He stated that discussions concerning better experimental designs are useful, however, there needs to be recognition that experimental designs are often specified in regulatory guidelines (e.g., the number of groups, the number of animals per group). Dr. Siglin stated the number and types of uncertainty factors applied could be a function of the confidence limits generated by the dose response (i.e., 95-percent confidence limits). Dr. Siglin asked the panel members whether application of the BMD should be restricted to experiments that lack a clear NOAEL. He pointed out that the BMD approach is not applicable to studies that show no effects at limit dose levels; however, this issue has not been specifically addressed in the guidance. Dr. Siglin concluded by

supporting the concept of keeping the BMD guidance as simple as possible so that it can be easily understood and applied by general toxicologists.

Amal Mahfouz of EPA's Office of Water expressed the opinion that the BMD is very useful for certain endpoints.

Hugh Pettigrew of EPA's Office of Pesticides Programs responded to a comment made during the panel discussions concerning the significance of selecting a factor of 3 for comparing different BMD estimates. He noted that it is the largest integer that is less than the square root of 10. EPA is always interested in orders of magnitude, he said; however, estimates that differ by less than the power of 3 only differ by half an order of magnitude. Concerning the use of the terms "extra risk" versus "additional risk," Dr. Pettigrew pointed out that EPA uses extra risk to regulate cancer risk and that under the new cancer risk assessment guidelines, EPA will still use extra risk. Dr. Pettigrew recommended that the authors of the draft guidance document make terminology compatible with other Agency guidance. Concerning the lower confidence limit on dose, Dr. Pettigrew was under the impression that this was a one-sided lower confidence limit on dose or a one-sided upper confidence limit on risk. He noted that existing software to make these calculations assumes a one-sided confidence limit, therefore, it does not make sense to present central estimates and upper and lower confidence limits because there is a conceptual difference between a lower one-sided confidence limit and a lower limit of a two-sided confidence interval.

William Marcus of EPA's Office of Science and Technology made the observation that EPA has been discussing dose-response curves for many years and that this workshop's discussions are nothing new. He also stated that workshop participants were discussing concepts in the absence of an understanding of what is really being addressed. Dr. Marcus expressed the opinion that cancer is not always the worst endpoint, and that cancer as an endpoint may not differ from any other toxicological endpoint. For example, lead is both dose related and dose responsive in bioassays. EPA decided not to regulate lead based on cancer because the biological significance of other effects (i.e., decreased mental aptitude in children) occurring at doses lower than those that cause cancer were more serious. Dr. Marcus posed the following questions to panel members:

- How do you measure decreased mental ability in children?
- Do you want to measure enzyme changes that result in nice statistical numbers, but may not have any biological significance?
- How do you decide which endpoint should be measured?
- Should you consider the biological significance, the statistical significance, or the endpoint that might provide you with information that alerts you to an unknown problem?

APPENDIX A

PEER CONSULTANTS/PRESENTERS

United States
Environmental Protection Agency
Risk Assessment Forum

Benchmark Dose Peer Consultation Workshop

Holiday Inn Bethesda
Bethesda, MD
September 10-11, 1996

Peer Consultants/Presenters

Bruce Allen
Project Manager
K.S. Crump Group
ICF Kaiser Engineers, Inc.
P.O. Box 14348
Research Triangle Park, NC 27709
919-547-1715
Fax: 919-547-1710

George Daston
Miami Valley Laboratories
The Proctor & Gamble Company
P.O. Box 398707
Cincinnati, OH 45239-8707
513-627-2886
Fax: 513-627-1908

Elaine Faustman
Department of Environmental Health
School of Public Health and Community Medicine
University of Washington
445 Roosevelt Way, NE - Suite 100
Seattle, WA 98105
206-685-2269
Fax: 206-685-4696

Jeff Fowles
Office of Environmental Health Hazard Assessment
Air Toxicology & Epidemiology Section
California Environmental Protection Agency
2151 Berkeley Way - Annex 11
Berkeley, CA 94704
510-540-3324
Fax: 510-540-2923

David Gaylor
Associate Director for Risk
Assessment Policy and Research
National Center for Toxicological Research
U.S. Food & Drug Administration
3900 NCTR Road (HFT-1)
Jefferson, AR 72079-9502
501-543-7001
Fax: 501-543-7576
E-mail: dgaylor@nctr.fda.gov

Daniel Guth*
National Center for Environmental Assessment
(MD-52)
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711
919-541-4930
Fax: 919-541-0245

William Hartley
Associate Professor
Toxicology and Risk Assessment
School of Public Health and Tropical Medicine
Tulane University
1501 Canal Street
New Orleans, LA 70112
504-588-5374
Fax: 504-584-1726
E-mail: hartley@mailhost.tls.tulane.edu

*Presenter

Rogene Henderson (Chair)
Senior Scientist
Inhalation Toxicology Research Institute
Building 9217 - Area Y
KAFB East
Albuquerque, NM 87115
505-845-1164
Fax: 505-845-1198
E-mail: rhenderson@lucy.tli.org

Carole Kimmel*
National Center for Environmental Assessment
U.S. Environmental Protection Agency
401 M Street, SW (8623)
Washington, DC 20460
202-260-7331
Fax: 202-260-8719

Abby Li
Toxicology Manager
Monsanto Company, Ceregen
645 South Newstead Avenue
St. Louis, MO 63110
314-694-7933
Fax: 314-694-7938
E-mail: aali@monsanto.com

Rashmi Nair
Monsanto Company
800 North Lindbergh Boulevard (A3NF)
St. Louis, MO 63167
314-694-8808
Fax: 314-694-8808
E-mail: rsnair@ccmail.monsanto.com

Bruce Naumann
Principal Toxicologist
Merck & Company, Inc.
One Merck Drive (WS2F-45)
Whitehouse Station, NJ 08889-0100
908-423-7908
Fax: 908-735-1496
E-mail: bruce_naumann@merck.com

James Olson
Professor
Department of Pharmacology & Toxicology
State University of New York at Buffalo
102 Farber Hall
3435 Main Street
Buffalo, NY 14214-3000
716-829-2319
Fax: 716-829-2801

Colin Park
The Dow Chemical Company
2030 Building
Midland, MI 48640
517-636-1159
Fax: 517-636-6451

William Pease
Environmental Defense Fund
Rockridge Market Hall
5655 College Avenue
Oakland, CA 94618
510-658-8008
Fax: 510-658-0630
E-mail: pease@uclink4.berkeley.edu

William Perry
Health Scientist
Directorate of Health Standards Program
Occupational Safety and Health Administration
200 Constitution Avenue, NW - Room N3718
Washington, DC 20210
202-219-7111
Fax: 202-219-7125

Christopher Portier
Acting Chief
Laboratory of Quantitative
and Computational Biology
National Institute for Environmental
Health Sciences
P.O. Box 12233 (MD-A306)
Research Triangle Park, NC 27709
919-541-4999
Fax: 919-541-1479
E-mail: portier@niehs.nih.gov

Lorenz Rhomberg
Harvard Center for Risk Analysis
Harvard School of Public Health
718 Huntington Avenue
Boston, MA 02115
617-432-0095
Fax: 617-432-0190
E-mail: rhomberg@hsph.harvard.edu

Woodrow Setzer*
Mathematical Statistician
Biometry Branch
Research and Administrative Support Division
National Health and Environmental
Effects Research Laboratory (MD-55)
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711
919-541-0128
Fax: 919-541-5394
E-mail: setzer.woodrow@epamail.epa.gov

*Presenter

Robert Sielken, Jr.
President
Sielken, Inc.
3833 Texas Avenue - Suite 230
Bryan, TX 77802
409-846-5175
Fax: 409-846-2671
E-mail: sielkeninc@aol.com

Thomas Starr
ENVIRON International Corporation
7500 Rainwater Road
Raleigh, NC 27615-3700
919-876-0203
Fax: 919-876-0201
E-mail: tbstarr@interramp.com

APPENDIX B

WORKSHOP AGENDA

United States
Environmental Protection Agency
Risk Assessment Forum

Benchmark Dose Peer Consultation Workshop

Holiday Inn Bethesda
Bethesda, MD
September 10-11, 1996

Agenda

TUESDAY, SEPTEMBER 10

8:00AM	Registration	
9:00AM	Welcome and Introduction	Carole Kimmel, National Center for Environmental Assessment (NCEA), EPA, Washington, DC
9:15AM	Workshop Structure and Objectives	Workshop Chair: Rogene Henderson, Inhalation Toxicology Research Institute (ITRI), Albuquerque, NM
9:35AM	Discussion of Simulation Studies	Woodrow Setzer, Jr. National Health and Environmental Effects Research Laboratory, EPA, Research Triangle Park (RTP), NC
9:55AM	Benchmark Dose Software Development	Daniel Guth, NCEA, EPA, RTP, NC
10:15AM	BREAK	
10:30AM	Selection of Studies and Responses for Benchmark Dose/Concentration Analysis	Discussion Leader: James Olson, State University of New York, Buffalo, NY
11:30AM	Selection of the Benchmark Response Level	Discussion Leader: Elaine Faustman, University of Washington, Seattle, WA
12:30PM	LUNCH	
1:30PM	(Continued) Selection of the Benchmark Response Level	Discussion Leader: Elaine Faustman

TUESDAY, SEPTEMBER 10 (CONT'D)

2:15PM	Model Selection and Fitting	Discussion Leader: Colin Park, The Dow Chemical Company, Midland, MI
3:15PM	BREAK	
3:30PM	(Continued) Model Selection and Fitting	Discussion Leader: Colin Park
4:30PM	Observer Comment Period	Facilitator: Rogene Henderson
5:00PM	Day One Closing Remarks	Rogene Henderson
5:15PM	ADJOURN	

WEDNESDAY, SEPTEMBER 11

8:30AM	Use of Confidence Limits	Discussion Leader: Lorenz Rhomberg, Harvard School of Public Health, Boston, MA
9:30AM	Selection of BMD/C to Use as the Point of Departure	Discussion Leader: Bill Pease Environmental Defense Fund and University of California, Oakland, CA
10:30AM	BREAK	
11:00AM	General Issues	Discussion Leader: Rogene Henderson
12:00PM	Observer Comment Period	Facilitator: Rogene Henderson
12:30PM	LUNCH	
1:30PM	(Continued) General Issues	Discussion Leader: Rogene Henderson
2:30PM	Closing Remarks/Chair's Summary	Rogene Henderson
2:45PM	ADJOURN	

APPENDIX C

CHARGE TO WORKSHOP PANEL MEMBERS

United States
Environmental Protection Agency
Risk Assessment Forum

Benchmark Dose Peer Consultation Workshop

Holiday Inn Bethesda
Bethesda, MD
September 10-11, 1996

CHARGE TO REVIEWERS

Our overall goal in developing this document is to have a procedure that is usable, that has reasonable criteria and defaults to avoid proliferation of analyses and model shopping, and that promotes consistency among analyses. Ultimately, we are trying to move cancer and noncancer assessments closer together, using precursor and mode of action data to extend and inform our understanding of risk in the range of extrapolation. We would like to have in one package something that is usable for cancer and noncancer assessments when endpoints are relevant to both.

Please review the technical points below. As you are preparing your technical comments, we would also like your advice on how best to achieve our goals as stated above. This should take the form of further points to be developed in the document or issues that should be clarified.

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

1. Selection of Studies and Responses for Benchmark Dose/C Analysis

   a. Is the selection of studies and endpoints for the BMD/C appropriate? for cancer? for noncancer?
   b. Should these be the same for cancer and noncancer data?
   c. Are there appropriate criteria for determining when data should be combined for analysis?

2. Selection of the Benchmark Response Level

   a. Is the use of biological significance or limit of detection an appropriate basis for the selection of the BMR?
   b. For the limit of detection, is the approach proposed in the document appropriate?
   c. Is information available to determine the appropriate power level? (Information on current simulation studies will be presented at the workshop.)
   d. Is the default for quantal and continuous data appropriate?

3. Model Selection and Fitting

   a. Is the order of model application for continuous and dichotomous data appropriate?
   b. Should other models be considered, or should the number of models applied be more restrictive?
   c. Are the parameters proposed as defaults for model structure appropriate?
      i. What should be the default approach for selecting the degree of the polynomial to use?
      ii. Is the default of not including a background parameter appropriate unless there is some indication of a background response level?
      iii. Is the use of extra risk as a default for quantal data appropriate?
      iv. Is the default of not including a threshold parameter appropriate?
      v. Is the default of modeling continuous data as such appropriate?
   d. Is the approach for determining the fit of the model appropriate? Are there additional or alternate criteria that should be used?

4. Use of Confidence Limits

   a. Should the lower confidence limit on dose be the definition of the BMD/C?
   b. Are the defaults for the method of confidence limit calculation appropriate?
   c. Is the default of 95 percent confidence limit appropriate?

5. Selection of the BMD/C To Use as the Point of Departure for Cancer and Noncancer Health Effects

   a. Comment on the determination of "equivalence" of models.
   b. Comment of use of the Akaike Information Criterion for comparing the fit of models.
   c. Is the default approach for selecting the BMD/C to use as the point of departure for cancer and noncancer dose-response analysis appropriate?

6. General Issues

   a. The discussions concerning the use of BMD/C approach in cancer and noncancer risk assessment.
   b. How understandable the document is for the general toxicologist/risk assessor.
   c. The overall organization of the document, further points to be developed or needing clarification.
   d. The examples of BMD/C analyses in Appendix D.

APPENDIX D

PREMEETING COMMENTS

United States
Environmental Protection Agency
Risk Assessment Forum

Benchmark Dose Peer Consultation Workshop

Holiday Inn Bethesda
Bethesda, MD
September 10-11, 1996

Premeeting Comments

PREMEETING COMMENTS

CONTENTS

Peer Consultants' Comments **Page**

Bruce Allen

REVIEW COMMENTS ON EPA'S BENCHMARK DOSE TECHNICAL GUIDANCE DOCUMENT

Bruce C. Allen
ICF Kaiser

The entire process of a risk assessment that potentially involves BMD calculation can be summarized as follows:

Step 1: Selection of appropriate studies and endpoints for use in the risk assessment.

Step 2: Determination, for each selected endpoint, whether a BMD estimate can and should be derived. If not, an alternative value may be determined.

Step 3: Calculation of the BMDs desired.

Step 4: Interpretation and use of the BMDs (or alternative values when BMDs have not been calculated).

The Agency has laid out clearly and succinctly (bottom of p. 11) the reasons why one would want to move away from complete reliance on NOAELs and LOAELs. In addition, the comment on p. 10 (lines 26-29), to the effect that a NOAEL and LOAEL characterize only one particular study (and even then, only a relatively small portion of the study) is a very important consideration. The guidance for application of the BMD approach should be judged in light of how well it appears to promote BMD analyses that improve the process of risk assessment, i.e., how well it eliminates or decreases the problems that have been identified with use of NOAELs.

In general, the guidance provides a reasonable and rational way of proceeding with BMD analyses. There are some particular restrictions and default choices that I would not have imposed, and there are some areas where more explicit guidance may need to be provided. These are presented and more fully discussed below, but my overall impression is that significant thought has been given and care has been taken in the development of the guidance.

My first concern relates to the comments in the introduction (p.12 line 5) that a BMD/C that is estimated will always (or should always) be in the observable range. Many situations will arise, and I believe the guidance does not necessarily rule these out, where a meaningful and useful BMD/C can be determined that is less than the doses used in the study from which it is derived. In fact, the boric acid example in Appendix D shows a case (for Study A of that example) where a BMD was less that the lowest positive dose (which happened to be a LOAEL by traditional thinking) when fetal weight was considered. Since one point of that example should be that the use of the BMD/C approach can obviate the need for additional testing when a NOAEL is not obtained, the statement on p.12 about the BMD/C being in the observable range appears to be inappropriate.

Similarly, on p. 17, following the definition of the BMD/C, the "requirement" that "at least one dose be near the range of the response level for the BMD/C" (lines 17-18) should not be a requirement at all. Clearly, there will be less uncertainty about the value of the BMD/C when that is the case (and that reduction in uncertainty will be reflected in tighter confidence limits used to define the BMD/C), but to make it a requirement that there be an experimental dose that gives a response about equal to the BMR would be too restrictive. In fact, if using a NOAEL or LOAEL is the alternative when this requirement is not met, this will lead to less consistency among points of departure, because in that case we know that there will not be a dose level to choose that will approximate the response level of interest. The advantage provided by dose-response modeling and

estimation via that modeling of doses that are associated with some predefined level of response is lost under this scenario.

Statements throughout the document (e.g., p. 17 line 12) that the BMD/C is not dependent on the doses used in a study should be toned down. Whereas, generally speaking, the choice of dose levels should have little effect on the estimation of the dose-response relationship overall, the choice of the doses will have some effect on the calculation of the bounds. It is precisely because the bounds reflect uncertainty about the dose level associated with a particular response, and that uncertainty depends on what response levels have been observed, that the BMD/C approach using lower bounds on dose is so powerful. Confidence limit calculations provide a natural way for one to express uncertainty about the value of the parameter (BMD/C) of interest.

With respect to the Data Array Analysis - Endpoint Selection process, more guidance may need to be given for cases in which there are a larger number of studies or endpoints considered relevant. In particular, how would one identify redundancy among endpoints (p. 18, line 29)? How does one know when one endpoint "represents others for the same target organ" (p. 19, line 4)? Moreover, it is not clear to me that a "smoothly increasing response" that allows good fit can or should be the driving factor for endpoint selection at this stage. Dealing with fit problems is an important consideration, but until the modeling is completed it may be difficult to determine whether good fits can be obtained. It is not clear that fit is an important consideration at the stage of endpoint selection.

Have there been any studies or work done to support the claim that having LOAELs differing by a factor of 10 (p. 19, line 13) will insure that the "critical" BMD/C will not be missed?

The subheading "1. Selection of Endpoints to be Modeled" can probably be eliminated. What needs to be emphasized is that the first stage of selection should be based on relevance, good experimental protocol, etc., without regard to the ability to derive BMD/C estimates. Secondarily, one wants to reduce the number of endpoints that need to be considered (redundancy, representativeness, sensitivity), and some of the endpoints chosen then may still not be amenable to dose-response modeling. Finally, one must pick which endpoints that remain can be modeled (data set requirements) and what one will do with the remaining ones that are considered relevant, representative, and potentially sensitive.

So, the section on minimum data set requirements (p. 19) should, first of all, include material from Appendix A about the data needs (or at least explicitly reference Appendix A here with a strong encouragement to look closely at the needs). Then the caveats about when one should not do dose-response modeling and estimate a BMD/C even when the data needs are satisfied can be provided in this section as well.

However, that being said, I do not think that all of the restrictions or constraints imposed (bottom of p. 19) are appropriate. I think a better way of characterizing the constraints would simply be to state that dose-response modeling should be done only when there is evidence of the shape of the dose-response relationship, if one exists. This would cover all of the really problematic cases that should be included in the list of constraints, but it leaves open the possibility of (appropriately) doing BMD/C estimation for some cases that would be excluded as the constraints are now stated. For example:

The statement of the first constraint (line 19, p. 19) is not very clear. For one thing, does a biologically but not statistically significant LOAEL count here? In general, what are the criteria

by which a LOAEL is determined to exist (authors statement, pairwise tests, trend tests?). If this statement implies that one should not model data sets for which a LOAEL was not actually observed (as opposed to the possibility that the number of doses and subjects could not have given a LOAEL because of design limitations), then this is equivalent to saying there is no evidence of dose-response for the data set under consideration. I would tend not to put as much emphasis on the existence of a LOAEL per se -- it is so dependent on sample size for one thing -- but rather the evaluation that something biologically and toxicologically "real" is happening. I can think of cases where a LOAEL may exist, but because of a large number of animals being tested, the differences that are statistically significant are not biologically meaningful and may just reflect differences that will inevitably exist among finite groups of "observations," even when those groups were drawn from the same population. On the other hand, the lack of a LOAEL may be caused by a small number of animals; perhaps in some cases lack of statistical significance may exist but it might be considered that evidence of a dose-response relationship may evident. I would explicitly allow consideration of other parts of the data base (other studies and/or other related endpoints for the chemical under consideration -- perhaps ones that are not being considered for modeling because of basic data deficiencies but which still carry information about dose-related effects; I would even consider related chemicals with similar effects) in order to make a "holistic" appraisal of the existence of dose-response relationships.

The consideration of the presence of dose-response relationships would also rule out modeling data sets with only one positive response level (lacking the ancillary evidence described in the preceding paragraph). A single positive response level (even when other doses have been tested and have exhibited background-level responses) should not be considered to provide evidence about the shape of the dose-response relationship and would therefore not be modeled.

On the other hand, the existence of only high levels of response (criterion starting at line 25 of p. 19) should not necessarily preclude modeling. There may be many cases in which responses are all above 50% but one still gets a clear picture of a dose-response relationship (what about cases where background starts out near 50%?). I would consider modeling those -- the uncertainty concerning the BMD/C corresponding to a lower level of response may be greater than might otherwise be the case, but that is covered by the calculation of confidence limits. Later (in the step when one interprets and chooses from among various BMD/Cs) this uncertainty may dictate that another BMD/C is used for regulatory purposes, but a BMD/C calculated from such data will carry information relevant to the final decisions to be made. If all responses are above 90%, however, then less (perhaps next to nothing) might be said about dose-response shape and such cases could be ruled out. The same would hold if there only existed a plateau of continuous responses (and no ancillary information).

The bottom line here is that it is the presence of evidence suggesting the existence and general shape of dose-response relationships that allows one to feel comfortable fitting dose-response models to the data set(s) under consideration. When that evidence is present, then the appropriate modeling can be done. When it is absent, then the constraints on the modeling are not well-defined and modeling should not be pursued.

Although this may require some additional elaboration, one might consider using a trend test on the positive doses (i.e., exclude the controls) as a way to determine if there exists sufficient information about a dose-response relationship for modeling and BMD/C estimation to be done.

There is a definite need to consider data combinations (p. 20). However, more guidance may be required for users to know what constitutes biological or statistical compatibility. Nevertheless,

the point (lines 5-7) that additional research can affect the BMD/C estimates (including increasing them, unlike a NOAEL estimate) is an important one that deserves to be emphasized.

Concerning the selection of the BMR level, I have one peripheral comment first (and this relates to many similar occurrences throughout the document). There is a discussion of biologically significant changes, with body weight as an example. One needs to be very careful with such examples, to be explicit about what it is that is being measured and considered to be biologically significant. In the case of body weight, a 10% change is suggested as a biologically significant change. Is this 10% change in the mean values of different groups (this interpretation is suggested on p. 58, line 12) or an individual drop in body weight that is 10% below an average (unexposed) level? To see what difference this could make, consider a test group that had 50 animals with 25 of them having body weights of 90g and 25 with body weights of 95g, whereas the controls averaged 100g. As a group the treated animals average 92.5g (only 7.5% below the controls and so not biologically significant?) whereas, individually, 50% of the animals exhibited a 10% decrease in body weight. The latter appears to be a serious effect if 10% decrease on an individual basis is important. The implications for setting BMRs are important also. If biological significance is on a group average basis, then the BMR should be defined in terms of changes in the mean value [(m(0)-m(d))/m(0) = .10]. On the other hand, one might want to set the BMR in terms of a relatively low probability (10%) that individual animals will experience a body weight that is 10% lower than background. In the former case (based on average change), the treated group in the example would not appear to have reached the BMR level. In the latter case (individual basis) the group would be considered to have greatly exceeded the BMR response (50% of the group members had 10% lower body weight!) and the dose for that group would appear to be well above the BMD (or even the MLE for the selected BMR).

Secondly, if the BMD is intended to replace a NOAEL, then does this assignment of the BMR to the biologically significant (LOAEL-like?) response tend to overestimate the NOAEL-like BMD? The same concern might apply to the BMRs based on detection limits, except that the relatively low power required (50%) might lessen the concern there.

It should be recognized (and perhaps explicitly stated somewhere in the document) that the decision to base BMRs on detection limits carries with it some implicit acknowledgments. First, that responses of a certain magnitude have no chance of being considered BMRs because of the sample sizes, background rate (or variability), and power choices made. This implies (as it has always done for NOAELs anyway) that one finds acceptable the high likelihood of "missing" changes of such magnitude. Have people considered sufficiently the basis for the determination of the standard sample sizes so that the Agency is willing to make such a statement? Second, when larger sample sizes than the standard are used, it is quite possible that the BMDs that result will be greater than LOAELs from those studies. The Agency needs to be willing to go on record as supporting use of such BMD estimates and not defaulting back to a LOAEL or NOAEL just for the sake of conservatism.

Because the manner in which the detection limits and BMRs are to be derived can only fix the power (50% by policy choice) and the sample size (standard size) *a priori* but the background rate may not be as well-determined, would the BMR be allowed to vary according to the observed background rate in any particular case? Would it be species and strain dependent and would historical control data be used to define it? More needs to be specified concerning the choices for background rates in the BMR derivations.

A set of tables or graphs that gives detection limits as a function of background rate and sample size might be very useful.

Starting in section III.C.2 (p. 23), issues of model fit are emphasized. It is important, therefore, to be clear about what constitutes good fit and how it will be measured. It is not clear that the standard techniques for assessing fit for continuous variables (e.g., F-tests) are adequate. Some computer-intensive (simulation) approaches can be used and might be recommended (or built into EPA's software). Such an approach would also alleviate another problem that is sometimes encountered, i.e., having no degrees of freedom available for formal, traditional tests. The simulation-based fit assessments do not need spare degrees of freedom, and I do not see a strong need to limit the flexibility of the models used to fit data just for the sake of getting those degrees of freedom.

Furthermore, the standard procedures for determining fit of continuous models look only at the predictions of the means as compared to the observed means. They do not directly consider the prediction of the variability. Yet, the estimates of variability are very important for BMD/C estimation to the extent that the BMRs are based either on (1) changes in the mean relative to the underlying variability or (2) a hybrid approach that depends on the predicted distributions around the mean values to derive probabilities of response.

It appears that the agency needs to do a bit more development and make some decisions about fit issues so that the guidance can be clear and explicit about how good or adequate fit will be determined.

Some questions/comments about the order of model application (p. 23, lines 12-20):

If a linear model is run first for continuous data, why not also for dichotomous data?

Why pick the polynomial model to run prior to the power model?

Do the choices for the continuous data models extend to the use of the hybrid approach? If so, then the Weibull model (which has the added advantage that it is the same model that can be applied to quantal data) should also be considered and explicitly listed.

Rather than picking an order for application, one might suggest the 2 or 3 models that should be considered and that they all be run initially. This does not put such a burden on the assessment of fit -- a barely acceptable linear model might be substantially improved by adding nonlinear terms, but this would not be determined in the step-wise procedure that is now specified. The standard goodness of fit assessments are not really very good at discriminating between model alternatives like that anyway.

With respect to the model structure (pp 23-25), the following comments are offered:

When (and if) one allows exponents on dose (and related parameters) to be less than one, some procedure should be described whereby the instabilities can be lessened. One way we have investigated is to do the fitting first with unrestricted exponents. Then, the exponent that is returned as the maximum likelihood estimate is used as the lower bound on the exponent for a second iteration, and it is only in the second iteration that lower bounds on dose are derived. Even this does not always eliminate some very small values for the lower bound.

I would recommend eliminating the restriction and discussion of the background parameter. I can think of no good reason for making a zero background the default. It is much easier to make the default be the inclusion of the background term and only allow it not to be estimated if the biological or toxicological data suggest that that is appropriate.

Similarly, I would allow consideration of the threshold parameter, even though it does not correspond to a biological threshold. Call it something else if desired, but there are instances where its inclusion is essential for obtaining adequate representation of the dose-response relationships. If alternative fit assessment procedures are implemented (see above) the loss of the degree of freedom is not crucial. Moreover, by explicitly allowing that parameter, the class of dose-response functions considered is increased. This is important because then the bounds that are calculated and which constitute the BMDs represent the uncertainty in the dose estimation for the larger family of curves. The resulting BMDs, even if they do not correspond to a curve that includes a threshold parameter, were allowed to have it if the optimization dictated that it was needed, and therefore one avoids criticism that the lower bounds are model-dependent in a way that excludes "threshold-like" behavior.

The section on selecting the BMD/C (pp. 26-27) needs to be substantially altered. First, there is a mixture of issues here that is not clearly delineated. The first issue is what to do with different BMD/Cs for the same endpoint (and study) resulting from different model predictions. The second issue is what to do with different BMD/Cs from different endpoints and/or studies. The first issue is adequately addressed by the first two bullet items in this section, although I would emphasize examination of the MLEs much more than has been done. The second issue is not addressed at all, except to the extent that one can infer from the discussion of the NOAEL/LOAEL and BMD/C comparison that the lowest BMD/C would be selected.

Even more important, that discussion of what to do with NOAELs/LOAELs when they look to be the most sensitive is not adequate. Especially because the guidance lays out several cases where BMD/C derivation should not be done, it is important to rethink how NOAELs and LOAELs can and should be used, and not just rely on old concepts. The whole idea is to get away from the problems of the NOAEL, not to exaggerate them by mixing them up pell-mell with BMD/Cs. As an example, if the critical effect was from a large study, and the LOAEL was a LOAEL because of the large sample size even though the observed response rate was less than the BMR, why would one choose that LOAEL (or the corresponding NOAEL) as the point of departure? The basic problems associated with use of NOAELs still plague this guidance if the last bullet item on p. 27 is all that is said about use of non-BMD/C results.

I have serious doubts that BMD/C approaches will prove to be very useful for cost-benefit assessments and I completely disagree with the statement that the BMD approach "provides a good starting point to develop benefits estimates for non-carcinogens" (p. 42, lines 22-23). NOAELs are no better, but this is not an improvement that is provided by BMD-like analyses.

The discussion on p. 57, lines 4-8 does not make sense to me. What is the point here?

On p. 59, line 1, do the authors mean "refined" or "defined." The interpretation of the level of effort needed before one can use the limit of detection approach for setting BMRs may depend on that distinction.

I also think that the statements about the low background rate in the EPA-sponsored work on developmental toxicity (p. 62, lines 1-3) is incorrect. For many of the endpoints that included resorptions, the background rate was quite large. If one considers the analyses of the developmental toxicity data sets that have been done to be the empirical equivalent of the limit of detection calculations proposed, then it appears that the results of that analysis should be interpreted as suggesting use of 5% additional risk when doing the recommended assessment of

such data. The Agency should explicitly state that, unless other comparable analyses of developmental toxicity data become available, the current information supports using 5% additional risk for developmental toxicity data, especially in light of statements that the choice of extra or additional risk is unimportant when a limit of detection approach is used.

The discussion on p. 68, starting with line 23 and continuing to p. 69, is not at all clear. Although lack of independence is a problem that needs to be (and has been) addressed in certain instances, what does the statement about "choice of the model form" mean?

On p. 72, lines 9-13, the discussion needs some attention. It is not that nonmonotonic data mean that typical models can not be used, it is only that the fits might suffer because of the nonmonotonicity. Careful consideration of the reasons for nonmonotonicity should be recommended. Even still, log-transforming doses will not do anything about nonmonotonic responses nor would it help much with abrupt increases in response.

The description of the figure on p. 79 may need some work. And why are the BMDs from the figure different from any of those in Table 2?

Other general comments include the following:

There is insufficient attention paid to pharmacokinetics and the use of "delivered dose" estimates in BMD analyses. The guidance should strongly encourage the use of such dose estimates for BMD derivation and make note of the fact that the dose conversions (for the test species) should be done prior to modeling, with "back-calculation" of human exposures associated with delivered dose versions of BMDs completed after the modeling and BMD estimation. It has been found (with vinyl chloride for example) that model fitting difficulties were largely resolved when appropriate delivered dose estimates were used. [As a minor point, it might be noted that the dose scalings referenced on p. 44 line 27 are for oral exposures -- inhalation concentration scaling is done using HEC considerations, right?).

The move to BMD approaches also offers and excellent opportunity to reassess the use of uncertainty factors (UFs). UFs have been touched on in the document, but more explicit and extensive discussion of how they might be considered and re-evaluated in light of the BMD/C method could be added.

I think that the document should include a flow-chart showing how the various considerations (mechanistic information, data set requirements, modeling constraints, etc.) come into play and direct the course of a noncancer risk assessment. What this document should be trying to do is describe the information inputs that point one in the direction of the type of analysis (BMD or otherwise) that ought to be done. A BMD analysis using the standard set of models, the default choices for BMRs, etc. may be the default of last resort, so to speak, in that one would do it only if other options that include more chemical-specific and mechanistic inputs can not be implemented or do not appear to be justified. A flow-chart would be very useful in summarizing that process.

