DRAFT
DO NOT CITE OR QUOTE

EPA/630/R-00/001
October 2000
External Review Draft

Benchmark Dose Technical Guidance Document

NOTICE

THIS DOCUMENT IS A PRELIMINARY DRAFT. It has not been formally released by the U.S. Environmental Protection Agency and should not at this stage be construed to represent Agency policy. It is being circulated for comment on its technical accuracy and policy implications.

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

DISCLAIMER

This document is an external draft for review purposes only and does not constitute U.S. Environmental Protection Agency policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

EXECUTIVE SUMMARY

The US EPA conducts risk assessments for an array of health effects that may result from exposure to environmental agents, and that require an analysis of the relationship between exposure and health-related outcomes. The dose-response assessment is essentially a two-step process, the first being the definition of a point of departure (POD), and the second extrapolation from the POD to low environmentally-relevant exposure levels. The benchmark dose (BMD) approach provides a more quantitative alternative to the first step in the dose-response assessment than the current NOAEL/LOAEL process for noncancer health effects, and is similar to that for determining the POD proposed for cancer endpoints (EPA, 1996). As the Agency moves toward harmonization of approaches for cancer and noncancer risk assessment, the dichotomy between cancer and noncancer health effects is being replaced by consideration of mode of action and whether the effects of concern are likely to be linear or nonlinear at low doses. Thus, the purpose of this document is to provide guidance for the Agency and the outside community on the application of the BMD approach in determining the POD for all types of health effects data, whether a linear or nonlinear low dose extrapolation is used.

This guidance document discusses the computation of BMDs and benchmark concentrations (BMCs), their lower confidence limits, data requirements, dose-response analysis, and reporting requirements that are specific to the use of BMDs or BMCs. The following convention for terminology has been adopted in this document: BMD is used generically to refer to the benchmark dose approach; in the more specific cases, BMD and BMC refer to the central estimates, for example the EDx or ECx for dichotomous endpoints (with x referring to some level of response above background, e.g., 5% or 10%). BMDL or BMCL refers to the corresponding lower limit of a one-sided 95% confidence interval on the BMD or BMC, respectively. This is consistent with the terminology introduced by Crump (1995) and with that used in the EPA's BMD software (BMDS) which is freely available on the Internet at http://www.epa.gov/ncea/bmds.htm. This terminology is a change, however, from that used in previous Agency documents (e.g., EPA, 1995), but has been adopted because it more clearly conveys the fact that the BMDL refers to the lower confidence limit on the dose that would result

in the required response.

As indicated above, the BMD approach is an alternative to the NOAEL/LOAEL approach that has been used for many years in dose-response assessment. The development of this approach has been pursued because of recognized limitations in the NOAEL/LOAEL approach. However, it is likely that there will continue to be endpoints that are not amenable to modeling and for which a NOAEL/LOAEL approach must be used. In some cases, there may be a combination of BMDs and NOAELs to be considered in the assessment of a particular agent, and the most appropriate value to use for dose-response assessment must be made by the risk assessor on the basis of scientific judgment and the modeling results.

This document addresses a number of issues that must be resolved in order to apply the BMD approach for dose-response assessment in a consistent manner:

1. Determination of appropriate studies and endpoints on which to base BMD calculations;
2. Selection of the benchmark response (BMR) value;
3. Choice of the model to use in computing the BMD;
4. Details surrounding computation of the confidence limit for the BMD (BMDL); and
5. Reporting requirements for BMD and BMDL computation.

Determination of appropriate studies and endpoints on which to base BMD calculations. Following the hazard characterization and selection of appropriate endpoints to use for the dose-response assessment, the studies appropriate for modeling and BMD analysis can be evaluated. All studies that show a graded monotonic response with dose likely will be useful for BMD analysis, and the minimum data set for calculating a BMD should at least show a significant dose-related trend in the selected endpoint(s). It is preferable to have studies with one or more doses near the level of the BMR to give a better estimate of the BMD, and thus, a shorter confidence interval. Studies in which all the dose levels show changes compared with control values (i.e., there is no NOAEL) are readily useable in BMD analyses, unless the lowest response level is much higher than the BMR.

There are at least three types of endpoint data: dichotomous (quantal), continuous, and categorical. This guidance provides definitions of these three types of data, and what information is needed in order to model the responses. For example, a dichotomous response may be

reported as either the presence or absence of an effect, a continuous response may be reported as an actual measurement, or as a contrast (absolute change from control or relative change from control). In the case of continuous data, when individual data are not available, the number of subjects, mean of the response variable, and a measure of response variability (e.g., standard deviation (SD), standard error (SE), or variance) are needed for each group. For categorical data, the responses in the treatment groups are often characterized in terms of the severity of effect (e.g., mild, moderate, or severe histological change). In general, endpoints that have been judged by the risk assessor to be appropriate and relevant to the exposure should be modeled if their LOAEL is up to 10-fold above the lowest LOAEL. This will help ensure that no endpoints with the potential to have the lowest BMDL are excluded from the analysis on the basis of the value of the LOAEL or NOAEL. Selected endpoints from different studies that are likely to be used in the dose-response assessment should all be modeled, especially if different uncertainty factors may be used for different studies and endpoints. As indicated above, the selection of the most appropriate BMDs and/or NOAELs (if some endpoints cannot be modeled) to use for determination of the POD must be made by the risk assessor using scientific judgement and principles of risk assessment, as well as the results of the modeling process.

Selection of the benchmark response (BMR) value. The calculation of a BMD is directly determined by the selection of the BMR. This guidance provides default criteria to be used for selecting the BMR in the case of quantal data and continuous data. For quantal data, an excess risk of 10% is the default BMR, since the 10% response is at or near the limit of sensitivity in most cancer bioassays and in some noncancer bioassays as well. If a study has greater than usual sensitivity, then a lower BMR can be used, although the ED_{10} and LED_{10} should always be presented for comparison purposes.

For continuous data, if there is an accepted level of change in the endpoint that is considered to be biologically significant then that amount of change is the BMR. Otherwise, if individual data are available and a decision can be made about what individual levels should be considered adverse, the data can be "dichotomized" based on that cutoff value, and the BMR set as above for quantal data. Alternatively, in the absence of any other idea of what level of response to consider adverse, a change in the mean equal to one control SD from the control

mean can be used. The control SD can be computed including historical control data, but the control mean must be from data concurrent with the treatments being considered. Regardless of which method of defining the BMR is used for a continuous dataset, the effective dose corresponding to one control SD from the control mean response, as would be calculated for the latter definition, should always be presented for comparison purposes.

Choice of the model to use in computing the BMD. The goal of the mathematical modeling in BMD computation is to fit a model to dose-response data that describes the data set, especially at the lower end of the observable dose-response range. In practice, this involves first selecting a family or families of models for further consideration, based on characteristics of the data and experimental design, and fitting the models using one of a few established methods. Subsequently, a lower bound on dose is calculated at the BMR. The guidance document introduces the topic of dose-response modeling and provides information on model selection for different types of data. In addition, model fitting, determining goodness-of-fit, and comparing models to decide which one to use for obtaining the POD are discussed. The guidance recommends that α=0.1 be used to compute the critical value for goodness of fit, instead of the more conventional values of 0.05 or 0.01, and that a graphical display of the model fit be examined as well. For comparison of models and selection of the model to use for BMDL computation, the use of Akaike's Information Criterion (AIC) is recommended.

Computation of the confidence limit for the BMD (BMDL). The guidance document discusses the computation of the confidence limit for the BMD, the fact that the method by which the confidence limit is obtained is typically related to the data type, and the manner in which the BMD is estimated from the model. Details for approaches to CI computation specific to particular data types (quantal, clustered, continuous, multiple outcomes) are provided in the document.

Reporting requirements from the BMD/BMDL calculations. The guidance document lists a number of reporting requirements for the BMD and BMDL. These are considered important for the risk assessor to judge whether or not the choice of studies and endpoints for modeling has been done appropriately and whether the most appropriate BMD and BMDL have been selected as the POD for low dose extrapolation.

In summary, the guidance document provides a decision tree that discusses step-by-step the process to be used in evaluating studies and endpoint types that are appropriate for modeling, selecting the BMR level, model fitting and BMD computation, judging the fit of the model, and the calculation of the BMDL. Finally, the document provides several examples of BMD and BMDL derivation using the EPA BMDS software.

I. INTRODUCTION

A. Purpose of This Guidance Document

The purpose of this document is to provide guidance for the Agency and the outside community on the application of the benchmark dose approach to determining the point of departure (POD) for linear or nonlinear extrapolation of health effects data. This guidance discusses computation of benchmark doses and benchmark concentrations (BMDs and BMCs) and their lower confidence limits, data requirements, dose-response analysis, and reporting requirements. The document provides guidance based on today's knowledge and understanding, and on experience gained in using this approach. The Agency is actively applying this methodology and evaluating the outcomes for the purpose of gaining experience in using it with a variety of endpoints. This document is intended to be updated as new information becomes available that would suggest approaches and default options alternative or additional to those indicated here and should not be viewed as precluding additional research on modified or alternative approaches that will improve quantitative risk assessment. In fact, the use of improved scientific understanding and development of more mechanistically-based approaches to dose-response modeling is strongly encouraged by the Agency.

Benchmark dose modeling is a highly technical exercise and this guidance is a technical document generally targeted at readers with sufficient background in this area. The document is not intended as a primer on modeling or risk assessment. The availability of software to facilitate the analysis can make the modeling appear deceptively simple, but often interpretation of the results is not trivial. It is recommended that BMD modeling be performed by or in collaboration with a statistician or someone familiar with the potential pitfalls of this type of analysis. Similarly, this document is not intended as a primer on toxicology; the procedures described herein do not replace the expert judgements of toxicologists and others who address the hazard characterization issues in risk assessment. Expert judgements on study quality, toxicological significance of observed effects, etc., are required independent of the use of BMD analysis and

are beyond the scope of this document. It is likewise beyond the scope of this document to provide guidance for RfC, RfD, or cancer potency computation, which are also more general risk assessment issues.

Since the methods for BMD computation require appropriate software, another purpose of this document is to provide enough information about preferred computational algorithms to allow users to make an informed choice in the selection of that software. The document does not advocate use of any particular software package, although it is recommended that software with well documented algorithms, such as the Agency's BMDS package, be used. Nor is this guidance intended to document any particular software package, although it will present examples for illustrative purposes that use the Agency's BMDS package. It is also expected that this guidance will inform the design of studies for the computation of BMDs and dose-response analysis, though this will not be covered explicitly.

B. Background

The US EPA conducts risk assessments for an array of health effects that may result from exposure to environmental agents. The process of risk assessment, based on the National Research Council paradigm (NRC, 1983), has several steps: hazard characterization, dose-response assessment, exposure assessment, and risk characterization. Hazard characterization includes a thorough evaluation of all the available data to identify and characterize potential health hazards. Dose-response assessment involves an analysis of the relationship between exposure to the chemical and health-related outcomes, and historically has been done very differently for cancer and noncancer health effects because of perceived differences between the mechanistic underpinnings of cancer and other toxic effects. As our understanding of the underlying biology of toxic effects has grown, however, the apparent differences between cancer and noncancer effects have lessened, to the point where it seems reasonable to develop quantitative methods based on similar considerations for all types of health effects, and to make approaches to dose-response assessment as consistent across health endpoints as our current mechanistic understanding allows. This section provides an overview of EPA's approaches to

dose-response assessment for cancer and non-cancer effects, and of the basis for developing more broadly applicable quantitative methods.

The primary distinction between characterizing risks of cancer and noncancer effects has been the expectation that cancer induction could result from even a single gene mutation in a single cell, while noncancer effects were generally assumed to occur only if a minimum, but possibly large, amount of damage had occurred. The practice for assessing dose-response for cancer effects has been to fit a statistical model (linearized multistage procedure) to tumor incidence data, and to assume low dose linearity to extrapolate risk at lower doses (USEPA, 1986). The modeling addresses variability in the data through an upper 95% bound on the slope of the relationship between exposure and risk at very low risk levels, typically 10^{-5} to 10^{-6}.

In contrast, the standard practice for the dose-response analysis of health effects other than cancer has been to estimate the minimum dose not to be exceeded, by identifying a lowest-observed-adverse-effect-level (LOAEL) and a no-observed-adverse-effect-level (NOAEL) from an appropriate study. The LOAEL is the lowest dose for a given chemical at which adverse effects have been detected, while the NOAEL is the highest dose at which no adverse effects have been detected. The NOAEL (or LOAEL, if a NOAEL is not present) is adjusted downward by uncertainty factors intended to account for limitations and uncertainties in the available data, to arrive at an exposure that is likely to be without an appreciable risk of deleterious effects in humans, that is, the reference dose (RfD) or reference concentration (RfC). Unlike cancer dose-response modeling, variability in the observed responses is not addressed.

It has been tempting to use the dose level below which no effects are observed in a study (sometimes called a "practical threshold") as an important point for describing a dose-response curve because of a presumed relationship between such a practical threshold and true thresholds (i.e., true no effect levels) in the dose-response. In fact, the practical threshold is really a consequence of the fact that any finite study has an inherent *limit of detection*, and is of little practical utility in describing toxicological dose-responses. In other words, the NOAEL does not represent a biological threshold and does not imply that lower exposure levels are without risk. Specific limitations of the NOAEL/LOAEL approach are well known and have been discussed extensively (Crump, 1984; Gaylor, 1983; Kimmel and Gaylor, 1988; Leisenring and Ryan, 1992;

EPA, 1986b, 1988a,b, 1989c; 1995c):

- The NOAEL/LOAEL is highly dependent on dose selection since the NOAEL/LOAEL is limited to being one of the doses included in a study.
- The NOAEL/LOAEL is highly dependent on sample size. The ability of a bioassay to distinguish a treatment response from a control response decreases as sample size decreases[1], so that the NOAEL for a compound (and thus the POD) will tend to be higher in studies with smaller numbers of animals per dose group.
- More generally, the NOAEL/LOAEL approach does not account for the uncertainty in the estimate of the dose-response which is due to the characteristics of the study design.
- NOAELs/LOAELs do not correspond to consistent response levels for comparisons across studies/chemicals/endpoints and for use as PODs for the derivation of RfCs.
- The slope of the dose-response curve is not taken into account in the selection of a NOAEL or LOAEL, and is not usually considered unless the slope is very steep or very shallow.
- A LOAEL cannot be used to derive a NOAEL when a NOAEL does not exist in a study. Instead, a tenfold uncertainty factor has been routinely applied to the LOAEL to account for this limitation.
- While the NOAEL has typically been interpreted as a threshold (no-effect level), simulation studies (i.e, Leisenring and Ryan, 1992) and reanalyses of developmental toxicity bioassay data (Allen et al, 1994a) have demonstrated that the rate of response above control at doses fitting the criteria for NOAELs, for a range of study designs, is about 5-20% on average, not 0%.

In an effort to address some of the limitations of the LOAEL and NOAEL, Crump (1984) proposed the benchmark dose (BMD) approach as an alternative (see section I.C. for more

[1]Note that for a study utilizing 6 animals per dose group, the 95% upper confidence limit (UCL) on an observed adverse response rate of 0% is 49%. That is, NOAELs chosen on the basis of no observed response in 6 animals could be too high a substantial proportion of the time. The 95% UCLs for groups of 10, 20 and 50 animals are 31% , 17%, and 7%, respectively, underscoring the importance of adequate sample sizes.

details). Benchmark dose modeling makes no particular assumption about the nature of toxicological dose-responses, other than that the change in response generally does not decrease with higher doses. In particular, there is no specific assumption of the relationship between a putative no-effect level in the dose-response and the benchmark dose. The goal of the BMD approach is to define a starting point of departure (POD) for the computation of a reference value (e.g., the RfD or RfC) or for linear low-dose extrapolation that is more independent of study design.

The BMD approach parallels the recommendations in EPA's Proposed Guidelines for Carcinogen Risk Assessment (1996a) regarding modeling tumor data and other (non-cancer) responses thought to be important precursor events in the carcinogenic process. The proposed guidelines promote the understanding of an agent's mode of action in determining the dose-response(s). Moreover, the dose-response extrapolation procedure follows conclusions in the hazard assessment about the agent's carcinogenic mode of action. The dose-response assessment under the proposed guidelines is a two-step process: (1) response data are modeled in the range of empirical observation -- modeling in the observed range is done with biologically-based, case-specific, or appropriate curve-fitting models; and then (2) extrapolation below the range of observation is accomplished by modeling if there are sufficient data or by a default procedure (linear, nonlinear, or both). For the default procedures, a point of departure (POD) for extrapolation is estimated from this modeling. The linear default is a straight-line extrapolation to the background response level from the POD, while the nonlinear default approach begins at the identified POD and provides either a margin of exposure (MOE) analysis or a reference value such as and RfD or RfC rather than estimating the probability of effects at low doses.

In the case of deriving reference values for noncancer effects, the POD is adjusted downward, to account for the uncertainty that is contributed by extrapolation from experimental animals to humans and to account for within human variability, as well as other limitations in the available data. Note that the NOAEL or LOAEL has been used as a default POD for low dose estimation or extrapolation, so that the primary difference between the two approaches is in how the starting point is determined. The POD for BMD modeling is the BMDL, or the lower 95% bound on the dose/exposure associated with the benchmark response, typically 10% above the

control response. Using the lower bound accounts for the uncertainty inherent in a given study, and assures (with 95% confidence) that the desired BMR is not exceeded (see section II.B. for a complete discussion of selecting the benchmark response).

As detailed above, the BMD approach is generally a preferable alternative to the NOAEL/LOAEL approach. For instance, a BMDL can be estimated even when all doses in a study are associated with a significant adverse response (i.e., when there is no NOAEL). Note, however, that there are some instances in which the NOAEL/LOAEL is the better choice. In particular, the available data may not be amenable to modeling, such as when all individuals in exposed groups respond. In such a case, BMD models may fail to fit the observed data, which provide very little resolution in the region of the benchmark response (usually 10%) anyway (although in such a case, the LOAEL is not very informative, either). Another circumstance may happen when an observed effect is so rare that it is not statistically significantly different from the control response, but may be found to be biologically meaningful (e.g., an increase in a rare malformation).

Note that the literature has used the terms BMD and BMDL in a confusing way (Crump, 1984, 1995). There is frequent need to refer to the central estimate and the lower confidence limit, as well as a more generically-defined point of departure in discussions of dose-response assessment. In this document, when talking in technical detail about the process of deriving benchmark doses, "BMD" or "BMC" will refer to the central estimate of the dose that is expected to yield the BMR, for example, the ED_{10}, or EC_{10}, and "BMDL" or "BMCL" will refer to the lower end of a one-sided confidence interval for that central estimate. "BMD" will be used to refer to the entire process. The POD for low dose extrapolation or for setting the RfD/C will be the BMDL or BMCL. To simplify further discussion in this document, we will use BMD and BMDL generically to mean oral or inhalation values, unless stated otherwise.

Illustrative Example: Using the BMD approach, the experimental data are modeled, and the benchmark dose (BMD) in the observable range is estimated (see Fig. 1). Unlike NOAELs and LOAELs, the BMD is not constrained to be one of the experimental doses, and the BMDL can thus be used as a more consistent POD than either the LOAEL or NOAEL. The BMDL accounts for the uncertainty in the estimate of the dose-response that is due to characteristics of

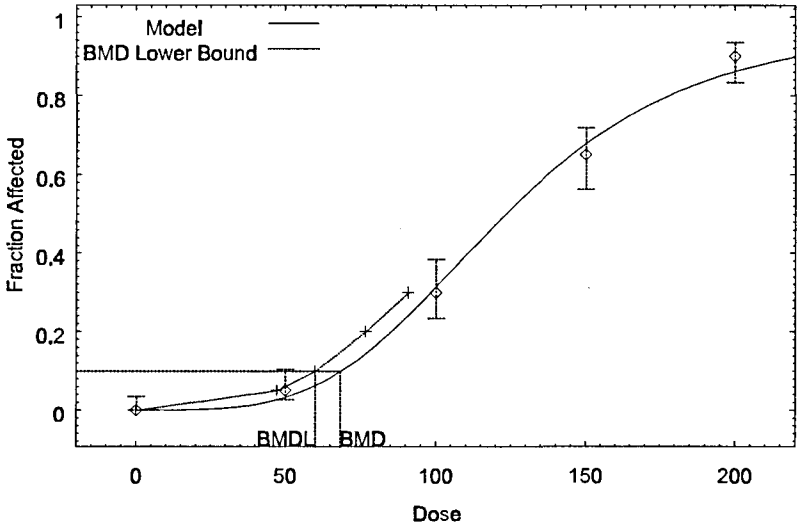

Figure 1 Sample of a model fit to dichotomous data, with BMD and BMDL indicated. The fraction of animals affected in each dose group is indicated by diamonds. The error bars indicate 95% confidence intervals for the fraction affected. The BMR for this example is an Extra Risk of 10%. The dashed curve indicates the BMDL for a range of BMRs. The dose labeled BMDL corresponds to the lower end of a one-sided 95% confidence interval for the BMD.

the experimental design. The BMD approach models all of the data in a study and the shape of the dose-response curve is integral to the BMDL estimation.

Since the benchmark dose procedure is quite general, a number of issues need to be addressed before benchmark doses can be used in a consistent manner for dose-response assessment:

1. how to select studies on which to base BMD calculations;
2. selection of endpoints on which to base BMD calculations;
3. selection of the benchmark response (BMR) value;
4. choice of the model to use in computing the BMD;
5. details surrounding computation of the confidence limit for the BMD (BMDL);
6. what information from the BMD calculation should be reported

These issues will be covered in some detail in the following chapters.

C. A Brief Review of Literature Relating to Benchmark Dose

1. Earlier uses of benchmark modeling in dose-response assessment

Benchmark dose-like approaches to dose-response assessment are not new. The procedure of Mantel and Bryan (1961) formerly was used widely for conservative low-dose cancer risk assessment. Their procedure calculated an upper confidence limit on the excess tumor incidence at the lowest experimental dose or an upper confidence limit on the tumor incidence at the dose estimated to produce a 1% tumor incidence, essentially a benchmark dose. Assuming a probit-log dose model, a conservative low-dose slope of one probit per factor of 10 reduction in dose below the upper limit on the benchmark dose was used to provide an upper bound estimate of cancer incidence at low doses. Gaylor and Kodell (1980), Van Ryzin (1980), and Farmer et al. (1982) proposed low-dose linear extrapolation to zero excess risk from the upper confidence limit on the excess incidence above background of an adverse effect at the lowest experimental dose or dose corresponding to a 1% incidence, again, a benchmark dose, to provide an upper bound on low-dose risks for convex (sublinear) dose-response curves. Gaylor (1983) and Krewski et al. (1984) compare linear extrapolation and safety factors for controlling low-dose risk. Crump (1984) first coined the term "benchmark dose," although variations of a benchmark dose procedure had been in use since the process developed by Mantel and Bryan (1961).