George Daston

George P. Daston
August 27, 1996

Comments on EPA's BENCHMARK DOSE TECHNICAL GUIDANCE DOCUMENT

My overall impression of this document is that it is a useful how-to manual on the application of BMD. It, along with the Risk Assessment Forum report entitled "The Use of the Benchmark Dose Approach in Health Risk Assessment" should be sufficient for toxicologists in the program offices to successfully and correctly apply this method for risk assessment. It is also worth noting that this document does a fine job of continuing EPA's efforts in harmonizing risk assessment for cancer and other forms of toxicity. My most significant suggestion for the application of BMD is that it be based on a central estimate of the BMR instead of on a lower confidence limit (LCL). While I agree with the use of a confidence limit in principle, its use becomes problematic when one tries to make comparisons across different toxic endpoints that are evaluated using study designs of varying group sizes, and varying statistical power. Using a central estimate will 1)make better use of the one area along the dose-response curve that we can model with some precision; and 2) facilitate the comparison of critical effects for different endpoints. It should be possible using the central estimateto have a single default level of response for the BMR, rather than the sliding scale "limit of detection" approach. My suggestion is fleshed out in my response to question 4. a below.

My answers to the specific questions are:

1. a. The selection of studies and endpoints are, and should be, the same as for NOAEL-based risk assessment. These studies, run according to regulatory guidelines, are widely regarded as satisfactory apical tests to detect hazards of all sorts. These studies should therefore be adequate bases for risk assessment, regardless of the method employed. However, the advent of the BMD should cause the Agency to suggest greater flexibility in study design, particularly in the number of dose groups and animals/group. It may be possible to design studies that better define the shape of the dose-response curve, especially its lower end, better than the standard 2-3 dose groups plus a control.

It is worth making explicit the distinction between endpoints that are dichotomous or quantal by nature (e.g., alive or dead, 5 or 6 fingers) than those that are quantal by fiat. The latter are best exemplified by the classifications of mild, moderate and severe that are widely used in histopathology. While these classifications are useful in providing the opinion of experts as to the severity and adversity of a finding, they obscure the fact that the observed responses are in reality part of a continuum. There may be some utility in de-quantalizing these types of data for the purposes of BMD-based risk assessment, particularly given the Agency's interest in using precursor and mose of action data to help understand the degree of risk in the range of extrapolation.

1. b. There is no reason to make a distinction between cancer and non-cancer endpoints.
1. c. These criteria are appropriate, and the example provided in Appendix D is a good one.

Other comments on Selection of Studies and responses: It appears that the guidance document indicates that the critical effect be selected as simply the lowest BMD. This is connoted by the statement that all endpoints whose LOAELs are within an order of magnitude of the lowest LOAEL shouldbe modelled. This seems to be inappropriate, as it does not take into account all of the other information that is part of expert judgement, such as the plausibility of the effect, its severity in comparison to other sensitive effects, etc., as well as the other information such as slope of the dose-response curve, that comes along with the calculation of the BMD, and may be very informative as to which effect should be selected as the basis for RfD calculation.

The first of the three criteria for a minimum data set (bullet points on p. 4 and p. 19) does not make sense to me. One of the real advantages of the BMD approach is that it would allow one to use a study that is statistically insufficient to generate a credible NOAEL or LOAEL but still conveys enough information such that there is clear evidence of a hazard. The second criterion in this section is also too restraining. While it is true that the choice of a mathematical model for a data set with only one positiveresponse group is arbitrary, it is no more arbitrary than the a priori choice of the dose level that ultimately becomes the NOAEL for the study. Given that the guidance for cancer risk assessment suggests a straight line as a default for low dose extrapolation, it seems appropriate to suggest a similar default for data within the experimental dose range.

2. a. For continuous variables, the biological significance determination is appropriate. The limit of detection approach is appropriate for those endpoints for which there is general agreement that some level of effect on that parameter is adverse, but there is insufficient information or consensus to pinpoint a specific level. In those instances, it seems to me that an 80% statistical power would be more comparable to currently employed limits of detection than a 50% level. For those continuous variables for which there is no generally agreed upon interpretation regarding adversity, it would be my opinion that these not be used for risk assessment, although they may still be useful as auxiliary information. For quantal endpoints, the decision as to whether something is adverse should also be made a priori and should be contained in the guidance given in regulatory risk assessment guidelines. I find a consistent level of response as the default for these endpoints to be more appealing than the limit of detection method, as this will facilitate comparison across endpoints.

2. b. It looks OK, although as noted in the response to 2.a. it is not my preferred option for quantal endpoints or some continuous ones.

2.c. The results of the simulation should prove useful. Others have made calculations on limits of detection based on the CVs for various endpoints commonly measured in screening studies with sample sizes recommended by regulatory testing guidelines. These may also be a good source of information.

2.d. See my response to 2. a, particularly regarding the need to first determine whether a response is adverse.

3. a. The order of model application is satisfactory. The log-logistic model for quantal data has been demonstrated to be flexible enough to handle most dose-response curves, and does not have the problems of the Weibull model in fitting the lower end of dose-response curves with very steep slopes. The recommendation that the curve from each model be graphically displayed and critically reviewed for its relevance to the data, especially the lower end of the dose-response, is appropriate and cannot be stressed too much.

3.b. The guidance document provides some flexibility in choosing additional models on an ad hoc basis as long as the choice is explicitly justified, so there is little need to include additional models. There is also no good reason to restrict further the number of models, at least until more experience is gained on the behavior of the models for a variety of toxic modalities.

3.c. These defaults apprear to be appropriate and are in line with what was recommendede by an expert group at the EPA/AIHC/ILSI workshop on benchmark dose. The only point that is not supported by that working group is the choice of excess risk over additional risk, a point on which that group could not reach consensus. The explanation for the decision not to include a threshold term is very well put: there is no relationship between this arbitrary contrivance and a biological threshold.

3. d. The approach is adequate, and as noted above, it is an excellent recommendation that the curve from each model be graphed. It should be stated in stronger terms that the exclusion of high dose data should be a last resort if none of the models fit. A preferred alternative would be to select other models that are not on the short list of recommended defaults. Furthermore, prior to excluding data, all of the data points should be graphed in a scattergram as an aid in determining the possible causes of lack of fit of any model (e.g., extreme non-monotonicity of the data).

4. a. EPA should consider using central estimates instead of lower confidence limits in calculating the benchmark dose, especially for data from studies that are conducted according to accepted regulatory guidelines. There are several reasons for this recommendation. First, one of the main reasons for using a LCL was to penalize studies with low sample size or other statistical deficiency as compared to standard guideline studies, which are widely regarded as adequate to detect hazard. However, as long as the studies that are used for BMD calculation meet the requirements of the guidelines, this reason for relying on confidence limits is obviated. Other means can be employed to handle substandard studies, such as reliance on confidence limits for those that do not meet the minimum requirements of the regulatory guidelines, employing additional uncertainty factors, or simply not considering them as the source of the critical effect for risk assessment. Second, use of the central estimate is a better, more precise use of the experimental data. These data represent the area of the dose-response relationship of which we are the most certain It is for this very reason that the potency comparisons in cancer risk assessment rely on central estimates of the TD10 rather than a LCL. Why then would we wish to arbitrarily

discard this small shred of certainty in an otherwise uncertain process? The third, very pragmatic reason for relying on a central estimate is that it greatly facilitates comparison of different endpoints. For a variety of reasons, regulatory guidelines for the detection of hazard of different endpoints rely on various numbers of animals per group and have different statistical power. A good example is the difference between developmental toxicity studies, where a 3-5% increase in risk for malformations or resorptions may be statistically discernable, vs. a neurotoxicity study where it may take a 30-40% decrement in a clinical parameter before it is discernable. It is clear that the consensus of the neurotoxicology community is that this design is satisfactory to detect a hazard; however, neither the NOAEL-based or BMD (calculated as a LCL) approach ould be adequate to assess risk from these studies. The former would be far too insensitive, and the latter would be overly sensitive. Furthermore, neither would be easily comparable with the developmental toxicity results. If we were to evaluate a set of chemicals that were equivalently neurotoxic and developmentally toxic using the NOAEL approach, all of the RfDs would probably be based on developmental toxicity; using the BMD, all would be based on neurotoxicity. This does not make sense. Therefore, I recommend that the BMD be calculated as a central estimate, and that other steps be taken to account for study insufficiency, etc. This recommendation also makes it easy to select a consistent level of response as the BMD for quantal endpoints, across all forms of toxicity. I find this to be preferable than the sliding scale approach that is now being taken in the Guidance Document.

4. b. If confidence intervals must be calculated, these defaults seem OK, at least to this non-expert.

4. c. As noted above, I do not advocate using confidence limits for studies that meet regulatory testing requirements. Should a confidence interval be used, I suggest that 1-1.5 standard deviations would be adequate.

5. a. These determinations of equivalency appear satisfactory.
5. b. I have no knowledge of the AIC.
5. c. Yes.

6. a. The EPA is to be congratulated for its continuing attempts to harmonize cancer and non-cancer risk assessment. The use of the BMD is one way of moving toward that goal. I think, however, that it should be recognized that the BMD is not a credible basis for low-dose extrapolation for either endpoint. While this is acknowledged for non-cancer endpoints, it is not for cancer. Statements like that on p. 11, line 18 need to be rethought and either qualified or removed.

6. b. The document appears to be right on target for the intended audience.

6. c. The organization is satisfactory.

6. d. The examples in Appendix D are excellent. I suggest that the remainder of the chemicals in IRIS for which the RfD was derived from a BMD also be included in Appendix D for illlustrative purposes.

Elaine Faustman

Elaine Faustman

School of Public Health and Community Medicine
Department of Environmental Health-Roosevelt
University of Washington
4225 Roosevelt Way NE, #100
Seattle, WA 98105-6099
Phone: (206) 543- 9711 FAX: (206) 685-4696

Comments on the USEPA Draft Benchmark Dose Technical Guidance Document (EPA/600/P-96/002A).

August, 1996

Page vii , line 4 & Page 3, lines 12-14. The first paragraph of document states that it is to be used in conjunction with EPA 1996c document. For usability, the more "stand alone" this document is, the easier it will be to use.

Page 3-4, lines 28, 29 and 1. Has the EPA conducted studies to determine that only LOAELs within 10X of other LOAELs need to be evaluated for BMD analysis? Is it true that no BMD/Cs would be less than 10 fold from the LOAEL. Add references or detail in later section.

Page 4, lines 2-14. Other criteria that could be evaluated, includes guidance on what to do with non-monotonic dose response relationships.

Page 4, lines 11-14. USEPA should explain in detail how they determined the criteria used in bullet item 3, under question 2. This reviewer had difficulty in determining how curves with all responses above 50% would all automatically be inappropriate to model. Additional criteria may be needed in this bullet item. For example, perhaps specifying maximal percent response change per doses evaluated or ratios of dose spacing compared to response change.

Page 10-12. This reviewer would suggest that the paragraph starting on Page 11 at line 21 should go on page 10 at line 30.

Page 11, lines 13-16. Important concepts are discussed here, yet it was unclear to this reviewer the rationale for choosing a linear default analysis. Will a further section answer this? More details need to be provided here, rather than just referring to other documents. [See comments on Page 16].

Page 11, lines 12-13. Please make the following changes in these lines:

page 11, lines 11-13.
".....with appropriate curve-fitting models; and then (2) extrapolation below the range of observation is accomplished by modeling if there are sufficient mechanistic data or approaches or by a default procedure (linear, nonlinear, or both) if no such models or mechanistic approaches exist."

Page 13, Figure 1. Is there a need to show the BMR and BMD at specific percentage response levels to clarify the figure? Also should the illustration show a BMR above the NOAEL, as well as below? Should the LOAEL be identified on this graph? Should statistical significance of data points be shown?

Page 14, lines 8-11. It is not user friendly to constantly refer to other documents that should be used in conjunction with this document. Write one benchmark document and provide enough details to be useful as a "stand alone" document.

Page 14, line 23. Add references to this sentence.

Page 14, line 29. What studies are referred to in this sentence? Only those by Faustman et al or is reference being made to earlier cited papers in lines 26-27.

Page 15, lines 12-14. These sentences give the impression that there was no biological rational for evaluating reduced fetal weight. These sentences should be modified to include this rational for these choosing these studies.

Page 15, line 14 & 17. Replace "cut off values" with response levels.

Page 16, line 9. This reviewer feels that caution should be used when mentioning the Baysian approaches because the only cited paper is not in a peer-reviewed journal. Please add additional specific references for this application or remove this concept.

Page 16, lines 12-18. Insert details from page 11, first and second paragraphs (lines 1-20) here. A brief paragraph without details (specifically details in lines 8-16) can be left on page 11 that discusses the loss of dichotomy between cancer and non-cancer approaches.

Page 16, lines 26-29 and page 17, lines 1-3. This willingness to continually incorporate new improvements in these processes would have more weight if the specific time of re-evaluation or the next re-evaluation was also specified or at least the process for re-evaluation was specified. (See also comments for page 40).

Page 17, line 17. How "near"?

Page 19, lines 13-14. Please provide rational for only looking at other endpoints if LOAEL is within 10 fold over the lowest LOAEL. Have studies been conducted that show that no other endpoints would result in lower BMD/Cs? For this reviewer, this point was not intrinsically obvious. (See earlier comments on Page 3-4, lines 28, 29 and line 1).

Page 19, lines 22-24. Specify what is done when only one responding group is present. Does the risk assessor use a NOAEL approach? Indicate what is done, not just what is not done.

Page 19, lines 25-28. This constraint needs to be explained. Why was 50% response chosen? This reviewer would suggest that more specific guidance could be given. (See earlier comments for page 4, lines 11-14).

Page 20, section 3 Combining data. Add reference here to peer-reviewed publication by Allen et al.

Page 21, lines 14-17. The guidance document should provide a few more details on what would be evidence of "biological significance". This reviewer would suggest listing example EPA risk assessment guidance for developmental toxicity here. Also, perhaps, referencing groups such as MARTA that publish guidance information. Would EPA accept "biological significance" only after peer review or consensus workshop concurrence?

Page 21, lines 18-22. Regardless of findings of simulation studies, this section needs to be expanded. Is an example given in Appendix B for each of these approaches? Adding a few more details here could help in understanding this approach. This was very confusing for reviewer.

Page 22, lines 1-4. Document should explain why "extra risk" should be used for BMR set on basis of biological significance. Also explain why it does not matter for limit of detection approaches. Need to add glossary so users truly understand extra versus additional risks.
Don't hide definitions in appendix.

Page 23, lines 7-8. Explain what "curve fitting in a manner similar to the EPA software" means. Reviewer needs additional details to understand these comments.

Page 23, lines 12-20. Provide a few details to justify the order which models are to be run on the data. This justification could be as simple as adding a few references or adding a few sentences that explain the order of model selection.

Page 23, lines 22-24. Again, this reviewer cautions the authors about referring to other documents for details that are needed in this document. Pull out key points and list here.

Page 24, lines 3-9. Add references to justify this approach.

Page 24, line 18. Add some examples of what additional information or "work" is needed.

Page 24, lines 19-20. To be user friendly, add complete thoughts here rather than just referring to other places in the document.

Page 24, lines 21-25. This reviewer agreed with the approach delineated for the "threshold" intercept term.

Page 25, line 25. Specify whether the EPA software will include this approach.

Page 25, lines 2-9. This reviewer would suggest that reference to studies showing loss of statistical powers should be added here. Also the reviewer would suggest that a simulation study would probably show a "reward" i.e. decrease in confidence limit and increase in BMD/C if data is kept as continuous data.

Page 25, lines 10-15. Authors should add a sentence or two that discusses likelihood theory.

Page 25, section 5. Where will the concept of "non convergence of models" be discussed? This reviewer suggests that this location might be appropriate.

Page 25, lines 24-29. This reviewer applauds the authors for their requirement of graphical displays of the data.

Page 26, lines 10-15. This reviewer cautions the dropping of "high doses" without producing some additional guidelines.

Page 26, lines 25-28. Additional details on the Akaike Information Criterion are need. Are these provided in the Appendix? If so, add note. This reviewer was surprised that no reference was given for this method. Authors must add peer-reviewed reference.

Page 27, lines 5-8. Add a few details on how risk assessors could use evaluation of the MLEs to determine patterns or add reference to example in appendix.

Page 27, lines 9-12. Author should specify what the risk assessor should do when there is a mixture of BMD/Cs and NOAEL/LOAEL values and the critical effect is a BMD/C

Page 28, lines 15-25. Good examples.

Page 29, lines 3-5. Would using a BMD/C approach possibly change the use of an extra 10 fold for inadequate experimental design (if the study could still meet the earlier criteria for allowing BMD/C use)?

Page 31, line 8. Insert the word "of" between the word "use" and the words "these approaches."

Page 31, Section V. This reviewer feels strongly that consistent nomenclature be used for all endpoints whether they are for non-cancer endpoints or cancer endpoints. Surely we can arrive at a consistent term for EDx, TDx, or BMDx.

Page 32, line 1. Please define lifetable methods and summary incidence methods.

Page 32, lines 3-9. Authors must provide some additional details here. The document describes these approaches for BMD/C calculations and it should also provide similar details for ED10s if this is really going to be a useful document. Identify and explain where there are differences in these approaches.

Page 33, line 18. Correct typographical error.

Page 34, line 17-18. Add details and reference to justify statement that "There might be modification other than DNA reactivity (e.g.. certain receptor-based mechanisms) that are better supported by the assumption of linearity."

Page 35, lines 26-27. Authors need to illustrate how the MOE analysis considers steepness of the slope. If these point are illustrated in the appendix, then please add reference here.

Page 36, lines 7-9. Authors need to explain what is meant by the statement that "... tumor data might support a greater MOE than a more sensitive precursor response...". Please give examples to illustrate what is meant by "greater MOE".

Page 36, lines 14-18. Authors again need to illustrate these points with examples and additional details.

Page 36, lines 19-28. Authors highlight problems in using different models for curve-fitting, yet they do not offer solutions. Will the new software that is being developed include multistage models for use? If so, cite here. How does the risk assessor resolve these differences between models? Authors must provide better, specific guidance.

Page 38, lines 21 and 22. What does this sentence mean? How can something be both more qualitative and quantitative? Explain.

Page 40, lines 3-12. See earlier comments about delineating a process for updating (page 16).

Pages 43-45. Authors do a good job at identifying research needs and inconsistencies that need to be addressed. Authors should describe a plan of how to address these critical needs.

Page 55, lines 15-26. Good discussion, but authors need to define what is meant by "poorest results". Does this refer to comparisons with NOAEL values, size of confidence levels, etc.? Please specify. Is this the reference that was used to set up criteria for acceptance of a BMD/C analysis? If so, please provide a few more details to substantiate these criteria.

Page 56, lines 14-17. Will the examples in the appendix show how continuous data is used when individual animal data are not available and only summary data with a measure of variability? Please do include in the examples

Page 59, lines 3 and 4. Do authors mean exposure versus dose in this sentence?

Page 59, lines 24-26. Why is the BMD/C : NOAEL ratio of one set as a goal? Please justify. Is the NOAEL set as a "gold standard" for comparison? This reviewer would suggest that this is inappropriate.

Page 60, lines 8-14. Add these definitions to the text as well, not just in appendix.

Page 64, Figure 3. Authors should prepare similar tables for standard experimental sizes for both cancer and non-cancer experiments. This would be especially interesting for the neurotoxicity behavioral study designs.

Page 65, lines 18-24. Are the "other requirements" for setting the BMR going to be developed by EPA? This reviewer would certainly encourage some additional agency work on this topic.

Page 66, lines 6-12. Please specify what are parameters a and b. What parameter is background in your equation?

Page 67, lines 4-6. Authors could give criteria for model selection.

Page 70, lines 2-8. Authors need to provide additional details on how to handle the identification of the "best" formats.

Page 71, line 9. Authors should explain what is a "correlation structure."

Page 71, lines 17-18. Will the EPA model package include a goodness of fit statistic program? This should be included.

Page 71, line 23. How large?

Page 73, line 13. Describe the likelihood ratio test and add a reference.

Page 73, lines 15-17. As this reviewer noted earlier, additional details and references on Akaike's Information Coefficient are needed.

Page 74, line 21. Add a few more details about the asymptotic normality approach for constructing confidence limits.

Page 75, lines 13-18. Could SAS macros be written and included as part of the EPA software?

Page 76-78. Authors need to provide additional details on how individual versus group data were used in the model.

Page 76, Figure 4. Was there a significant trend test for these data? What was the GOF statistic? Please add these details.

Page 78, lines 21-28 and Page 79, Figure 4. The explanation for Figure 4 needs improvement. A key to identify line types is needed and possibly a larger range of line styles maybe necessary to clarify responses. For example, line 23 and 24 refer to the lower solid line which this reviewer could not identify.

Page 80, Table 1. Add what incidence is evaluated to the table heading.

Page 81, lines 4-6. What provision is available in the guidelines?

Page 81, Table 2. Were these values obtained using Fleiss, 1981 approach? If so, please cite this reference.

Page 82, lines 6-8. How did the author assess the "excellent model fit?" What are the GOF statistics?

Page 83, Table 4. Authors should carry this assessment to a conclusion. Illustrate how the databases could be combined.

Page 85, Table 1. What are the units listed for fetal wt.?

Page 86, Table 2. What incidence is given in this table? Please label.

Page 91, Example 4. It appears to this reviewer that a Fleiss, 1981 based table for N=50 is needed to obtain the values in Table 1. If this is so, please add.

Page 92, line 14. Authors state that the data could not be adequately fit. How do the authors know this? Was the GOF statistic rejected?

Page 92, Example 4. Authors need to give "bottom line". What BMD/C will be used for risk management?

Page 94, Appendix E. Model development looks great but give increased indication throughout text what models and features will be included in the model package . Will any GEE approaches be included?

Appendix: In general, the examples were set-up well and this reviewer liked the inclusion of a summary of the main points that were to be illustrated in each example. Overall however, the examples need additional details, perhaps even example data input in a format that will be compatible with the model package that you are putting together. Also provide basic statistical information and trend analysis for each of these examples. Include statistics for GOF and likelihood ratio tests if applied. All examples should have NOAEL and LOAEL values given for comparison. It was not always clear to me that a risk assessor had enough details from these pages to identify and independently determine all of these values. Authors should insure that each example stands alone and no other resources are needed to do any of these calculations.

Glossary. This technical document needs a glossary. Key words that need to be included in the glossary (not all inclusive list):

Akaike Information Criterion
Asymptotic Normality Approach
Lifetable Methods
Likelihood Theory
Likelihood Ration Statistics
Goodness of Fit Statistics
Summary Incidence Methods

Jeff Fowles

Comments on the USEPA external review draft document: Benchmark Dose Technical Guidance Document. August 9, 1996. EPA/600/P-96/002A.

1) *Selection of Studies and Responses for Benchmark Dose/C Analysis.*

It is important to acknowledge that, particularly for acute non-cancer toxicity studies, the data are often comprised of small sample groups (i.e. 5 or 6 animals per group). This means that observing responses in the range of the BMR will not be possible for many of these studies. Therefore, it seems that the criteria for data used in BMD/C calculations for acute toxicity studies requires some flexibility. This concern would probably only apply to non-cancer data sets.

There are methods for determining appropriateness of combining data sets for analyses. One example of such an approach is given in the appendix under "combining data sets".

2) *Selection of Benchmark Response Level*

The use of biological significance for continuous data seems a reasonable approach. This approach would presumably supersede any lack of statistical significance if the data are expressed as dose-related changes in the mean. The transformation of continuous to quantal data for the analyses is a straightforward and logical process.

The explanation of the limit of detection needs further detail. As it is currently written, there is little actual guidance provided, and much is left to the risk assessor's understanding of relatively sophisticated statistics. For example, according to the guidelines, a power level must be chosen for each species and endpoint. The risk assessor must then decide on an appropriate incidence of detecting a difference from background (50% is given as an example in the document - is this a default recommendation?). How should one go about selecting this distinguishing incidence?

The authors should make some clarifications in certain other areas as well. On page 59, lines 25-26 there is a statement that there is a goal of achieving an "average BMD/C:NOAEL ratio of one." This "goal" has not been discussed in any previous portion of the document and should be explained.

The defaults for the quantal and continuous data, in general, seem reasonable. I did not understand the percentages given on page 60, lines 22-23. It seems that the percentage for extra risk in the example should be 50%, not 10%.

In our experience, there are arguments supported in this guideline for use of a 5% BMR and the log-normal probit model for acute responses (single exposure studies of membrane irritation, neurological disturbances, frank effects, or lethality). The data for this default are presented in the attachment provided.

3) *Model selection and fitting:*

The models presented appear to be adequate for the majority of chronic and developmental data sets. However, the log-normal probit model has been historically the dominant model used in acute toxicity studies and our preliminary analysis indicates that the probit model compares favorably with the Weibull model using proximity to NOAELs and distance between maximum likelihood estimate and 95% lower confidence limits as evaluation criteria. Since, on the last page of the document, the probit model is included in the software planned for release, it should be added as an acceptible default method for acute toxicity data sets.

4) *Use of confidence limits:*

This section of the document seems complete.

5) *Selection of the BMD/C to Use as the Point of Departure for Cancer and Non-cancer Health Effects*

There should be a preliminary analysis comparing models in order to determine if the default factor of 3 for differences in BMD/C results is appropriate for eliminating concerns about model selection.

A discussion of the theory behind the Akaike Information Criterion, including its major assumptions would be helpful.

6) *General comments*

The document is written in an informal easy-to-read style, which is useful for the general readership. There is a good deal of very useful information conveyed in the document. However, in certain areas, the document relies on a good deal of implicit knowledge and contains unsupported default assumptions. The decisions to use defaults in key areas are not clearly explained (e.g. why 10% is now the recommended default for quantal data sets, when previously the benchmark dose workshop concluded 5% or 10% were adequate (Barnes et al., 1995)). A more detailed discussion about the mathematical assumptions that are integral in the different models would be helpful (e.g. Weibull versus quantal polynomial models, versus probit model). Risk assessors need to be educated about the assumptions they are making when using these models to avoid their reliance on "black-box" software outputs.

Other specific comments:

Section on uncertainty factors (page 29): The document states that the only change in uncertainty factors with BMD/C methodology would be the absence of a LOAEL to NOAEL uncertainty factor. However, the document also states that there is a "goal" of achieving a BMD/C to NOAEL ratio of 1, on average (page 59, lines 25-26). In other words, the BMD/C is simply trying to approximate as closely as possible, a NOAEL. The problem with these two simultaneous statements is that the basic premise of conducting BMD/C analyses is to improve the considerations of dose-response and sample size in our estimations of a threshold. If we have confidence in our methodology, then a) it is not necessary to judge the results by their proximity to the NOAEL, and b) uncertainty in the estimate has been reduced. We have proposed that the uncertainty would be reduced in intra-animal variability in response, which likely has some bearing on inter-individual variability.

Typographical and/or grammatical errors:

Page 27, line 5: "a" should be changed to "of".

Attachment A. Comparisons of BMC and NOAELs for Probit and Weibull Models Using Acute Toxicity Data.

Comparison of Benchmark Concentrations with NOAELs and LOAELs for Acute Toxicity Endpoints

Chemical	MLE_{01} (95% CI)	MLE_{05} (95% CI)	NOAEL (ppm)	LOAEL (ppm)	NOAEL/ BC_{01}	NOAEL/ BC_{05}	Slope	Endpoint	Study
Acrolein	16.5 (9.1)	18.8 (12.0)	11.2	23.3	1.2	0.9	12	lethality (hamster)	USEPA (1992)
Ammonia	13.4 (7.8)	20.1 (13.6)	30	50	3.8	2.2	3.83	respiratory irritation (human)	MacEwen et al. (1970)
Ammonia	67.0 (11.6)	77.8 (20.8)	ND 5*	50	ND 0.4*	ND 0.2*	10.5	eye and respiratory irritation (human)	Verberk et al. (1977)
Ammonia	6124 (4591)	7100 (5733)	ND 876*	8758	ND 0.2*	ND 0.15*	10.6	lethality (mice)	Silver and McGrath (1948)
Ammonia	18236 (14247)	20609 (17287)	20950	23380	1.5	1.2	12.8	lethality (rats)	Appleman et al. (1982)
Benzene	5536 (4539)	6548 (5650)	4980	7490	1.1	0.9	9.3	lethality (mice)	Svirbely et al. (1943)
Carbon tet	8338 (6813)	10,137 (8724)	8900	10,300	1.3	1.0	8.03	lethality(rats)	Adams et al. (1952)
Chlorine	211 (169)	232 (197)	213	268	1.3	1.1	16	lethality (rats)	MacEwen and Vernot (1972)
EGBE	1180 (954)	1371 (1170)	1032	1482	1.1	0.9	10.49	lethality (mice)	Werner et al. (1943)
EGEE	2196 (1546)	3307 (2223)	ND 299*	2990	ND 0.2*	ND 0.1*		lethality (mice)	Werner et al. (1943)
EGME	2151 (1663)	2548 (2089)	2461	3439	1.5	1.2	9.27	lethality (mice)	Werner et al. (1943)
Formaldehyde	0.504 (0.253)	0.715 (0.435)	0.5	1.0	2.0	1.1	4.50	eye irritation (human)	Kulle et al. (1987)
HCl	1464 (946)	1772 (1271)	1793	2281	1.9	1.4	8.21	lethality (rat)	Hartzell et al. (1985)
HCl	1941 (1148)	2122 (1410)	2078	2678	1.8	1.5	17.6	lethality (rat)	Darmer et al. (1974)

Chemical	MLE_{01} (95% CI)	MLE_{05} (95% CI)	NOAEL (ppm)	LOAEL (ppm)	NOAEL/ BC_{01}	NOAEL/ BC_{05}	Slope	Endpoint	Study
HCl	671 (344)	1003 (609)	410	1134	0.8	0.7	3.9	lethality (mouse)	Darmer et al. (1974)
HCN	56.4 (17.6)	75.8 (32.5)	83.2	107.2	4.7	2.6	5.3	lethality (mouse)	Bhattacharrya et al. (1991)
HF	216 (166)	242 (204)	263	278	1.6	1.3	11.69	lethality (mouse)	Wohlslagel et al. (1976)
Methyl bromide	769 (601)	871 (726)	875	956	1.5	1.2	12.56	lethality (mouse)	Alexeeff et al. (1985)
Methyl isocyanate	0.052 (0.036)	0.055 (0.041)	0.045	0.062	1.2	1.1	33	respiratory irritation (human)	Mellon Inst. (1963)
Methyl isocyanate	22 (16)	27 (22)	22.7	33.5	1.4	1.0	7.3	lethality (rat)	Rhone-Poulenc (1992)
Methyl isocyanate	10.6 (5.9)	13.2 (8.5)	13	19	2.2	1.5	7.0	lethality (guinea pigs)	Ferguson and Alarie (1991)
Methyl isocyanate	8.7 (2.8)	11.3 (4.8)	8	16	2.9	1.7	6.0	lethality (rat)	Geil et al. (1987)
Methylene chloride	17,847 (14,836)	18,144 (16,082)	17,250	18,500	1.2	1.1	95	lethality (rat)	NTP (1986)
Phosgene	0.8 (0.3)	1.3 (0.7)	1	2	3.3	1.4	3	lethality (mouse)	Kawai et al. (1973)
Styrene oxide	5.98 (2.24)	9.49 (4.63)	ND 1.5*	15	ND 0.67*	ND 0.32*	3.4	lethality (rats)	Sikov et al. (1986)
Vinyl chloride	5877 (2968)	7345 (4340)	4000	8000	1.3	0.9	7.50	CNS effects (human)	Lester et al. (1963)
Vinyl chloride	236,000 (227,000)	253,000 (246,000)	250,000	275,000	1.1	1.0	22.8	lethality (mouse)	Prodan et al. (1975)
Vinyl chloride	524,000 (424,000)	545,000 (466,000)	500,000	575,000	1.2	1.1	40	lethality (rabbit)	Prodan et al. (1975)
Vinyl chloride	502,000 (410,000)	527,000 (453,000)	500,000	575,000	1.2	1.1	33	lethality (guinea pig)	Prodan et al. (1975)

* No NOAEL was evident. Not included in the analysis, shown for comparison only.

Comparison of Benchmark Concentrations from Probit and Weibull Models

Chemical	MLE_{01} (95% CI) Probit	MLE_{01} (95% CI) Weibull	MLE_{05} (95% CI) Probit	MLE_{05} (95% CI) Weibull	NOAEL (ppm)	LOAEL (ppm)	Endpoint	Study
Acrolein	16.5 (9.1)	13.0 (6.0)	18.8 (12.0)	17.0 (10.2)	11.2	23.3	lethality (hamster)	USEPA (1992)
Ammonia	67.0 (11.6)	15.4 (1.1)	77.8 (20.8)	30.8 (5.7)	ND 5*	50	eye and respiratory irritation (human)	Verberk et al. (1977)
Ammonia	18236 (14247)	12662 (7857)	20609 (17287)	17342 (12816)	20950	23380	lethality (rats)	Appleman et al. (1982)
Benzene	5536 (4539)	4085 (3007)	6548 (5650)	5775 (4731)	4980	7490	lethality (mice)	Svirbely et al. (1943)
Chlorine	211 (169)	179 (119)	232 (197)	219 (168)	213	268	lethality (rats)	MacEwen and Vernot (1972)
EGBE	1180 (954)	872 (608)	1371 (1170)	1207 (955)	1032	1482	lethality (mice)	Werner et al. (1943)
EGEE	2196 (1546)	1435 (781)	3307 (2223)	2309 (1562)	ND 299*	2990	lethality (mice)	Werner et al. (1943)
EGME	2151 (1663)	1472 (930)	2548 (2089)	2161 (1586)	2461	3439	lethality (mice)	Werner et al. (1943)
Formaldehyde	0.504 (0.253)	0.356 (0.130)	0.715 (0.435)	0.659 (0.346)	0.5	1.0	eye irritation (human)	Kulle et al. (1987)
HCl	1464 (946)	982 (453)	1772 (1271)	1509 (893)	1793	2281	lethality (rat)	Hartzell et al. (1985)
HCl	1941 (1148)	1300 (750)	2122 (1410)	1693 (1169)	2078	2678	lethality (rat)	Darmer et al. (1974)
HCl	671 (344)	387 (154)	1003 (609)	829 (454)	410	1134	lethality (mouse)	Darmer et al. (1974)
HCN	56.4 (17.6)	35.5 (5.8)	75.8 (32.5)	63.6 (20.0)	83.2	107.2	lethality (mouse)	Bhattacharrya et al. (1991)
HF	216 (166)	156 (94)	242 (204)	213 (154)	263	278	lethality (mouse)	Wohlslagel et al. (1976)
Methyl bromide	769 (601)	628 (425)	871 (726)	813 (629)	875	956	lethality (mouse)	Alexeeff et al. (1985)

Comparison of Benchmark Concentrations from Probit and Weibull Models (continued)

Chemical	MLE_{01} (95% CI) Probit	MLE_{01} (95% CI) Weibull	MLE_{05} (95% CI) Probit	MLE_{05} (95% CI) Weibull	NOAEL (ppm)	LOAEL (ppm)	Endpoint	Study
Methyl isocyanate	0.052 (0.036)	ND	0.055 (0.041)	ND	0.045	0.062	respiratory irritation (human)	Mellon Inst. (1963)
Methyl isocyanate	22 (16)	15 (9)	27 (22)	24 (17)	22.7	33.5	lethality (rat)	Rhone-Poulenc (1992)
Methyl isocyanate	10.6 (5.9)	5.81 (2.1)	13.2 (8.5)	10.0 (5.2)	13	19	lethality (guinea pigs)	Ferguson and Alarie (1991)
Methyl isocyanate	8.7 (2.8)	5.8 (0.74)	11.3 (4.8)	9.84 (2.5)	8	16	lethality (rat)	Geil et al. (1987)
Phosgene	0.8 (0.3)	0.4 (0.09)	1.3 (0.7)	1.1 (0.4)	1	2	lethality (mouse)	Kawai et al. (1973)
Vinyl chloride	5877 (2968)	6574 (2308)	7345 (4340)	8717 (4331)	4000	8000	CNS effects (human)	Lester et al. (1963)

Distance Between NOAEL and Benchmark Concentration Depends on the Endpoint Examined and Model Used

	$NOAEL/BC_{01}$	$NOAEL/BC_{05}$
USEPA Workshop (Developmental endpoints only) *Weibull model*	**29 ± 44**	**5.9 ± 8.4**
Cal/EPA (Acute endpoints not including developmental toxicity) *Probit model*	**1.8 ± 0.9 n = 25 studies**	**1.2 ± 0.4 n = 25 studies**
Cal/EPA *Weibull model*	**4.3 ± 3.8 n = 18 studies**	**1.8 ± 0.9 n = 18 studies**

Comparison of Models

Probit Model

Parameter	Mean	SD
$NOAEL/BC_{05}$*	1.2	0.4
$NOAEL/BC_{01}$*	1.8	0.9
MLE_{05}/95% LCL**	1.5	0.6
MLE_{01}/95% LCL**	1.8	1.0
BC_{05}/BC_{01}**	1.4	0.3

* n = 25 studies (4 studies did not contain NOAELs)
** n = 29 studies

Weibull Model

Parameter	Mean	SD
$NOAEL/BC_{05}$*	1.8	0.9
$NOAEL/BC_{01}$*	4.3	3.8
MLE_{05}/95% LCL**	2.0	1.1
MLE_{01}/95% LCL**	3.2	3.1
BC_{05}/BC_{01}**	2.3	1.1

* n = 18 studies (2 studies did not contain NOAELs)
** n = 20 studies

Use of the Probit model in acute toxicity BMD/C calculations

The log-probit model is among the most widespread models used in acute toxicity testing and has traditionally been used extensively for determination of acute lethality and other dichotomous responses (Finney et al., 1971; Rees and Hattis, 1994). Furthermore, because the model is normally distributed, it is biologically plausible and accounts for some degree of inter-individual variability (Rees and Hattis, 1994).

For a toxic response with a specific threshold, a 1 to 5 percent response approaches the margin of useful extrapolation for acute noncarcinogenic data due to the limited number of animals used in most experiments. Use of the 95 percent lower confidence limit on concentration takes into account some variability of the test population and is dependent on the number of subjects in the study.

Attachment B. Example of a method for evaluating the combination of data sets

Combining Data Sets

The approach proposed by Stiteler et al. (1993) illustrates one method to evaluate the statistical validity in combining data sets based on differences in the maximum likelihood estimates. The Ishinishi et al. (1988) study on the non-cancer health effects of diesel exhaust in rats provides data that can be used with this approach. Rats (male and female groups) were exposed to light or heavy duty diesel exhaust for a full-lifetime (30 months). The exposures were for 16 hours/day, 6 days/week. The sample sizes were 59-61 female and 64 male rats/group. The endpoints examined were sensitive measures of pulmonary epithelial and alveolar damage. A benchmark dose analysis was performed using a log-normal probit model with the female rat data (Tox-Risk software for the IBM-PC). Heavy duty diesel was determined to be more potent than the LD diesel for induction of hyperplastic lesions. The dose-response relationship of the males exposed to HD was not well modeled by the log-normal probit relationship, particularly at extrapolated low doses. For example, the 95% lower confidence limit on the BD_{01} using the male HD data was 2.3E-3 $\mu g/m^3$, or a factor of 4.4E5 below the MLE. This is partially due to the shallow dose-response slope from the male rat data. The test for the acceptance of combining data sets using maximum likelihood estimates from the log-normal analysis indicated that the male and female data sets should not be combined. The test for combining the data sets is shown below:

Data Set	Maximum Log Likelihood
Combined	-153.961
Female	-85.129
Male	-65.297

The natural logarithm of the Generalized Likelihood Ratio (GLR) is
[-153.961 - (-85.129 + -65.297)] = -3.535.

The likelihood ratio test statistic is calculated as -2 (ln GLR) = 7.07

The chi-square test for one degree of freedom results in p = 0.008, indicating that there are significant differences between the 2 data sets such that they should not be combined.

For the above reasons, the female rat HD diesel data were used to determine the benchmark dose for diesel exhaust. Similar sex-dependent responses were observed in both the light and heavy duty diesel experiments.

Rat Lung Hyperplasia Data Following Diesel Exposure (Ishinishi et al., 1988)

Diesel type	**Diesel Concentration**				
Heavy duty	**0 mg/m^3**	**0.46 mg/m^3**	**0.96 mg/m^3**	**1.84 mg/m^3**	**3.72 mg/m^3**
Male	0/64	2/64	4/64	4/64	8/64
Female	1/59	1/59	3/61	10/59	17/60
Light duty	**0 mg/m^3**	**0.11 mg/m^3**	**0.41 mg/m^3**	**1.18 mg/m^3**	**2.32 mg/m^3**
Male	4/64	3/64	2/64	4/64	38/64
Female	0/59	1/59	4/61	8/59	49/60

Attachment References

Adams, E.M., Spencer, H.C., Rowe, V.K., McCollister, D.D., and Irish, D.D. 1952. Vapor toxicity of carbon tetrachloride determined by experiments on laboratory animals. *Archives of Industrial Hygiene and Occupational Medicine* 6:50-66.

Alexeeff, G.V., Kilgore, W.W., Munoz, P, and Watt, D. 1985. Determination of acute toxic effects in mice following exposure to methyl bromide. *Journal of Toxicology and Environmental Health* 15(1):109-123.

Appelman, L.M., Ten Berge, W.F., Reuzel, P.G.J. 1982. Acute inhalation toxicity study ammonia in rats with variable exposure periods. *American Industrial Hygiene Association Journal* 43:662-665.

Barnes, D.G., Daston, G.P., Evans, J.S., Jarabek, A.M., Kavlock, R.J., Kimmel, C.A., Park, C., and Spitzer, H.L. 1995. Benchmark dose workshop: Critieria for use of a benchmark dose to estimate a reference dose. *Regulatory Toxicology and Pharmacology* 21:296-306.

Bhattacharya, R., Vijayaraghavan, R. 1991. Cyanide intoxication in mice through different routes and its prophylaxis by α-ketoglutarate. *Biomedical and Environmental Science* 4:452-459.

Darmer, K.I., Kinkead, E.R., Di Pasquale, L.C. 1974. Acute toxicity in rats and mice exposed to hydrogen chloride gas and aerosol. *American Industrial Hygiene Association Journal* 35:623-631.

Ferguson, J.S., Alarie, Y. 1991. Long term pulmonary impairment following single exposure to methyl isocyanate. *Toxicology and Applied Pharmacology*, 107:253-268.

Geil, R.G. 1987. Six-hour acute inhalation toxicity study in rats - Methyl isocyanate and hexamethylene diisocyanate. Prepared by International Research and Development Corporation for Dow Chemical. EPA/OTS New Doc. ID: #86-870002225.

Hartzell, G.E., Packham, S.C., Grand, A.F., Switzer, W.G. 1985. Modeling of toxicological effects of fire gases: III. Quantification of post-exposure lethality of rats from exposure to HCl atmospheres. *Journal of Fire Science* 3:195-207.

Ishinishi, N., Kuwabara, N., Takaki, Y. et al. 1988. Long term inhalation experiments on diesel exhaust. In: Diesel Exhaust and Health Risk. Results of the HERP Studies. Entire Text of Discussion. Research Committee for HERP Studies. Japan Automobile Research Institute, Inc. Tsukuba, Ibaraki 305, Japan.

Kawai, M. 1973. Inhalation toxicity of phosgene and trichloronitromethane (chloropicrin). *J. Sangyo Igaka* 15(4):406-407.

Kulle, J.T., L.R. Sauder, J.R. Hebel, D. Green, and M.D. Chatham 1987. Formaldehyde dose-response in healthy nonsmokers. *Journal of the Air Pollution Control Association* 37:919-924.

Ledbetter, A.D., Adkins, B.,Jr. 1992. Acute 1-hour inhalation toxicity study of methyl isocyanate in rats. Prepared by ManTech Environmental Technology, Inc. for Rhone-Poulenc Ag Company, Research Triangle Park, NC, Project No. 6030-003.

Lester, D., Greenberg, L.A., and Adams, W.R. 1963. Effects of single and repeated exposures of humans and rats to vinyl chloride. *American Industrial Hygiene Association Journal* 3:265-275.

MacEwen, J., Theodore, J., and Vernot, E.H. 1970. Human Exposure to EEL concentration of Monomethylhydrazine. SysteMed Corp, Wright-Patterson Air Force Base, Ohio. AMRL-TR-70-102,23.

MacEwen, J., and Vernot, E. 1972. Annual Technical Report: Aerospace Medical Research Laboratory, Toxic Hazards Research Unit, Wright-Patterson Air Force Base, Ohio AMRL-TR-72-62 (NTIS AD755-358).

National Toxicology Program (NTP). 1986. Toxicology and carcinogenesis studies of dichloromethane (methylene chloride) in F344/N rats and B6C3F1 mice. U.S. Dept. of Health and Human Services. NIH Publication 86-2562. Reasearch Triangle Park, NC.

Pozzani, U.C., Carpenter, C.P. 1963. The feasibility of using methyl isocyanate as a warning agent in liquid carbon monoxide. In: Compilation of Toxicology on Methyl Isocyanate. Prepared by the Mellon Institute for Union Carbide Corporation, Pittsburgh, PA, Report No. 26-23, pp. 151-155.

Prodan, L., Suciu, I., Pislaru, V., Ilea, E., and Pascu, L. 1975. Experimental acute toxicity of vinyl chloride (monochloroethene). *Annals of the New York Academy of Sciences*. 154-158.

Rees, D.C., and Hattis, D. 1994. Developing quantitative strategies for animal to human extrapolation. in, A.W. Hayes (ed): Principles and Methods of Toxicology, 3rd Ed. Raven Press, Ltd., New York.

Sikov, M.R., Cannon, W.C., Carr, D.R., Miller, R.A., Montgomery, L.F., and Phelps, D.W. 1981. Teratologic assessment of butylene oxide, styrene oxide, and methyl bromide. NIOSH Technical Report, Publication No. 81-124, U.S. Government Printing Office, Washington, D.C. 20402.

Silver, S.D. and McGrath, F.P. 1948. A comparison of acute toxicities of ethylene imine and ammonia in mice. *Journal of Industrial Hygiene and Toxicology* 30(1):7-9.

Stiteler, W.M., Knauf, L.A., Hertzberg, R.C., and Schoeny, R.S. 1993. A statistical test of compatibility of data sets to a common dose-response model. *Regulatory Toxicology Pharmacology* 18:392-402.

Verberk, M.M. 1977. Effects of ammonia in volunteers. *International Archives of Occupational Health* 39:73-81.

Werner, H.W., Mitchell, J.L., Miller, J.W., and Von Oettingen, W.F. 1943. The acute toxicity of vapors of several monoalkyl ethers of ethylene glycol. *Journal of Industrial Hygiene and Toxicology* 25:157-163.

Wohlslagel, J., DiPasquale, L.C., and Vernot, E.H. 1976. Toxicity of solid rocket motor exhaust: effects of HCl, HF, and alumina on rodents. *Journal of Combustion Toxicology* 3:61-70.

David Gaylor

Comments on
"Benchmark Dose Technical Guidance Document"
EPA/600/P-96/002A, August 9, 1996

David W. Gaylor, Ph.D.
National Center for Toxicological Research
Food and Drug Administration

Following are my comments to the questions raised in the "Charge to Reviewers."

1. Selection of Studies and Responses.

   a. No comment.

   b. Yes.

   c. No comment.

2. Selection of the Benchmark Response Level.

   a. Defining biological significance as some specified change in the average is of little value. This does not indicate how many individuals may be at risk. For example, a shift in average body weight of 10% might put a small percent of individuals at risk or half of the individuals at risk, depending on the standard deviation.

      The "limit of detection" basically mimics the NOAEL. As such, it retains most of the bad properties of the NOAEL and is harder to compute. It is counter to the benchmark dose approach and reverts back to the NOAEL.

      For continuous data, the so-called hybrid methods should be used (e.g., Gaylor and Slikker, 1990). Where an adverse level for an individual cannot be established, an abnormal range can be used, e.g., below the first percentile or above the 99th percentile. Then the proportion (risk) of animals in the abnormal range can be estimated as a function of dose. This requires choosing an appropriate distribution (e.g., log-normal) and an estimate of the standard deviation. This approach is then compatible with that used for cancer risk assessment.

   b. No. See comment for 2a.

c. No. See comment for 2a.

d. Okay for quantal. Not for continuous (see comment for 2a).

3. Model Selection and Fitting

a. No comment.

b. For continuous data, the choice of biologically-based models should be recommended, e.g., the Michaelis-Menten or Hill equation for receptor mediated processes. At least, saturation-type models with asymptotes should be considered.

c.i. No comment.

c.ii. No. Almost every biological effect occurs spontaneously, although it might be rare. It is not realistic to set the background at zero. In fact, this may create poor fits. The model estimate of the background often will be better than the estimate based on only the control animals.

c.iii. Additional risk is equally good.

c.iv. Yes.

c.v. No comment.

d. No. Goodness-of-fit tests should be performed with higher P-values, e.g., $P < 0.20$, A $P < .05$ will allow very poor fits. At this rejection level, only the worse fits will be discarded. The purpose of a Goodness-of-fit test is not to keep all but extremely poor fits, rather it is to keep only the better fits.

Further, a chi-square Goodness-of-fit test can be performed on just the data from the lower doses, as this is the region of interest.

4. Use of Confidence Limits.

a. Yes.

b. No comment.

c. Yes.

5. Selection of the BMD/C to use as the Point of Departure.

   a. No comment.

   b. A brief description of the Akaike Information Criteria would be useful.

   c. Yes.

6. General Issues.

   a. No comment.

   b. Okay.

   c. Okay.

   d. No comment.

Further Comments:

A. Most biological measurements appear to be described by a log-normal distribution. This is particularly true as most measurements are greater than zero with an occasional high reading. The default for distributions of measurements should be the log-normal. The possible exceptions are organ and body weights where the normal distribution is appropriate.

B. Page 34. A major argument for low dose linearity, additivity to background, should be included. Chemicals that augment an ongoing toxic process will produce low-dose linearity. The appearance of adverse effects in control animals indicate that threshold doses have already been surpassed by endogenous or other exogenous sources. Hence, the addition of even a small dose of a chemical to such a process will produce an effect, albeit small, unless there is total homeostatic control.

C. Page 37, line 14. Van Ryzin (1980) used 1% as a benchmark response level.

D. Page 69, lines 21-23. This statement is not true. In fact, this could be the worst approach. Variables that are affected by dose create colinearities among variables that make it very difficult to establish cause and effect.

William Hartley

August 26, 1996

William R. Hartley, Sc.D.
Associate Professor
Tulane Medical Center
School of Public Health and Tropical Medicine
New Orleans, LA 70112

Review: Benchmark Dose Technical Guidance Document

GENERAL Review comments are provided in the order of the issues detailed in the Charge to Reviewers.

SPECIFIC COMMENTS

1. Selection of Studies and Responses for Benchmark Dose/C (BMD/C) Analysis:

a. The proposed guidance for selection of studies and endpoints for the BMD/C is appropriate for noncancer and cancer assessment. The process for selection of studies involves evaluation of data for which modeling is feasible, so the BMD/Cs can be estimated. More discussion would be useful on potential for a multivariate analysis approach. The focus on endpoints that are relevant to humans and the most sensitive effect is where extensive toxicological knowledge is required on the part of the risk assessor. As the BMD/C concept is used by the risk assessment community, there will be considerable discussion and hopefully growing consensus on identification of relevant endpoints. I generally agree that endpoints should be modeled if their LOAELs are up to 10-fold above the lowest LOAEL.

b. The same general criteria regarding data quality and potential relevance to human health should be used for the selection of studies and responses for cancer and noncancer assessment using the BMD/C approach. Generally cancer studies will always be tumor incidence data unless there are data on precursor events in the carcinogenic process such as physiological disturbances and other organ toxicity. The document correctly notes that the straight-line extrapolation from the LED10 and the LMS procedure result in similar estimations of potency. This will probably be the most common procedure for cancer assessment. The BMD/C approach for noncancer effects will still result in a "safe dose" calculation.

c. There are appropriate criteria for determining when data should be combined for analysis. Data sets should be both statistically and biologically compatible before being combined for dose-response modeling. The document states the advantages of combining appropriate data sets for risk assessment and research direction.

2. Selection of the Benchmark Response (BMR) Level

a. The use of biological significance and limit of detection are an appropriate basis for the selection of the BMR. The document provides a clear rationale for the two bases for specifying the BMR: a biologically significant change in response for continuous endpoints, or the limit of detection for either quantal data or continuous data. Default decisions are clearly established.

b. To find the magnitude of response just detectable, a default power level of 50% and a one-sided test with a Type I error of 0.05 is suggested. The explanation of selection the default power level of 50% needs further discussion. Data (cases) to support this approach should be discussed.

c. I am not aware of extensive data to determine the appropriate power level. Perhaps the simulation studies will provide a basis for a decision.

d. Default decisions for quantal and continuous data are appropriate. For quantal data, an increase (10%) in extra risk is the best default approach for public health protection.

3. Model Selection and Fitting

a. The order of model application for continuous and dichotomous data is appropriate. It is important that the guidelines, in the final form, continue to allow the risk assessor to use other models.

b. The number of models allowed should not be more restrictive.

c. Comments on parameters proposed as defaults for model structure.

i. Recommend the degree of polynomial equal k-1 (number of exposure groups minus one) by convention. This is identified as step 1 in the document. I know of no reason to deviate from this approach.

ii. The background parameter should not be included unless there is some indication of a background response level. There should be some discussion of the use of background data from concurrent controls versus historical control data. This issue is important in cancer risk assessment and possibly in noncancer assessment where background data may exist on some test procedures in laboratory animal species.

iii. The use of extra risk as a default for quantal data is appropriate and protective of public health.

iv. I agree that a threshold parameter is not a biologically meaningful parameter in models for BMD/C analysis. Therefore, it should not be included.

v. I agree that in BMD/C analysis, continuous data should be modeled directly without conversion to dichotomous format to avoid the loss of valuable information.

d. The document is clear that the criteria for final model selection will be based on how well the various models describe the data, conventions regarding the biological endpoint being evaluated, and model application to multiple data sets. Some discussion regarding the conventions regarding the biological endpoint being evaluated would be useful. I believe this refers to both toxicological and statistical consensus for analysis of particular types of data. Growth and development of consensus regarding currently difficult endpoints to evaluate (for example - immunological data) will be useful and hopefully become more important in model selection.

4. Use of Confidence Limit

a. The lower (95%) confidence limit on dose should be the definition of the BMD/C. This is protective of public health and is similar to using the upper confidence limit on risk as currently practiced in quantitative cancer risk assessment. Considering the uncertainties, it is the only choice from a public health standpoint.

b. Defaults for the method of confidence limit calculation are appropriate.

c. The default to the lower 95% CL on dose is appropriate by convention and protective of public health given the uncertainties.

5. Selection of the BMD/C to Use as the Point of Departure for Cancer and Noncancer Health Effects

a. Centering the selection of equivalent models using the goodness-of-fit (GOF) statistic ($p>0.05$) is appropriate. However, as stated, it is important that the GOF criteria be applicable to the low dose end of the dose-response curve. I agree with the criteria presented for elimination of high dose groups from the data set.

b. Not enough information is presented for me to specifically evaluate the Akaike Information Criterion (AIC). Based on the general information presented the approach of a rank based value measuring deviance of the model fit seems reasonable.

c. Generally agree that the default approach for selecting the BMD/C to use as the point of departure for cancer and noncancer dose-response analysis is appropriate. However, more supporting documentation is needed to support the factor of three approach for determination of model dependence. Generally, selection of the lowest BMD/C will be protective of public health. However, if there are outlier values for the BMD/C consider a geometric mean value or other statistic. For example, under the current cancer guidelines, geometric means may be used for determination of potency/slope factors.

6. General Issues

a. The discussion concerning the use of BMD/C approach in cancer and noncancer risk assessment is adequate provided that the risk assessor is familiar with other supporting USEPA risk assessment guidance documents which detail quantitative procedures. Although it is stated that this document supplements existing guidance documents, a brief discussion of uncertainty factor application to the BMD/C should be included. Procedures for application of uncertainty factors for interspecies variation (animal-to-human dose extrapolation - cancer vs noncancer approach) and intrahuman variability (sensitivity) should be included.

b. The document is understandable to the general toxicologist/risk assessor. However the proposed procedures are more complex than current methods. Providing the opportunity to attend workshops and use USEPA software will be useful for most risk assessors. Many standard quantitative risk assessment procedures have been established over the years, and there may be some resistance to changes to overcome.

c. The overall organization of the document is good with a logical progression of issues and guidance. Implementation of the procedures will require more statistical support for toxicologists conducting risk assessments. Software with "help screens" will reduce some but not all the requirement for additional support from statisticians. Established approaches for selection of the BMR for various endpoints or types of endpoints need to be added to the document before release. Comments on cost-benefit analysis needs are interesting but I do not expect that this issue will be resolved any time in the near future with out extensive input from risk managers and others. I suggest deletion of this issue from the document.

d. Examples provided in Appendix D are useful/essential. However, a generic case studies approach similar to previous USEPA workshops on quantitative risk assessment should be developed and included with the document (perhaps as an appendix).

Abby Li

Abby Li 8/29/96

Comments on the Draft of the BMD Technical Guidance Document

1. It is premature to apply the BMD approach until toxicologically-based criteria for use of a BMD can be established empirically.

The Benchmark Dose (BMD) approach is a potentially useful tool for risk assessment. It is important for the EPA to encourage the use of quantitative risk assessment approaches that might improve the current risk assessment process. However, there is very little practical experience in applying the BMD approach to most data sets, with the exception of developmental toxicity. This makes it very difficult to set criteria for determination of the BMD that would be biologically based. For this reason, I believe that it is premature to replace the currently used NOAEL/safety factor approach with the BMD approach for deriving RfC's and RfD's for non-cancer endpoints. At most, it could be used in conjunction with the currently used NOAEL/safety factor approach, which the majority of the participants at the EPA/ILSI/AIHC benchmark workshop agreed has been sufficiently protective (Barnes, et al, 1995, p.305). This will allow us to collect the experimental data needed to determine the most suitable conditions under which the BMD approach might be applied rather than rely on arbitrary defaults.

At present the benchmark dose approach has only been rigorously applied to the developmental toxicity endpoints of dead implants, malformed fetuses, and fetal weight. There is very little experience in applying this approach to the vast majority of non-cancer endpoints typically used to evaluate the toxicity of regulated chemicals. For example, the benchmark dose has not been systematically applied to any of the behavioral data which make up the bulk of the data generated in conducting the EPA's neurotoxicity screening battery.

The limited experience in applying the benchmark dose approach to non-cancer endpoints requires an increased reliance on default criteria that are not empirically based. This is especially evident in the guidance on the criteria for selecting the benchmark response level (BMR). Although this section tries to offer methods of selecting the BMR that take into account different endpoints and experimental designs, they are not practical alternatives because there is insufficient experience. For example, the guidance document acknowledges that for most continuous endpoints there is no consensus on what a "biologically significant" effect is. So, for all practical purposes, either the statistical default on Limit of Detection or the default on 10% increase in extra risk will be used. Without further empirical evidence, the guidance on using limit of detection at a default power level of 50% is essentially an arbitrary decision that has no clear biological or even sound statistical basis. Without further empirical evidence, the default guidance of using 10% increase in risk is an arbitrary criteria for many non-genotoxic endpoints based on developmental toxicity data sets which may or may not be relevant to other non-cancer endpoints and study designs. **Thus, the benchmark dose method could be reduced to becoming a highly sophisticated method of deriving an arbitrary result.** Even more dangerous, the sophisticated statistical methods employed gives the illusion that there is scientific legitimacy to the approach when in fact very little scientific judgement is used in establishing the BMD.

2. <u>The BMD approach may not be well-suited for neurotoxicity data sets..</u>

The proposed EPA's neurotoxicity risk assessment guidelines emphasizes that behavioral data (approximately 32 endpoints) should be evaluated in terms of patterns of effects, not individual endpoints. At present, the NOEL could occur at a dose level that produces behavioral effects if there is no

pattern of effect consistent with a neurotoxic effect. The BMD approach as outlined in the guidance document does not provide a way to evaluate different endpoints in a manner consistent with the EPA's neurotoxicity risk assessment guidelines. In addition, there is increasing emphasis on the use of quantitative measurements rather than subjective evaluations in neurotoxicity testing. Statisticians/toxicologists are still exploring different methods to apply the BMD to continuous data sets and it seems premature to recommend any method for analysis of continuous data.

3. The Confidence Limit should NOT be included in the definition of the BMD/C.

The guidance document states on page 3 that "the BMD/C accounts for variability in the data since it is defined as the lower confidence limit on the dose estimated to produce a given level of change." Although it is true that the lower confidence limit takes into account animal to animal variation and experimental variation (# animals, methods of evaluation), it also is dependent upon the type of model selected. All other things being equal, model selection appears to have great influence on the LEDx and much smaller impact on the EDx. Thus, an additional uncertainty is introduced into the analysis in a non-transparent manner when the lower confidence limit is used. For this reason, the EDx and NOT the LEDx should be the point of departure for further risk assessment. Using the EDx will make the estimate for the point of departure for risk assessment a much more precise estimate than if the LEDx is used. The risk manager could then assess the experimental variation and determine if additional safety factors need to be added.

4. The limit of detection (LOD) with the 50% power level is arbitrary and confusing.

The LOD sounds like a statistically elegant method to determine the BMR. However, in the absence of any empirical evidence to support the 50% power or any other power, it is an arbitrary method. If one wishes to arbitrarily set criteria for the BMR, then it would be much more straightforward to set some value (like a 10% or 20% change) and acknowledge that this is an arbitrary science policy decision. The LOD with the power level has less intuitive meaning to the average toxicologist and the issue is confused by the selection of a 50% power level. Statistical methods are guiding the selection of the BMR, instead of biological criteria.

5. Should there be a selection of BMD/C if there is model dependency?

If the BMD/C estimate exhibits a high degree of model dependence, then why would one conclude that the most conservative estimate should be used. Wouldn't it be more appropriate to NOT use this approach?

6. Can some operational limits be defined that would limit extrapolation from going below the observable range?

At the beginning of the guidance document, it is clearly stated that the BMD is estimated in the observed range. Is there a way to define operational limits that would prevent extrapolation from going below the observable range.

Rashmi Nair

Comments on the Draft of the BMD Technical Guidance Document

1. Selection of Studies and Responses for Benchmark Dose/C Analysis

   a. Is the selection of studies and end points for the BMD/C appropriate? for cancer? for noncancer?

   The Agency's approach to review the overall database on a given compound to identify and characterize the hazards of the compound is indeed appropriate. The guidance document appropriately discusses that the selection of the critical study should be based on the human exposure situation that is being addressed, the quality of studies in question and the relevance and adequacy of the endpoints. The guidance document further recommends that representative endpoints that show smoothly increasing response with increasing dose should be selected in order to obtain a good fit of the dose response model. This is where we believe the guidance needs to be modified. It is important to point out that use of BMD/C approach is "an alternative" to the NOAEL/LOAEL approach. What this means is that the BMD/C approach is "a tool" in the overall tool box of a risk assessor. Therefore, if the data on critical endpoint from the best quality study are not conducive to model fitting but provide an adequate point of departure, i.e., an appropriate NOAEL, for extrapolation to derive an acceptable exposure level then there's no need to use alternative endpoints to determine a BMD/C.

   The emphasis should be on the appropriate endpoint which is biologically significant and has an adequate point of departure from which to extrapolate. To illustrate our concern, a situation which is not uncommon in toxicological data available for deriving RfDs is presented. Suppose a given compound produces hepatocellular hyperplasia and some focal necrosis at the high dose but no histological changes are observed at the lower dose levels. Since the effects are observed only at the high dose, the data cannot be appropriately modeled for BMD/C. Based on the current Agency guidance alternative endpoints like liver weight or liver enzyme changes could be modeled and a BMD/C obtained. If this BMD/C is an order or two lower than the actual NOAEL for histological changes, then there is no justification for using the lower BMD/C because it is quite possible that liver enzyme changes are more of a marker for exposure rather than being mechanistically linked to the histological changes in the liver.

   b. Should these be the same for cancer and noncancer data?

   The scientific judgment that is used for selecting the appropriate study for both cancer and non-cancer endpoints is similar. If multiple studies are available, the choice of the study is determined by the overall quality of the study in terms of use of appropriate protocols, adequate number of animals, appropriate route of exposure, adequate number of tissues for histopathological evaluation etc. The big difference between cancer and non-cancer studies is that for non-cancer data, generally either a critical endpoint or a few endpoints

can be selected based on biological plausibility from the many endpoints which may show statistical significance for obtaining the point of departure for extrapolation to human hazards. For cancer on the other hand if tumors are observed at multiple sites in different sexes of animals very seldom does knowledge exist on the appropriate tumor to use for extrapolation to human risk and generally the responses are modeled to determine the one providing either the lowest cancer slope factor or as proposed in the new cancer guidelines (EPA, 1996) the lowest LED_{10} is selected as the point of departure. Thus different endpoints and somewhat different criteria should be used for cancer versus non-cancer studies. The issue of use of non-tumor data for identifying the point of departure for extrapolation for cancer data is still being addressed and therefore the use of these type of data should await additional analysis of this issue.

c. Are there appropriate criteria for determining when data should be combined for analysis?

The current guidance document provides the advantages of combining "appropriate" data but provides no guidance on when it is appropriate to combine data. An example is provided where data on the same endpoint are combined from two different studies which were conducted using the same protocol. Guidance is needed on both when different datasets on a given endpoint from different studies can be combined and when data from different endpoints can be combined from a given study because the same mechanism of action is responsible for the changes in these endpoints.