2. Properties of the Benchmark Dose

A number of research efforts, many of which have dealt with reproductive and developmental toxicity data, have provided extremely useful information for application of the BMD approach (e.g., Alexeeff et al., 1993; Catalano et al., 1993; Chen et al., 1991; Krewski and Zhu, 1994, 1995; Auton, 1994; Crump, 1995; Fowles, et al., 1999). In a series of papers by Faustman et al. (1994), Allen et al. (1994a and b), and Kavlock et al. (1995), the BMD approach was applied to a large database of developmental toxicity studies. In brief, the results of these studies showed that when the data were expressed as the proportion of affected fetuses per litter (nested dichotomous data), the NOAEL was on average 0.7 times the BMDL for a 10% probability of response, and was approximately equal, on average, to the BMDL for a 5% probability of response. When data were expressed as counts of dichotomous endpoints (i.e.,

number of litters per dose group with resorptions or malformations), the NOAEL was approximately 2-3 times higher than the BMDL for a 10% probability of response above control values (approximately 20 animals per dose group), and 4-6 times higher than the BMDL for a 5% probability of response. Expressing the data as the proportion of affected fetuses per litter is the more appropriate way to analyze developmental toxicity data. However, the results of the quantal data analysis also may apply to using the BMDL approach with other quantal data, and suggest that the NOAEL in these cases may be at or above the 10% true response level, depending on sample size and background rate.

Since reduced fetal weight in developmental toxicity studies often shows the lowest NOAEL among the various endpoints evaluated, the application of the BMD to these continuous data also was evaluated (Kavlock et al., 1995). A variety of cutoff values was explored for defining an adverse level of weight reduction below control values. In some cases, data were analyzed using a continuous power model, and in other cases, the data were transformed to dichotomous data. Comparisons with the NOAEL showed that several cutoff values could be used to give values similar to the NOAEL. These analyses suggest ways in which BMDLs may be developed for continuous data from a variety of endpoints.

Fowles, et al. (1999) examined acute inhalation lethality data, and compared NOAELs to benchmark doses corresponding to 1%, 5%, and 10% response incidences. Sample sizes averaged around10 – 20 animals per dose group. Similarly to the "quantal" parts of the results of the Allen *et al.* (1994, a and b) studies, BMDLs based on 10% incidence corresponded approximately to NOAELS. However, because the dose-response for lethality is so steep, BMDLs for 5% and 1% incidences were very close to those for 10% incidence. As a result, the BMDLs for a 1% incidence were on average only about 1.6 or 3.6 times smaller than a NOAEL, depending on whether a log-probit or Weibull model was used.

A simulation study by Kavlock et al. (1996) examined various aspects of study design (number of dose groups, dose spacing, dose placement, and sample size per dose group) for two endpoints of developmental toxicity (incidence of malformations and reduced fetal weight). Of the designs evaluated, the best results (that is, those with the shortest confidence intervals) were obtained when two dose levels had response rates above the background level, one of which was

near the BMR. In this study, there was virtually no advantage in increasing the sample size from 10 to 20 litters per dose group. When neither of the two dose groups with response rates above the background level was near the BMR, satisfactory results were also obtained, but the BMDLs tended to be lower. When only one dose level with a response rate above background was present and near the BMR, reasonable results for the maximum likelihood estimate and BMDL were obtained, but in this case, there were benefits of larger dose group sizes. The poorest results were obtained when only a single group with an elevated response rate was present, and the response rate was much greater than the BMR.

3. Approaches to BMD Computation

Many noncancer health effects are characterized by multiple endpoints that are not completely independent of one another. Lefkopoulou et al. (1989), Chen et al. (1991), Ryan (1992), Catalano et al. (1993), Zhu et al. (1994), Krewski and Zhu (1995), and Fung et al. (1998) have worked on this issue using developmental toxicity data, and have shown that, in general, the BMDL derived from a multinomial modeling approach is lower than that for any individual endpoint. This approach has not been applied to other health effects data, but should be kept in mind when multiple related outcomes are being considered for a particular health effect.

Dose-response modeling for continuous endpoints is made more difficult because there is not a natural probability scale in which to characterize risk. Of course, one approach is to explicitly dichotomize such continuous endpoints, and then model the explicitly dichotomized endpoints as any other quantal endpoint. In separate 1995 papers, Crump and Kodell et al. detailed a new approach to deriving a BMDL for continuous data based on a method originally proposed by Gaylor and Slikker (1990). This approach makes use of the distribution of continuous data, estimates the incidence of individuals falling above or below a level considered to be adverse or at least abnormal, and gives the probability of responses at specified doses above the control levels. This results in an expression of the data in the same terms as that derived from analyses of quantal data, that is, it *implicitly* dichotomizes the data while retaining the full power of modeling the continuous data while allowing direct comparison of BMDs and BMDLs derived from continuous and quantal data. Gaylor (1996) compared benchmark doses computed for continuous endpoints directly to those computed after first explicitly dichotomizing the data,

and found that, even for moderate sample sizes, substantial precision was lost upon explicitly dichotomizing the data. West and Kodell (1999) compared such an implicit method for continuous data to the result of modeling explicitly dichotomized endpoints. They found that, for sample sizes in the range of 10 to 20 animals per dose group, the implicit approach gave substantially better results than did the approach of modeling explicitly dichotomized data. Thus, when it is possible to do, it is generally better to derive BMDs and BMDLs for continuous data from models of the continuous data (perhaps using the hybrid approaches described by Gaylor and Slikker, 1990, Crump, 1995 or Kodell et al., 1995).

Most approaches to benchmark dose modeling have focused on modeling a single or multiple responses from a single study. Categorical regression modeling (Dourson et al., 1985; Hertzberg, 1989; Hertzberg and Miller, 1985; Guth et al, 1997; Simpson et al, 1996ab) allows the results for multiple endpoints across studies to be used to make an overall assessment of the toxicity of a compound, based on a larger data base. Although so far this method has not been widely used for benchmark dose computation, it shows promise as a way to more quantitatively and rigorously combine information from a rich database.

Bayesian approaches to benchmark dose calculation express the uncertainty in the benchmark dose estimate with a probability distribution (in Bayesian parlance, the *posterior* distribution), in contrast to the confidence limits used by the more commonly used frequentist approach (Hasselblad and Jarabek, 1995). Although the Bayesian approach has not been widely used so far, it has some potentially useful features. It would be relatively easy to combine results from different data sets to provide a more robust estimate, along with an evaluation of the uncertainty in that estimate that would take into account the variability among studies. This would be a clear improvement over the more widely used methods, which only quantify the uncertainty inherent in a single study.

Gaylor, et al. (1998) reviewed statistical methods for computing benchmark doses, and Murrel et al. (1998) discussed some consequences of basing the benchmark dose on a confidence limit and suggested an approach for setting benchmark response levels for continuous endpoints.

4. General Discussions of Standards for the Benchmark Dose

Several workshops and symposia have been held to discuss the application of the BMDL and appropriate methodology (Kimmel et al., 1989; California EPA, 1993; Beck et al., 1993; SRA Symposium, 1994; Barnes et al., 1995). The participants at the EPA/AIHC workshop (Barnes et al., 1995) generally endorsed the application of the BMD approach for all quantal noncancer endpoints and particularly for developmental toxicity, where a good deal of research has been done. Less information was available at the time of the workshop on the application of the BMD approach to continuous data, and more work was encouraged. A number of other issues concerning the application of the BMD approach were discussed. The guidance and default options set forth in the current document are based in part on the outcome of this workshop, the background document (EPA, 1995c), and on more recent information and discussions, including those at a peer consultation workshop on the 1996 draft of this report (USEPA, 1996).

II. BENCHMARK DOSE GUIDANCE

This section describes the proposed approach for carrying out a complete BMDL analysis. It is organized in the form of a decision process including the rationale and defaults for proceeding through the analysis, and follows a similar framework to that outlined in the background document (EPA, 1995c). The guidance here imposes some constraints on the BMDL analysis through decision criteria, and provides defaults when more than one feasible approach exists.

A. Data Evaluation and Endpoint Selection

The first step in the process of hazard characterization is a complete review of the toxicity data available on an agent to identify and characterize the hazards related to a particular compound or exposure situation. This involves the determination of adverse effects or precursors of adverse effects from all available data and the most appropriate endpoints, the so-called "critical effect(s)," on which to base the NOAEL or BMD. Guidance on review of endpoint data for hazard characterization can be found in a number of EPA publications (EPA, 1991a, 1994c, 1995f, 1996a and b). This process is essentially the same whether using a BMD or a NOAEL approach. The following discussion summarizes some of the more important issues related to study design and data reporting when using the BMD approach. This guidance does not change the way in which hazard characterization is done, particularly regarding the determination of adversity and selection of endpoints. It does discuss the types of data and study designs most amenable to dose-response modeling, but allows for the possibility that NOAELs will continue to be used for some endpoints, and that in some cases there will be a combination of BMDs and NOAELs to be considered in the assessment of a particular agent.

1. Data Evaluation

a. Design

In general, studies with more dose groups and a graded monotonic response with dose will be more useful for BMD analysis. Studies with only a single dose showing a response different from controls may not be appropriate for BMD analysis, though if the one elevated response is near the BMR, adequate BMD and BMDL computation may result (see Kavlock, et al, 1996). Studies in which responses are only at the same level as background or at or near the maximal response level are not considered adequate for BMD analysis. It is preferable to have studies with one or more doses near the level of the BMR to give a better estimate of the BMD, and thus, a shorter confidence interval . Studies in which all dose levels show changes compared with control values (i.e., no NOAEL) are readily useable in BMD analyses, unless the lowest response level is much higher than that at the BMR.

b. Aspects of Data Reporting

In many cases, the risk assessor must rely on published reports of key toxicological studies in performing a dose-response assessment. Reports from the peer-reviewed literature may contain summary information which can vary in completeness *vis-a-vis* the data requirements of the BMD method. The optimal situation is to have information on individual subjects, but this is unlikely in published reports. It is more common to have summary information (group level information, e.g., mean and standard deviation) concerning the measured effect, especially for continuous response variables, and it must be determined whether the summary information is adequate for the BMD method to proceed.

Dichotomous data are normally reported at the individual level (e.g., 11/50 animals showed the effect). Occasionally a dichotomous endpoint will be reported as being observed in a group with no mention of the number of animals showing the effect. This usually occurs when the incidence of the endpoint reported is ancillary to the focus of the report. For BMD modeling of dichotomous data, both the number showing the response and the total number of subjects in the group are necessary.

Continuous data are reported as a measurement of the effect, such as body weights or enzyme activity in control and exposed groups. The response might be reported in several

different ways, e.g., as an actual measurement, or as a contrast (absolute change from control or relative change from control). To model continuous data when individual animal data are not available, the number of subjects, mean of the response variable, and a measure of variability (e.g., standard deviation, SD; standard error, SE; or variance) are needed for each group. The lack of a numerically reported SD or SE precludes the calculation of the BMD. In some cases, a measure of variability is presented for the control group only and this information can be used for modeling by making an assumption, for example, that the variance in the exposed groups is the same as the controls. However, the modeling of data and calculation of the confidence limits will not be as precise as when the variance information is available for individual groups.

Categorical data are defined as a type of quantal data in which there is more than one defined category in addition to the no-effect category and the responses in the treatment groups are characterized in terms of the severity of effect (e.g., mild, moderate, or severe histological change). Results may be classified by reporting an entire treatment group in terms of category (group level reporting), or by reporting the number of animals from each group in each category (individual level reporting). For example, a report of epithelial degenerative lesions might state that an exposed group showed a mild effect (group level) or that in the exposed group there were 7 animals with a mild effect and 3 with no effect (individual level reporting). In the latter case, the BMD can be calculated using a quantal model after combining data in severity categories (e.g., model all animals with a particular severity of effect or all with greater than a mild effect). Dichotomous data can be viewed as a special case in which there is one effect category and the possible response is binary (e.g., effect or no effect). Information may also be treated as categorical in cases where an endpoint is inherently a dichotomous or continuous variable, but because the endpoint is reported only descriptively, and the number affected and total number exposed are not reported, it cannot be treated quantitatively. Modeling approaches have been discussed for categorical data with multiple categories (Dourson et al., 1985; Hertzberg, 1989; Hertzberg and Miller, 1985) and for group level categorical data (Guth et al., 1997, Simpson et al., 1996a,b). These regression models can also be used to derive a BMD, by estimating the probability of effects of different levels of severity.