2. Selection of the Benchmark Response Level

a. Is the use of biological significance or limit of detection an appropriate basis for the selection of the BMR?

While in theory this sounds like an interesting idea, there is no justification for placing this yet untried approach in a regulatory guidance document. Devlopment of health-based criteria have a significant impact on society in monetary terms and in light of these consequences, only well established procedures which are scientifically defensible should be used for developing these criteria. From our perspective, application of the benchmark dose methodology has only been tested widely on development toxicity data and this exercise led to the need for a more appropriate model which takes into account within litter variance. Evaluation of other types of datasets will have additional learnings and therefore before benchmark dose is recommended as the default methodology for both cancer and all non-cancer endpoints, additional evaluation of this methodology is appropriate.

b. For the limit of detection, is the approach proposed in the document appropriate?

As per our statistician, the approach provided in the guidance document is a "statistical overkill".

c. Is information available to determine the appropriate power level? (Information on simulation studies will be presented at the workshop.)

d. Is the default for quantal and continuous data appropriate?

The default appears to be appropriate for quantal data but for continuous data the default can only be considered to be appropriate for developmental endpoints. As stated in the guidance document, as of the time the ILSI/EPA workshop, the participants of that workshop were reluctant to recommend the use of continuous data for deriving appropriate points of departure for extrapolation. In our evaluation of continuous data from subchronic toxicity studies (Nair et al., 1995a), no consistent relationship between the NOAEL and the continuous endpoint BMDs was found. Also, as stated in the guidance document, a new approach has been proposed by Crump (1995) but this approach has not been applied to many actual datasets because the software based on this approach has not been available until very recently. Thus, the default for continuous data should await evaluation of this methodology with actual datasets.

3. Use of Confidence Limits

a. Should the lower confidence limit on dose be the definition of the BMD/C?

Instead of using the lower confidence limit on dose as the definition BMD/C and the point of departure, the ED10 should be the default point of departure for extrapolating to acceptable levels. Two reasons why we support the use of ED10 versus the lower confidence limit are (1) the values for ED10 appear to be model independent while the LED10 values appear to be model dependent. The controversy that arises when different values are obtained based on different models from the same dataset can be avoided by use of the ED10. (2) As has been suggested by others at the ILSI workshop, the apparent precision that is added with the use of the statistical confidence limit in the first step of developing regulatory health standards is minuscule, compared to the uncertainty that is introduced through use of multiple conservative uncertainty factors for deriving an RfD. Recent evaluations (Nair et al. 1995b and Sherman et al. 1995) of a substantive number of subchronic and chronic study datasets show the conservative nature of the 10X uncertainty factors that are currently used to extrapolate to an RfD. So there are significant conservative assumptions to make up for the animal variability. Also, use of consistent modern day regulatory protocols and international harmonization will help reduce the variability in the animal data. Thus, the inconsistency in the two steps of the procedure for deriving a health based criteria provides a false sense of precision in the procedure and has a chance of misinterpretation.

b. Are the defaults for the method of confidence limit calculation appropriate?

See above

c. Is the default of 95% confidence limit appropriate?

See above

5. Selection of the BMD/C to Use as the Point of Departure for Cancer and Noncancer Health Effects

a. Comment on the determination of "equivalence" of models.
b. Comment of use of the Akaike Information Criterion for comparing the fit of models.

As per our statistician, the Akaike Information Criterion is an appropriate criterion for comparing the fit of models.

c. Is the default approach for selecting the BMD/C to use as the point of departure for cancer and noncancer dose-response analysis appropriate?

The benchmark dose methodology appears to be a potential alternative to the NOAEL approach for developing health based exposure criteria but cannot be presently recommended as the default approach for regulatory use. As discussed in Nair et al. (1995a), and the present guidance document the BMD method is an acceptable alternative only when good dose-response data are available. In our evaluation we found that only 30 of the 51 randomly selected studies in the Monsanto database had sufficient dose response data to even consider modeling for obtaining a benchmark dose. Of these only, 16 studies had appropriate quantal data. We feel our data is representative of the overall database that is currently available in toxicology. In addition additional validation of the methods is needed with different endpoints. Use of this methodology should await further validation for different endpoints and different datasets.

References

Nair, R.S., Stevens, M.W., Martens, M.A. and Ekuta, J. (1995a). Comparison of BMD with NOAEL and LOAEL Values Derived from Subchronic Toxicity Studies. Archives of Toxicology (Suppl. 17). Toxicology in Transistion pp 44-54, Springer-Verlag, New York.

Nair, R.S., Sherman,J.H., Stevens, M.W. and Johannsen, F.R. (1995b). Selecting a More Realistic Uncertainty Factor: Reducing Compounding Effects of Multiple Uncertainties. Human and Ecological Risk Assessment 1, 576-589.

Sherman, J.H., Stevens, M.W., Johannsen, F.R. and Nair, R.S. (1995). Quantification of Inter-Species Variability in Response to Systemic Toxicants: Analysis of Pesticide NOELs Defined by an Expert Committee. Society of Risk Analysis and the Japan Section of SRA, Annual Meeting and Exposition. Abstract No. C8.02.

Bruce Naumann

Comments on EPA's Benchmark Dose Technical Guidance Document

Comments on general issues are provided first, followed by responses to the questions raised in the "Charge to Reviewers" document. Additional technical comments are also provided on specific sections of the technical guidance and organized by the order they appear in the document.

General Issues

a. The discussions concerning the use of the BMD/C approach in cancer and noncancer risk assessment were generally well written and easily understood. Specific comments on technical issues are provided below.

b. In general, the document should be understandable to the general toxicologist/risk assessor. If training materials are available, perhaps selected slides/examples would be a useful addition to the document within the main text or appendices. There are a number of sections, particularly in Appendix C, that require a strong statistical background and experience in the use of mathematical models and will be of limited use to the generalist.

c. The Executive Summary should be rewritten to be a true "executive summary", and to present the key points discussed in the document in, at most, one or two pages. As it is now written it is redundant with long passages from the main body of text but does not include the detail needed to understand the rationale for the choice of defaults and constraints. It does, however, represent a good summary of the technical guidance document for possible use elsewhere. The text in Appendices A and B should be incorporated into the main body of the document in the appropriate sections. As mentioned above, consideration should be given to simplifying or deleting a number of the sections in Appendix C because the information provided, while well prepared, is too technical and will be of little practical use to the risk assessor using the BMD/C method. It is doubtful that terms such as "asymptotic normality", beta-binomial" and "correlation

structure" will have much meaning to the non-statistician. In Appendix C, Section 4 "Assessing how well the model describes the data" and Section 5 "Comparing models" should be incorporated into the main body of the document. Of particular importance are the discussions of the criteria for evaluation of goodness-of-fit and the need for graphical displays. The separate listing of models available in the new software in Appendix E clearly identifies the range of models proposed now and will facilitate revisions to the technical guidance document in the future.

A number of points could be developed further, including: basis for choice of BMD/C, criteria for assessing goodness-of-fit, importance and use of graphical displays, examples of BMD/C derivation, calculation of power for "typical" study designs, and influence of BMD/C on choice of appropriate uncertainty factors and margins of exposure.

d. The examples in Appendix D should be retained and expanded to include more detail on how to step through the BMD procedure. This can be done without being too prescriptive. It might be helpful to include sample input screens and the output generated by the EPA's new software. Providing greater detail in the examples will result in more consistent application of the method and fewer deviations (misuses).

1. Selection of Studies and Responses for Benchmark Dose/C Analysis

a. As a general comment, guidance of the type offered in this document and the constraints recommended are definitely needed to ensure proper use of the BMD method and limit its misuse and application to poor or inappropriate data sets. The discussion on selection of studies and endpoints is appropriate although several minor points should be considered, including the possible bias introduced when selecting a "representative" endpoint for a particular target organ (see also specific comments on p.18-19 of guidance document below).

b. There should be little difference with respect to the selection of cancer and noncancer endpoints for BMD analysis - both should be based on an evaluation of the quality of studies, relevance to humans and reporting adequacy. However, the application of UFs or criteria for acceptability for MOEs may differ between noncancer, cancer and surrogate-cancer endpoints. This also holds true for the NOAEL. This section tends to suggest that studies should be selected to allow BMD derivation rather than presented as qualification criteria for whether BMD should be calculated at all with the implicit default to the use of the NOAEL.

c. The following are some possible criteria for determining when studies could be combined:

1. Statistical evidence that the study attributes are not different (e.g., population variance, group mean responses at the same (or very similar) dose levels).

2. Similarities in the conduct of the studies (e.g., species, strain, group size, protocol, laboratory).

3. Similarities in endpoints and data reporting (e.g., individual values vs. summary statistics).

4. Congruence in modeling results between individual and combined data sets, i.e., does the combined model yield similar values for goodness-of-fit, MLE and lower bound on dose at the benchmark response level (BMR).

5. Ability to clearly state the rationale for combining studies.

The combining and weighting of different endpoints within studies, as with the boron example (#3) in Appendix D, should be approached with caution because, despite the use of expert judgment, it is still subjective. There is a need to avoid any appearance of

manipulating the data and to be able to explain the rationale for weighting endpoints. Transparency is very important to avoid the "black box" aspect that is inherent to the BMD approach (and mathematical modeling in general).

2. Selection of the Benchmark Response Level

a. I agree that the first choice should be the level associated with a biologically significant effect as determined by expert judgment. I also agree that, in the absence of a biologically significant level, the limit of detection should be used rather than a fixed level of response (e.g., 10%). Since this is such an important aspect of the BMD method, the information provided in Appendix B should be incorporated into the section that addresses this issue in the document.

b. The flexible approach proposed for determining the limit of detection is appropriate and will probably be the single most important aspect of the procedure to ensure that it is not applied to poor or inappropriate data sets.

c. I am not aware of any readily available databases or summary information that would enable a quick estimate of the power for typical study designs. However, information should be readily available to the Agency to assign a power to typical guideline studies of various types based upon past submissions and the literature. I recognize that his would take some work but the Agency should move in this direction. A single default power level should be adopted and reflected in this document and the examples in Appendix D. A power level of 80% is consistent with the convention for study design that attempts to minimize the Type II error (β) of rejecting the null hypothesis when it is really true. This is usually restricted to between 10 and 20% yielding a power ($1-\beta$) of 80 to 90%. Use of a power of 50% results in an additional level of conservatism that is unnecessary considering the other health-conservative aspects already built into the risk assessment process. Using a power level of 50% means that the Agency is willing to incorrectly state that a response is above the limit of detection 50% of the time with, the attendant

economic burden, and without providing additional protection of health beyond what is necessary.

d. The defaults for selecting a BMR for quantal and continuous data are appropriate. The risk assessor should make a specific determination that the endpoint evaluated using a quantal model actually represents a biologically significant effect.

3. *Model Selection and Fitting*

a. There is not enough practical experience with BMD to state, with confidence, the "proper" order of model application for continuous data. If a scientific rationale does exist, it should be explicitly stated here. It is likely to be preferable to encourage use of all three continuous models and choose the one with the best fit (statistically and visually). This is the approach proposed for quantal data.

b. Any models that can fit the data well should be considered, although I agree that it would be nice to limit the number of models in the future as experience dictates.

c. Comments on parameter defaults:

i. The default approach for the degree of polynomial should be to use the value that gives the best fit.

ii. A background term should be included. If background is zero, won't it drop out of the calculation?

iii. Use of extra risk as a default implies that the incidence of response is independent in the absence of data to indicate otherwise. This would be inappropriate for compounds that cause an increase in the incidence of a spontaneous lesion when this was not known or suspected beforehand.

iv. The decision not to use a threshold is desirable so as not to "force" the curve and to *let* the model fit the data.

v. Continuous data should be modeled directly so that no information is lost due to conversion to dichotomous data. Either way, the risk assessor still has to decide on an appropriate BMR (biological significance or limit of detection).

d. More detail on the Akaike Information Criteria (AIC) should be provided in the text and appendices with examples of how it can (and should) be used to evaluate how well a model fits the data. I support removal of the high-dose group if it unduly influences the behavior of the model in the low-dose region. A visual "sanity" check on the fit in the low-dose region should always be an integral component of the evaluation of model fit. Additional guidance should be given on when a model does not provide an adequate fit visually.

Is it possible to provide additional guidance on the use of goodness-of-fit? Are there cut-off values for the F statistic (continuous data) or p-value for the Chi-Square analysis (quantal data) that could be used to indicate how well the model fit the data (even though it was statistically adequate)? Could the MLE/LED_{10} ratio be used as a measure of fit (or variability) in the low-dose region, with ratios > 3 triggering additional scrutiny to determine whether the BMD should be used in place of the NOAEL?

In the absence of a clear biological or statistical basis to pick one (equivalent) model over another, the range of BMD/Cs or the geometric mean BMD/C should be provided to the risk manager rather than the lowest BMD/C. Comparison of the BMD/C(s) with the NOAEL (or LOAEL) in every case will help detect problems and will aid in deciding on the appropriate MOE.

4. Use of Confidence Limits

a. The central estimate rather than the lower confidence limit on dose should be the definition of BMD as it was recently used by its originator (Crump, 1995). The central estimate (MLE) is superior to the LED_{10} (BMDL) for the following reasons:

1. The MLE is a more stable and precise estimate of the response at a given dose level (especially in the low-dose region) and is a closer representation of the experimental data.

2. The MLE is almost always within the range of observation while the LED_{10} may not be.

3. The MLE doesn't require a downward adjustment to the uncertainty factor (UF) for intraspecies variability that use of the LED_{10} should include to account for using the lower tail of the distribution of doses for a particular BMR.

4. Adjustments in the uncertainty factors and margin of exposure (MOE) can be made when the MLE is used. It is more transparent to make adjustments at the risk management stage than to incorporate the conservatism in the dose-response assessment.

5. Many other health-conservative steps are already built into the risk assessment process.

Future analyses and simulations should evaluate which of the possible MLEs (05, 10, biologically significant or limit of detection), for a wide range of toxicological endpoint, will approximate the NOAEL best.

b. The default for the method of confidence limit calculation (using likelihood theory) seems appropriate. It should be clearly stated how the confidence limits are calculated by the new EPA software so that other models can be compared and to avoid the "black box" criticism.

c. If it is determined that the LED_{10} should be used, the 95% confidence limits is appropriate because it has been the convention in risk assessment and statistical analysis (except in rare situations when a greater level of significance is needed). Presentation of the 95% and 99% confidence limits could provide the risk manager with a better perspective on the variability of the data in the low-dose region.

5. Selection of the BMD/C to Use as the Point of Departure for Cancer and Noncancer Health Effects

a. Comments on determination of "equivalence of models" is provided above in "Model Selection and Fitting".

b. Other than the reference to the Akaike Information Criterion in the document, there was no information provided to evaluate how it it provides a measure of fit. I am not familiar with this parameter.

c. The BMD/C can be used as a point of departure for noncancer risk assessment as long as the choice of the BMD/C (central estimate of lower bound on dose) is a good estimator of the NOAEL. The use of BMD to extrapolate beyond the range of observation for cancer risk assessment is inconsistent with the original intent and current use of BMD for noncancer endpoints. As it is used now, the BMD is likely to be valuable as a tool for hazard ranking and either the central estimate or the lower confidence limit could be used. However, if BMD is used for low-dose extrapolation (in the absence of biologically-based or case-specific models) it should be recognized, and communicated, that this use has little or no biological basis and is only being done as a matter of science policy.

The use of BMD/C for cancer endpoints will be more transparent than the use of the LMS model and the technical guidance document clearly states what is being done, i.e., drawing a straight line from a point in the range of observation (where we have data) to zero. Some of the key issues are: 1) what the point of departure should be, 2) what biological endpoints should be used (e.g., preneoplastic changes or tumor incidence, and 3) what level of risk should be used. The latter should depend on the former two.

As with noncancer endpoints, the point of departure for cancer risk assessment should be the central estimate (MLE) rather than the lower bound on dose for the same reasons cited earlier. Health-protective steps introduced as a matter of science policy should be incorporated separately by using an appropriate choice for acceptable risk or MOE, depending on the endpoint modeled.

Since it is likely that many chemicals will be evaluated using, at least, the linear default, additional guidance will be needed on the appropriate risk levels, UFs or MOEs to apply to subtle biological changes (e.g., adduct levels, mutation rates or cell proliferation rates) that may be mechanistically related to tumor formation, but have no quantitative relationship yet defined. Presumably, these effects and their response rates will be treated differently than tumor incidence data. The use of non-tumor data may be more appropriate for determination of the MOE and should be used with caution when extrapolating to an acceptable risk due to the uncertain relationship between these effects and tumor formation. For example, DNA repair mechanisms may produce non-linearities in the low-dose region for surrogate endpoints (or tumor data for that matter) resulting in an overestimation of the risk predicted by linear extrapolation.

TECHNICAL COMMENTS ON SPECIFIC SECTIONS OF THE DOCUMENT

I. EXECUTIVE SUMMARY

pp. 1-9. The Executive Summary should be rewritten to be a true "Executive Summary" and to present the key points discussed in the document in, at most, one or two pages. Much of it as it is written now is merely cut-and-pasted from the main body of the document. Presumably, those that need to use the document will read it from cover to cover. The existing executive summary could still be useful in other communications about the BMD method.

II. INTRODUCTION

p. 15, lines 1-11. Application of the benchmark dose (BMD) method to developmental toxicity endpoints and has shown that the BMD (LED_{10}) is a relatively imprecise estimator of the NOAEL. The BMDs calculated to date for quantal developmental toxicity endpoints have generally been 2-3 times lower than the NOAEL unless special models for nested data are used. Continuous data yield better predictions but the goal to have an overall average BMD/NOAEL ratio of 1 has not yet been achieved. The imprecision in estimating the NOAEL may actually exceed the imprecision of the experimental NOAEL (even if derived statistically) in corresponding to the "true" NOAEL. This newly introduced uncertainty can be addressed by using a central estimate instead of a lower bound on dose and/or by adjusting the MOE.

p. 16, lines 25-26. The NOAEL and BMD should be used in parallel until sufficient experience is gained on how and when to use the BMD. It is premature at this time to advocate general use of BMD for endpoints other than developmental toxicity.

p. 17, lines 6-8. The definition should be more general and should not specifically advocate use of the 95% confidence level but rather specify a "statistically derived dose", leaving the door open for use of the MLE.

p. 17, lines 10-11. Stating that the BMD is intended to be used for "low-dose extrapolation" is not consistent with the original intent of the method which was to characterize the dose-response relationship within (or near) the observable range. Using this language blurs the distinction between "risk" and the margin of exposure (MOE) which reflects the uncertainties when estimating a safe dose in humans.

p. 17, lines 15-16. The BMD/C approach may actually increase the uncertainty in NOAEL estimation due to the imprecision in this estimation and use of the LED_{10}. This suggests that a modification in the UF for intraspecies variation should be considered because some of the variation is accounted for by use of the 95% confidence limit.

III. BENCHMARK DOSE GUIDANCE

A. Data Array Analysis - Endpoint Selection

p. 18, lines 21-23. The reader should be cautioned here that, for some chemicals, use of a study that provides a NOAEL from a quality study for a relevant, sensitive endpoint is preferable to a BMD (which may be higher) from a study where it *can* be calculated.

p. 18, lines 28-29, p. 19, lines 1-6. There may be a conservative bias introduced when selecting one of several endpoints representing the same target organ. The most sensitive may be one of lower severity which needs to be reflected in the acceptability of the MOE or choice of UFs (and MF in particular).

1. Selection of Endpoints to be Modeled

p. 19, lines 13-14. Inclusion of endpoints if the LOAEL is 10 times the lowest LOAEL seems reasonable, but what is the scientific basis for this statement? If empirical data are available, they should be mentioned here.

B. Criteria for Selecting the Benchmark Response Level (BMR)

p. 21, line 21. If a power level of 80-90% is used in the design of a typical study then we should use a value consistent with this as a default because it is presumed that a quality study will be used for the critical endpoint. Include the section on discussion of power from Appendix B here.

p. 59, lines 21-22. Use of 0.80 for power should be reflected here (and elsewhere in the document) if this is determined to be the default in the absence of a power for a "typical" study design.

C. Mathematical Modeling

2. Order of Model Application

p. 23, lines 12-14. Explanation should be given to the rationale for the proposed order. Is the preference based upon the desire for simplicity? If there is no sound scientific basis, then all three models should be run to see which gives the best fit.

3. Determining the Model Structure

p. 24, lines 3-9. Why use the simplest model with an adequate (statistical) fit? Why not use a stepwise reduction in polynomial and see which gives the best fit visually and statistically (F statistic, p-value for Chi-Square analysis and/or AIC).

p. 24, lines 10-18. A background term should be included in the model until more work is done to justify its exclusion as a default.

p. 24, lines 19-20. The reference to selection of the BMR is unclear. This section should be expanded briefly and clarified.

p. 24, lines 21-25. An explanation should be given for why "threshold", within the context of BMD modeling, is not a biologically meaningful parameter. This is important because of the expected orientation of the reader who is considering BMD as a tool for evaluating noncancer endpoints (and some cancer endpoints) that are recognized as having biological thresholds. It might be helpful to point out that the BMR and the BMD that is calculated actually describe the biological threshold (one that is not discernible from background).

p. 24, lines 26-29. Modeling of continuous data directly is appropriate because of the loss of information (and precision) by converting to quantal data. Use of a hybrid approach should be mentioned as a possible future enhancement, but should not be advocated (e.g., line 29 "can be used") until it has been validated and there has been sufficient experience with its use for a variety of toxicological endpoints.

IV. Using the BMD/C in Noncancer Dose-Response Analysis

A. Introduction

p. 28, lines 4-25. The level of experience with the use of the BMD approach is accurately depicted in this section and is illustrated by the relatively few chemicals that have been evaluated using this method. The risk assessor should be advised to use the methodology with caution with all endpoints, but particularly with endpoints other than developmental toxicity where most of the experience has been.

B. Effect of the BMD/C Approach on Use of Uncertainty Factors

p. 28, lines 28-29, p. 29, lines 1-5. This very brief paragraph states the obvious advantage of the use of the BMD method when only a LOAEL is available (circumventing the need for the LOAEL-to-NOAEL conversion and attendant conservatism of applying an extra 10X uncertainty factor). However, it fails to identify other opportunities for its use with respect to the choice of uncertainty factors. For example, use of the BMD could impact the choice of the uncertainty factor for intraspecies variability if the LED is used instead of the MLE, since some of the variability is already accounted for by use of the lower 95% confidence limit for the distribution of doses (as predicted by the curve fitting model) at the BMR. Use of the LED and applying the usual UF results in "double dipping" with respect to experimental and intraindividual variability.

It might be worth mentioning that the BMD method can be used as a tool to evaluate data sets to determine the appropriateness of default (10X) UFs. For example, the ratio of the MLE/LED might be a good surrogate for a data-derived factor to describe intraindividual variability (which contributes, at least, some extent to the susceptibility of a subset of the general population) and could be applied directly in human studies for a specific compound or collectively for many studies to develop a distribution of values. This distribution could be used to replace the default distribution currently being developed for probabilistic RfDs.

Recent analyses by Allen et al. (1994) might also shed some light on the current default UF for LOAEL-to-NOAEL conversion. For example, the QLOAEL/$QBMD_{10}$ (7.4) to QNOAEL/$QBMD_{10}$ (2.9) ratio of 2.6 (7.4/2.9) is significantly less than 10. The equivalent ratio for continuous data was 2.2.

C. Dose-Response Characterization

p. 29, lines 8-14. An important distinction is made here between "level of risk" and "degree of protection" which has implications on the use of the BMD as a point of departure. The BMD only provides information about the response at a given dose, which is usually a response in animals. Perhaps with human studies the percent response can be equated to risk within the range of observation, but not below this range, and not for the general population. The MOE reflects the attendant uncertainties in estimating a dose that is unlikely to be without effect in the general population. This is not a description of *risk*, but rather a statement of *safety*.

p. 29, lines 26-29. This proposed wording should be modified, as appropriate, to reflect the consensus position(s) on the use of central estimates vs. lower confidence limits and flexible determination of BMR based upon biological significance and limit of detection as already mentioned.

V. Use of Benchmark-Style Approaches in Cancer Risk Assessment

p. 35, lines 28-29. Why suggest that the value for addressing human differences in sensitivity should be greater than 10 when the current default *is* 10. Besides speculation based upon theoretical considerations, there are no sound data to indicate that this default has not been adequately protective.

p. 36, lines 1-6. Why constrain an adjustment for interspecies differences to "no less than 1/10" if available toxicokinetic data for a specific compound might suggest a lower value.

p. 37, lines 8-26. No real explanation or scientific justification is given on why the LED_{10} was chosen as the point of departure. If anything, the discussion of ED_{10}s, etc., suggests that use of the central estimate is more appropriate.

C. Dose-Response Characterization

p. 38, lines 8-14. As mentioned earlier, several other procedural choices and defaults inherent to the BMD method compound the conservatism already built into cancer risk assessment. EPA should supplement the analysis of Krewski (1990) to demonstrate that a linear extrapolation from the LED_{10} gives "similar" results than from the TD_{50} and lower points. This analysis might justify the use of the MLE which is superior for the reasons stated earlier.

Respectfully Submitted,

Bruce D Naumann

Bruce D. Naumann, Ph.D., DABT
Principal Toxicologist
Merck & Co., Inc.

27 Aug 96

Date

James Olson

James R. Olson (8/28/96)

Comments on the Benchmark Dose Technical Guidance Document (EPA/600/P-96/002A)

1. Selection of Studies and Responses for Benchmark Dose/Concentration (BMD/C) Analysis

The document states (starting on line 22, page 3 and line 13, page 18) that selection of the appropriate studies and endpoints is discussed in Appendix A and in various EPA publications (U.S. EPA, 1991a, 1994c, 1995f, 1996a and b). The selection of the appropriate studies is based on the human exposure situation that is being addressed, the quality of the studies, and relevance and reporting adequacy of the endpoints. The ultimate goal is to identify studies that can be successfully modeled, so that BMD/Cs can be calculated and used in risk assessment. It would be helpful to reorganize this section with this ultimate goal in mind. If only high quality, peer reviewed studies are to be considered, this needs to be stated directly, rather than citing previous EPA documents. If human studies are given priority over animal studies, this also must be clearly stated. Adequate exposure assessment is a key issue in human studies and it is an issue which also needs to be considered in animal studies. Pharmacokinetic considerations, including physiologically based pharmacokinetic (PB-PK) models, tissue dosimetry, body burden, should be considered along with exposure assessment for a given study.

Selection of the appropriate endpoints for the BMD/C analysis is the next important consideration. The document (line 9, page 19) was somewhat vague, stating that the endpoints to model should focus on endpoints that are relevant or assumed relevant to humans and potentially the "critical" effect (i.e., the most sensitive). The document indicates that multiple endpoints can be modeled, but are there sensitive responses (biological vs toxic) which are not appropriate to model? If so, this should be stated clearly. For cancer risk assessment, the document states that it is customary to adjust for background tumor rates, combine fatal and incidental tumors, and correct for early mortality (line 24, page 31). The statement that nontumor data may actually be used instead of tumor data for determining the point of departure for the MOE analysis (line 20, page 35) needs further clarification. Which nontumor data are being considered? It would be helpful to discuss cancer and noncancer data in the same part of the document addressing endpoints.

The section on the Minimum Data Set for Calculating a BMD/C is presented, starting on line 16, page 19). This section is most helpful in identifying data that are appropriate for Modeling and BMD/C analysis. However, more information is needed in this section. The statement that "the number of dose groups and subjects should be sufficient to allow determination of a LOAEL," needs to be defined further. Perhaps the criteria could be more rigorous and recommend that there be more than two exposure groups with a response different than control. An ANOVA analysis would also be appropriate to assess differences between groups.

The section on Combining Data for a BMD/C Calculation (page 20) needs further development. Define statistically and biologically compatible data sets. Should the data sets be combined only when the same species and strain of animal were used? The issue of combining data sets prior to modeling has significant advantages which are briefly discussed in the document.

2. Selection of the Benchmark Response (BMR) Level

Selecting the appropriate BMR level is critical in establishing a BMD/C. The document indicates that there are two bases for specifying the BMR: 1) biologically significant change in response for continuous endpoints, or 2) the limit of detection for either quantal or continuous data (page 21). These general approaches seem appropriate. Biological significance in most cases is not defined for a given response. There is a need to document with references specific cases (endpoints) where a biologically significant changes have been accepted/recommended. Are there cases where a 5% change is considered biologically significant (line 13, page 20)? Appendix B is helpful in providing a more in depth discussion of these issues, particularly with regard to setting the limit of detection for specifying the BMR. Terms such as extra and additional risk need to be defined in the body of the document as well as in Appendix B. The limit of detection approach proposed in the document seems appropriate, however further information is needed to determine the appropriate power level. The default for quantal data appears adequate, however it is not clear what the default is for continuous data.

3. Model Selection and Fitting

Once again, this section could benefit from reorganization, with some of the more critical discussion in Appendix C included in the main body of the document. The order of model application for continuous and dichotomous data presented on page 23 appear appropriate, however there still appears to be doubt regarding the rational selection of models for dichotomous data. Since many end users may not have a background in modeling, it may be useful in the future to limit the number of models. Adequate, rigorous evaluation of various models will be necessary prior to restricting the application of these models. The following comments are directed at proposed defaults for model structure: i) Approach for selecting the degree of the polynomial used appears appropriate, but I have limited background in this area. ii) The default approach for including or excluding background parameters appears to be somewhat in doubt. This parameter may vary for continuous and dichotomous data. In many cases a background response rate is present with quantal data, however this is not generally the case for continuous data. It appears that in all cases the BMD will be reduced when background parameters are excluded from the model. iii) Line 19, page 24 should clearly state and justify that extra risk will be used as a default for quantal data. iv) Not including a threshold parameter appears appropriate. v) The default of modeling continuous data as such is appropriate and is well justified in the document.
A section which addresses the approach for determining the fit of the model needs to be included in the body of the document.

4. Use of Confidence Limits

A new section on confidence limits should be incorporated into the body of the document, with limited reference to appendix C.

5. Selection of the BMD/C to Use as the Point of Departure for Cancer and Noncancer Health Effects

Lines 16-18, page 26 need to be reworded to provide to provide a more clear concluding

statement regarding the determination of "equivalence" of models.

The use of the Akaike information Criterion for comparing the fit of models appears appropriate but should be presented in greater detail in the body of the document. The section on the default approach for selecting the BMD/C to use as the point of departure for analysis needs to be more clearly presented, including a discussion of cancer and noncancer data. Lines 4 and 5, page 27 are not clear. Should the statement in parenthesis read , "(eg more than a factor of 3 compared to others)"?

6. General Issues

The discussions regrading the use of BMD/C approaches in cancer and noncancer risk assessment are useful , but appear to be somewhat out of order in the overall organization of the document. Issues directly relevant to cancer and noncancer risk assessment should be incorporated earlier in the document. For example, a good definition of BMD/C is not given until page 17 of the document. Following this concise description of the BMD/C it would be helpful to describe benefits and application of this approach for cancer and noncancer risk assessment.

Some reorganization of the document would be most helpful to make it more useful for the general toxicologist/risk assessor. The main body of the document should be understandable without extensive reference to Appendix material and other EPA Documents, such as the frequent reference to the background document on this topic (EPA, 1995c).

The examples of BMD/C analyses in Appendix D are helpful and should remain in the appendix section.

Colin Park

PREMEETING COMMENTS

Benchmark Doses: opinions based upon a sample size of 1.

The use of a Benchmark Dose for many cancer and noncancer endpoints is likely a more rational point of departure than the current methodologies.

Replacing the LMS or NOEL procedure is an improvement in the default process, even though the results may be essentially the same. The use of extrapolation factors from an LED or ED more accurately describes the uncertainties in the process and makes it more clear what the process really entails.

I believe, however, that the methodology in the proposed guidance document represents statistical overkill. The precision and effort incorporated in estimating a point of departure must be tempered by the knowledge that the next step in the process is the incorporation of uncertainty factors of between 100 and 100,000. The appropriate factor(s) needed to appropriately protect public health are largely uncertain. Setting up complex methodologies to more precisely estimate the point of departure may fall into the category of killing a fly with a sledge-hammer. In particular, I believe that power calculations, sequential model fitting, and confidence limits are all unwarranted. As a statistician, if I believe we have overdone the estimation procedures, I can only imagine what toxicologists might think about this approach.

Specifics

1. **Selection of Response level.**

   The response level (e.g. 1%, 5%, 10%), should be the same for most applications, otherwise the interpretation of the result will not be consistent, and/or different uncertainty factors will then be applied. I believe that an ED05 or ED10 is the most appropriate point of departure in most situations. Chris Portier has pointed out that an ED01 might be most appropriate for some epidemiological applications since an ED05 often represents an extrapolation upwards.

2. **Model selection.**

   If an ED is used rather than a lower confidence level on it (see later), the model selection issue will be of much less concern. For an ED10 or ED05, a purely linear interpolation between the two responses bracketing the ED value will be sufficiently accurate in most cases and does not require a computer. If there is strong nonlinearity in the data, a nonlinear model will be required but estimation of the ED will be much less model dependent than the calculation of the lower confidence interval. Thus, the output is less model dependent and reflects a biological endpoint rather than a statistical endpoint.

3. ED or LED

I strongly believe that an ED (05 or 10), not a lower confidence on it, should be the starting point for whichever extrapolation process is used. There are a number of reasons for this recommendation:

The primary reason for this recommendation is the apparent precision that is inferred if a statistical bound is placed on one portion of the process. Under the default procedures, risk or Margins of Exposure are generally calculated by taking results from high dose rodent studies, estimating a specific point on the dose response curve (e.g. ED10), extrapolating the rodent results to man, then applying largely empirical extrapolation factors or uncertainty factors. The methodology is useful for setting regulatory limits, but there are large uncertainties, for example, "spanning an order of magnitude", in where the regulatory limit should be.

To apply a complex statistical procedure which reduces the ED10 by about 25% to 50% when extrapolation/uncertainty factors in the range of 100 - 100,000 are then applied does not make sense for the following reasons:

a. The apparent precision added to the answer is not scientifically warranted.

b. There is more than one statistical methodology that could properly be used to calculate the confidence limits, resulting in different answers. For example, the LMS upper bounds as opposed to the upper statistical confidence limits on the MS model.

c. The ED10 is conceptually a very simple endpoint. For example, it can be reasonably estimated in most cases by interpolation between actual data points using graph paper and a ruler. Replacing this simplistic endpoint with one which requires a computer program, resulting in a "black box" answer loses the transparency and simple logic of the process. I think we are losing focus on what the question really is.

d. It is not clear what the resulting interpretation should be. Some may think that since an upper 95% confidence limit is being calculated, the resulting low dose risks, e.g. one in a million, have a 5% chance of being exceeded. It is clear from a publication resulting from a number of expert workshops (ILSI, 1996), that this is not felt to be a valid interpretation by the experts. The total procedure is felt to be more conservative than the above interpretation. But the use of a statistical confidence limit on one small part of the total extrapolation process will mislead risk managers and the public as to the precision and interpretation of the results.

It has been argued that the use of confidence limits rather than the central estimate will reward better experimentation. From a theoretical point of view this is true to a small extent, but from a practical view it will have little or no effect. Bioassays and FIFRA/TSCA tests must meet minimum standards, for instance on

number of animals, to be acceptable to the agency. The question then is what would be gained by doing additional work beyond the minimum?

Simulation work that I have done has shown that doubling the number of animals in each dose group will increase the LED10, but by only a minimal amount; generally by 20 to 35%. For example, the LED10 is often on the order of ED10/2 or ED10 - ED10/2. Under conditions of almost perfect goodness of fit, the upper bound on the theoretical increase in the LED10 resulting from doubling the number of animals may be approximately

$$ED10 - \frac{(ED10/2)}{\sqrt{2}}$$

or 0.65 • ED10, as compared to 0.5•ED10.

I cannot imagine that we would ever double the number of animals for such a small gain.

Another argument in support of the LED as compared to the ED is that the LED would allow optimal experimental designs to increase sensitivity, and thereby increase the LED. Simulation results have also shown this to be only of academic interests; the gains are minimal.

In summary, the use of an LED over an ED has numerous drawbacks with almost no redeeming value.

I would also remind the agency that they convened an expert peer review group to examine, among other things, the question of using an ED as compared to an LED (EPA, 1994). The work group came back with a strong recommendation that the ED be used, for most of the reasons stated above. I strongly urge EPA to follow the advice that they sought and received.

I can understand that the EPA wishes to harmonize the cancer and noncancer methodologies, but I believe that they are harmonizing in the wrong direction. If the agency feels that the ED10 is, on average, an overestimate of a NOEL (noncancer), I believe that using the ED05 would be more appropriate than using an LED10.

I look forward to discussing these issues in the upcoming workshop.

Colin Park
517-636-1159
USDOWYLC@IBMMAIL.COM

William Pease

COMMENTS ON RAF DRAFT BENCHMARK DOSE TECHNICAL GUIDANCE DOCUMENT

William S. Pease, Ph.D.
Environmental Defense Fund

1. Selection of Studies and Responses

The document has a thorough, adequate discussion of the selection of studies and responses that should be subjected to Benchmark Dose (BMD) analysis.

EPA should consider emphasizing "severity" of impact (in contrast with sensitivity) as the principal consideration, as risk characterizations for different compounds (based on the margin of exposure between current exposure and the BMD) will be more comparable if they employ a common level of adverse impact as their point of departure. This is a feature that is (at least rhetorically) possessed by the current NOAEL/UF approach: standard-setting begins for all compounds from a level observed to have no adverse impacts. Uncertainty factors are then applied to derive a reference dose (RfD) that is likely to be below the population's threshold for any adverse impact. The hazard indices that are conventionally used to conduct noncancer risk assessment (the ratio of current exposure to RfD) are therefore interpreted as an exposed population's distance from a "safe" level of exposure. The BMD approach complicates this interpretation of hazard indices because it replaces the NOAEL with a starting point that represents doses associated with different percentage increases in responses of differing severity for different compounds.

EPA has not proposed to address these potential variations in the seriousness of impacts observed at the point of departure with an additional uncertainty factor based on severity, although the nature of the response is a consideration when evaluating the adequacy of calculated margins of exposure or hazard indices (36). To maintain the integrity of an exposure/RfD ratio as a risk characterization tool, it would be ideal if BMD starting points for different compounds could be selected to be roughly comparable in terms of potential impact on human health. Further effort should be devoted to issues raised by attempting to define a "common level

of adverse impact" as a point of departure for standard setting. One alternative would involve scoring observable endpoints in terms of their severity (ranging from measurable modulation of biological function through frankly adverse) and attempting to select the critical endpoint in terms of biological significance. Another option would involve using a sliding percentage incidence scale for different categories of endpoint severity.

2. Selection of Benchmark Response Level

The guidance document identifies three approaches to selecting a BMR level (a biologically significant change in response, the limit of detection, or, as a default for quantal data, a 10% increase in extra risk.) While a biologically significant change is the most conceptually appealing basis for a BMR level (and could be defined to ensure a common level of adverse impact as a starting point), this would require substantial further work by the EPA to obtain public input and scientific consensus.

There are also substantial problems with the proposed limit of detection approach: significant resources and considerable debate may surround the agency's effort to simulate the sensitivity of detecting various endpoints using standard protocols, and the resulting BMR levels may be very difficult to interpret in the standard-setting process. From the example of HFC-134a provided in the guidance document (80-83), it is clear that this approach can produce points of departure that are associated with between 30-50% incidence of adverse effects (Leydig cell hyperplasia, in this example). It is unclear why existing limitations in the sensitivity of standardized tests should be rewarded with the prospect of much less stringent RfD/Cs. The document notes that the BMC using the limit of detection approach is substantially higher than that based on a default of 10% incidence, and that "this difference ... would be translated directly into the derivation of the RfC or other health exposure limit because there would be no difference in the application of uncertainty factors for the two approaches" (83). This approach appears to provide the wrong incentive for adequate testing, rewarding current limitations in sensitivity to detect some endpoints, rather than stimulating improved testing to detect biologically significant alterations in these endpoints.

Adopting all three potential approaches to defining a BMR level will foster confusion in the interpretation of BMD-derived health standards. EPA should rely more strongly on use of a science policy default at this point in the implementation of the BMD method: unless the BMR can be specified on the basis of a biologically significant change in response, a 10% increase in extra risk should be used to specify the BMD/C. (The Gaylor and Slikker approach to converting continuous data to quantal data should also be adopted as a default). EPA should initiate a process to define "biologically significant change" for endpoints that are most frequently reported in standardized testing.

3. Model Selection and Fitting

No comment.

4. Use of Confidence Limits

No comment.

5. Selection of BMD/C to Use as Point of Departure

The statistical criteria proposed for selecting among potential models to derive a BMD/C are reasonable and utilize plausible policy defaults (factor of 3 equivalence, use of Akaike Information Criterion to select among equivalent models, selection of lowest BMD/C as final health protective default) in order to ensure that EPA does not become mired in proliferating analyses and model shopping debates.

6. General Issues

a. Rationale for moving to a BMD approach

The arguments in favor of improving the underlying scientific foundation of dose-response assessment by reducing reliance on No Observed Adverse Effect Levels are persuasive and have achieved consensus in the scientific community. However, these arguments may not be sufficient to justify the transition problems likely to accompany such a substantial change in conventional approaches to noncancer risk assessment. It is unlikely that the science benefits alone (of increased precision in dose-response assessment and improved testing incentives)

are likely to justify the resource costs required to revise existing RfD/Cs or reissue regulatory standards based on the NOAEL/UF approach.

The guidance document presents two policy rationales for moving to a BMD approach to help strengthen the scientific case for change. Unfortunately, these policy arguments have not been sufficiently developed to be persuasive. First, EPA maintains that "we are trying to move cancer and noncancer assessments closer together" (vii) and notes that the recently proposed Carcinogen Risk Assessment Guidelines adopt a modification of the BMD approach (using the LED_{10} as a point of departure for default low dose extrapolation). However, there is no discussion of how the use of similar points of departure will facilitate the comparison of cancer and noncancer risk assessment results. The Presidential Commission on Risk Assessment and Risk Management recently recommended adoption of a margin of exposure (MOE) approach to enable direct comparison of cancer and noncancer risks (presumably, one would compare the magnitude of the ratios of current exposure to the BMD for noncancer effects and carcinogenic effects, and smaller ratios would indicate greater potential public health concerns). This risk characterization use of BMD/Cs could contribute useful information to the risk management process, but it requires that the point of departure for carcinogens and noncarcinogens be roughly equivalent in terms of severity (so that only the ratios of current exposures to common levels of adverse impact need to be compared).[1] This potential policy use imposes restrictions on the selection of the BMR (favoring a fixed probability of similar adverse impacts over the use of the limit of detection method), but these considerations are not addressed in the guidance document.

Second, EPA maintains that the benchmark dose approach "provides a good starting point to develop benefits estimates for non-carcinogens" for use in cost-benefit analyses (42), while acknowledging that the approach

---

[1] The document's example statement "of what can used to communicate [MOEs or reference values] based on BMD/Cs" (29-30) illustrates that the EPAs' current approach is unlikely to improve risk communication and will certainly confuse the general public: "The BMD/C corresponds to a dose level which yields (with 95% confidence) a level of effect in a test species of, for example, 10% or less for quantal data, or that represents a change from the control mean of, for example, 5% for continuous data. This is about the lowest level of effect that can be detected reliably in an experimental study of this design .... Overall, the BMD/C will be a more consistent point of departure than the NOAEL and will not be constrained by the doses used in a particular study."

cannot generally be used to indicate the incidence of morbidity or mortality in exposed populations. There is a growing demand from the policy process that input from health risk assessment be capable of economic valuation for incorporation into cost-benefit balancing, and a real risk that input that cannot be monetized (even if it is indicative of a potential public health problem) will be ignored in decision-making. In this context, selection of a non-cancer risk assessment method that avoids generating incidence estimates and produces output (incomparable MOE ratios of current exposure to different degrees of adverse impacts) that is practically impossible to monetize is unlikely to improve the risk management process. It will certainly not alter the current regulatory system bias towards emphasizing low dose cancer risks at the expense of noncancer effects, because only cancer risks will be assessed using methods compatible with cost-benefit analysis.

Use of the MOE is one of two possible approaches to using BMD methods to place cancer and noncancer effects on a common scale and to generate incidence estimates for cost-benefit analyses. The alternative would be to estimate upper bounds on possible noncancer risks using linear extrapolation below the BMD, as is done for cancer risks. The guidance document does acknowledge that BMD methods could be used to determine the number of cases expected for various noncancer endpoints "when exposure levels are in or near those in the experimental data range" (42), but it dismisses use of BMD methods to generate lower dose risk estimates without discussion.[2] The document does not address existing low dose risk assessments that have been done using the BMD method, [3]

---

[2] The document states only that "Although the BMD/C is associated with a defined level of risk in the study population from which the BMD/C was calculated, it would be misleading to translate that to a level of risk at the MOE or RfD/C or other reference value" (29).

[3] See W. Pease, J. Vandenberg and K. Hooper (1991). Comparing alternative approaches to establishing regulatory levels for reproductive toxicants: DBCP as a case study. Environmental Health Perspectives 91:141-155.

M. Meistrich (1992). A method for quantitative assessment of reproductive risks to the human male, Fundamental and Applied Toxicology, 18:479-490.

D. Hattis et al.,(1988). Male fertility effects of glycol ethers: A quantitative analysis, Center for Technology, Policy and Industrial Development, Massachusetts Institute of Technology.

D. Hattis (1991). Use of biological markers and pharmacokinetics in human health risk assessment, Environmental Health Perspectives, 90:229-238.

nor does it make an organized presentation of the issues involved (both strengths and weaknesses) in this type of application. It does not present potential applications that have been important components of the basic scientific papers by Kimmel and Gaylor promoting the BMD approach. [4] This approach should be presented in the guidance document as one option for fulfilling demands from the policy process to apply common methods to cancer and noncancer effects, and reasons for its rejection should be provided.[5]

b. Discussion of other impacts of BMD approach

In the political arena, a primary question about EPA's new proposed methodology will be its impact on existing standards that have been established using the NOAEL/UF approach. There is a good discussion (15) of the scientific evaluations that have been done to date on the relationship between NOAELs and BMD response levels, but this information is not used to make any general statements about whether existing RfDs will be more or less stable if EPA makes a transition from NOAELs to BMDs as a starting point for deriving health-based standards. This issue should be addressed explicitly, using examples of current RfDs. How many compounds are likely to require smaller UFs in standard derivation due to the elimination of the need for factors accounting for LOAEL to NOAEL extrapolation or modifying factors based on data quality?

K. Silver et al. (1991). Methodology for quantitative assessment of risks from chronic respiratory damage: Lung function decline and associated mortality from coal dust, Center for Technology, Policy and Industrial Development, Massachusetts Institute of Technology.

[4] Kimmel and Gaylor illustrate how the BD approach can be used to provide upper bound risk estimates on exposure levels deemed allowable by the conventional NOAEL/UF approach.

[5] The 1995 RAF report on The Use of the Benchmark Dose Approach in Health Risk Assessment provides two principal arguments against using the BD approach for low dose noncancer risk estimation: 1) BMD models are statistical models that do not incorporate biological *mechanisms, so predictions may be in error at low doses (7, 61). This argument* can as easily be raised about linear extrapolation applied for low dose risk estimation on carcinogens. Most BMD models can be given a biological interpretation and support alternative assumptions about the existence of a toxicological threshold. Some of the no-threshold models discussed (e.g., Weibull) do not require linearity at low doses. We actually may have more variety in assessment models available to us for non-carcinogen risk estimation than for cancer QRA, providing more opportunity to avoid error . 2) Linear extrapolation at low doses is inappropriate for non-carcinogens, so BMD models will give us erroneous results. While BMD models must be applied with care to non-carcinogenic risk estimation, there are a number of *toxicants and endpoints where linear extrapolation may be reasonable. Moreover, Crump* (1976) provides a general rationale supporting low dose linearity if a toxicant's damage is additive to background mechanisms of disease, which also clearly supports some uses of BD methods for low dose risk estimation

How many RfDs are likely to change by a substantial factor, and how does EPA propose to handle requests from interested parties to review and revise existing RfD/Cs?

William Perry

Comments on Benchmark Dose Technical Guidance Document - U.S. EPA External Review Draft August 9, 1996

Adam Finkel and William Perry
Occupational Safety and Health Administration

I. SUMMARY

This Draft is far too involved statistically for what it tries to do - estimate a BMD or point of departure in the observable range (10% increase in general) for the application of safety or uncertainty factors. We suggest that if all EPA wants to do is to replace the NOAEL approach with a BMD, then a simple polynomial of degree k-1, with unrestricted parameters and without step-up or step-down procedures, can be used to fit both quantal and continuous data. Nested models for reproductive risk assessments can also be put into polynomial form. If EPA wants to be more conservative in the BMD range, other simple procedures for polynomial fitting by step-up procedures with the low-dose portions of the dataset could used.

Of greater concern to us is EPA's emphasis on a methodology which employs the BMD_{10} as a point of departure for the use of uncertainty factors. We believe that the uncertainty factor approach should be limited and that modeling should be relied upon more frequently for human risk extrapolation. We believe that the types of rules and guidance presented in the Draft can be used to extend the range of extrapolation down to at least 10^{-2} and in many cases to the 10^{-4} level for noncarcinogens and even lower for carcinogens (see, for example, Baird et al., 1996. Noncancer risk assessment: A probabilistic alternative to current practice. Human and Ecological Risk Assessment, 2:78-99). To have confidence in extrapolation down to the 10^{-4} range requires assumptions which would substitute a probabilistic structure for EPA's planned continued use of uncertainty factors. Even extending the range to 10^{-3} after all would enable other regulatory agencies, such as OSHA, to use the results directly in regulatory decisions. We suggest below, for example, that the critical dataset be modeled in

"human equivalent doses" as a substitute for the "animal-to-human" uncertainty factor. Likewise, the lower confidence limit on the ED_{10} could be considered the "incomplete data base" uncertainty factor. We have also offered a reference for building inference structure into the "human variability" and "subchronic to chronic" uncertainty factors. We believe such an approach offers much more promise especially to the risk manager, who can be presented with a reduced range of uncertainty and a distribution of human threshold doses, instead of RfD's based on uncertainty factors.

In short, we believe that the Draft is too complex for its stated purpose of BMD_{10} estimation mainly from animal studies, and that its focus should be on methods for low-dose extrapolation for human risk assessment.

II. INTRODUCTION

The Draft document seeks to provide general rules and procedures for the "modeling" of data in the observable range, the main purpose being to derive a standardized measure of dose as a point of departure for use of uncertainty factors to derive an RfD. That standardized measure is the lower confidence limit on the 90% confidence interval (derived by the asymptotic distribution properties of the likelihood ratio statistic) of the estimated dose level which produces a 10% response in the selected study or studies. The rules for use of the procedures are general enough for them to be recommended for both continuous and quantal data, for cancer, reproductive, developmental, and neurotoxicology risk assessments, as well as for all other effects for which EPA wants to estimate RfDs or RfCs. Several models are prescribed with general freedom for additional curve-fitting allowed. The document promises that the EPA will make the necessary data-fitting software available - 13 models (7 for quantal data, 3 for nested quantal data, 3 for continuous data), listed on page 94. Supposedly, the continuous data models could be fit to individual data, but without adjustments for either time or covariates.

Our comments will be divided into two sections - general and specifics. The GENERAL section will cover the Draft's approach to risk assessment modeling and the SPECIFICS sections will be comments on the individual pages, plus the limit-of-detection concept.

III. GENERAL

The Draft is long on technical details for performing all kinds of modeling but is short on purpose. EPA's proposed cancer guidelines remarks that EPA wants to scale back the approach for estimating cancer risks so that risk estimates do not appear to derive from an overly sophisticated approach that is not warranted. However, the BMD approaches outlined by this document seem to fall into the same trap. If the purpose is really to estimate a point of departure for applying the standard safety factors or generating a margin of exposure, why go to all the trouble to examine fits of several models and select ones based on complicated statistical criteria? Why not just employ a few of the most flexible models and take it from there since EPA acknowledges that the proposed models have no biological significance and the primary purpose of BMD analysis is to provide consistent points of departure across the board?

In general, although this Draft has many good procedures and is quite well written and documented in most areas, we believe that it misses on two major points; 1) that the purpose of modeling is to make the most use of the data and knowledge about the biological process of the disease or condition; and 2) that the purpose of risk assessment is to predict effects in humans at the environmental levels they are exposed to, not effects in animals at experimental exposure levels.

With respect to the use of modeling to make best use of the available data, we believe that this Draft fails in this regard, primarily because it has already chosen to depart at a 10% response level, rather than seek a model which can be used to extrapolate to lower response levels. Modeling is done for either or both of two main

purposes - prediction of adverse effects and risks related to exposure to hazardous substances and determination of analytical factors that affect estimation of risks. The Draft has chosen to emphasize only the former, with the possible exception of discussing the order of model applicability on page 5 prescribing that "continuous data: A linear model should be run first." (Supposedly, the reason for fitting a linear model is to test the slope parameter, but this appears not to be the case here, since for dichotomous data, on pg. 3, a linear, or one-hit, model is not even mentioned as a possibility). The Draft then prescribes many statistical and other procedures, which will be discussed below, to get to the BMD_x, but in the end the reader is left wondering whether any of this actually makes any difference. Our feeling is that if all one wants to do is to predict a BMD_{10}, then why not use an unrestricted (parameter) polynomial of degree k-1 to fit the data, using likelihood fitting procedures for both quantal and continuous data? Procedures could be used which would put higher weights on the responses in the 1% to 20% range.

Although we do not advocate the use of an unrestricted polynomial for estimating the BMD_{10} as a general approach for risk assessment, we believe it is superior to the suggested procedures for BMD_{10} because it: a) provides a consistent approach for both quantal and continuous data; b) generally provides a better prediction in the region of interest; c) provides a lower confidence limit more related to sample size and less reliant on model structure or fit; and d) establishes the exercise as a pure curve-fitting procedure for prediction, which it is meant to be - a "truth in advertising" if you will.

With respect to the importance of predicting effects in humans, the Draft actually starts off on the right track by stating on page 19 (also pg. 3) that "the selection of endpoints to model should focus on endpoints that are relevant or assumed relevant to humans and potentially the 'critical' effect (i.e., the most sensitive)". It further recognizes that different experimental designs will produce different LOAELs and NOAELs, and so suggests modeling datasets "if their LOAEL is up to 10-fold above the lowest LOAEL" (pg.3). What the Draft misses here is that the suggested modeling will

still be done with animal doses, not human equivalent doses. A consequence of calculating a BMD based on the animal data is that, where there are differences in pharmacokinetics between humans and animals, the wrong endpoint may be selected for modeling.

An example of this is seen in case study Example #4, 1,3-butadiene (BD) and ovarian atrophy, pgs. 91-93. In this case study female mice chronically exposed to 6.25 or 625 ppm BD via inhalation for up to 2-years exhibited ovarian, and uterine atrophy (males similarly exposed exhibited testicular atrophy). The draft selects ovarian atrophy as the most sensitive endpoint. However, Sipes et al in a series of papers presented strong evidence that the BD metabolite diepoxybutane (BDE) is the agent responsible for ovarian atrophy, and other authors have presented papers showing that the mouse metabolizes BD to BDE at a much faster rate than does the human or rat, and has much higher blood and tissue BDE levels. This suggests that a) the most sensitive reproductive endpoint for humans is not ovarian atrophy or uterine atrophy (both of which may be correlated because ovarian atrophy is most likely caused by loss of ovarian hormones), but probably testicular atrophy whose links with BDE have not been established, and b) that risk assessment modeling should be based on human equivalent dose, wherever possible, with a default $(\text{body weight})^{3/4}$ species conversion factor. Therefore, we suggest the following wording be added to pg. 3 line 25:

"Similarly, whenever possible, the dose metric should be modeled as human equivalent dose. Experimental studies or PBPK modeling may be used to establish human equivalent doses, or, as a default, dose/ $(\text{body weight})^{3/4}$ may be used for animal-to-human dose conversion. Such a species conversion factor should be interpreted to convert from the typical animal to the typical human, but should not be interpreted to account for inter-individual variation among humans.

Likewise, to be consistent with this thought , on page 18, line 20 add the sentence:

"Studies based on humans should be given priority as critical effect studies wherever possible and appropriate."

In a similar vein we suggest including a paragraph with respect to applying the BMD method to internal dose such as found in the middle paragraph page 28 in reference EPA 1995C.

Most important with respect to human risk assessment is the EPA concept that modeling should not be used to extrapolate below the observable range. We believe that EPA has chosen this wrong road to travel for many, if not most, of these suggested applications. We believe that the models fit to the data should be used for extrapolation, how far to be determined by both biological and statistical considerations. At a minimum, we would extrapolate to the 0.01 level and seek to go to the 0.0001 level for noncarcinogens (lower for carcinogens) depending on rules for application. We also believe the methodology which should be developed should seek to minimize the use of uncertainty factors, much like the above suggested use of "human equivalent dose" in the modeling does for the animal-to-human uncertainty factor. For example, Baird et al (1996) use the slope parameter of the log-Probit to develop distributions for the other uncertainty factors which relate to the adverse effect. The result is a probability distribution of human threshold doses, from which the risk manager could estimate the likelihood that various exposures were below the human population threshold.

IV. SPECIFICS

A. Limit of Detection.

The concept of limit of detection (LOD) is discussed on pages 17, 20, and 21 as an alternative to biological significance of an effect. The statistical primer on power and sample size is given on pages 59-62, and an example of its use is presented as example #2, HFC-134a, on pages 80-84. However, it is still unclear when and why it should be used in BMD calculations. Power and sample size considerations are

important when designing a study or when statistical significance has not been found. However, in the example, there is both a NOAEL and LOAEL based on statistical tests of Leydig cell hyperplasia, and the LOAEL represents a statistically significant extra risk of 22%. The calculated BMC_{10} for both the polynomial and Weibull models (no threshold) is 11000 ppm. We were unable to understand the use of LOD here at all. The authors also use the LOD for illustration in the 1,3 BD example, #4, pg. 92, also without any enlightenment. If there is any usefulness of LOD in BMD applications, then maybe the authors can clarify the explanation. Perhaps it fits under example #1, carbon disulfide, whose biological significance of BMD_{10} is questioned, but whose BMR is at a higher level of response than the highest dose group in the study.

Also, we question the significance of using a limit of detection for an effect as a benchmark response in cases where the risk assessor cannot determine a biologically significant benchmark response. Why should the size of a typical animal study be the basis for defining a BMR? If there is uncertainty about whether an effect is biologically significant, or how much of a change is deemed to be significant, what is the point of modeling the response since we can't translate the result into anything that is meaningful in terms of risks to people?

B. Specific line citings

Pg. 1 lines 16-18. The LOAELs and NOAELs are usually based on statistical significance. Add something like "while the NOAEL is the highest dose at which adverse effects are not significantly increased over controls."

Pg. 2. Line 6. There is no **nonlinear** extrapolation procedure proposed as a default procedure. Furthermore, we thought the proposed cancer guidelines did not recommend a point of departure for case-specific or biologically-based models. We thought the point of departure is strictly a default notion of departure to a line through zero.

Pg. 2. Line 20. The Tukey NOSTASOT approach for determining a NOAEL

considers the slope of the dose-response curve. See reference EPA 1995C pg. 29.

Pg. 4. Line 5. "At a minimum, the number of dose groups and subjects should be sufficient to allow determination of a LOAEL." It seems to us that as long as a BMD is being used, the lower confidence interval will reflect the small sample size, and that this restriction is not necessary.

Pg. 4. Lines 7-10. For a BMD_{10} we believe that one exposure group with a response in the 10% to 20 % increase range should be enough for an estimate.

Pg 4. Lines 11-14. The Draft states that modeling is not appropriate if all responding groups show greater than 50% response, or where there is a clear plateau. If the first dose group shows a response on the order of 50% or so, why should BMD be rejected since we are only extrapolating down to a 10% risk level? The decision whether BMD analysis is appropriate should be based on consideration of the whole data set. It is easy to foresee a sigmoid dose-response with a very steep slope (e.g. as seen with certain hormones), so that there is very little difference in dose between a BMD_{10} and a BMD_{50}. We don't see why this can't provide good information if the BMDs are close.

Pg. 5. Line 2. **Also, Example #1 Carbon disulfide pgs 76-78.** The example given for >10% decrease in nerve conduction velocity (NCV), actually represents a decrease from a control *population*, <u>not</u> an *individual* decrease of 10%. Thus, the question of biological significance becomes a problem of a dose of carbon disulfide which shifts the whole population one standard deviation to the left. While 4.5 m/s decline in NCV won't harm the average person (45 m/s), it will be much more critical to the person already on the lower tail. Thus, the question should be rephrased as at which dose will NCV become an impairment and what percent of the population will be affected at that dose. We question whether the 10% reduction is the correct BMR. Finally, since cumulative exposure was found to fit the data better than workplace exposure concentrations (pg. 78 lines 2-3), the BMD should account for this in its suggested. A BMC of 12 ppm or 20 ppm is not an adequate representation.

Pg. 5. Line 20. "A linear model should be run first." Why? Furthermore, the sequence is odd, since in the following sentence,. "If the fit of a linear model is not adequate, the polynomial model should be run --". But, if the polynomial (k-1) doesn't fit,

then the advice (pg 6 lines 14-17) is to drop the highest dose together with the highest degree in the polynomial. This doesn't quite fit, however, with the statement on pg 6 line 16-17 "to select the model with the fewest parameter that still achieves adequate fit.", or that on page 73, line 9 that "Finally, it is generally considered preferable to use models with fewer parameters, when possible." If this is the objective, then the polynomial procedure should be a step-up rather than a linear followed by a step-down. For risk assessment extrapolation modeling we would support a step-up since it provides more conservative low-dose risk estimates, but for BMD_{10} estimation alone as a point of departure, a step-down procedure is preferable, since prediction within the observable range is the primary objective.

Pg 5. Line 23. To be consistent with line 20, if for no other reason, a linear (one-hit) model should be run first. Also, Probits and Log-Probits have much more history in bioassay experiments than does the logistic, and the Probit gives nearly identical results in the observable range. We would substitute Probit for logistic and include log-Probit on page 94.

Pg 6. Line 5. "Guidance" should be changed to "rules" or "commandments" as long as "will" is the verb mood on lines 7, 12,15, and 16.

Pg 7. Line 2. (also pg 24 lines 21-24). "Because (threshold) is not a biologically mneaningful parameter". If the threshold term is not a biologically meaningful parameter, then what in this exercise is? We can understand EPA's reluctance to use a threshold term based on a) EPA's desire to protect public health, b) EPA's concern that declaring a threshold for some compounds and not others would be complicating, and c) the limited applicability of the threshold concept in the BMD calculations, but EPA should be candid about it. Also, since it is already in both the THC and THRESH software, from which EPA will be basing its software, why not include it (line 5)? Crump's original paper in 1984 contained a few very nice examples where it seemed clear from the data that there was a likely threshold, at least for all practical purposes. It would be difficult to defend using a non-threshold model in such circumstances, or at least estimating a threshold directly from the empirical data if not from the models themselves. Finally, if you don't include threshold, remove it from the example discussion on page 82.

Pg 8. Line 5. "Models should be eliminated that *do not adequately describe* the dose-response in the range near the BMR". This should be defined better. We can define a statistical concept of low-dose goodness-of-fit in, say the 0 to 20% range. We believe this is an important concept, especially since the authors claim that the threshold parameter is unimportant (pg.7 line 3-4).

Pg. 9. Line 6. Can there be a reference for the Akaike Information Criterion, if not in the summary then in the main text?

Pg 9. Line 4. If two or more BMD model forms yield BMD estimates within a factor of 3, EPA recommends ranking models according to the Akaike Information Criterion. This is perhaps the best example that EPA's approach tends towards being unjustifiably sophisticated (even more so than for cancer risk assessment in this particular case). Are factors of three really distinguishable, and what is wrong with presenting BMDs as ranges, which is more honest?

Pg. 10, Lines 12-13. The LMS procedure can produce <u>any</u> desired estimate (MLE, mean, 95th percent UCL, 99th percent UCL, etc.) or an entire probability distribution for the q_1 term. The fact that EPA has always ignored all estimates other than the 95th percent UCL does not mean that this is a constraint inherent in the LMS procedure.

Pg. 11, Lines 7-8. We think that it is inappropriate to refer to "conclusions" about mode of action. Except in rare cases, these will be "suppositions."

Pg 29. Lines 9-14. EPA states that it would be misleading to translate a BMD for a defined level of risk from an animal study into a level of risk at the RfD or MOE for humans, primarily, it seems, because the average level of effect at the BMD is less than the BMR (of 5 or 10%) given that the BMD is defined as a lower confidence limit. But the whole point of using the LCL for the BMD is because the average level of risk may be as high as the BMR given the statistical limitations of the underlying study. Thus, we feel that this statement provides no justification for not extrapolating down to moderately low risk levels on the order of 1% or 0.1%.

Pg. 36, Lines 1-6. This should be clarified to explain whether the 10-fold factor is to be applied above and beyond the conversion via $BW^{0.75}$ (itself a factor of nearly 10 for

mice) or whether the $BW^{0.75}$ conversion is meant to correct for "interspecies sensitivity." If the latter is meant, the document ought to make clear that this is not a phenomenon related to inherent susceptibility, but merely to allometry.