2. Selection of Studies to be Modeled

Following a complete review of the toxicity data, the risk assessor must select the studies appropriate for benchmark dose analysis. The selection of the appropriate studies is based on the human exposure situation being addressed, the quality of the studies, and the relevance and reporting adequacy of the endpoints.

The process of selecting studies for benchmark dose analysis is intended to identify those studies for which modeling is feasible, so that BMDLs can be calculated and used in dose-response assessment. In most cases, the selection process will identify a single study or very few studies for which calculations are relevant; all studies considered relevant should be modeled. Cases in which there are a number of studies, or studies with a number of endpoints reported may require a large number of BMD calculations. In these cases, it may be possible to select a subset of endpoints as representative of the effects in the target organ or the study. This selection can be made on the basis of sensitivity or severity, which may be more easily compared within a single study in the same target organ than across studies.

3. Selection of Endpoints to be Modeled.

Once studies have been evaluated with regard to their appropriateness for BMD modeling, the selection of endpoints to model should focus on the dose-response relationships. For example, differences in slope (at the BMR) among endpoints could affect the relative values of the BMDLs to the corresponding LOAELs/NOAELs. Thus, selection of endpoints should not be limited to only the one with the lowest LOAEL. In general, endpoints within a study that have been judged by the risk assessor to be appropriate and relevant to the exposure should be modeled if their LOAEL is up to 10-fold above the lowest LOAEL. This will help ensure that no endpoints with the potential to have the lowest BMDL are excluded from the analysis on the basis of the value of the LOAEL or NOAEL. Selected endpoints from different studies that are likely to be used in determination of the POD should all be modeled, especially if different uncertainty factors may be used for different studies and endpoints. The selection of the most appropriate BMDs to use for determining the POD must be made by the risk assessor using scientific judgement and principles of risk assessment, as well as the results of the modeling process.

4. Minimum Data Set for Calculating a BMD

Once the critical endpoints have been selected, data sets are examined for the appropriateness of a BMD analysis. The following constraints on data sets to use for BMD calculations should be applied:

- There must be at least a statistically or biologically significant dose-related trend in the selected endpoint.
- The data set should contain information relevant to dose-response for modeling. A determination of the amount of information about the dose-response that is available need not be quantitative or technical. For example, a data set in which all non-control doses have essentially the same response level provides limited information about the dose-response, since the complete range of response from background to maximum must occur somewhere below the lowest dose: the BMD may be just below the first dose, or orders of magnitude lower. When this situation arises in quantal data, especially if the maximum response is less than 100%, it is tempting to use a model like the Weibull with no restrictions on the power parameter, because such models reach a plateau of less than 100% and most modeling programs do not include other models for quantal data that have this property. This situation can result in seriously distorted BMDs, because the model predictions jump rapidly from background levels to the maximum level. In principle, other models could be found that force the BMD to be anywhere between that extreme and the lowest administered dose. Thus the BMD computed here depends solely on the model selected, and goodness of fit provides no help in selecting among the possibilities. (see the quantal data examples in the appendix for a worked example of this situation). The sad reality in such situations is that the data provide little useful information about dose-response; the ideal solution is to collect further data in the dose-range missed by the studies in hand.

  When there is a jump between non-control doses between no response and maximal response, there is still limited information about dose-response, but the dose-spacing may ameliorate the situation, since the BMD is effectively bracketed between the two doses that determine the jump.

5. Combining Data for a BMD Calculation

Data sets that are statistically and biologically compatible may be combined prior to dose-response modeling, resulting in increased confidence, both statistical and biological, in the calculated BMD. In addition, the use of combined data sets may encourage further studies if the additional data can affect the BMD estimate. Allen et al. (1996) provided an example of a case where data on boron developmental effects could be combined for the BMD analysis. The simplest approach to combining datasets is simply to treat the data as if they were all collected simultaneously. If it is plausible that the multiple datasets represent a homogeneous picture of the dose-response (for example, the responses at doses common to two or more datasets are essentially the same, and statistically undifferentiable), then this is an appropriate approach. More likely, there will be some variability among datasets which will require more elaborate modeling to include properly. There is as yet too little practical as well as theoretical experience with this situation to allow specific guidance in the matter, other than to say that statistically appropriate methods must be used and justified if data sets are combined for modeling. An example of statistically accommodating variability among studies is the model for categorical regression developed by Simpson, et al. (1996, a and b).

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

At the time of this writing, the Agency is developing guidance for the selection of the appropriate response level, or BMR, for use with BMD modeling. In the interim, this document will describe BMR selection as it has typically been done to date.

The major aim of benchmark dose modeling is to model the dose-response data for an adverse effect in the observable range (i.e., across the range of doses for which toxicity studies have reasonable power to detect effects) and then select a "benchmark dose" at the low end of the observable range to use as a "point of departure" for deriving quantitative estimates below the range of observation and to use as a basis for comparison of effective doses corresponding to a common response level across chemicals or endpoints. Because different study designs have different sensitivities to observe adverse effects (i.e., limits of detection), the low end of the observable range

will correspond to different response levels for different study designs. A 10% response level is conventionally used (at least for dichotomous endpoints) to define effective doses (i.e., ED_{10}s and LED_{10}s) for comparing potencies across chemicals or endpoints (e.g., for chemical rankings). This response level is used for such comparisons because it is at the low end of the observable range for many common study designs, although for some designs the limit of detection is above the 10% level and for others it is below. For the POD, on the other hand, it is not critical that a common response level be used for all chemicals or endpoints, and for the purposes of deriving quantitative estimates at doses below the observable range, it may be desirable to use response levels below 10%, if possible, in order to minimize the degree of low-dose extrapolation required. Thus, while it is important to always report ED_{10}s and LED_{10}s for comparison purposes, the actual "benchmark dose" used as a POD may correspond to response levels below (or sometimes above) 10%, although for convenience standard levels of 1%, 5%, or 10% have typically been used rather than a floating level dependent on the actual limit of detection of the relevant study.

For continuous data, there are various possibilities for selecting the BMR (see below); however, regardless of which of the options is used, it is recommended that the BMD (and BMDL) corresponding to a change in the mean response equal to one control standard deviation from the control mean always be presented for comparison purposes (see below, third bullet for continuous data). This value would serve as a standardized basis for comparison, akin to the ED_{10} for dichotomous data.

The following describes the criteria conventionally used currently for selecting the BMR. For quantal (dichotomous) data, the conventional approaches are fairly straight forward. For continuous data, on the other hand, there is less historical precedence to draw upon, however some reasonable options are presented. Once a BMR is selected and the dose-response data are modeled, the BMD is explicitly determined.

- Quantal data:
    - An excess risk of 10% has generally been the default BMR for quantal data. The 10% response is at or near the limit of sensitivity in most cancer bioassays and in some noncancer bioassays as well.

- If a study has greater than usual sensitivity, then a lower BMR can be used, although the ED_{10} and LED_{10} are always presented for comparison purposes. For example, reproductive and developmental studies having nested study designs often have greater sensitivity, and for such studies a BMR of 5% has typically been used. Similarly, epidemiology studies often have greater sensitivities and a BMR of 1% has typically been used for quantal human data.

- Continuous data:
  - If there is a minimal level of change in the endpoint that is generally considered to be biologically significant (for example, a change in average adult body weight of 10%, or the doubling of average level for some liver enzyme), then that amount of change can be used to define the BMR. (The BMD [and BMDL] corresponding to a change in the mean response equal to one control standard deviation from the control mean should also be presented for comparison purposes [see third bullet].)
  - If individual data are available and a decision can be made about which individual levels can be reasonably considered adverse (perhaps based on a quantile of the control distribution, for example), then the data can be "dichotomized" based on that cutoff value, and the BMR can be set as above for quantal data. (The BMD [and BMDL] corresponding to a change in the mean response equal to one control standard deviation from the control mean should also be presented for comparison purposes [see third bullet].)
  - In the absence of any other idea of what level of response to consider adverse, a change in the mean equal to one control standard deviation from the control mean (see Section II C2e) can be used. The control standard deviation can be computed including historical control data, but the control mean must be from data concurrent with the treatments being considered (Crump, 1995). This gives an excess risk of approximately 10% for the proportion of individuals below the 2^{nd} percentile or above the 98^{th} percentile of controls for normally distributed effects.

C. Modeling the Data

1. Introduction

The goal of the mathematical modeling in benchmark dose computation is to fit a model to dose-response data that describes the data set, especially at the lower end of the observable dose-response range. The fitting must be done in a way that allows the uncertainty associated with parameter estimates to be quantified and related to the estimate of the dose that would yield the benchmark response. In practice, this procedure will involve first selecting a family or families of models for further consideration, based on characteristics of the data and experimental design, and fitting the models using one of a few established methods. Subsequently, a lower bound on dose is calculated at the BMR. This section is too brief to do more than introduce the topic of modeling. Some references for further reading are: Chapter 10 of Draper and Smith (1981), Gallant (1987), Bates and Watts (1988), McCullagh and Nelder (1989), Seber and Wild (1989), Ross (1990), Clayton and Hills (1993), Davidian and Giltinan (1995).

Dose-response models are expressed as functions of dose, possibly covariates, and a set of constants, called parameters, that govern the details of the shape of the resulting curve. They are fitted to a data set by finding values of the parameters that adjust the predictions of the model for observed values of dose and covariates to be close to the observed response. Dose-response models for toxicology data are usually of the type called "nonlinear" in mathematical terminology. In a linear model, the value the model predicts is a linear combination of the parameters. For example, in a linear regression of a response y on dose, the predicted value is a linear combination of a and b, namely, $a \times 1 + b \times dose$ .Note that, even a quadratic or other polynomial is a linear model, in this sense: $y = a + b \times dose + c \times dose^2 + d \times dose^3$ is a third degree polynomial (a cubic) equation, but is still a linear combination of the parameters, a, b, c, and d. In contrast, in a nonlinear model, for example the log-logistic with background,

$$p = P_0 + \frac{1 - P_0}{1 + e^{-[a + b\log(dose)]}}$$

the response is not a linear combination of the parameters (here, P_0,

a, and *b*). The distinction is important, because nonlinear models are usually more difficult to fit to data, requiring more complicated calculations, and statistical inference is more typically approximate than with linear models. Note that this definition of "linear" is in contrast to the way the term is used in reference to cancer dose-response assessment, in which the phrase "low-dose linear" refers to models in which the linear coefficient on dose is positive.

At the present, although biological models may often be expressed as nonlinear models (e.g., Michaelis-Menten curves), nonlinear models do not necessarily have a biological interpretation. 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. Since it is preferable to use special purpose modeling software, EPA is in the process of developing software which includes several models and default processes as described in this document (http://www.epa.gov/ncea/bmds.htm).

2. Background for Model Selection

This section provides some basic statistical background and guidance on how to go about choosing a model structure appropriate to the data being analyzed, selection of "equivalent" models, and confidence limit calculation to derive the BMDL to use as the point of departure.

a. Selecting the Model

The initial selection of a group of models to fit to the data is governed by the nature of the measurement that represents the endpoint of interest and the experimental design used to generate the data. In addition, certain constraints on the models or their parameter values sometimes need to be observed, and may influence model selection. Finally, it may be desirable to model multiple endpoints, at the same time. The diversity of possible endpoints and shapes of their dose-responses for different agents precludes specifying a small set of models to use for BMD computation. This will inevitably lead to the need for judgement and occasional ambiguity when selecting the final model and BMDL for dose-response assessment. It is hoped that, as experience using benchmark dose methodology in dose-response assessment accumulates, it will be possible to narrow the number of acceptable models.

i. Type of endpoint

The kind of measurement variable that represents the endpoint of interest is an important consideration in selecting mathematical models. Commonly, such variables are either continuous, like liver weight or the activity of a given liver enzyme, or discrete, commonly dichotomous, like the presence or absence of abnormal liver status. However, other types are common in biological data; for example: ordered categorical, like a histology score that ranges from 1-normal to 5-extremely abnormal; counts, such as counts of deaths or the numbers of cases of illness per thousand person-years of exposure to a given exposure condition; waiting time, such as the time it takes for an illness to appear after exposure, or age at death, or multiple endpoints (such as survival, weight, and malformations in a developmental toxicity study) considered jointly (see, references in section I.C.2). It is beyond the scope of this document to consider all possible kinds of variables that might be encountered, so further discussion will concentrate on dichotomous and continuous variables.

Dichotomous variables. Data on dichotomous variables are commonly presented as a fraction or percent of individuals that exhibit the given condition at a given dose or exposure level. For such endpoints, normally we select probability density models like logistic, probit, Weibull, and so forth, whose predictions lie between zero and one for any possible dose, including zero.

Continuous variables. Data for continuous variables are often presented as means and standard deviations or standard errors, but may also be presented as a percent of control or some other standard. From a modeling standpoint, the most desirable form for such data is by individual. Unlike the usual situation for dichotomous variables, summarization of continuous variables results in a loss of information about the distribution of those variables.

The preferred approach to expressing the BMR will determine the approach to modeling continuous data. Two broad categories of approach have been proposed: 1) to express the BMR as a particular change in the mean response, possibly as a fraction of the control mean, a fraction of the range of the response (when there is a clear maximum response), a fraction of the standard deviation of the measurement from untreated individuals, or a level of the response that expert opinion holds is adverse; or 2) to decide on a level of the outcome to consider adverse, and treat the proportion of individuals with the adverse outcome much as one would a dichotomous variable.

Typical models to use in the first situation include linear and polynomial models, and power

models or other nonlinear models such as Hill models. In the second situation, one approach is to classify each individual as affected or not, and model the resulting variable as dichotomous.

An alternative is to use a so-called "hybrid" approach, such as that described by Gaylor and Slikker (1990), Kodell et al. (1995), and Crump (1995), which fits continuous models to continuous data, and, presuming a distribution of the data, calculates a BMD in terms of the fraction affected. Using this approach, the probability (risk) of an individual with an adverse level is estimated directly as a function of dose in four steps (Gaylor and Slikker,1990). In the first step, the probability distribution among individuals of the continuous measure is established for the control group. Often this distribution may be approximately log-normal, i.e., the logarithm of the values of the biological measure are normally distributed. Since most biological effects do not assume negative values, the log-normal distribution satisfies this condition. If high values are adverse, a large percentile (e.g., 99th percentile) of the distribution may be selected as a cutoff value for normal levels with larger values considered adverse. Conversely, if low values are adverse, a small percentile (e.g., first percentile) may be selected to classify individuals with lower values as adverse.

In the second step, a dose-response curve is fit to the data to establish how the average value changes as a function of dose. In the third step, the variability of individuals about the average dose-respose curve is calculated. Often this can be expressed simply by the standard deviation about the dose-response curve. It is common for the standard deviation of biological measurements to be proportional to their average value, i.e., a constant coefficient of variation. Again, this is a property of the log-normal distribution. However, the coefficient of variation may change with dose which leads to a more complicated analysis of the data. In this case, it is often useful to model the variance as proportional to the mean raised to a power. This model includes the case where the coefficient of variation is constant, where the variance is proportional to the square of the mean, and the coefficient of variation is the square root of the constant of proportionality.

From the average values estimated from the dose-response curve in step 2 and the variability of values about the curve estimated in step 3, it is possible in the 4th step to estimate the probability, for any dose, that an individual is in the adverse range established in the 1st step. Hence, the BMD can be estimated for a specified BMR. The BMDL can then be calculated for use as a POD for low dose risk assessment.

ii. Experimental design

The aspects of experimental design that bear on model selection include the total number of dose groups used and possible clustering of experimental subjects. The number of dose groups has a bearing on the number of parameters that can be estimated: the number of parameters that affect the overall shape of the dose-response curve generally cannot exceed the number of dose groups.

Clustering of experimental subjects is actually more of an issue for methods of fitting the models than for choice of the model form itself. The most common situation in which clustering occurs is in developmental toxicity experiments, in which the agent is applied to the mother, and individual offspring are examined for adverse effects. Another example is for designs in which individuals yield multiple observations (repeated measures). This can happen, for example, when each subject receives both treatment and control (common in studies with human subjects), or each subject is observed multiple times after treatment (e.g., neurotoxicity studies). The issue in all of these examples is that individual observations cannot be taken as independent of each other. Most methods used for fitting models rely heavily on the assumption that the data are independent, and special fitting methods need to be used for data sets that exhibit more complicated patterns of dependence (see, for example, Ryan 1992; Davidian and Giltinan, 1995).

iii. Constraints and covariates

An obvious constraint on models for dichotomous data has already been been discussed: probabilities are constrained to be positive numbers no greater than one. However, biological reality may impose other constraints on models. For example, most biological quantities are constrained to be positive, so models should be selected so that their predicted values, at least in the region of application, conform to that constraint. In models in which dose is raised to a power which is a parameter to be estimated (such as a Weibull model), if that parameter is allowed to be less than 1.0, the slope of the dose-response curve becomes infinite at a dose of zero. This often results in numerical problemss in calculating the confidence interval. This is an undesirable situation, and the default is to constrain these parameters to be at least 1.0 (see Example 1).

In quantal models, often a background parameter quantifies the probability that the outcome being modeled can occur in the absence of exposure. It may be tempting to reduce the number of parameters to be estimated by fixing the value of the background parameter to be zero. However,

only when it is clear that an outcome is *impossible* in the absence of the exposure is it permissible to fix the value of the background to zero.

It is preferred that a so-called "threshold" term not be included in the models used for BMD/C analysis because, while it is not an estimate of a biological threshold, it is easily confused with such because of confusing terminology, and because most data sets can be fit adequately without this parameter and the associated loss of a degree of freedom. The software currently distributed by EPA does not currently include this parameter. However, occasionally, it may happen that the increase in a response is so precipitous that including a threshold parameter facilitates the dose-response modeling, and in such cases it is acceptable to include the parameter.

It is sometimes desirable to include covariates on individuals when fitting dose-response models. For example, litter size has often been included as a covariate in modeling laboratory animal data in developmental toxicity studies. Another example is in modeling epidemiology data, when certain covariates (e.g., age, parity) are included that are expected to affect the outcome and might be correlated with exposure. In continuous models, if the covariate has an effect on the response, including it in a model may improve the precision of the overall estimate by accounting for variation that would otherwise end up in the residual variance. In any kind of model, any variable that is correlated (non-causally) with dose, and which affects outcome, would need to be included as a covariate.

b. Model Fitting

The goal of the fitting process is to find values for all the model parameters so that the resulting fitted model describes those data as well as possible; this is termed "parameter estimation." In practice, this happens when the dose-group means predicted by the model come as close as possible to the data means. One way to achieve this is to write down a function (the objective function) of all the parameters and all the data, with the property that the parameter values that correspond either to an overall minimum (or, equivalently, an overall maximum) of the function, or that result in function values of zero, give the desired model predictions.

The actual fitting process is carried out iteratively, and starts with an initial guess for the parameter values. This guess is iteratively updated to produce a sequence of estimates that (usually) converge. Many models will converge to the right estimates for most data sets from just about any

reasonable set of initial parameter values; however, some models, and some data sets, may require multiple guesses at initial values before the model converges. It also happens occasionally that the fitting procedure will converge to different estimates from different initial guesses. Only one of these sets of estimates will be "best". It is always good practice when fitting nonlinear models to try different initial values, just in case.

There are a few common ways to construct objective functions: the methods of nonlinear least squares, maximum likelihood, and generalized estimating equations (GEE). The choice of objective function is determined in large part by the nature of the variability of the data around the fitted model. The method of nonlinear least squares, where the objective function is the sum of the squared differences between the observed data values and the model-predicted values, is a common method for continuous variables when observations can be taken as independent. A basic assumption of this method is that the variance of individual observations around the dose-group means is a constant across doses. When this assumption is violated (commonly, when the variance of a continuous variable changes as a function of the mean, often proportional to the square of the mean, giving a constant coefficient of variation), a modification of the method may be used in which each term in the sum of squares is weighted by the reciprocal of an estimate of the variance at the corresponding dose. This method is especially appropriate when the data to be fitted can be supposed to be at least approximately normally distributed.

Maximum likelihood is a general way of deriving an objective function when a reasonable supposition about the distribution of the data can be made. Because estimates derived by maximum likelihood methods have good statistical properties, such as asymptotic normality, maximum likelihood is often a preferred form of estimation when that assumption is reasonably close to the truth. An example of such a situation is the case of individual independently treated animals (e.g., not clustered in litters) scored for a dichotomous response. Here it is reasonable to suppose that the number of responding animals follows a binomial distribution with the probability of response expressed as a function of dose. Continuous variables, especially means of several observations, are often normal (gaussian) or log-normal. When variables are normally distributed with a constant variance, minimizing the sum of squares is equivalent to maximizing the likelihood, which explains in part why least squares methods are often used for continuous variables. In developmental toxicity

data, the distribution of the number of animals with an adverse outcome is often taken to be approximately beta-binomial. This particular likelihood is used to accommodate for the lack of independence among littermates.

A third group of approaches to estimating parameters are the related quasi-likelihood method (McCullagh and Nelder, 1989) and the method of GEE (see Zeger and Liang, 1986), which require only that the mean, variance, and correlation structure of the data be specified. GEE methods are similar to maximum likelihood estimation procedures in that they require an iterative solution, provide estimates of standard errors and correlations of the parameter estimates, and estimates are asymptotically normal. Their use so far has primarily been to handle forms of lack of independence, as in litter data, and would be useful in any of a number of kinds of repeated measures designs, such as occur in clinical studies and repeated neurobehavioral testing.

c. Assessing How Well the Model Describes the Data

An important criterion is that the selected model should describe the data, especially in the region of the BMR. Most fitting methods will provide a global goodness-of-fit measure, usually providing a P-value. These measures quantify the degree to which the dose-group means that are predicted by the model differ from the actual dose-group means, relative to how much variation of the dose-group means one might expect. Small P-values indicate that it would be unlikely to achieve a value of the goodness-of-fit statistic at least this extreme if the data were actually sampled from the model, and, consequently, the model is a poor fit to the data. Since it is particularly important that the data be adequately modeled for BMD calculation, it is recommended that α=0.1 be used to compute the critical value for goodness of fit, instead of the more conventional values of 0.05 or 0.01. P-values cannot be compared from one model to another since they assume the different models are correct; they can only identify those models that are consistent with the experimental results. When there are other covariates in the models, such as litter size, the idea is the same, just more complicated to calculate. In this case, the range of doses and other covariates is broken up into cells, and the number of observations that fall into each cell is compared to that predicted by the model.

It can happen that the model is never very far from the data points (so the P-value for the goodness-of-fit statistic is not too small), but is always on one side or the other of the dose-group

means. Also, there could be a wide range in the response, and the model predicts the high responses well, but misses the low dose responses. In such cases, the goodness-of-fit statistic might not be significant, but the fit should be treated with caution. One way to detect such situations is with tables or plots of residuals: measures of the deviation of the response predicted by the model from the actual data. If the residuals are scaled by an estimate of their standard deviation, then residuals that exceed 2.0 in absolute value warrant further examination of the model.