Pg. 38, Line 6. Despite the flippant and excessive use of this term outside EPA, there are no "biased" defaults, only defaults that properly serve different purposes. The $BW^{0.75}$ can be properly described as a "central tendency" default, but it should not be given a superior moral footing over truly health-conservative assumptions.

Pg. 39, Line 7. The document should make clear that, at present, EPA's cancer risk assessment methods make no allowance whatsoever for inter-individual variations in human susceptibility. If newer approaches involving the BMD do succeed in "addressing the principal sources of human variability," this will be a first for the Agency (and will then require more explanation of how to compare the newer, biologically-plausible methods with the implausible ones currently in use). However, we are skeptical whether such a fundamental (but critical) change will actually occur.

Pg. 41, Lines 1-3. Only a stout utilitarian would claim that the sole purpose of a benefits analysis is to estimate the adverse effects in a population. Despite the views of "mainstream economists", "benefits to society" are nothing more than a collection of benefits to individuals; if some individuals face unacceptably high risks, their welfare must be factored in to the social analysis.

Christopher Portier

DEPARTMENT OF HEALTH & HUMAN SERVICES
Public Health Service

National Institutes of Health
National Institute of
Environmental Health Sciences
Division of Intramural Research
Laboratory of Quantitative and
Computational Biology
P. O. Box 12233, MD A3-06
Research Triangle Park, NC 27709
T:(919)541-4999/F:(919)541-1479
e-mail: portier@niehs.nih.gov

Eastern Research Group, Inc.
110 Hartwell Avenue
Lexington, MA 02173-3134
Attn.: Kate Schalk

RE: EPA/600/P-96/002A Benchmark Dose Technical Guidance Document

In general, I strongly support the concept of moving away from the use of NOAEL's and LOAEL's when estimating doses to which uncertainty factors can be applied. The current draft guidelines are a clear improvement over the existing methodology and will clearly be an aid in moving this debate forward and discontinuing the use of LOAEL's and NOAEL's. However, I feel that, in several areas, additions and clarifications in the document will greatly enhance it's utility and function. Of major interest to me would be augmentation of the document to support "mechanistic linkage" in the evaluation of benchmark doses, better guidance on why a certain measure of risk and BMR is chosen, proper use of information contained in confidence bounds, and less stringent choices on the types of statistical analyses to be done. These are detailed in my comments below.

On the first point, the guidelines make no mention of the use of mechanistic models for calculating effect doses. The document concentrates on the analysis of single data sets and fails to recognize that many of the endpoints being studied which do not reflect morbidity or mortality (mechanistic data) are "mechanistically linked" to a morbidity or mortality endpoint. In many cases, the mechanistic data can be obtained at lower doses effectively stretching the dose-response curve for the morbidity/mortality endpoint. This oversight could well lead to reduced use of mechanistic models in favor of analyses of single data sets even though there is clear, "mechanistic linkage". That issue aside, there is little motivation in this document for researchers to pursue the underlying cause of a specific response, since, the criteria for inclusion, exclusion and presentation of the dose-response analysis does not identify or suggest uses for presumed "mechanistic linkage".

From a statistical perspective, the document lacks balance. Certain aspects of the analytical procedures appear to be extensively detailed (e.g. use of likelihood-based confidence bounds), some are ignored (e.g. choice of statistical tests for goodness-of-fit), and some appear to be poorly developed (e.g. what measure to use for continuous data). There needs to be better balance in the document concerning the degree to which statistical procedures are outlined. In general, I feel the authors should use much broader terms focusing on process more than on technical merit of one method over another.. For example, it would be appropriate to require a sensitivity analysis of model choice on the resulting BMD/C but inappropriate to restrict Weibull models to a shape parameter of ≥ 1 without any concern for the data, the endpoint and the effect of the restriction. Finally, the document fails to concern itself with assumptions (and their impact) which go into the calculation of confidence bounds. Just as effective doses are sensitive to model choice, confidence bounds can also be sensitive to assumptions which are generally ignored.

I also feel it is inappropriate to present the confidence bound on an estimate of effective dose without presentation of the point estimate. There is a suggestion in this document that this confidence bound represents population variance rather than variance on the estimated mean for the population derived solely from the single data set studied. This should be clarified. In general, I would prefer that the point of departure for extrapolation be the point estimate of the effective dose for the BMR and that the confidence bound be presented as partially informative of the uncertainty in this estimate. The reasons for this opinion are:

- point estimates represent our central tendency estimate of what is correct (especially for mechanistic models); as such, they provide our best understanding of the response and accurately portray this to the risk manager

- confidence bounds are subject to (generally) hidden assumptions, do not generally cover all sources of variation and often are misinterpreted; confidence bounds simply represent the range of uncertainty in an estimator based upon the data at hand and a set of assumptions concerning distribution of these data and how we intend to derive the bound

- confidence bounds are less sensitive to the trend in the data than are the point estimates; as such, focusing on the confidence bounds alone can cause you to miss important consistencies in the point estimates which may alter your opinion about the quality of the full set of estimates

- there is no other area in which confidence bounds are interpreted as the estimate for policy decisions; in this case, it is simply used as a tool to add unknown conservativeness to the estimate of the dose for departure to extrapolation.

- when uncertainty factors are used, the appropriate place for variance in the estimate of the effective dose is as an uncertainty factor; as such, when consistent trends are observed across numerous data sets, the uncertainty factor could be altered to reflect stronger belief in the effective dose chosen

It appears to me that the choice of the BMR is not based upon the proper objective. In this document, the method detailed for use of a 10% BMR is tied to the ability to detect statistically significant findings in a routine (or specific when power calculations can be done) experimental protocol. However, the main focus should be on the trend in the data, the consistency of the observed patterns of dose-response and the degree to which one must extrapolate to estimate the effective dose. At the least, if statistical significance was insisted upon as a criteria for calculating the point-of-departure for extrapolation, why not find the highest estimated effective dose which includes zero in it's lower confidence bound. I do not advocate this concept, but at least it is consistent with the stated purpose of the approach.

I would prefer a BMR which is at or below the lowest dose used in the study; in this case I feel a little bit of extrapolation would be good as it would be sensitive to curvature in the data. None of these issues are considered.

Below are specific comments on the Executive Summary. These comments also relate to the appropriate sections of the document; the comments are not repeated.

Page 2, line 24: Is this the proper reference to the first use of this procedure or similar procedures? I seem to recall an Interagency Regulatory Liaison Group publication from the late 70's with a similar theme. David Gaylor was part of that group and may be able to locate the reference.

Page 3, line 3: This statement, concerning variability, is in common use in this area and yet I am uncertain what is meant here. Does this imply accounting for all known sources of variance? Does this include bias? Is it implying population variance or variance on the estimate of the dose relating to the BMR?

Page 3, line 25: "critical" needs a better definition.

Page 3, line 28: What is the justification for this 10-fold rule? It would be easy to define a study design which violates this rule but provides the excellent information for a BMD/C analyses. For example, a study with 30 dose groups with 5 animals per dose may be superior to a study with 3 dose groups of 50 animals per group, but would surely yield a higher LOAEL. Why does the LOAEL even enter the discussion here? Should not the inclusion of a study be based upon its relevance, quality and design rather than overall effect in pair-wise tests?

Page 4, line 3: "critical endpoints" has not been defined.

Page 4, line 5: Why is the determination of a LOAEL so important? What is the role of trend in the data? What is the role of data from several studies with the same general trend and lacking pairwise statistical significance? Again, I feel it would be more appropriate to base inclusion upon scientific merit of the study rather than statistical significance in pairwise tests. However, if a statistical criteria is needed, it should be tied to the ability to estimate model parameters rather than a test of significance.

Page 4, line 7: This statement is unsupported and, in my best statistical opinion, incorrect. The example above could clearly be devised to provide no pairwise test significance, yet very sound data for estimation of a BMD. This bullet should be removed or modified to reflect an assessment of the quality of the data for modeling, not testing.

Page 4, line 11: This point is also too vague. All three points in this section could be handled in a better way. To do a BMD/C analysis on any data set, you must have a design which results in sufficient data to have estimable model parameters (a clearly defined statistical term to allow for dose-response analysis) and sufficient degrees-of-freedom to allow for variance estimation. Once you have decided the data are scientifically valid and the endpoints is relevant to public health, that covers it.

Page 4, line 28: "Biological significance" is undefined. Do you mean of clinical importance to a healthy individual (such as your example suggests) or of public health importance to a population? There is a considerable difference in the two perspectives; a halving of the PFC response in a healthy individual may be no problem, in an immune challenged individual, it could be devastating. Since populations contain both, one must account for distributional shifts in the population. Maybe there needs to be a guideline document specifically addressing population risks for clinical measures.

Page 5, line 3: Why use the ability of a statistical test to determine an effect (limit of detection) as the main criteria in an estimation process? Should you not instead focus on where extrapolation begins and interpolation ends? Or maybe on signal-to-noise on prediction of doses corresponding to a certain BMR (see general comment above)? While testing and estimation are indeed linked as statistical procedures, the linkage used here is inappropriate.

Page 5, line 8: 10% is too high; in some cases, this would be an extrapolation upward. I would prefer 1% but might not complain too much if it were 5%.

Page 5, line 15: This statement discourages a variety of alternative methods not "sanctioned" in this document. Included in this list would be a number of well studied concepts such as mechanistic modeling, Hill models for receptor-ligand binding, bootstrapping data, other confidence region procedures, and different risk measures such as life expectancy. Adequate software for analysis is absolutely required for calculating the types of quantities described her, but it has never been (at least to my recollection) EPA's policy to have software be the defining factor in their guidelines and I would hope this will not be the initial case.

Page 5, line 20 vs. line 23: There is no justification for the discrepancy in approach to continuous vs. quantal data. So, why the difference in linear first, etc.? Would it not be better to recommend that any models used be subject to certain rules concerning estimability, variance, goodness-of-fit, etc. rather than these seemingly arbitrary restrictions?

Page 6, line 6: This restriction is unacceptable from both a statistical and a biological point-of-view. Instability in data results in instability in a model. An arbitrary constraint on the model will result in a misrepresentation of the quality of the data for fitting the model. In addition, some biological processes could result in powers of < 1. Finally, this represents a restriction in how data should be analyzed that has little to do with the individual data set being studied. As mentioned several times earlier, data analysis should be flexible and reflect the quality of the data; not subject to restrictions based upon a requirement of "stability". This bullet should be dropped.

Page 6, line 19: This bullet could be considerably improved. Within the context of the model, it is always possible to use a statistical test to evaluate the importance of the background parameter in the model. Even without a formal test, a sensitivity analysis the importance of background to the BMD/C calculations could be performed. The subjective strategy presented here does not seem to follow sound data analysis rules.

Page 7, line 1: Because none of the models used provide for a protective effect, you will have problems with certain data sets if you fail to allow for a flat region. An example would be the effect of cyclophosphamide on mortality from listeria infection (Luster et al. FAAT 21, 71-82, 1993). Here, because of a demonstrated protective effect at low-doses, there is an effective threshold for this response. In immunotoxicology, this pattern of response occurs often enough to warrant concern in any guidance you propose.

Page 7, line 6: Most (if not all) induced biological effects have a theoretical maximum. For quantal data, that maximum is generally assumed to be 1. For continuous endpoints, such as gene expression, it would be some maximal expression rate. The issue of what to use as a measure seems clear; it should always be extra risk where the measure is BMR=(response at effective dose - background response)/(maximum response - minimum response). The question then becomes one of proper design to allow for estimation of the maximum response and the use of models which can account for a maximum level of effect. In cases where the data are insufficient to define the maximum effect, historical information, common sense or an alternative measure would need to be used. This is a point in the document where the authors could provide some guidance on the best designs for effective dose calculation.

Page 7, line 20: Why is it necessary to be so specific on the methodology for confidence bounds? What role would bootstrapping play here? This would certainly be difficult to apply to mechanistic models. Finally, as per my general comments, I do not like the concept of portraying a confidence bound as a point estimate.

Page 8, line 2-24: There are no biological considerations given here (the previous line suggests they would be here). While I agree that G-O-F and graphical presentations are important, they should not be the only considerations. Finally, there is no discussion of the use of models with more parameters rather than discarding data.

Page 9, line 1-17: In keeping with the concepts of the new cancer guidelines, it would be appropriate to drop this section or simply refer to ways in which the results of all analysis could be presented to the risk manager allowing them to choose an appropriate dose from which to extrapolate.

Thank you for the opportunity to review this document.

Sincerely,

Christopher J. Portier, Ph.D.
Chief, Laboratory of Quantitative and Computational Biology, and
Associate Director for Risk Assessment, Environmental Toxicology Program
National Institute of Environmental Health Sciences

Lorenz Rhomberg

Premeeting Comments

General Comments

I am in favor of EPA's adoption of the benchmark dose method, and I applaud the agency for laying out a considered, thorough guidance document on the technical aspects. Clearly, a lot of work has gone into this document. The document is generally well written (although a few parts need some attention) and most often it steers a good course between allowing case-by-case flexibility and being sufficiently prescriptive to spell out how to execute the available choices in thought-out, concrete guidance.

Frankly, in this regard it far surpasses the recently proposed revision to the carcinogen assessment guidelines. Those guidelines, in prompting case-by-case flexibility in approach, often mandate general analytical paths and then leave the analyst without much in the way of a practical technical framework to follow. I think the present BMD document demonstrates how involved and difficult it is to put generally phrased ideas and principles (even laudable principles that are widely agreed-upon) into concrete technical practice. The general idea of the BMD approach can be stated in a sentence or two, but the methods to implement that simple notion have been under concerted development for several years now. Their explication in this document needs nearly 100 pages of technical discussion which, while making admirable progress, nonetheless leave some key technical points and methodological particulars insufficiently defined or explored. One wonders whether the several dozens of such two-sentence, generally phrased ideas in the proposed carcinogen guidelines will receive similarly rigorous thought in their implementation processes.

Inevitably, a document such as this must compromise among the needs of several audiences; some will seek rigorous and thorough technical discussion of the analytic methods and their derivation, others will wish for an elucidation of the underlying rationale and reasoning behind the technical approaches, and still others ask for a didactic approach that explains the operation, basis, and meaning of the methods in generally accessible terms. By and large, the document succeeds in balancing these. As noted further below, however, some attempts in the body of the document at non-technical explanations of key statistical ideas could be improved. (The treatment of these in the appendices is much better, however.) These key ideas include confidence limits, statistical power, independence and additivity, and risk above background.

I personally am in the camp that would like to see more attention paid to rationale and reasoning behind the chosen methods. An explicit statement of such rationale is important for several reasons: 1) it expresses the analyst's goal or intent in applying the analytical procedures; 2) it therefore forms the basis for judgment in choice of appropriate procedures and methods—one can only judge the soundness of a procedure when there is a clear statement of what one intends to accomplish by it; 3) it provides a framework for advancing the methodology upon the advent of new kinds of data—for instance data on mechanisms of toxic action and pharmacokinetics; and finally, 4) a clear statement of rationale behind risk assessment methods is essential to effective communication of the risk characterization to risk managers and the public. For these reasons—especially the last—the NAS report Science and Judgment in Risk Assessment recommended that

the agency articulate the intended basis, rationale, and major assumptions behind its risk assessment methodology.

In view of this, I feel that the present document needs a clearer statement regarding what the BMD/C is intended to represent; i.e., not just how it is determined, but the rationale for determining it in that way. To a large degree, the BMD/C (at least in this document) is presented just as the outcome of particular procedures—the document seems to skirt characterizing it as having any specific toxicologic meaning. The NOAEL, in contrast, is not just the outcome of an experiment; it has an intended meaning (albeit a somewhat vague one) as an observable point related to the "bottom" of the distribution of individual susceptibilities to the endpoint in question. One can understand the way of determining the NOAEL as a means for accomplishing the intended estimation—and one can judge the success or failure of the procedure in reference to this intended end. (And the NOAEL does indeed have such failures, some of which prompted the development of the BMD approach).

Too often in risk assessment, we agree on procedures to apply without ensuring that there is a common understanding of the intent, meaning, and scientific rationale of the analyses. As soon as such intent and meaning become issues (as in application of judgment or use of novel data) a debate begins to rage—not about the case-specific data, but about conflicting interpretations of the implicit assumptions inherent in the defaults one hopes to supersede. Instead of productive discussion about the agent's toxicological properties, the debate then focuses on conflicting theories about reconstruction of the default methods' original rationale.

There is a lot more to risk assessment than benchmark doses and NOAELs. The document presumes that other aspects of the risk assessment process and the basic framework for its conduct are set out elsewhere and basically agreed upon. I have taken this as given in my comments. It is worth noting, however, that BMD-determination methods interact with other outstanding questions in risk assessment. These include the use and interpretation of uncertainty factors, the division of extrapolations into pharmacokinetic and pharmacodynamic components, the additivity-to-background issue and the assumptions about dose thresholds, contrasting dose-scaling methods and rationales for cancer and non-cancer endpoints, estimation of risks above the RfD/C, dose-rate effects, and risks from less-than-lifetime exposure.

Defining the BMR

I am concerned about the shift from a consistent risk level for the BMR (be it 10% or 1%, extra or additional risk) to an emphasis on limit of detection and/or "biological significance." A consistent risk level makes the meaning of the BMD comparable across applications and fosters consistency of application, use, and interpretation. One of the drawbacks of the NOAEL is its dependence on experimental design—although all NOAELs should be somewhere at the bottom of the distribution of individual susceptibilities (presuming the existence of a threshold), exactly where it is vis-*a*-vis the distribution (i.e., what percentile it represents) is dependent on sample sizes, dose placements, etc. that are properties of the particular experiment, not of the toxicology of the compound. In other words, reliance on the NOAEL hampers the generalization of the

particular experimental results to inform other situations largely because of questions about the comparability of "no effect" determinations across studies.

A key advantage of an ED-10 (or of any risk-specific dose approach) is that it focuses on the toxicological properties one is hoping to characterize and generalize with the experiment rather than on the experiment itself. One can then separate (or hope to separate) the issues of the toxicological properties imperfectly revealed by experiments and the degree of imperfection of that revelation. That is, one can keep separate the issues surrounding the meaning of an ED-10 for the risk assessment process and those regarding potential bias, uncertainty, and pitfalls in the estimation of the ED-10 with the data at hand. Going to a BMD that is based on limit of detection abandons a principal advantage that this method has over the NOAEL.

I don't like the idea of EPA getting into the business of making rulings on "typical" experimental designs, sample sizes, or background rates for different kinds of toxicological tests. There are several reasons: 1) it will be a lot of work that the agency doesn't need and that delays the implementation of application of BMD methods until the relevant characterizations have been done; 2) it will be the source of lots of unproductive argument as interested parties squabble about what design and background findings should be enshrined in official policy (perhaps with an eye to whether BMDs will go up or down as a result); 3) it will be hard for the agency to avoid making pronouncements as to valid or sufficient experiments generally, leading to the necessity of EPA experimental guidelines for virtually all toxicological testing that might result in BMDs—i.e., it won't stop at getting input for level-of-detection (LOD) determination but will lead to de facto standards for toxicological testing; 4) it will therefore tend to freeze experimental design; 5) it will create an incentive for poor experiments with high LODs so as not to allow the agency to say that a lower LOD is achievable in "typical" experiments.

All of this is in addition to the fact that a determination of BMR by the LOD is ill advised, for reasons already named.

In essence, the limit-of-detection method for determining the BMR changes the question from one of estimating a toxicologically meaningful point near the lower end of the distribution of susceptibilities to one of estimating the dose level at which a NOAEL is expected in future experiments (given a definition of NOAEL in statistical significance terms). That is, the LOD approach embraces the very problems with the NOAEL that the BMD approach seeks to overcome! If such manipulation and specification of experimental design is to be undertaken, why not just apply it to the NOAEL procedure? This would overcome many of the objections to the dependence on experimental design—you just specify the design to be used!

"Biological Significance" in Defining the BMR

The term "biological significance" is used as a key notion without being defined. This is critical, since a number of meanings have been attached to the term in different settings, and some interpretations would be quite harmful to the BMD methodology.

In most cases, "biologically significant" seems to mean "adverse," as when the BMR for continuous endpoints is defined as a change in the mean that is "biologically significant." In these cases, stating the matter more straightforwardly in terms of adversity seems far preferable. The notion of adversity is problematic (especially for continuous endpoints, as discussed below), but giving it another name doesn't help.

In other settings, "biologically significant" is sometimes used to refer to the evidence of an agent's tumorigenicity provided by the appearance among exposed individuals of a few tumors of a type that is historically rare; the notion is that even if such responses are not "statistically" significant (as judged by pairwise comparison against concurrent controls), they are nonetheless valid evidence. This is really still a statistical notion of significance, but the "test" is done informally, giving weight to the prior experience with control groups from earlier studies. If this sense of "biologically significant" were to be applied to BMR determination, it would suggest that any observation of an historically rare response should be taken as demonstration of the agent's ability to cause the response at that dose, regardless of statistical criteria. This seems ill advised.

If a "biologically significant" response means one that has biological consequences, then any response can be ruled significant depending only on the sensitivity of one's ability to detect biological sequelae.

As I have said elsewhere in these comments, I feel that the use of an "adverse" degree of change in the population mean of a continuous measure to define a BMR is burdened with difficulties of interpretation. This is owing to problems with defining adversity at the population level, discussed below, and my preference for a BMD based on a standardized level of risk, discussed above.

BMD for Continuous Endpoints

The methods for continuous data still need a lot of thought and development. I think it is critical that methods for such endpoints be as consistent as possible in toxicological and risk assessment meaning from one case to another. But as now proposed, such meaning is very dependent on the method of determination. I can offer no solution to the problems I raise at present. It seems difficult to construct a scheme in which two notions that ought to be fundamental—"Adversity" and "Risk"—have logically consistent meanings across methods.

These need to have conceptions that are consistent across applications of BMD to continuous data and consistent with their use for dichotomous data. Failing this, the BMD procedures (for they will be multiple) will always have an arbitrary element, with the meaning of the BMR level dependent on the methodology for getting there as much as on toxicological properties or implications for risk assessment. When there are several methods that could be applied, there must be a clear basis—either in explicit policy goal (such as public health protection in the face of uncertainty) or (preferably) in meaning, accuracy of estimation, or biological and toxicological interpretation—for the choice of one method over others to be seen as other than arbitrary. In the present document, the choice seems to be based mostly on "guidance," i.e., on choice rules and defaults that

are set out without a lot of justification in a consistent framework of meaning of the resulting BMD and risk assessment rationale. The concepts of risk and adversity do not have consistent meanings across methods, and they must form the basis of this framework.

"Adversity" and Continuous Endpoints

The concept of adversity of an endpoint is difficult enough for quantal data in current non-cancer risk assessment, but at least it is separable from the process of characterizing NOAELs, and could be separate from the characterization of BMDs. Any observable endpoint that can be scored as present or absent can produce dichotomous data. The adversity question is separate—for each such endpoint one must decide whether presence of the effect should be considered an impact on health or a mere accommodation of the body to the change of circumstances provided by the dose. There are arguments about whether this judgment (for it is a judgment) should be part of risk assessment or of risk management. It is possible to have this argument because the adversity issue is separate from the assessment issue; one can readily make a NOAEL for a change in enzyme levels, tooth discoloration, or whatever, and later decide whether that endpoint is to be considered as an unacceptable impact on health.

With continuous data, there are several proposed courses of action to define adversity, and they tend to get bound up in the BMD derivation process. Each method has problems, and the differing solutions tend to make the meaning of adversity dependent on the BMD calculation process.

Dichotomization—this preserves compatibility of meaning with dichotomous endpoints. It has two problems: the definition of a cut-off and the implicit valuation of changes in the measured feature in different parts of its range.

The cut-off problem is that the choice of cut-point introduces an arbitrary element; if it were clear at what point in the measurement scale adversity begins, the endpoint would already be a dichotomous one.

The implicit valuation problem is that, by introducing the cut-point, one is saying that only changes in the measure that result in dropping below the cut-point are of concern; an individual that was already below the cut-point (and hence "affected") becoming still lower is not considered adverse, while an individual well above the cut-point that experiences a drop in value that just fails to bring him below the cut-point is similarly supposed to have suffered no ill effect. If, for example, the continuous variable were IQ and exposure to lead dropped the whole distribution of IQs by 5 points, the adversity would be determined only by the number of additional people who fell below a certain level (say 85)—i.e., from 87 to 82, from 86 to 81, etc. Drops from 110 to 105, from 100 to 95, or from 80 to 75 would play no role in the determination of the adversity of the lead exposure. That is, there is an implicit valuation of changes in different parts of the range, with health concern being focused only on changes in the region of the cut-point. (This is not just "loss of information" in the statistical sense, which the document mentions, but also a constraint on what kinds of impacts get attended to.)

In a sense, the dichotomization process fails to acknowledge a distribution of values in the unexposed population. If adversity is conceived of as an unacceptable degree of deterioration, attributable to exposure, in the value of the variable that one would have had had one not been exposed, then a single cut-off value for the whole population misses the point. It might be possible in some circumstances to measure the value of interest before and after the exposure and then to make a cut-off for degree of change, with affected individuals being those who have unacceptable degrees of perturbation of their original condition. But it will often be impossible to ascertain a meaningful pre-exposure condition. In any case, the problem of attending only to values around the cut-point remains, it is just recast on an individual level.

The alternative to dichotomization presented in the document is to model the change in mean value in the population as a function of exposure. In a way, this goes to the opposite extreme—only the general shift in the distribution is attended to, while the individual fates of members of the population are ignored. That is, to the degree that there is an absolute level of the variable that is compatible with individual health, the extent to which a fall in the general distribution puts individuals into the realm of ill health (by pushing them too low on the value of the variable in question) is not measured *per* se. For example, if we lower the nerve-conduction velocity of a *population* by 10%, how many people are put into a state that can be considered ill health? Are those already low in the distribution at especial risk? How many do we place into the category of being a risk from further lowering by eroding their reserve capacity?

That is, the two approaches, dichotomization and modeling the mean response, pay attention (imperfectly in each case) to different aspects of adversity, both of which seem potentially relevant. Dichotomization implies an absolute level of the variable that is necessary for health; modeling the mean response implies that lowering from one's initial value by some amount is adverse, i.e., it stresses relative change and measures impact only at the population level. Both are crude in that they make the distinctions sharper than they really are and give no weight to impacts that are qualitatively the same, but not of sufficient magnitude to trigger the cut-off criterion.

"Risk" and Continuous Endpoints

The above considerations about "adversity" interact with the notion of "risk." For dichotomized variables, risk is the probability that an individual will be moved from the unaffected to the affected class as a result of exposure. If the initial states are unknown, then this risk is the same for everyone with similar exposures—some fraction of the population will be affected, but not knowing who, everyone's prospects for being among them is a matter of chance. But if we can know initial states, then those already near the cut-off are clearly at much greater risk than those who are not.

The notion of "risk" is quite different for continuous variables modeled as a change in the mean with exposure. An unacceptable degree of population change is defined (either "biologically" or by limit of detection), but at the dose a which this level of change is achieved there is no identification of individuals who are or are not personally affected. Thus, there is no real meaning to individual risk. Below a BMR defined in this way, everyone is a member of a population that collectively has insufficient impact for concern,

and above it, everybody is a member of the population that is collectively considered affected, regardless of their personal state.

Specific Comments

The following comments apply to specific points in the document, identified by page and line number in the format [pagenumber - linenumber].

[1-14] - It should state that BMD methods apply to chronic noncancer toxicity.

[1-17] - These LOAEL and NOAEL definitions are not precise. The LOAEL is where "adverse effects have been detected." How about statistical significance? The NOAEL is " the highest dose at which not adverse effects have been detected." Again, the role of significance not is clear, and it should apply to detection at that or lower doses.

[2-18] - "The NOAEL does not account for variability in the data." This is a very poor and misleading way to address experimental uncertainty (which appears to be what was intended). The NOAEL addresses, indeed is based on, variability in the data, but the use of the NOAEL as a measure of the compound's toxicity neither addresses the different responses at other doses nor the experimental uncertainty (not variability) in the no-effect level owing to limited sample size.

[3-1] - the BMD is only more consistent than the NOAEL if it refers to the same risk level across studies, but this is no longer so given the switch to multiple forms of BMR determination, as discussed above.

[3-3] - See [2-18].

[3-25] - The notion of "sensitive" needs to be defined.

[4-5] - "At a minimum, the number of dose groups and subjects should be sufficient to allow determination of a LOAEL." This criterion makes a precise definition of LOAEL essential. I don't agree that the criterion should be applied. A hazard may be clearly detected by trend tests or other methods without a strict LOAEL being definable.

[4-7] - This seems to be equivalent to the previous criterion. How would one identify a group statistically different from controls with a trend test?

[4-9] - "With only one responding group, there is inadequate information." Does "responding" mean at a significant level? Then two elevated doses are necessary! I disagree strongly with this criterion, which is not justified. ("Too arbitrary" is too vague.)

[4-28] - As discussed above, "biological significance" must be defined more precisely than as something that is "biologically significant.."

[5-4] - It is not clear how statistical significance is to be used as a criterion for biological significance.

[5-8] - This definition of defaults is not clear. Appendix B has a much better explanation.

[6-10] - A better reason to avoid parameter values less than unity is that the biological and risk assessment interpretation of infinite slopes at low doses is untenable.

[6-12] - Why a top-down approach for choosing the degree of the polynomial? A bottom-up approach would seem preferable, increasing the degree until there was no further significant explanation of variance.

[6-19] - The term "background parameter" is not defined. The claim that excluding such a parameter is conservative is surprising, and I suspect this conclusion depends on how such a parameter appears in the model—as additive or independent background. It is conservative to omit independent background, but I doubt it is generally so to omit additive background.

[7-1] - "Threshold parameter" also needs definition.

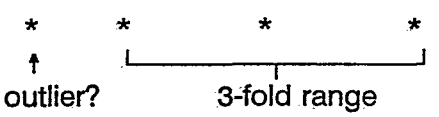
[7-21] - Why are asymptotic methods preferred, especially given the small sample sizes typical of many non-cancer experiments. Why are bootstrap methods not used?

[8-11] - Quantal models based on the tolerance distribution idea don't assume equal probability of response among individuals, but only that one doesn't know the individual tolerances ahead of time—a subtle but important distinction. It is important for the underlying logic of the BMD approach to be clear whether one is relying on underlying variation in tolerance or underlying stochasticity.

[8-13] - Continuous responses may be lognormal too.

[9-9] - Should a criterion for model preference be that no dose-dropping to achieve fit has been done?

[9-13] - There is danger in the provision for dropping "outliers" among indistinguishable models. It is easy to imagine a case where the total span of models is greater than a factor of 3 and that dropping an extreme leaves the remaining models within 3-fold of one another, but that the "outlier" is closer to its nearest neighbor than the other models. "Outlier" needs to be more clearly defined if this provision is not to be used as an argument why all models except those in the top (or bottom) 3-fold range should be ignored.



[10-16] - Linear cancer low-dose extrapolation is also based on concerns for additivity to background processes and the stochastic nature of carcinogenesis.

[10-20] - See [1-17].

[10-21] - Saying the NOAEL is "adjusted downward" is a misleading description of the RfD/C process. Nothing is adjusted, but a lower level of exposure is chosen as being able to be claimed safe.

[10-26] - Criticizing the LOAEL and NOAEL as being merely operational and without toxicological meaning is ironic, given that the BMD is being set up to be exactly that. See my earlier general comments.

[11-25] - See [2-18].

[12-24] - See comments on "biological significance." The issue of defining a degree of quantitative change that is adverse needs a lot of further development before it can be meaningfully used in this process.

[13-Figure 1] - The ordinate should be labeled as risk over background.

[13-Figure 1] - The curve labeled "Lower statistical limit on dose" illustrates how it is necessary to be precise in how one discusses confidence limits. As I understand it, the asymptotic likelihood ratio-based methods for confidence-limit generation specified in the text actually lead to alternative sets of model parameters, specifying alternative curve fits, that provide the intended lower bound on dose only for a specific risk level. For a given point, this alternative curve will generally be farthest from the MLE at that risk level, but nearer to it, and even crossing it, at other risk levels. The dotted line shown can only be achieved by making a composite curve out of local pieces of a lot of these risk-specific alternative curves. That is, the dotted line corresponds to no single bounding curve, but to linked up pieces of a lot of mutually incompatible different ones.

[15-27] - See [68-11].

[17-11] - The phrasing could imply that one will usually extrapolate a dose-response curve from the BMD/C.

[17-12] - The BMD/C is dependent on the doses in the study in the sense that its determination depends on the data.

[18-20] - The criteria for selecting appropriate studies should include the reliability with which the endpoint can be detected or measured.

[18-24] - It should say "all such studies should be modeled," i.e., one does not have to model rejected studies.