Another way to detect the form of these deviations from fit is with graphical displays. Plots should always supplement goodness-of-fit testing. It is extremely helpful that plots that include data points also include a measure of dispersion of those data points, such as confidence limits.

In certain cases, the typical models for a standard study design cannot be used with the observed data as, for example, when the data are not monotonic, or when the response rises abruptly after some lower doses that give only the background response. In these cases, adjustments to the data (e.g., a log-transformation of dose) or the model (e.g., adjustments for unrelated deaths) may be necessary. In the absence of a mechanistic understanding of the biological response to a toxic agent, data from exposures that give responses much more extreme than the BMR do not really tell us very much about the shape of the response in the region of the BMR. Such exposures, however, may very well have a strong effect on the shape of the fitted model in the region of the BMD. Thus, if lack of fit is due to characteristics of the dose-response data for high doses, the data may be adjusted by eliminating the high dose group. The practice carries with it the loss of a degree of freedom, but may be useful in cases where the response plateaus or drops off at high doses. Since the focus of the BMD analysis is on the low dose and response region, eliminating high dose groups is reasonable. Alternatively, an entirely different model could be fit.

d. Comparing Models

It will often happen that several models provide an adequate fit to a given data set. These models may be essentially unrelated to each other (for example a logistic model and a probit model often do about as well at fitting dichotomous data) or they may be related to each other in the sense that they are members of the same family that differ in which parameters are fixed at some default value. For example, one can consider the log-logistic, the log-logistic with non-zero background, and the log-logistic with threshold and non-zero background to all be members of the same family

of models. Goodness-of fit statistics are not designed to compare different models, so alternative approaches to selecting a model to use for BMDL computation need to be pursued.

Generally, within a family of models, as additional parameters are introduced the fit will appear to improve. This general behavior is due solely to the increase in the additional parameters. Likelihood ratio tests can be used to evaluate whether the improvement in fit afforded by estimating additional parameters is justified. Such tests cannot be applied to compare models from different families, however. Some statistics, notably Akaike's Information Criterion (AIC) (Akaike, 1973; Linhart and Zucchini, 1986; Stone, 1998; AIC is $-2L + 2p$, where L is the log-likelihood at the maximum likelihood estimates for the parameters, and p is the number of model degrees of freedom) can be used to compare models with different numbers of parameters using a similar fitting method (for example, least squares or a binomial maximum likelihood). Although such methods are not exact, they can provide useful guidance in model selection.

When other data sets for similar endpoints exist, an external consideration can be applied. It may be possible to compare the result of BMDL computations across studies if all the data were fit using the same form of model, presuming that a model can be found that describes all the data sets. Another consideration is the existence of a conventional approach to fitting a kind of data. In this case, communication with specialists in that type of data is eased when a familiar model is used to fit the data. Neither of these considerations should be seen as justification for using ill-fitting models. Finally, it is generally considered preferable to use models with fewer parameters, when possible.

e. Using Confidence Limits to Get a BMDL

Confidence limits express the uncertainty in a parameter estimate that is due to sampling and/or experimental error. The interval between two confidence limits is called a *confidence interval.* Confidence intervals can be two-sided, that is, localize their corresponding parameter on both sides, or one-sided, that is, localize their corresponding parameter on only one side. It may be convenient to think of a one-sided confidence interval as one limit of a two-sided interval goes to either infinity or minus infinity. For example, a one-sided 95% confidence interval for a parameter would share a limit with the two-sided 90% confidence interval for the parameter, and have plus or minus infinity (or, perhaps, 0, for a parameter such as the BMD that must be non-negative) as its

second limit. Confidence limits bracket those values which, within a particular model family, are consistent with the data, but do not account for or assume any correspondence between the modeled animal data and the human population of concern. The "confidence" or "coverage" associated with an interval indicates the percent of repeated intervals based on experiments of the same sort that are expected to include the parameter being estimated, for example, the BMD. With rare but important exceptions, calculated confidence intervals are approximations, in the sense that the actual coverage of the interval usually diverges somewhat from the desired. The choice of confidence level represents tradeoffs in data collection costs and the needed data precision. Just as 0.05 is a convenient (but not necessarily good for all data) level for tests, 95% is a convenient choice for most limits and is the default value recommended in this guidance.

A lower confidence limit is placed on the BMD to obtain a dose (BMDL) that assures with high confidence (*e.g.*, 95%) that the BMR is not exceeded. This process rewards better experimental design and procedures that provide more precise estimates of the BMD, resulting in tighter confidence intervals and thus higher BMDLs. Some procedures and examples for calculating BMDLs or BMCLs are given by Gaylor *et al.* (1998).

The method by which the confidence limit is obtained is typically related to the manner in which the BMD is estimated from the model. When parameters are estimated using the method of maximum-likelihood, confidence intervals (CIs) may be based on the asymptotic distribution of the likelihood ratio or on the asymptotic distribution of the maximum likelihood estimates (MLEs). While both can give problems in ranges where the assumptions needed to use asymptotic theory begin to weaken (e.g., as sample sizes decrease), in general it is preferred to base CIs for parameters estimated by maximum likelihood on the asymptotic distribution of the likelihood ratio, owing to their tendency to give better coverage behavior (Crump and Howe, 1985).

To compute a CI for a model parameter based on the distribution of the likelihood ratio, first compute the maximum likelihood estimate of all the parameters in the model. Next, separate the model parameters into one parameter whose CI is being computed (call it μ) and all the other parameters. Then find the value of μ such that, when the other parameters are adjusted to maximise the likelihood, the log-likelihood is reduced from that at the maximum likelihood estimates by exactly $\chi^2_{(1,1-\alpha)}/2$, where $\chi^2_{(1,1-\alpha)}$ represents the quantile of the χ^2 distribution corresponding to 1 degree

of freedom and an upper tail probability of α (see, for example, Crump and Howe, 1985; Venzon and Moolgavkar, 1988). When the value of interest cannot be expressed as a model parameter, a similar, but more complicated, approach is used.

Details for other approaches to CI computation specific to particular data types follow:

Quantal Data. For quantal data each individual is classified according to whether or not it exhibits a particular adverse effect, *e.g.*, death or cancer. Quantal data provide the simplest case for estimating BMDs. Consider an experiment consisting of animals exposed to several doses of a substance, and suppose that the number of animals exhibiting a particular adverse effect is binomially distributed at each dose level. After a suitable dose-response curve has been fit to the experimental data, the BMDL is defined as a lower confidence limit on the exposure level that corresponds to a specified excess risk (*e.g.*, 10%) above background. The exposure level itself is the effective dose, or the BMD_{10}. There are several ways to calculate a lower confidence limit. One is to apply standard statistical theory (specifically, the delta method, see for example Gart *et al.*, 1986) to approximate the variance of the estimated BMD. This estimated variance can then be used as the basis for constructing a lower confidence limit on the BMD. The logarithm of doses can be used to ensure a positive BMDL. A second approach is to calculate an upper confidence limit on the excess proportion (risk) of animals possessing an adverse effect as a function of dose. The BMDL is the dose where the upper confidence limit for the estimate of risk equals the specified level of risk, *e.g.*, 10%, desired for the BMD (see *e.g.*, Kimmel and Gaylor, 1988).

Clustered Data: Reproductive and Developmental Effects The issue of litter effects for reproductive and developmental experiments complicates the calculation of a confidence limit. The pregnant mother is the experimental unit and statistical methods must account for the tendency of littermates to respond similarly. Chen and Kodell (1989) and Williams (1975) have proposed methods based on the assumption that the number of affected individuals in a litter follows a beta-binomial distribution. The probability of an affected individual increases with dose of a toxic agent. To fit this model, maximum likelihood estimates can be obtained from the beta-binomial log likelihood (Chen and Kodell, 1989).

One disadvantage of the beta-binomial distribution and other correlated binomial distributions is their computational complexity. A second disadvantage is a lack of robustness if the

assumed distribution is incorrect. Alternative analyses can be based on quasi-likelihood, or more generally, generalized estimating equations. Liang and Zeger (1986) and Liang (1986) describe a general approach for the analysis of correlated data. This approach is referred to as Generalized Estimating Equations (GEE). Ryan (1992) discusses the use of this approach for developmental toxicity. The GEE approach requires specification of only the mean and variance functions of the data. To estimate dispersion parameters, a separate equation is required. A simple example is the moment estimates. An important addition in the GEE method is the inclusion of an empirical variance "fix-up" that relaxes the distributional assumptions so that the model parameters and their variances will be estimated correctly, even if the variance function is misspecified. There is still incentive to correctly specify the variance function since it improves statistical efficiency (Liang and Zeger, 1986).

Continuous Data Different techniques for calculating a BMD are required for continuous measurements. Examples of continuous endpoints are body weights, organ weights, and hematological and clinical chemistry measurements. For such data measured on a continuum, there generally is no sharp demarcation between normal and adverse values. In the absence of a clinical definition of an adverse level, a low or high percentile (*e.g.*, the first and 99th percentile) could be used to define an abnormal observation. For values that are normally distributed, these percentiles are estimated by the mean $\pm$ 2.33 standard deviations from the control animals. Such extreme values might be considered adverse or, at least, undesirable and can be classified as abnormal.

Crump (1995) shows the relationship between a change in the mean response, relative to the standard deviation, and the excess risk. For example, if values beyond the 98th to 99th percentile of control animals are considered abnormal, a dose that causes a shift in the average of one standard deviation results in approximately an excess risk of 10% of the animals in the abnormal range. This provides a very simple method for establishing a BMD associated with a risk of approximately 10%. A lower confidence limit on this BMD can be calculated using standard regression procedures.

Multiple Outcomes In addition to the clustering or litter effect, multivariate outcomes are often encountered. This is particularly true of developmental and reproductive toxicity data because exposure to agents can affect many different stages in the reproductive process. Once implantation has occurred, exposures to toxicants can result in early pregnancy loss, malformation, low fetal

weight, and/or subsequent developmental problems. The BMD can be based on the risk of being abnormal. Abnormality is defined as exhibiting any of several selected abberative endpoints. Several authors have discussed the development of dose-response models for multivariate data (Chen *et al*, 1991; Ryan *et al.*, 1991; Catalano and Ryan, 1991; Ryan, 1992b; Catalano *et al.*, 1993; Zhu *et al.*, 1994; Krewski and Zhu, 1994, 1995).

Thus, the BMDL is determined by 1) selecting an endpoint(s), 2) identifying a BMR (a predetermined level of change in response relative to controls), 3) establishing, by an appropriate estimation procedure, a model that fits the data adequately, and 4) calculating a confidence limit at the BMR using the model and the same estimation procedure.

At the time of this writing, commercial software is available that is designed specifically for carrying out steps 3) and 4) by maximum likelihood or GEE methods. EPA has software for this purpose (using maximum likelihood methods) that is widely-available to all potential users.

f. Selecting the model to use for POD computation

To summarize the preceeding sections, it is recommended that the following steps be followed to select the model(s) to use for computing the POD:

- Assess goodness-of-fit, using a value of α=0.1 to determine a critical value.
- Further reject models that apparently do not adequately describe the relevant low-dose portion of the dose-response, examining residuals and graphs of model and data.
- As the models remaining have met the default statistical criteria for adequacy and visually fit the data, any of them theoretically could be used for determining the BMDL. The remaining criteria for selecting the BMDL are necessarily somewhat arbitrary, and are adopted as defaults.
- If the BMDL estimates from the remaining models are within a factor of 3, then they are considered to show no appreciable model dependence and will be considered indistinguishable in the context of the precision of the methods. Models are ranked based on the values of their Akaike Information Criterion (AIC), a measure of the deviance of the model fit adjusted for the degrees of freedom, and the model with the lowest AIC is used to calculate the BMDL. If this is not unique, the simple average or geometric mean of the BMDLs with the lowest AIC is used.

- If the BMDL estimates from the remaining models are not within a factor of 3, some model dependence of the estimate is assumed. Since there is no clear remaining biological or statistical basis on which to choose among them, the lowest BMDL is selected as a reasonable conservative estimate. If the lowest BMDL from the available models appears to be an outlier, compared to the other results (e.g., there are several other results, all within a factor of 3), then additional analysis and discussion would be appropriate. Additional analysis might include the use of additional models, the examination of the parameter values for the models used, or an evaluation of the BMDs to determine if the same pattern exists as for the BMDLs. Discussion of the decision procedure should always be provided.
- In some cases, relevant data for a given agent are not amenable to modeling and a mixture of BMDLs and NOAEL/LOAELs results. When this occurs, and the most biologically relevant effect is from a study considered adequate but not amenable to modeling, the NOAEL should be used as the point of departure.

D. Reporting Requirements

Any computation of a BMD or BMDL should include the following elements:

- Study or Studies Selected for BMD Calculation(s)
  - Rationale for study selection
  - Rationale for endpoints (effects)
  - List dose response data used
- Dose-Response Model(s) Chosen for each Case
  - Rationale
  - Estimation procedure (e.g., maximum likelihood, least squares, generalized estimating equations)
  - Estimates of model parameters with standard errors
  - Goodness-of fit test statistics
  - Standardized residuals (observed minus predicted response/standard deviation)

- Choice of BMR for Each Case
  - Rationale
  - Procedure used if for continuous data
- Computation of the BMD
  - List the BMD Value.
- Calculation of the Lower Confidence Limit for the BMD (BMDL) for Each Case
  - Confidence limit procedure (e.g., likelihood profile, delta method, bootstrap)
  - List BMDL Value(s)
- Graphics for Each Case
  - Plot of data points with error (standard deviation) bars
  - Plot of fitted dose-response
  - Plot of confidence limits for the fitted curve (optional; if included, the narative should describe the methods used to compute them.)
  - Identify BMD and BMDL
- BMDs and BMDLs for Default BMRs
  - For dichotomous data, the BMD and BMDL for an extra risk of 0.10
  - For continuous data, the BMD and BMDL corresponding to a change in the mean response equal to one control standard deviation from the control mean.

E. Decision Tree

The following decision tree depicts the general progression of steps in a BMD calculation. A separate BMD calculation should be conducted for each endpoint/study combination that is a reasonable candidate for becoming the basis for a final quantitative risk estimate. Unlike comparing NOAELs or LOAELs across endpoints or studies, the relative values of potential BMDs are not readily transparent until after the modeling has been completed.

For each candidate endpoint/study combination:

1. Select the appropriate BMR based on the type of data (i.e., quantal vs. continuous), sensitivity of study design, toxicity endpoint, and judgements about the adversity of the

endpoint if continuous (see Section II.B).

2. Model the dose-response data, using appropriate model structures for the type of data (i.e., quantal vs. continuous, depending on how the BMR is defined) and study design (e.g., nested) (see Section II.C.2.a). For modeling cancer bioassay data, a specific default algorithm is generally used except for case-specific situations in which an alternate model may be superior (e.g. a time-to-tumor model, a biologically-based model). For other types of experimental animal data, curve-fitting can be attempted with any appropriate models. Human data are modeled in a case-specific way which may need to account for covariates, competing causes of mortality, etc.

3. Assess the fit of the models (see Section II.C.2.c). Retain models that are not rejected using a p-value of 0.1. Examine the residuals and plot the data and models; check that the models adequately describe the data, especially in the region of the BMR. (Sometimes it may be necessary to transform the data in some way or to drop the highest exposure group(s) (e.g., if the behavior at high exposures can be attributed to early mortality or enzyme saturation effects) and repeat the modeling in order to get a good fit.)

4. Calculate 95% lower confidence limits on the candidate BMDs (i.e., BMDLs) using the models which adequately fit the data (see Section II.C.2.e).

5. Select from among the models which adequately fit the data (see Section II.C.2.f). If the BMDL estimates from these remaining models are within a factor of 3 they are considered indistinguishable, and the model with the lowest AIC can be selected to provide the BMDL. If the BMDL estimates are not within a factor of 3, some model dependence is assumed, and the model with the lowest BMDL estimate should be selected unless it appears to be an outlier, in which case further analysis may be appropriate.

6. Document the BMD analysis as outlined in Section II.D. on reporting requirements.

REFERENCES

Akaike, H (1973) Information theory and an extension of the maximum likelihood principle. in *Proceedings of the Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki, eds. Akademiai Kiado, Budapest, pp. 267-281.

Alexeeff, G.V.; Lewis, D.C.; Ragle, N.L. (1993) Estimation of potential health effects from acute exposure to hydrogen fluoride using a 'benchmark dose' approach. Risk Analysis, 13(1):63-69.

Allen, B.C.; Kavlock, R.J.; Kimmel, C.A.; Faustman, E.M. (1994a) Dose-response assessment for developmental toxicity: II. Comparison of generic benchmark dose estimates with NOAELs. Fund. Appl. Toxicol., 23:487-495.

Allen, B.C.; Kavlock, R.J.; Kimmel, C.A.; Faustman, E.M. (1994b) Dose-response assessment for development toxicity: III. Statistical models. Fund. Appl. Toxicol., 23:496-509.

Allen, B.C., and P.L. Strong, C.J. Price, S.A. Hubbard, and G.P. Daston (1996) Benchmark dose analysis of developmental toxicity in rats exposed to boric acid. Fund. Appl. Toxicol. 32: 194-204.

Auton, T.R. Calculation of benchmark doses from teratology data (1994). Regulatory Toxicology and Pharmacology: 19: 152-167.

Barnes, D.G.; Daston, G.P.; Evans, J.S. Jarabek, A.M.; Kavlock, R.J.; Kimmel, C.A.; Park, C.; Spitzer, H.L. (1995) Benchmark dose workshop: Criteria for use of a benchmark dose to estimate a reference dose. Regulatory Toxicol. Pharmacol., 21:296-306.

Bates, D. M. and Watts, D. G. (1988) *Nonlinear Regression Analysis and Its Applications*. Wiley. New York.

Beck, B.D.; Conolly, R.B.; Dourson, M.L.; Guth, D.; Hattis, D.; Kimmel, C.; Lewis, S.C. (1993) Symposium overview: improvements in quantitative noncancer risk assessment. Fund. Appl. Toxicology 20:1-14.

California Office of Environmental Health Hazard Assessment. (1993) Safety assessment for non-cancer endpoints: The benchmark dose and other possible approaches. Summary report.

Catalano, P. J., and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Am. Stat. Assoc.* **87**, 651-658.

Catalano, P.J.; Scharfstein, D.O.; Ryan, L.M.; Kimmel, C.A.; Kimmel, G.L. (1993) Statistical model for fetal death fetal weight and malformation in developmental toxicity studies. Teratology 47:281-290.

Chen, C.; Farland, W. (1991) Incorporating cell proliferation in quantitative cancer risk assessment: approaches, issues, and uncertainties. In: Butterworth, B.; Slaga, T.; Farland, W.; McClain, M., eds. Chemical induced cell proliferation: implications for risk assessment. New York: Wiley-Liss, pp. 481–499.

Chen, J. J., and Kodell, R. L. (1989). Quantitative risk assessment for teratologic effects. J. *Am. Stat. Assoc.* **84**, 966-971.

Chen, J.J.; Kodell, R.L.; Howe, R.B.; Gaylor, D.W. (1991) Analysis of trinomial responses from reproductive and developmental toxicity experiments. Biometrics 47:1049-1058.

Clayton, D.; Hills, M. (1993) Statistical Models in Epidemiology. Oxford University Press, Oxford.

Cogliano, V.J. (1986) The U.S. EPA's methodology for adjusting the reportable quantities of potential carcinogens. Proceedings of the 7th National Conference on Management of Uncontrolled

Hazardous Wastes (Superfund '86). Washington: Hazardous Materials Control Research Institute, pp. 182–185.

Collins, M.A., G.M. Rusch, F. Sato, P.M. Hext, and R.-J. Millischer. 1995. 1,1,1,2-Tetrafluoroethane: Repeat exposure inhalation toxicity in the rat, developmental toxicity in the rabbit, and genotoxicity in vitro and in vivo. Fund. Appl. Toxicol. 25: 271-280.

Cox, D.R.; Hinkley, D.V. (1974) *Theoretical Statistics,* chapter 7, Chapman and Hall, London.

Crump, K.S. (1984) A new method for determining allowable daily intakes. Fundamental and Applied Toxicology 4:854-871.

Crump, K. S. (1995) Calculation of benchmark doses from continuous data. Risk Analysis 15: 79-89.

Crump, K. S.; R. Howe. (1985) Chapter 9. in *Toxicological Risk Assessment.* D.B. Clayson, D. Krewski, I. Munro, eds. Boca Raton: CRC Press, Inc.

Davidian, M.; Giltinan, D. M. (1995) Nonlinear Models for Repeated Measurement Data. Chapman and Hall, London. Farmer, J.H.; Kodell, R.L.; Gaylor, D.W. (1982) Estimation and extrapolation of tumor probabilities from a mouse bioassay with survival/sacrifice components. Risk Analysis 2(1):27–34.

Dourson, M.L., Hertzberg, R.C., Hartung, R., Blackburn, K. (1985) Novel methods for the estimation of acceptable daily intake. Toxicology and Industrial Health 1:23-41.

Draper, N.; Smith, H. (1981) Applied Regression Analysis, Second Edition. Chapter 10. Wiley, New York.

Farmer, J.H.; Kodell, R.L.; Gaylor, D.W. (1982) Estimation and extrapolation of tumor probabilities from a mouse bioassay with survival/sacrifice components. Risk Analysis 2(1):27–34.

Faustman, E.M.; Allen, B.C.; Kavlock, R.J.; Kimmel, C.A. (1994) Dose-response assessment for developmental toxicity: I. Characterization of data base and determination of NOAELs. Fund. Appl. Toxicol., 23:478-486.

Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions*. Second Edition. Wiley. New York.

Fowles, J. R., Alexeeff, G. V., Dodge, D. (1999) The use of benchmark dose methodology with acute inhalation lethality data. Regulatory Toxicology and Pharmacology 29: 262-278.

Fung, K. Y., Marro, L., and Krewski, D. (1998) A comparison of methods for estimating the benchmark dose based on overdispersed data from developmental toxicity studies. Risk Analysis 18: 329-342.

Gallant, A. R. (1987) *Nonlinear Statistical Models*. Wiley. New York.

Gart, J. J., Krewski, D., Lee, P. N., Tarone, R. E., and Wahrendorf, J. (1986). Statistical Methods in Cancer Research, Vol. 3. *The Design and Analysis of Long-Term Animal Experiments*. International Agency for Research on Cancer, Lyon, France.

Gaylor, D.W. (1983) The use of safety factors for controlling risk. Journal of Toxicology and Environmental Health 11:329-336.

Gaylor, D. W. (1996) Quantalization of continuous data for benchmark dose estimation. Regulatory Toxicology and Pharmacology 24: 246-250.

Gaylor, D.W.; Kodell, R.L. (1980) Linear interpolation algorithm for low dose risk assessment of toxic substances. J. Environ. Pathol. Toxicol. 4:305–312.

Gaylor, D.; Slikker, W., Jr. (1990) Risk assessment for neurotoxic effects. NeuroToxicology 11: 211-218.