[19-5] - This phrase appears to argue that one should select data that make the models work. This is the tail wagging the dog.

[19-10] - See [3-25].

[19-19] - See [4-5].

[19-21] - See [4-9]. Also, one must define what it means to be "different" than controls. Is it statistical significance?

[20-2] - It needs to be more clearly spelled out what makes data sets "statistically and biologically compatible," at least in terms of what judgments must be reached.

[20-2] - Toxicology and risk assessment have always been very wary of combining data from separate studies, and with good reason. There are serious issues of comparability of experimental animals, lack of randomization across studies, consistency of husbandry, consistency of endpoint diagnosis, and so on. If this long-standing practice is to be changed, it needs much more justification than provided.

[21-14] - See [4-28].

[21-23] - See [5-8]. The discussion of additional and extra risk is not at all clear here. It is much better in Appendix B, which I had to read to make out what was intended here. The limit of detection of a study is expressed in degree of response, and this degree in turn can be expressed either as extra risk or additional risk. Thus [22-2] is confusing in its implication that the limit of detection somehow involves extra or additional risk.

[24-1] - See [6-10].

[24-7] - See [6-12], but here there is the added puzzlement of why a linear model (i.e., degree of 1) is chosen first, and then a polynomial with a top-down degree determination. Why start with bottom-up, then switch to top-down? The impetus to start with a linear model shows the value of the bottom-up approach in my view.

[24-11] - See [6,19]. Also, the criteria for deciding whether to use a background parameter need to be clearer. Actual presence of a response in the control group seems a poor criterion, given the small sample sizes. The description of monotonicity should make clear if "same" and "lower" refer to statistically significant differences or not.

[25-22] - "$p>0.05$" should be "$p<0.05$"

[26-15] - See [9-9].

[27-5] - See [9-13].

[28-11] - I disagree that basing some BMD/Cs on continuous variables and others on discrete ones "is not a problem using the approach advocated in this document." As discussed above, I think it is still a big problem and should greatly concern the agency.

[29-22] - "[c]ommunicating...the central estimates" of what? Of dose associated with the BMR?

[34-4] - Does a "response" below the 10% level need to be significant to be a point of departure?

[34-14] - The "unfounded sophistication" of the LMS procedure is not an implication of that procedure but rather an unfounded inference on the part of certain critics.

[35-28] - It should be clearer that this is addressing variability among humans in sensitivity. The need to define sensitivity, noted several times earlier, arises again. It is essential to realize that one must have a clear stance on the partition of sensitivity between pharmacokinetic and pharmacodynamic aspects to do this sensibly. Also, there is not such thing as defining sensitivity independently from the dose units that are being considered.

[35-29] - Is the 10-fold factor for human variability in sensitivity to be calculated in terms of administered dose? If an internal dose measure is used that is not proportional to administered dose, 10-fold in such units may be much more or less than 10-fold in administered dose units. What is The rationale for the 10-fold?

[36-2] - See [35-28], [3,25]. This issue is particularly acute for cross-species extrapolation. If a human is 10-fold more sensitive than a mouse to a mg/kg/day dose, then he is 200-fold *less* sensitive on a mg/day basis and 7,000-fold less sensitive on a total lifetime mass basis, assuming chronic exposure.

[36-17] - It is not clear how using both linear and nonlinear default procedures may be used to "distinguish between the events operative at different portions of the dose-response curve and consider the contributions of both phenomena." How does having two alternatives help judge their combined effect?

[39-2] - "...risk estimates at lower doses from a nonlinear model can be substantially lower than those from a linear model." But the nonlinear approach that is suggested, the margin-of-exposure, has no risk estimates at lower doses, and indeed it specifically rejects the idea that such risk estimates should be attempted.

[39-7] - The application of the BMD methodology to the case of cancer risk assessment raises a number of questions that need to be explored:

- For tumors analyzed as secondary to toxicity, must the experiments on toxicity be for chronic exposure? Must the toxicity be chronic toxicity?
- What happens to the lifetime dose-adjustment defaults when partial lifetime exposure can engender such toxicity?
- How is cross-species scaling to be done when the immediate endpoint is noncancer toxicity but leading to tumors? Is carcinogen dose scaling or noncancer dose scaling used?
- If RfC-style dosimetry is used, how is route extrapolation to be done?
- What dose units are appropriate for judging the margin of exposure? What units are used for the 10-fold factors? How is nonlinearity between administered dose and internal dose to be incorporated in MOE analysis? How do such considerations interact with the rationale for the size of the 10-fold factors and any empirical justification they may have?
- How does one stop consideration of the low-dose extrapolation of fitted models beyond the point of departure?

[42-14] - "These estimates potentially can be used to determine the number of cases expected for various non-cancer endpoints when exposure levels are in or near those in the experimental data range." What about the effects of the uncertainty factors for animal-to-human extrapolation and for human variability in sensitivity?

[44-26] - I'm not sure that current carcinogen assessment methods address variability among humans at all, even implicitly.

[47-25] - A major typo.

[57-18] - This is a very important issue. I'm not sure what the right approach is. To do route extrapolation before BMD analysis incorporates nonlinearity into the BMD calculation, but is this a good thing? What does it say about the shape of the curve, the size of margin of exposure, the difference between the ED-10 and LED-10, etc.? This needs discussion.

[59-26] - "...the goal of overall average BMD/C:NOAEL ratio of one." This is an obscure place to state such a goal. Why is there such a goal? Why should not the BMD methods stand on their own merits? Clearly, it helps with acceptance if the change does not undermine precedents or lead to markedly different conclusions. But the BMD methods should not be tailored to make as little difference in the assessment of risks as possible. If a sound approach dictates taking different views of risks, so be it.

[67-20] - Experimental animals don't "present with" conditions, and we seek to protect human health impacts whether or not people seek medical attention. This is jargon misused.

[67-25] - See [24-11].

[68-11] - Between the choices of modeling the continuous data directly and dichotomizing the responses, there is the semi-parametric, rank-based approach to BMD/C analysis developed by Ron Bosch, David Wypij and Louise Ryan. [Bosch, R.J., Wypij, D., and Ryan, L.M. (1995). A semiparametric model for quantitative risk assessment with continuous responses. ASA Proceedings of the Biometrics Section, 112-117; Bosch, R.J., Wypij, D., and Ryan, L.M. (1996). A semiparametric approach to risk assessment for quantitative outcomes. Risk Analysis, to appear.] The basic idea here is analogous to the two-sample comparison, where one could compare two groups by a t-test or by a 2x2 chi-square test if one dichotomizes the response. "In between" these two options (say, if the data are not normally distributed) lies the Mann-Whitney rank-sum test which is the nonparametric alternative. This approach is exactly the generalization of the Wilcoxon or Mann-Whitney test to the BMD/C dose-response problem.

[69-27] - I'm not sure that coming close to the data means is the object of fitting when the method is not least squares.

[71-23] - The reasons given for rejecting use of p-values for choice among models sound very much like the logic behind maximum likelihood estimation when one is estimating *parameters within a model.*

[72-10] - Monotonic data should be good for using typical models. There is presumably a typo here somewhere.

[73-16] - The Akaike Information Index needs a reference, even in a draft.

[73-22] - "Confidence limits bracket those models..." It is really one model with different parameter values. The usage here has shifted from elsewhere.

[73-28] - "...to cover some reasonable amount of the distribution of the source of the modeled data." This phrase is entirely unclear to me.

[74-1] - Here is where a good non-technical explanation of confidence limits, their derivation and meaning, is needed. The existing passage is much too technical for the uninitiated, and old hat for those who can read it.

[77-18] - Is this an example of using a 10% default level or of using "biological significance" as a BMR-defining method?

[87-17] - In this example, the data sets can be combined if they are left as continuous, but the criteria for combining reject the combination if the data are dichotomized. This should illustrate some more general qualms about either dichotomization or the criteria for combining studies.

==========

Lorenz Rhomberg Harvard Center for Risk Analysis 617-432-0095
(rhomberg@hsph.harvard.edu)

==========

Robert Sielken, Jr.

Benchmark Dose Peer Consultation Workshop

Benchmark Dose Technical Guidance Document
August 9, 1996, External Review Draft

Comments by Robert L. Sielken Jr.
August 28, 1996

0. General Comments / Personal Perspective

- The major point of debate for me and most people is the risk assessment for doses that are below dose levels known to have a substantial probability of causing an adverse human health effect. I shall refer to these doses as low doses regardless of the absolute magnitude of the doses and refer to their associated probabilities of an adverse effect (whether zero or greater than zero) as low-dose risks. Hopefully, high-dose risks are more identifiable and quantifiable, and the issues in their management more clear.

- My focus is on the issue of how to assess low-dose risks.

- There are currently two major approaches to assessing low-dose risks. One approach is that currently used for noncancer health effects which is based on the idea that there is a dose (e.g., a Reference Dose (RfD)) that is likely to be without appreciable risk of deleterious effects. For the sake of simplicity, I shall refer to this approach as the reference-dose approach. The reference-dose approach does not explicitly quantify low-dose risks (for doses either below or some distance above the reference dose). The other approach is that formerly/currently used for cancer health effects which is based on the idea that all doses greater than zero have risks greater than zero and utilizes high-to-low-dose extrapolation to characterize those risks. For the sake of

simplicity, I shall refer to this latter approach as the high-to-low-dose-extrapolation approach. The high-to-low-dose-extrapolation approach explicitly quantifies the low-dose risks based on dose-response modeling or assumed form of high-to-low-dose extrapolation.

- For quantitative cost/benefit or benefit/cost analyses, some quantitative values need to be assigned to low-dose risks.

The reference-dose approach does not explicitly assign numerical values to low-dose risks. Even if one were to interpret the risks for doses below the reference dose as zero, the risks for doses greater than the reference dose would still need to be quantified. If the starting point for the determination of the reference dose were a BMD determined via dose-response modeling, then presumably the fitted dose-response model could be used to quantify the risks for dose greater than or equal to the BMD. However, that still leaves the risks for doses between the RfD and the BMD not quantified -- and this dose region may well be the region of paramount concern for risk management. Thus, the reference-dose approach does not generally provide the quantitative characterizations of risks at different doses needed for cost/benefit analyses.

The high-to-low-dose-extrapolation approach does quantify the risks at all doses. However, the major problem with the dose-response modeling of low-dose risks, especially those well below the experimental/observed dose region, is that the results of any one dose-response model are subject to doubt/disbelief/uncertainty. Nevertheless, a good uncertainty analysis of the risks in the low-dose region that would reasonably characterize the relative likelihood of different risks in the low-dose region would greatly alleviate this problem. Thus, the high-to-low-dose-extrapolation approach coupled with good quantitative uncertainty analyses does generally provide the quantitative characterizations of risks at different doses needed for cost/benefit analyses.

If **probabilistic risk assessment** is interpreted to mean high-to-low-dose extrapolation using dose-response models accompanied by uncertainty analyses that incorporate all/more of the available relevant information including explicit evaluation of the alternatives for the components of the dose-response modeling and incorporate the current state of knowledge about the relative likelihood of these alternatives, then probabilistic risk assessment seems to be the best approach to doing risk assessments in support of cost/benefit analyses.

- People ought to be more optimistic rather than pessimistic about uncertainty analyses. Uncertainty analyses can be more than just qualitative; they can also be quantitative. Furthermore, uncertainty analyses can do more than provide a range of possibilities. New tools and techniques enable uncertainty analyses to incorporate all/more of the available relevant information including explicit evaluation of the alternatives for the components of the dose-response modeling and incorporate the current state of knowledge about the relative likelihood of these alternatives.

- The public and defense lawyers do not like to deal with risks greater than zero. If a dose is less than a Reference Dose (RfD) or other dose "that is likely to be without appreciable risk of deleterious effects," then that dose is interpreted by the public and the lawyers as "zero risk" and is okay/acceptable. On the other hand, if cancer is the health effect of concern, then currently any dose greater than zero is interpreted as having a risk greater than zero and that is not okay, not acceptable, or at least a problem for both the public and lawyers.

The problem lies in the public and litigators perception/reaction to any non-zero risks and their treatment of very small probabilities of an adverse health effect and de minimis/negligible risks.

- It is my belief that there are two major reasons why people want to calculate a benchmark dose (BMD) and use it as the basis for a reference dose (RfD) or margin of exposure (MOE) calculation. The first reason is that they believe that (for many modes of action associated with cancer and noncancer effects) there are dose levels where the risks are de minimis (roughly, the risks are either zero or so small that they ought not be of concern). The second reason is that they believe that the only way to reflect low-dose de minimis risks is via an RfD or MOE analysis. The second reason is supported by the belief that up to this point in time the quantitative characterization of low-dose risks has mostly been done so poorly that it is better to not quantify low-dose risks at all.

The unhappiness with past and frequently current attempts to quantify low-dose risks stems from many failures. For instance, the linearized multistage model and other low-dose linear extrapolations have frequently failed to reflect available biological data, use biologically based dose-scales, adequately reflect interspecies differences, etc. Furthermore, the obstacles to using anything other than default assumptions have been so great that additional research and data collection have not been encouraged. In addition, there has been little agency acceptance or encouragement of analyses that consider multiple alternatives and more of the available data -- alternative high-to-low-dose-extrapolation models, alternative dose scales, alternative characterizations of interspecies differences, alternative data sets, etc.

Many of the attempts to include more science in the quantification of low-dose risks have been plagued not by poor science but rather by the poor way the science has been implemented. The "implementation" or "what is done in practice" in the name of a good idea or scientific principle is what is critical to the usefulness of the result. For example, lots of mischief has occurred in the name of the idea that any curve is roughly linear over a short distance -- such as using straight lines to estimate curves over a long

distance, and using straight-line interpolations over long distances to estimate slopes over a short distance at a low dose.

In the spirit of improving quantitative dose-response modeling and extrapolations rather than abandoning them entirely, many efforts are ongoing to try to improve existing methods of quantifying low-dose risks -- physiologically-based pharmacokinetic (PBPK) modeling, expanded uncertainty analyses (considering multiple alternatives, tree analyses, distributional characterizations, and other probabilistic techniques), etc. However, progress in many people's minds seems to be perceived as too slow and/or too difficult. The consequence seems to be a fall back to an improved version of the reference-dose approach (based on a BMD or ED_{10} or LED_{10}) which incorporates the idea of a low-dose region where the risks can be characterized as de minimis. However, the price for this fall back is the loss of quantified risks in the region between the area of de minimis risks and the region of the BMD or the ED_{10}. This price will be paid when we try to do relevant cost/benefit analyses without quantified risks in the region between the area of de minimis risks and the region of the BMD or the ED_{10}.

- If a benchmark dose (BMD) is going to be calculated and used as the basis for a reference dose (RfD) or margin of exposure (MOE) calculation, then how should the BMD be calculated? This is the question that is addressed in the following comments.

1. *Selection of Studies and Responses for Benchmark Dose/C Analysis*

a. **Is the selection of studies and endpoints for the BMD/C appropriate? for cancer? for noncancer?**

- In the spirit of incorporating more of the available data and explicitly exploring the quantitative impact of alternatives, it may be more useful and informative to evaluate and report the BMD for each of the available relevant studies and endpoints. Tree analyses may be a useful tool in organizing, presenting, and communicating these multiple alternative calculations of a BMD. A plausibility distribution could also be used to reflect the relative likelihood of these calculations being relevant to humans.

- One of the most important issues associated with BMD/C analyses is the issue of what level of adversity or severity should be associated with the endpoint for a BMD/C.

  This issue is raised repeatedly throughout the document but never satisfactorily resolved. In fact the last sentence in the body of the report points to an awareness of this key missing ingredient -- "These and other issues, such as how to deal with severity of effect, need further consideration by the Agency." (page 45, lines 2-3)

  This issue has been raised in the past -- and not really fully answered -- when the distinction was made between a NOEL and a NOAEL; that is, the difference between an effect and an adverse effect.

- BMDs based on different effects or effects of different severity or different adversity are not comparable and thus counter to one of the main reasons advanced for calculating BMDs in the first place.

- "Biological significance" for different substances may not be associated with the same levels of adversity or severity of the resultant health effects and hence does not guarantee comparability.

- All cancer effects do not have the same adverse health consequences (death, disability, discomfort, disfigurement, etc.). All noncancer effects do not have the same adverse health consequences (death, disability, discomfort, disfigurement, etc.). Similarly, cancer effects are not necessarily more severe or less severe than noncancer effects.

- Cost/benefit analyses and comparability of BMDs need for the BMDs to be associated with or accompanied by some specified measure of the adversity of the health effect upon which they are based.

  The benefit of a risk reduction surely depends on the adversity/severity of the health effect avoided.

- Either all BMDs must refer to some minimum level of adversity/severity in their corresponding human health effects or each BMD must be accompanied by a label/index indicating the adversity/severity in its corresponding human health effect. The latter may be more useful in comparisons and cost/benefit analyses.

- Time to response is rightfully a component of adversity/severity. Yet time to response is not mentioned in the document. It certainly matters to most individuals when they have a response (an adverse health effect). For example, mortality at age 85 is better than mortality at age 55. Adversity/severity could include the amount of time an individual could expect to be free from the response.

- Time to response can also be a useful component of cost/benefit analyses.

- The adversity/severity of a substances effect might also rightfully include the likelihood/probability that the effect can be cured/repaired/treated.

- The specification of the endpoint for a BMD should reflect the ultimate use/uses that are to be made of BMD based analyses. One of the proposed uses is certainly cost/benefit analyses. I believe several people engaged in such analyses have said that risk assessors should determine the needs of the users of risk assessments in order to determine how the risk assessment should be done.

  For example, Lester Lave of Carnegie Mellon University is paraphrased in Risk Policy Report, July 19, 1996, in connection with Commission on Risk Assessment & Risk Management proposal as emphasizing "the crucial role of good risk assessments as a foundation for benefit-cost analyses, noting that underlying risk data -- not economic valuations -- are always the greatest source of contention", being "highly critical of the common practice of risk analysts presenting data to economists and asking for an after-the-fact benefit-cost analysis," and suggesting "that it is essential for risk analysts and economists to communicate from the start about what data are needed to ensure that risk assessments are usable by economists, as recommended under the commission's new management framework."

b. Should these be the same for cancer and noncancer data?

- The level of adversity/severity associated with a BMD is of equal importance to both cancer and noncancer data.

c. Are there appropriate criteria for determining when data should be combined for analysis?

- There are certainly many pitfalls associated with combining data for analysis. If the data were analyzed both separately and combined, the different results could be communicated in the form of a distribution indicating the relative plausibility of the different results. In any specific instance, the combining of data for analysis may not be right or wrong but somewhere in between -- with the relative plausibility of the combined result depending on the specifics of the particular instance.

2. Selection of the Benchmark Response Level

a. Is the use of biological significance or limit of detection an appropriate basis for the selection of the BMR?

- BMDs based on biological significance are not comparable to BMDs based on limit of detection.

b. For the limit of detection, is the approach proposed in the document appropriate?

- The approach proposed in the document for BMDs based on the limit of detection is not appropriate. The approach proposed in the document for BMDs based on the limit of detection attempts to have the BMD mimic a NOAEL. However, I thought that what we were seeking to do was to replace the NOAEL by a quantity that had a more well defined risk -- like a dose that has a 10% added risk.

c. Is information available to determine the appropriate power level? *(Information on current simulation studies will be presented at the workshop.)*

d. Is the default for quantal and continuous data appropriate?

3. Model Selection and Fitting

- Dose-response modeling should incorporate biologically based dose scales. Delivered doses based on PBPK modeling and other methods of determining the dose delivered to the target tissue should be considered. Biologically effective doses reflecting not only the amount delivered to the target tissue but also the net amount of health effect related activity (DNA adducts formed and not repaired, cell proliferation, cell toxicity, etc.) should also be included as an alternative dose scale for dose-response modeling.

a. Is the order of model application for continuous and dichotomous data appropriate?

b. Should other models be considered, or should the number of models applied be more restrictive?

- Uncertainty analyses can indicate and reflect the explicit evaluation of the quantitative impact of alternative models. The family of models should not be unduly restricted *a priori.*

c. Are the parameters proposed as defaults for model structure appropriate?

I. What should be the default approach for selecting the degree of the polynomial to use?

- The degree of the polynomial should be sufficient to fit the curvature in the dose-response data. The degree of the polynomial should be increased until further increases do not significantly improve the fit. Analogously, the degree of the

polynomial could be decreased until further decreases significantly lessen the goodness of the fit.

The degree of the polynomial should not be limited by the number of experimental dose levels. Furthermore, if the parameters are restricted, then the number of parameters need not be limited by the number of experimental dose levels either.

- The selection of the degree of the polynomial (or more generally the structure of the dose-response model) interacts with the selection of restrictions on the possible values of the model parameters; therefore, an "alternative" is not just the degree, but an ordered pair [degree, {parameter restrictions}]. Because the possible role of hormesis is often linked with parameter restrictions, the selection of the structure of the model family (e.g., degree of the polynomial and parameter restrictions) involves a very controversial sequence of decisions. This sequence of decisions has a very real potential to build into the analysis a preconceived result (or predetermined nature of the result), overwhelming all of the information in the data. Excessive reliance on technical aspects of an "AIC theory" will only tend to divert attention from the real, decisive issues such as "Does hormesis exist in the case at hand?"

  The EPA document only discusses parameter restrictions for the log-logistic model (page 23).

ii. **Is the default of not including a background parameter appropriate unless there is some indication of a background response level?**

- No. The default should be that the background parameter is included unless there is clear evidence that the background response level is zero.

  Explicitly modeling the background dose and/or the background response probability can be very important when there are interspecies differences in the background doses and/or the background response probabilities.

  An example of this problem occurs when the Agency uses its standard procedure for converting animal dose levels to "human equivalent doses" (HEDs) and then uses these HEDs during the fitting of the cancer dose-response model to the animal response frequencies. The problem with this use of HEDs is not widely understood.

- For a finite-size data set with a true, nonzero background rate, there is some probability that there are no responses in one experiment.

iii. **Is the use of extra risk as a default for quantal data appropriate?**

- There are at least two reasons why the use of added risk is more appropriate than the use of extra risk. The first reason is that added risks are comparable from one substance to another. The added risks for two substances indicate which substance has the greater increase in the probability of a response -- without having to incorporate the probability of that person not being a victim of the background response rate. The second reason is that added risk is more easily and accurately interpreted by the public. If there are N people

exposed, then the expected number of additional responses is N times the added risk and is not equal to N times the extra risk. Unfortunately, when risk is characterized in terms of extra risk, the expected number of additional responses is usually falsely reported/communicated as N times the extra risk.

iv. **Is the default of not including a threshold parameter appropriate?**

- No. The objective is to estimate the BMD. In the absence of a biologically-based model, the objective is to estimate the BMD with as simple (parsimonious) a model as possible. Threshold parameters can facilitate the parsimonious fit of curve-fitting models to data. In the context of the estimate of the BMD, it is the properties of the estimate of the BMD that are important -- not the properties of the estimate of the threshold parameter -- and the estimation of the BMD can be improved in some cases when the model contains a threshold parameter.

  The cases where the inclusion of a threshold parameter are especially useful are when the response-frequency data at zero and the lower experimental/observed doses are flat (non-increasing) and the response frequencies beyond some positive dose follow a simple pattern. In such cases polynomials and other models based on powers of dose struggle to compromise between the initial flat dose-response relationship and the eventual increasing dose-response relationship.

v. **Is the default of modeling continuous data as such appropriate?**

d. Is the approach for determining the fit of the model appropriate? Are there additional or alternate criteria that should be used?

- The fitting criterion should reflect the objective. The primary objective of the fitted models is to provide an estimate of the BMD. Given the objective of estimating the BMD, the fit is most important in the neighborhood of the BMD and less important elsewhere. Therefore, the fitting criterion should be most heavily influenced by the fit in the neighborhood of the BMD. The situation is similar to the difference between using least squares versus weighted least squares as the estimation criterion. Weighted least squares can be used to more heavily weight the fit in the region of concern and give less weight to the region further away. When the objective is to estimate the BMD, weighted least squares (with greater weights near the BMD) would be preferable to least squares (which has equal weights everywhere).

  Maximum likelihood estimation is generally a good estimation criterion and is a good fall back here. On-the-other-hand, maximum likelihood estimation does not emphasize (weight) the fit in the neighborhood of the BMD. However, maximum likelihood estimation could be modified to incorporate weights.

  Comparisons of different fitted models should emphasize the fits in the neighborhood of the BMD.

  Graphical evaluations of fits, regardless of the fitting criterion, are worthwhile and should be encouraged.

4. Use of Confidence Limits

a. Should the lower confidence limit on dose be the definition of the BMD/C?

- No. I believe that the best estimate from the fitted model should be the definition of the BMD/C. Thus, if maximum likelihood estimation is the fitting criterion, then the maximum likelihood estimate from the fitted model should be the definition of the BMD/C.

  I believe that the uncertainty in the estimate of the BMD should be reflected in the uncertainty analysis for the BMD and not be in the definition of the BMD.

  Uncertainty analyses can be more than just qualitative; they can also be quantitative. Furthermore, uncertainty analyses can do more than provide a range of possibilities. New tools and techniques enable uncertainty analyses to incorporate all/more of the available relevant information including explicit evaluation of the alternatives for the components of the dose-response modeling and incorporate the current state of knowledge about the relative likelihood of these alternatives.

  One of the reasons given in the past for basing the BMD on the lower confidence limit was to encourage the generation/collection of more data. This goal should be encouraged through the effect of sample size, etc. on the uncertainty analysis.

  I also favor the use of the best estimate of the BMD as the definition of the BMD/C because it facilitates comparisons between substances. Best estimates indicate where we think the BMDs are and their relative locations. Confidence limits indicate where we think the

BMDs might be (instead of where we think they are) and confounds the evaluation of relative locations.

I think that the "old" argument that best estimates are too unstable to use for comparisons was greatly over-simplified and not correct (or at least not always correct) and is largely irrelevant for benchmark response (BMR) levels near 10% added/extra risk.

I also favor the use of the best estimate of the BMD as the definition of the BMD/C because it is most appropriate for cost/benefit analyses. Of course, I also believe that cost/benefit analyses should reflect not only the best estimate of the BMD but also the uncertainty analysis which indicates the relative likelihood of different values of the BMD. In other words, I believe that cost/benefit analyses should be based on the plausibility distribution for the BMD. The plausibility distribution indicates how likely the BMD is to be different values based on the available relevant information and explicit consideration of the dose-response modeling alternatives.

Another reason I favor the use of the best estimate of the BMD as the definition of the BMD/C is that a statistical confidence limit on the BMD does not adequately reflect the uncertainty about the value of the BMD -- for example, a statistical confidence limit does not reflect the uncertainty in the choice of the family of models to be fit to the data. Thus, a confidence limit on the BMD gives a false impression that it is an accurate reflection of the real uncertainty and encompasses all elements of uncertainty.

b. Are the defaults for the method of confidence limit calculation appropriate?

- I do not believe that confidence limits are equivalent to an uncertainty analysis and do not believe that confidence limits should replace uncertainty analyses.

  Uncertainty analyses can/should include alternative data, alternative dose-response models, alternative assumptions, etc. Thus, uncertainty analyses include much more that just experimental variability. (By experimental variability, I mean, for example, that if an experiment were repeated, the outcome in the replicate experiment might not be exactly the same, say 28 responses out of 100 animals at risk, as in the original experiment).

- I believe that the impact/uncertainty introduced by "experimental variability" is better reflected by procedures like those based on bootstrapping than by confidence limits.

- Confidence limits and confidence intervals do not necessarily have the properties that most people attribute to them. In the nice cases, like confidence intervals on the mean of a normal distribution, the probability that the 95% confidence interval will contain the true value of the mean (θ) is 0.95 <u>for all values of θ</u>. However, this characteristic (<u>for all values of θ</u>) is not true in all cases. In fact, a confidence limit procedure gives only a <u>lower bound</u> on the probability that the procedure will perform successfully. For example, if the procedure is a confidence interval on a parameter, then the procedure performs successfully if on a trial the numerical confidence interval contains the true value of the parameter. A 95% confidence interval on a parameter θ, means that <u>at least</u> 95% of the time the confidence interval will contain the true value of θ. In cases like those involving dose-response modeling, the confidence interval may contain θ 99.999% of the time for some values of θ, contain θ 99% of the time for some other values of θ, and contain θ 95% of the time for still

some other values of θ. All that the 95% means is that at least 95% of the time the confidence interval will contain the true value of θ regardless of the value of θ. This means that, for the value of θ corresponding to the situation one is actually in, the 95% confidence limit procedure will contain θ at least 95% of the time but it may be big enough to contain θ 97.5%, 99%, 99.99%, etc. of the time. Thus, for some values of θ, the 95% confidence limit may be much bigger than it really needs to be and hence contain many more values than it needs to --thereby giving a false impression (an overstatement) of the range of likely values.

Unfortunately, in the case of dose-response modeling, the requirement that a confidence limit procedure perform successfully at least 95% of the time for all possible models in the family of models often means that it performs successfully 95% of the time for one subfamily of models (like linear models) and performs successfully much greater than 95% of the time for another subfamily of models (like nonlinear models) -- thereby exaggerating the range of model possibilities when the true model is a member of this latter subfamily of models.

A simple example of the behavior of 95% confidence limits in the case of dose-response modeling is the linearized multistage model. If the true multistage model is quadratic in dose at low doses, then the 95% upper confidence limit on the added/extra risk at one of these low doses exceeds the true added/extra risk at that dose much more often than 95% of time.

- In general, in the context of dose-response modeling and especially nonlinear dose-response modeling, I have had fewer problems with likelihood ratio based confidence limit techniques than confidence limit techniques based on asymptotic normality.

c. **Is the default of 95% confidence limit appropriate?**

- A common misinterpretation (and correct interpretation) of the 95% in a 95% confidence limit is discussed in my last comment under 4.b.

- Although 95% is a frequent choice, the appropriateness of the choice actually depends on the relative magnitude of the cost of the limit not containing the target value and the cost of the limit containing values other than (beyond) the target value.

5. **Selection of the BMD/C to Use as the Point of Departure for Cancer and Noncancer health Effects**

a. **Comment on the determination of "equivalence" of models.**

- The defaults of retaining (especially for the purposes of a quantitative uncertainty analysis) all models that are not rejected because of a bad fit, evaluating graphical representations of the fitted models, and focusing on the fit in the region of the BMR (possibly eliminating one or more high dose groups) are reasonable defaults.

b. **Comment on use of the Akaike Information Criterion for comparing the fit of models.**

- The Akaike Information Criterion (AIC) is not ideal for comparing the fit of models when the primary objective of these fitted models is to provide an estimate of the BMD. Given the objective of estimating the BMD, the fit is most important in the neighborhood of the BMD and less relevant elsewhere. The situation is similar to the difference between using least squares versus weighted least squares as the estimation criterion. Weighted least squares can be used to more heavily weight the fit in the region of concern and give less weight to

the region further away. Comparisons of different fitted models should emphasize the fits in the neighborhood of the BMD.

It is true that the Akaike Information Criterion (AIC) has been successfully used to determine the number of parameters in autoregressive time series models and other models like polynomials where there is a natural sequence of additional parameters (increasing powers in the case of polynomials). However, this use of AIC is not the same as comparisons across different families of models with entirely different structures (like multistage, Weibull, probit, and log-logistic families of models).

c. **Is the default approach for selecting the BMD/C to use as the point of departure for cancer and noncancer dose-response analysis appropriate?**

- In the spirit of incorporating more of the available data and explicitly exploring the quantitative impact of alternatives, it may be useful and informative to evaluate and report the BMD for each of the available relevant studies and endpoints. Tree analyses may be a useful tool in organizing, presenting, and communicating these multiple alternative calculations of a BMD. A plausibility distribution could also be used to reflect the relative likelihood of these calculations being relevant to humans.

- For the purposes of cost/benefit analyses and risk management decisions, quantitative uncertainty analyses may be equally important or even more important than the selection of a single number to represent the BMD/C.

Uncertainty analyses can be more than just qualitative; they can also be quantitative. Furthermore, uncertainty analyses can do more than

provide a range of possibilities. New tools and techniques enable uncertainty analyses to incorporate all/more of the available relevant information including explicit evaluation of the alternatives for the components of the dose-response modeling and incorporate the current state of knowledge about the relative likelihood of these alternatives.

- Uncertainty analyses and tree analyses explicitly presenting the BMD/C values corresponding to alternatives can provide the risk manager with more information and lessen the censoring of information in the risk assessment stage. This would be consistent with the goal of the risk assessor providing useful information to the risk manager rather than the risk assessor partially usurping the role of the risk manager.

- For the purposes of cost/benefit analyses and risk management decisions, the information in the dose-response models (and accompanying uncertainty analyses) concerning the dose-response relationship and the added/extra risks at different doses should not be lost or unreported. That is, information about the value of the BMD/C is not the only useful information generated during dose-response modeling and uncertainty analyses.

6. General Issues

- In general, I thought that the document was well written, well organized, and understandable.

- There are some significant issues that are not mentioned in the document and some others that were not thoroughly discussed. Several of these issues are raised in my comments.

a. **The discussions concerning the use of BMD/C approach in cancer and noncancer risk assessment;**

- Biological data from PBPK modeling and other sources which provide information on the shape of the dose-response relationship below the BMD/C should not be ignored in the evaluation of risks below the BMD/C.

- The hazard ranking for a specific BMR (e.g., 10%) does not imply that the same ranking would hold for a different BMR (1%, 5%, 20%, etc.). The real hazard ranking is determined by the shapes of the dose-response relationships for the different substances.

  There is a real danger that the hazard ranking for a specific BMR will be misinterpreted/misused as an absolute ranking for all BMRs.

- The personal computer software being developed for EPA will no doubt be useful. However, there is a real danger associated with limiting analyses solely to that software or discouraging the development of additional software.

  Any software package embodies limitations, restrictions, assumptions, and specific methods which may be beneficially explored in uncertainty analyses. Additional software should facilitate this information gathering.

  Differences between software in the insignificant digits of a calculation are insignificant and should not be used as a reason to discourage software development.

b. How understandable the document is for the general toxicologist/risk assessor;

- The document does a good job of introducing and explaining most concepts. This helps make the document understandable for the general toxicologist and risk assessor.

c. The overall organization of the document, further points to be developed or needing clarification;

- The Akaike Information Criterion (AIC) was never clearly defined.

  The Encyclopedia of Statistical Sciences defines the criterion as choosing the m, number of parameters, to minimize

  $$[(n+m+1)/(n-m-1)] \times \text{(residual mean square with m predictor parameters)}$$

  where n is the sample size.

  In other references, AIC implies minimizing with respect to k:

  $$-2 \times \text{(maximum log-likelihood with k parameters)} + 2k$$

- The GEE methods were never clearly defined.

- Background doses and background response rates are important in route-to-route extrapolations as well as in interspecies extrapolations.

d. The examples of BMD/C analyses in Appendix D.

- The inclusion of examples is a good idea. Several important issues are not addressed in these examples (e.g., the use of biologically relevant dose scales in dose-response modeling and the importance of adversity/severity).

Thomas Starr

Preliminary Comments on EPA's *draft Benchmark Dose Technical Guidance Document*

Thomas B. Starr, Ph.D.
ENVIRON Corporation
28 August 1996

1:21-24. *The LOAEL and NOAEL represent an operational definition of quantities that can characterize a study, and do not necessarily have any consistent association with underlying biological processes nor with thresholds.*

This statement applies equally well to BMD/C's (as defined in the document) and also to estimated upper confidence limits on risk at a given dose.

2-18. *The NOAEL also does not account for variability in the data ...*

A NOAEL must fail to produce a significant response, so whatever the observed response was, it must have been below the null hypothesis rejection cut point for the study, and the latter is determined, at least in part, by the variability of the data. Furthermore, LOAELs must produce a statistically significant response, so they do not share this "limitation" with NOAELs.

2-18-19. *..., a study with a limited number of animals will often result in a higher NOAEL than one which has more animals.*

I know this assertion is "common knowledge", but what is the factual basis for it? When have real studies identical except in sample size been conducted to examine the potential effects of sample size on NOAEL behavior?

2:20-21. *In addition, the slope of the dose-response curve is not taken into account in the selection of a NOAEL ...*

If Tukey's NOSTASOT trend test procedure were used, as in Faustman et al., this limitation would not apply. Note that BMD/Cs also are dependent on study design, *in particular on dose selection and spacing.*

2:27. *The notation: BMD/C*

LED10 or something equivalent would be preferred because it is more explicit about what it actually is.

3:28-29. *In general endpoints should be modeled if their LOAEL is up to 10-fold above the lowest LOAEL. This will insure that no endpoints with the potential ...*

The only way to **insure** is to analyze them all, not just those with LOAELs within a factor of 10 from the lowest LOAEL.

4:8-10. *With only one responding group, there is inadequate information about the shape of the dose-response curve, and mathematical modeling becomes too arbitrary.*

It is interesting, and perhaps ironic, that this situation, with only the highest dose group responding significantly, and, let's say, with lots of response-free doses really close to the highest dose tested, would provide the best possible empirical support for the existence of a sharply thresholded dose-response.

5:3-7. *Limit of Detection: ...*

I will argue strongly against this option. I believe that toxicologists, perhaps with some limited assistance from biostatisticians, must decide what a biologically significant level of response is **prior** to conducting experiments to be used for risk assessment purposes. Then biostatisticians can design the experiments to detect those effects with 80% or greater power. It is very serious mistake, in terms of senseless expenditure of valuable resources, to encourage the conduct of experiments which will fail (by design) to detect biologically significant effects as often as half of the time.

5:8-13. *Defaults: ...*

Why 10% extra risk? Is 10% extra risk typically detectable with any reasonable amount of power? If it isn't, then the studies to be used for this purpose will need to be strengthened. I believe very strongly that more stringent minimal criteria need to be established for studies to be used for risk assessment purposes. If strict criteria are not established now, we will be stuck with the same inadequate designs forever!

5:20. *A linear model should be run first.*

Why? Is this a conservative policy decision? No rationale is provided.

5:23-24. *Dichotomous data: ...*

Drop the log-logistic model and add the probit model. The logistic model was suggested by Fisher to simplify analysis of odds ratios which are themselves just approximations to relative risks. The models considered should be structured so as to provide "easy access" to an underlying cumulative hazard function, which might, at some future date, be "explained" or accounted for with mechanistic data. I also strongly recommend that a one-hit model *with intercept* be included in the family. This would provide a fairer test of the utility and adequacy of a threshold model than has been previously undertaken. Models flexible enough to resemble a

thresholded dose-response even without intercepts don't provide a fair null hypothesis test of the existence of a non-zero intercept.

6:8-9. *Unconstrained models may be applied if necessary to fit certain data.*

If this means that exponents smaller than 1 would be permitted in certain circumstances, I vigorously oppose it. No model should be entertained which produces nonsensical results, such as near-infinite slope near zero dose. They are just too easy to misuse.

6:17-18. *(based on p>0.05 from the goodness-of-fit statistic).*

A rationale needs to be provided for this cut point.

6:19-27. *Background parameter: ...*

I believe that a background parameter should be included in the models *unless there is good reason not to.* Non-monotonicity is not a good reason to include a background parameter. Historical control information is particularly important in this regard, yet is not mentioned here.

7:1-5. *Threshold Parameter: ...*

I believe strongly that a one-hit model *with intercept* should be included in the software. See my comments above re *Dichotomous Data.*

7:19-24. *Confidence limit calculation*

With the very small sample sizes of most studies of non-cancer endpoints, I find it difficult to accept that *asymptotic* properties of the likelihood ratio statistic are truly relevant. Small sample behavior needs to be considered; if little is known, then this should be a high priority.

Options other than 95% lower bounds should be made available. The 95% lower bounds are, on average, linear through zero even in the 10% response range. Thus, they do not reflect adequately any curvature in the underlying data. The studies are generally too weak to reject linear lower 95% bounds most of the time even at 10% response rate levels. I strongly encourage making bounds as "soft" as 80% part of the package, if only to reveal more clearly how insensitive the more stringent bounds are to curvature in the data.

8:4-11. *GOF in the range near the BMR*

This is very important, especially since what happens at the high dose end of the dose-response curve may very well be totally irrelevant to the issue of human risk.

Consideration should be given to weighting deviations inversely as some function of their *distance* from the low dose end of the observable range.

9:2-17. *somewhat arbitrary default criteria*

The Akaike Information Criterion is not described in any detail, and its performance characteristics with small samples and nearly equivalent models are probably not well-established. I am skeptical that this is a good idea. Furthermore, I believe that great care should be taken not to *trivialize* differences as small as factor a 3. This factor could be enormously important in terms of the costs of compliance. If the models are truly indistinguishable according to legitimate statistical criteria, then any of them could be used, and from the compliance cost side, the one with the largest estimate would obviously be preferred.

I don't understand why the desire for reasonable conservatism should come into play only with models whose estimates differ by more than a factor of 3. This is itself an arbitrary cut point. Furthermore, factors greater 3 are likely to be even more significant in terms of differentials in the costs of compliance. The desire for *reasonable conservatism* has made its way into too many places in this document. It needs to be balanced against the potentially unreasonable costs of compliance, (c.f., *The Perils of Prudence* by Nichols and Zeckhauser).

10:23-24. *Margin of Exposure (MOE)*

I am deeply concerned about how this concept will be employed in comparing risks across endpoints and substances. I greatly prefer the toxicologic concept of Margin of Protection (MOP). In contrast to MOE, which is a simple ratio of exposure levels, the MOP is a relative risk, i.e, a ratio of the estimated risks associated with two exposure levels.

The critical idea here is that Margin of Exposure is **not** synonymous with Margin of Protection (MOP) **except** when the relationship between exposure and the likelihood of a toxic response is linear. When a linear dose-response relationship prevails, a 100-fold Margin of Exposure will confer a 100-fold Margin of Protection, all other factors being equal. In contrast, when a threshold-like nonlinear dose-response prevails, a 100-fold Margin of Exposure could confer a 100,000-fold or even greater, perhaps infinite, Margin of Protection.

How then can one decide which of two materials to use in a particular application, even if they produce exactly the same toxicity? Suppose, for example, that material A offers a 10-fold MOE, but has a nonlinear (cubic, for example) dose-response, so that its MOE of 10 actually confers a $10^3 = 1000$-fold Margin of Protection. Material B, on the other hand, offers a 100-fold MOE, but has a shallow, nearly linear dose-response, so it confers only a 100-fold Margin of

Protection. Material B would be preferred by a factor of 10 based on the MOEs, but in fact, it is material A that provides the 10-fold greater Margin of Protection! This simple example illustrates two points: that 1) MOE calculations are not necessarily conservative; and 2) poor decisions may result from their use in comparative risk settings if one focuses on ratios of exposure levels rather than on the nature of the underlying dose-response relationships.

At a public meeting on the Commission on Risk Assessment and Risk Management's draft report held on 11 July in Washington DC, Dr. Adam Finkel (OSHA) described the net impact on toxicity assessment of the MOE approach as "Breaking What's Not Broken". Dr. Steven Bayard (EPA on loan to OSHA) agreed with Finkel that use of MOEs for comparative risk assessments is not a good idea. And, perhaps surprisingly, I agreed wholeheartedly with both of them! Use of MOE to conduct comparative risk analyses is most definitely **not** a good idea.

MOE use for comparative risk purposes presumes that exposure is a perfect surrogate for response; thus, its use is predicated on the assumption, albeit implicit, that all dose-response relationships are linear. While the MOE approach is simpler and more transparent than the linearized multi-stage model cancer risk assessment methodology it would replace, simpler is not necessarily better. Although the MOE approach would give both cancer and noncancer endpoints a common metric, it provides an essentially linear metric, the worst possible metric to be using for noncancer endpoints. I believe that public health would best be served by expending significant efforts to raise the level of consciousness and sophistication of the public, risk assessors and managers, and all other interested parties to the point where they can begin to appreciate the very real complexities and uncertainties surrounding these extremely difficult issues. I am concerned that the MOE approach works in exactly the opposite direction: it serves only to trivialize, by linearizing, all of toxicology.

10:26-29. *The LOAEL and NOAEL represent ...*

Exactly the same statement applies to BMDs. It's not fair to single out NOAELs and LOAELs as suffering these limitations when BMDs also possess them.

11:24-25. *... and is dependent on study design, in particular on dose selection and spacing.*

The same limitation applies to BMDs. Say so.

11:25-26. *... a limited number of animals will often result in a higher NOAEL than one which has more animals.*

This would only be true if the true response at the NOAEL were nonzero, and large enough to be detectable with the larger number of animals, but also too small to be detected with the smaller number of animals. How *often* do these particulars apply?

11:29. *Additionally, a LOAEL cannot be used to a derive a NOAEL ...*

The Faustman et al. papers show how the average LOAEL to NOAEL ratio in the studies they analyzed was about 2.5. This empirical factor could be used to estimate a NOAEL from a LOAEL in cases where one was not obtained.

12:5-7. *The BMD/C ... can be used as a more consistent point of departure than either the LOAEL or NOAEL.*

I disagree. I think LOAELs are every bit as good, if not better, especially in relation to lower bounds on ED10s. Maybe the problem is semantic. What exactly is meant by *more consistent*?

12:8-10. *The BMD/C accounts for variability in the data ...*

LOAEL responses must meet or exceed the least statistically significant response of the study design, and thus they do *account for variability in the data.* NOSTASOT NOAELs do also (see my earlier discussion on this point). It is simply not true that confidence limits must be employed as estimates to account for data variability. I believe that this is just one more manifestation of *reasonable conservatism.*

12:25-26. *The level of significance can be based on biological significance, or on statistical significance.*

The latter is a very poor second choice, and I am opposed to it as a default. Biological significance, established *a priori,* is the only truly legitimate criterion. Anything else is susceptible to accusations of *data dredging* or *cheating.* Minimal criteria for adequacy for risk assessment purposes need to be established, and determination of biological significance should be one of them. Otherwise, experimental designs can be manipulated to give practically any desired result.

16:16-18. *Thus, guidance is provided in this document on the use of the BMD/C as the point of departure for low dose extrapolation of both cancer and noncancer health effects.*

I thought that low dose extrapolation of noncancer health endpoints was to be avoided. If this is coming, I object strenuously.

17:5-8. *Definition of the Benchmark Dose ...*

Strike the words ... *the statistical lower confidence limit on* ... . The BMD should be a central estimate. The confidence interval can be given, but an unbiased (at least not intentionally biased) starting point for risk assessment should be employed. I have made and will continue to make the same argument in regard to the new draft cancer guidelines.

If we do not move off the lower bounds, then assessments of noncancer endpoints will be linearized in much the same way as cancer endpoints have been since the advent of the linearized multi-stage model. Studies of noncancer endpoints are generally too weak to reject linearity of the lower 95% confidence bound for risk-specific doses even as large as those producing 10% responses. Use of such lower bounds will likely introduce *additional* conservatism into the risk assessment process, particularly for dichotomous data, because these bounds are little different from the doses that would be obtained from straight-line-through-zero extrapolation from upper bounds on the responses at the NOAEL, LOAEL, or the maximum dose tested.

17:15-16. *The BMD/C approach does not reduce uncertainty inherent in extrapolating from animal data to humans (except for that in the LOAEL to NOAEL extrapolation) ...*

I don't understand meaning of the parenthetical phrase. Has the uncertainty in extrapolation from the LOAEL to the NOAEL been quantified objectively? If so, by how much has the BMD/C approach reduced it?

18:22. ... *studies for which modeling is feasible, ...*

Feasible is too weak a restriction. Only those studies for which modeling is a *reasonable exercise* should be considered.

19:21-24. *There must be more than one exposure group with a response different than controls ...*

See my earlier comments about this situation which may actually represent a sharp threshold.

20:27-29. *For endpoints for which there is no agreed upon biologically significant change, ...*

A range of potentially biologically significant changes should be explored. This would be far preferable to falling back on detection limits, which according to sound experimental design practice, were derived from estimates of minimum

biologically significant changes.

21:2-4. *In this proposal, there are two bases for specifying the BMR: a biologically significant change in response for continuous endpoints, or the limit of detection for either quantal or continuous data.* ...

Why is biological significance not also relevant to quantal endpoints? Elsewhere in the document, (I'm not sure exactly where) I believe there is a statement that any quantal response is biologically significant. Does EPA really believe this? I hope not. I believe that biological significance if every bit as important a consideration for quantal endpoints as it is for continuous ones.

21:9. *The limit of detection is based on ... and whether extra or additional risk is used in the model.*

I don't understand how the choice between extra or additional risk could influence the limit of detection of a study.

21:18-22. *Limit of Detection*

50% power is way too low. See my previous discussion.

21:23-24. *Defaults*

10% increase in extra risk is probably below detection limits (with reasonable power) of most noncancer study designs.

22:8-10. *The goal of mathematical modeling ... is to fit a model ... that describes the data set, especially at the lower end of the observable dose-response range.*

This is a laudable goal, but I do not see how the goodness of fit criteria, the AIC, or the maximum likelihood estimation process have been tailored to meet this goal. Actually, the fitting process finds a "best" model in some overall sense; it gives no special emphasis at all to the lower end of the observable dose-response range. See my earlier discussion for how one might approach this problem via inverse distance weighting.

22:23-26. *Thus, criteria for final model selection will be based solely on whether various models describe the data, conventions for the particular endpoint under consideration, and, sometimes, the desire to fit the same basic model form to multiple data sets.*

These criteria seem to have little relevance to the above-stated goal of the

modeling exercise.

23:6-20. *Order of Model Application*

No rationale is provided for running a linear model first. Is this reasonable conservatism again?

I recommend dropping the log-logistic model and adding the probit and one-hit with intercept models. See my earlier discussion.

23:28-29. *Unconstrained models may be applied if necessary to fit certain data.*

I disagree strongly. Weibull models with shape parameters less than one are simply biologically nonsensical near zero dose, and they should not be allowed into consideration for that reason alone.

24:3-9. *Degree of Polynomial ...*

The principal of parsimony seems to be playing a role in this discussion. This principal has proved useful when the plausibility of alternative but differentially complex mechanistic explanations of phenomena is considered. But the BMD approach is not mechanistically based. It is strictly empirical, and simpler is thus not necessarily to be preferred.

24:10-18. *Background Parameter ...*

A background parameter should be included unless there is good reason not to.

24:21-25. *Threshold parameter ...*

Justification for exclusion of a threshold term includes the statement that it is not a biologically meaningful parameter. Are any of the parameters of the proposed models *biologically meaningful*? I think not. As discussed previously, a one-hit model with intercept should be included in the suite of models to be included in the software.

25:5-9. *Conversion to dichotomous data could be considered ... in cases where the need for a probabilistic estimate of response outweighs the loss of information.*

A specific example of such a situation would be very helpful.

25:10-15. *Confidence Limit Calculation*

As indicated in my previous discussion, methods that account for the small sample characteristics of the likelihood ratio statistic would be preferable to those based on its asymptotic behavior, because the situations we will be considering are not likely to anywhere near asymptotic. I also strongly encourage the production of softer confidence limits (such as 80% and central estimates) for comparison purposes, because these are likely to be more responsive to any curvature present in the data. **In every case, the model parameters that correspond to the confidence limits as calculated should be provided.** This is mentioned nowhere in the document but is crucially important to gaining insight into the low dose behavior of those limits.

26:8-9. *For example, a smooth change of slope may be deemed more reasonable for a given response than an abrupt change.*

Or vice versa.

26:23-28. *... estimates from the remaining models are within a factor of 3, ...*

A better case needs to be made the AIC is really a useful ranking tool. I am not convinced. Also, why not use the AIC all the time? Finally, see my previous discussion of the importance of costs of compliance.

27:5-7. *Additional analysis might include the use of additional models, the examination of the parameter values for the models used, ...*

Examination of the parameter values for the models used is extremely important and deserves much heavier emphasis in the document.

29:11. *The dose at the MOE ... but the exact degree of protection is unknown, ...*

This is quite an understatement, but I agree in principle that proper interpretation of MOEs is impossible absent information regarding mechanism that has direct relevance to low-dose response.

29:26-29. *The BMD/C corresponds to a dose level which yields (with 95% confidence) ...*

Replace this with *The BMD is a lower 95% confidence bound on the dose level estimated to yield ...*

30:6-8. *Overall, the BMD/C will be a more consistent point of departure than the NOAEL*

This part of the assertion has not been established to my satisfaction.

32:29. *In contrast, EPA uses ED10s from the lower end of the experimental range because of its focus on human environmental exposure.*

Keep up the good work. Don't fall back onto LED10s.

34:2. *The LED10 is proposed as the point of departure.*

There is a kind of bootstrapping going on between the cancer guidelines and the BMD guidance documents. I understand that the draft cancer guidelines, which originally recommended ED10s as the starting point, replaced them with LED10s in part to be consistent with the BMD approach that had previously proposed using lower bounds on effect doses.

I'm not sure which came first, but I am vigorously opposed to the bounding procedure. Because the study designs for noncancer endpoints are so weak, the focus on bounds serves only to linearize noncancer endpoints in essentially the same manner as has already been accomplished with cancer. Risk managers deserve to see best estimates, in fact, entire distributions of estimates, not just particular estimates that are biased in many subtle ways, most unquantifiable, in the interests of public health.

34:17-18. *There might be modes of action other than DNA reactivity (e.g., certain receptor-based mechanisms) better supported by the assumption of linearity.*

I don't understand this sentence at all. If it is being suggested that certain **hypothetical** receptor-based mechanisms might **imply** linear dose responses, I would agree. However, I do not believe that linearity *per se* can be proven. It is just the other side of the threshold argument.

35:23-24. *... thus, the key objective of the MOE analysis is to describe for the risk manager how rapidly response may decline with dose.*

MOEs are just exposure ratios. I don't see how they can be used to say anything about how rapidly **response** may decline with dose, except in the special case where the dose-response is linear; in that case, MOEs are synonymous with MOPs (Margins of Protection).

37:16-17. *For an upper bound on linear extrapolation, a straight line from an upper bound on risk at the lower end of the experimental range had been proposed by Gaylor and Kodell (1980).*

If this were to be compared with the BMD/C approach proposed in the guidance document, it would differ very, very little. I prefer it to the BMD/C approach,

because it emphasizes that the real uncertainty in toxicology studies arises in the response data, not the treatment levels. Unlike the BMD/C approach, it does not confuse the issue by converting that response uncertainty to some corresponding uncertainty in the dose associated with a pre-specified, and likely never tested, dose. The doses employed in toxicology studies are by far the best defined characteristics of them.

38:13. *... slope factor would retain the use of statistical bounds, ...*

As noted previously, I am vigorously opposed to this approach. The bounds (both LED10s and BMD/Cs) are too insensitive to curvature in the data. They thus serve to linearize all of toxicology. A shift to central estimates alleviates this problem, and more stringent minimal experimental design criteria for risk assessment purposes can ameliorate any residual concerns regarding anti-conservatism.

41:25-27 *To estimate benefits, the numbers ... are obtained by multiplying individual risk times the size of the population exposed.*

How will individual risk be characterized? By upper bound risk estimates and/or upper bound exposure estimates? The opportunities for overstatement of benefit would seem to be virtually limitless.

42:22-23. *The benchmark dose approach thus provides a good starting point to develop benefits estimates for non-carcinogens.*

I disagree strongly. If the approach were to be implemented using central estimates rather than lower bounds, it would be far better. However, the main difficulty is the acknowledged absence of mechanistic information that is relevant to extrapolation, both across species and from high to lower doses. Absent case-specific mechanistic information relevant to these extrapolations, we are stuck with empirical curve-fitting, which can be trusted (at least to some extent) only for **interpolation** between data points; it cannot and should not be trusted at all for extrapolation.

I believe that the document must place far greater emphasis on the primacy of mechanistic data in guiding mathematical dose-response modeling. Additional approaches should be explored that might aggressively encourage, if not mandate, the collection of such data with experimental designs adequate for risk assessment purposes. If such progress is not demanded, we will be forever stuck analyzing inadequate studies with inadequate models.

APPENDIX E

FINAL OBSERVER LIST

EPA United States
Environmental Protection Agency
Risk Assessment Forum

Benchmark Dose Peer Consultation Workshop

Holiday Inn Bethesda
Bethesda, MD
September 10-11, 1996

Final Observer List

Sue Anne Assimon
Toxicologist
Center for Food Safety & Applied Nutrition
Contaminants Branch
U.S. Food & Drug Administration
200 C Street, SW (HFS-308)
Washington, DC 20204
202-205-8705
Fax: 202-260-0498

Monica Barron
Toxicologist
Office of Solid Waste
U.S. Environmental Protection Agency
401 M Street, SW (5307-W)
Washington, DC 20460
703-308-0483
Fax: 703-308-0509
E-mail: barron.monica@epamail.epa.gov

Steven Bayard
Statistician
U.S. Occupational Safety and Health Administration
200 Constitution Avenue, NW - Room N3718
Washington, DC 20120
202-219-7075 Ext: 131
Fax: 202-219-7125
E-mail: spbayard@aol.com

John Bowers
Mathematical Statistician
U.S.Food & Drug Administration
200 C Street, SW (HFS-708)
Washington, DC 20204
202-205-4782
Fax: 202-205-5069
E-mail: jbowers@vax8.csfan.fda.gov

William Burnam
Chief, Health Effects Division
Office of Pesticides
U.S. Environmental Protection Agency
401 M Street, SW (7509)
Washington, DC 20460
703-305-6193
Fax: 703-305-5147

Dan Byrd
President
CTRAPS
Suite 1150
1225 New York Avenue, NW
Washington, DC 20005
202-371-0603
Fax: 202-484-6019
E-mail: ctraps@radix.net

Clark Carrington
Toxicologist
Center for Food Safety & Applied Nutrition
Contaminants Branch
U.S. Food & Drug Administration
200 C Street, SW (HFS-308)
Washington, DC 20204
202-205-8705
Fax: 202-260-0498

Arthur Chin
Associate Toxicologist
EXXON Biomedical Sciences, Inc.
Mettlers Road (CN-2350)
East Millstone, NJ 08875
908-873-6255
Fax: 908-873-6009

David Cragin
Senior Toxicologist
ELF Atochem
2000 Market Street
Philadelphia, PA 19103
215-419-5880
Fax: 215-419-5800

Vicki Dellarco
Senior House Scientist
National Center for Environmental Assessment
U.S. Environmental Protection Agency
401 M Street, SW (8601)
Washington, DC 20460
202-260-7336
Fax: 202-260-0393
E-mail: dellarco.vicki@epamail.epa.gov

Karen Ekelman
Senior Review Toxicologist
Center for Food Safety & Applied Nutrition
U.S. Food & Drug Administration
200 C Street, SW (HFS-227)
Washington, DC 20204
202-418-3052
Fax: 202-418-3126

Hisham El-Masri
National Institute of Environmental
Health Sciences
P.O. Box 12233 (MD-A306)
Research Triangle Park, NC 27709
919-541-4432
Fax: 919-541-1479
E-mail: helmasri@wayout.niehs.nih.gov

Ellen Faria
Merck & Company, Inc.
1 Merck Drive (WS-2F-45)
Whitehouse Station, NJ 08889-0100
908-423-7907
Fax: 908-735-1496

Adam Finkel
Director, Health Standards Programs
Occupational Safety and Health Administration
200 Constitution Avenue, NW - Room N 3718
Washington, DC 20204
202-219-7075
Fax: 202-219-7125
E-mail: afinkel@dol.gov
epamail.epa.gov

Lynne Haber
Senior Associate
ICF/Clement, Inc.
9300 Lee Highway
Fairfax, VA 22031-1207
703-934-3126
Fax: 703-218-2668
E-mail: lhaber@icfkaiser.com

Jim Ivett
Principal Scientist
Corning Hazleton, Inc.
9200 Leesburg Pike
Vienna, VA 22182-1699
703-893-5400 Ext: 5496
Fax: 703-759-6947
E-mail: jli@cho.com

Annie Jarabek
Toxicologist
National Center for Environmental Assessment
U.S. Environmental Protection Agency (MD-52)
Research Triangle Park, NC 27711
919-541-4847
Fax: 919-541-1818
E-mail: jarabek.annie@epamail.epa.gov

Cindy Jengeleski
Manager, Scientific Programs
American Industry Health Council
2001 Pennsylvania Avenue, NW - Suite 760
Washington, DC 20006
202-833-2183
Fax: 202-833-2201
E-mail: cjengeleski@aihc.org

Alan Katz
Executive Director
TAS, Inc.
1000 Potomac Street, NW
Washington, DC 20007
202-337-2625
Fax: 202-337-1744

Arnold Kuzmack
Office of Water
U.S. Environmental Protection Agency
401 M Street, SW (4301)
Washington, DC 20460
202-260-5821

Susan Lewis
Panel Manager
Chemical Manufacturers Association
1300 Wilson Boulevard
Arlington, VA 22209
703-741-5635
Fax: 703-741-6091
E-mail: susan_lewis@mail.cmahq.com

Ron Lorentzen
Strategic Manager, Risk Assessment
U.S. Food & Drug Administration
200 C Street, SW (HFS-16)
Washington, DC 20204
202-205-8753
Fax: 202-401-2893
E-mail: rjl@fdacf.ssw.dhhf.gov

Amal Mahfouz
Senior Toxicologist
Office of Water
U.S. Environmental Protection Agency
401 M Street, SW (4304)
Washington, DC 20460
202-260-9568
Fax: 202-260-1036
E-mail: mahfouz.amal@epamail.epa.gov

William Marcus
Office of Water
U.S. Environmental Protection Agency
401 M Street, SW (4301)
Washington, DC 20460
202-260-5400

Elizabeth Margosches
Chief, Epidemiology & Quantitative
Methods Section
Health & Environmental Review Division
Office of Pollution, Prevention & Toxics
U.S. Environmental Protection Agency
401 M Street, SW (7403)
Washington, DC 20460
202-260-1511
Fax: 202-260-1279
E-mail: margosches.elizabeth@epamail.epa.gov

Yogi Patel
Toxicologist
Health and Ecological Criteria Division
Office of Water
U.S. Environmental Protection Agency
401 M Street, SW (4304)
Washington, DC 20460
202-260-5849
Fax: 202-260-1036

Hugh Pettigrew
Mathematical Statistician
Health Effects Division
Office of Pesticide Programs
U.S. Environmental Protection Agency
401 M Street, SW (7509C)
Washington, DC 20460
703-305-5699
Fax: 703-305-5147

Kathleen Raffaele
Toxicologist
Division of Health Effects Evaluation
Center for Food Safety & Applied Nutrition
U.S. Food & Drug Administration
200 C Street, SW (HFS-227)
Washington, DC 20204
202-418-3056
Fax: 202-418-3126

Chad Sandusky
Director, Special Projects
Technical Assessment Systems, Inc.
1000 Potomac Street, NW
Washington, DC 20007
202-337-2625
Fax: 202-337-1744
E-mail: tas@tasinc.com

Don Saunders
Director, Toxicology
CIBA Crop Protection
P.O. Box 18300
Greensboro, NC 27419-8300
910-632-2322
Fax: 910-632-2997
E-mail: donald.saunders@usgr.mhs.ciba.com

Val Schaeffer
Toxicologist
Epidemiology & Health Sciences Division
Consumer Products Safety Commission
4330 East West Highway - Room 600-15
Washington, DC 20207
301-504-0994
Fax: 301-504-0025

Ceinwen Schreiner
Toxicology Consultant
Mobil Oil Corporation
P.O. Box 310
Paulsboro, NJ 08066
609-222-2703
Fax: 609-222-3064
E-mail: caschrei@pau.mobil.com

Jennifer Sead
Scientist
439 East Luray Avenue
Alexandria, VA 22301
703-683-7465
Fax: 703-683-7465
E-mail: jseed89770@aol.com

Joseph Siglin
Senior Toxicologist
EXXON Biomedical Sciences, Inc.
Mettlers Road (CN 2350)
East Millstone, NJ 08875-2350
908-873-6289
Fax: 908-873-6009

Bonnie Stern
Senior Scientist
EA Engineering Science & Technology
8401 Colesville Road - Suite 500
Silver Spring, MD 20910
301-565-4216
Fax: 301-587-4752
E-mail: bs@eaeng.mhs.compuserv.com

Shirley Tao
Toxicologist
Center for Food Safety & Applied Nutrition
Contaminants Branch
U.S. Food & Drug Administration
200 C Street, SW (HFS-308)
Washington, DC 20204
202-205-8705
Fax: 202-260-0498

Abraham Tobia
Manager, USTOX
BASF Corporation
26 Davis Drive
P.O. Box 13528
Research Triangle Park, NC 27709
919-547-2172
Fax: 919-547-2421
E-mail: tobiaa@basf.com

Richard Williams
Chief
Center for Food Safety & Applied Nutrition
Economics Branch
U.S. Food & Drug Administration
200 C Street, SW (HFS-724)
Washington, DC 20204
202-401-6088
Fax: 202-260-0794
E-mail: rxw@fdacf.ssw.dahhs.gov

Jeanette Wiltse
Associate Director
National Center for Environmental Assessment
U.S. Environmental Protection Agency
401 M Street, SW (8601)
Washington, DC 20460
202-260-7317
Fax: 202-260-0393
E-mail: wiltse.jeanette@epamail.epa.gov

William Wood
Executive Director
Risk Assessment Forum
U.S. Environmental Protection Agency
401 M Street, SW (8103)
Washington, DC 20460
202-260-6743
Fax: 202-260-3955

☆U.S. GOVERNMENT PRINTING OFFICE: 1997 - 549-001/60111