Gaylor, D.W.; Kodell, R.L.; Chen, J.J.; Springer, J.A.; Lorentzen, R.J.; Scheuplein, R.J. (1994) Point estimates of cancer risk at low doses. Risk Analysis 14(5):843–850.

Gaylor, D. W., Ryan, L., Krewski, D., and Zhu, Y. (1998). Procedures for calculating benchmark doses for health risk assessment. *Regulatory Toxicol. Pharmacol.* **28**, 150-164.

Gerrity, T.R.; Henry, C.J., eds. (1990) Summary report of the workshops on principles of route-to-route extrapolation for risk assessment. In: Principles of route-to-route extrapolation for risk assessment, proceedings of the workshops; March and July; Hilton Head, SC and Durham, NC. New York, NY: Elsevier Science Publishing Co., Inc.; pp. 1-12.

Gold, L.S.; Sawyer, C.B.; Magaw, R.; Backman, G.M.; de Veciana, M.; Levinson, R.; Hooper, N.K.; Havender, W.R.; Bernstein, L.; Peto, R.; Pike, M.C.; Ames, B.N. (1984) A carcinogenic potency database of the standardized results of animal bioassays. Environ. Health Perspect. 58:9–319.

Gold, L.S.; Bernstein, L.; Kaldor, J.; Backman, G.; Hoel, D. (1986a) An empirical comparison of methods used to estimate carcinogenic potency in long-term animal bioassays: lifetable vs summary incidence data. Fund. Appl. Toxicol. 6:263–269.

Gold, L.S.; de Veciana, M.; Backman, G.M.; Magaw, R.; Lopipero, P.; Smith, M.; Blumenthal, M.; Levinson, R.; Bernstein, L.; Ames, B.N. (1986b) Chronological supplement to the carcinogenic potency database: standardized results of animal bioassays published through December 1982. Environ. Health Perspect. 67:161–200.

Gold, L.S.; Slone, T.H.; Backman, G.M.; Magaw, R.; Da Costa, M.; Lopipero, P.; Blumenthal, M.; Ames, B.N. (1987) Second chronological supplement to the carcinogenic potency database: standardized results of animal bioassays published through December 1984 and by the National Toxicology Program through May 1986. Environ. Health Perspect. 74:237–329.

Gold, L.S.; Slone, T.H.; Backman, G.M.; Eisenberg, S.; Da Costa, M.; Wong, M.; Manley, N.B.; Rohrbach, L.; Ames, B.N. (1990) Third chronological supplement to the carcinogenic potency database: standardized results of animal bioassays published through December 1986 and by the National Toxicology Program through June 1987. Environ. Health Perspect. 84:215–286.

Gold, L.S.; Manley, N.B.; Slone, T.H.; Garfinkel, G.B.; Rohrbach, L.; Ames, B.N. (1993) The fifth plot of the carcinogenic potency database: results of animal bioassays published in the general literature through 1988 and by the National Toxicology Program through 1989. Environ. Health Perspect. 100:65–168.

Gold, L.S.; Manley, N.B.; Slone, T.H.; Garfinkel, G.B.; Ames, B.N. Rohrbach, L.; Stern, B.R.; Chow, K. (1995) Sixth plot of the carcinogenic potency database: results of animal bioassays published in the general literature 1989–1990 and by the National Toxicology Program 1990 to 1993. Environ. Health Perspect.103 Suppl 8:3–122.

Guth, D. J. Carroll, R. J. Simpson, D. G. and Zhou, H. (1997) Categorical regression analysis of acute exposure to tetrachloroethylene. Risk Analysis 17: 321-332.

Haas, C.N. (1994) Dose-response analysis using spreadsheets. Risk Analysis 14(6):1097–1100.

Haber, L. T., Allen, B. C., and Kimmel, C. A. (1998) Non-cancer risk assessment for nickel compounds: Issues associated with dose-response modeling of inhalation and oral exposures. Toxicological Sciences 43: 213-229.

Hasselblad, V.; A.M. Jarabek (1995) Dose-response analysis of toxic chemicals. In: Bayesian biostatistics. D.A. Berry, D.K. Stangl, eds. Marcel Dekker, Inc. New York.

Heindel, J.J., C.J. Price, E.A. Field, M.C. Marr, C.B. Myers, R.E. Morrissey, and B.A. Schwetz (1992). Developmental toxicity of boric acid in mice and rats. Fund. Appl. Toxicol. 18:266-.

Hertzberg, R.C. (1989) Fitting a model to categorical response data with application to species extrapolation of toxicity. Health Physics 57:405-409.

Hertzberg, R.C., Miller, M. (1985) A statistical model for species extrapolation using categorical response data. Toxicology and Industrial Health 1:43-57.

Hext, P.M. and R.J. Parr-Dobrzanski, 1993. HFC 134a: 2 Year inhalation toxicity study in the rat. ICI Central Toxicology Laboratory, Alderley Park, Macclesfield, Cheshire, UK. Report No. CTL/P/3317.

Hoel, D.G. (1990) Assumptions of the HERP index. Risk Analysis 10(4):623–624.

Hoover, S.M.; Zeise, L.; Pease, W.S.; Lee, L.E.; Hennig, M.P.; Weiss, L.B.; Cranor, C. (1995) Improving the regulation of carcinogens by expediting cancer potency estimation. Risk Analysis 15(2):267–280.

Howe, R.B.; Crump, K.S.; Van Landingham, C. (1986) Global 86: a computer program to extrapolate quantal animal toxicity data to low doses. Prepared for U.S. EPA under contract 68–01–6826.

Jarabek, A.M., Hasselblad, V. (1992) Application of a Bayesian statistical approach to response analysis of noncancer toxic effects. Toxicologist 12:98.

Johnson B.L., J. Boyd, J.R. Burg, S.T. Lee, C. Xintaras, and B.E. Albright. 1983. Effects on the peripheral nervous system of workers' exposure to carbon disulfide. Neurotoxicology 4(1): 53-66.

Kavlock, R.J., B.C. Allen, C.A. Kimmel, E.M. Faustman. (1995) Dose-response assessment for developmental toxicity: IV. Benchmark doses for fetal weight changes. Fund. Appl. Toxicol., 26:211-222.

Kavlock, R.J.; Schmid, J.E.; Setzer, R.W., Jr. (1996) A simulation study of the influence of study design on the estimation of benchmark doses for developmental toxicity. Risk Analysis 16:391-403.

Kimmel, C.A., Gaylor, D.W. (1988) Issues in qualitative and quantitative risk analysis for developmental toxicity. Risk Analysis 8:15-20.

Kimmel, C.A.; Wellington, D.G.; Farland, W.; Rose, P.; Manson, J.M.; Chernoff, N.; Young, J.F.; Selevan, S.G.; Kaplan, N.; Chen, C.; Chitlik, L.D.; Siegel-Scott, C.L.; Valaoras, G.; Wells, S. (1989) Overview of a workshop on quantitative models for developmental toxicity risk assessment. Environmental Health Perspectives 79:209-2150.

Kimmel, C.A., M. Siegel, T.M. Crisp, and C.W. Chen (1996) Benchmark concentration (BMC) analysis of 1,3-butadiene (BD) reproductive and developmental effects. Fund. Appl. Toxicol. (Suppl., no. 1, part 2) 30: 146.

Kodell, R.L.; Chen, J.J.; Gaylor, D.W. (1995) Neurotoxicity Modeling for Risk Assessment. Regulatory Toxicology and Pharmacology 22:24-29.

Krewski, D. (1990) Measuring carcinogenic potency. Risk Analysis 10(4):615–617.

Krewski,D ,Brown,C ,and Murdoch,D. Determining "safe" levels of exposure: safety factors or mathematical models? Fund Appl Toxicol 4: S 383- S 394 (1984).

Krewski, D., and Zhu, Y. (1994). Applications of multinomial dose-response models in developmental toxicity risk assessment. *Risk Anal.* **14**, 613-627.

Krewski, D.; Zhu, Y. (1995) A simple data transformation for estimating benchmark doses in developmental toxicity experiments. Risk Analysis 15:29-39.

Krewski, D.; Szyszkowicz, M.; Rosenkranz, H. (1990) Quantitative factors in chemical carcinogenesis: variation in carcinogenic potency. Regul. Toxicol. Pharmacol. 12:13–29.

Krewski, D.; Gaylor, D.; Szyszkowicz, M. (1991) A model-free approach to low-dose extrapolation. Environ. Health Perspect. 90:279–285.

Kupper, L.L.; Hafner, K.B. (1989) How appropriate are popular sample size formulas? The American Statistician 43:101-105.

Lefkopoulou, M.; Moore, D.; Ryan, L. (1989) The analysis of multiple binary outcomes: Application to rodent teratology experiments. Journal of the American Statistical Association 84: 810-815.

Leisenring, W. and Ryan, L. (1992) Statistical properties at the NOAEL. *Regulatory Toxicology and Pharmacology* 15: 161-171.

Liang, K-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Linhart, H. and Zucchini, W. (1986). *Model Selection*. Wiley, New York.

McCullagh, P.; Nelder, J.A. (1989) Generalized Linear Models, Second Edition. Chapman and Hall, London.

Mantel,N and Bryan,WR. Safety testing of carcinogenic agents. J Natl Cancer Instit 27: 455-470 (1961).

Moolgavkar, S.H.; Knudson, A.G. (1981) Mutation and cancer: a model for human carcinogenesis. J. Natl. Cancer Inst. 66:1037–1052.

Murrell, J.A., Portier, C. J., and Morris, R. W. (1998) Characterizing dose-response: I: Critical assessment of the benchmark dose concept. Risk Analysis 18: 13-26.

National Research Council (NRC) (1983) Risk Assessment in the Federal Government: Managing the Process. Prepared by: Committee on the Institutional Means for Assessment of Risks to Public Health, Commission on Life Sciences. Washington, DC.

National Research Council (NRC) (1993) Issues in risk assessment. Washington: National Academy Press, pp. 115–116.

National Research Council (NRC) (1994) Science and Judgment in Risk Assessment, Committee on Risk Assessment of Hazardous Air Pollutants, Board on Environmental Studies and Toxicology, Commission on Life Sciences, National Academy Press, Washington, DC.

National Toxicology Program (NTP) (1991) Technical report on the toxicology and carcinogenesis of 1,3-butadiene (CAS No. 106-99-0) in B6C3F$_1$ mice (inhalation studies). U.S. Department of Health and Human Services, Publc Health Service, National Institutes of Health, National Toxicology Program. NTP TR434, NIH Publ. No. 92-3165.

Peto, R.; Pike, M.C.; Bernstein, L.; Gold, L.S.; Ames, B.N. (1984) The TD_{50}: a numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments. Environ. Health Perspect. 58:1–8.

Price and Berner, 1995. A benchmark dose for carbon disulfide: Analysis of nerve conduction velocity measurements from the NIOSH exposure database. Report to the Chemical Manufacturers Association Carbon Disulfide Panel.

Research Triangle Institute (RTI) (1994) Determination of the no-observable-adverse-effect-level (NOAEL) for developmental toxicity in Sprague-Dawley (CD) rats exposed to boric acid in feed on gestational days 0 to 20, and evaluation of postnatal recovery through postnatal day 21. RTI Identification Number 65C-5657-200.

Ross, G. J. S. (1990) *Nonlinear Estimation*. Springer-Verlag. New York.

Ryan, L. M., Catalano, P. J., Kimmel, C., and Kimmel, G. (1991). Relationship between fetal weight and malformation in developmental toxicity studies. *Teratology* **44**, 215-223.

Ryan, L. M. (1992a). The use of generalized estimating equations for risk assessment in developmental toxicity. *Risk Anal.* **12**, 439-447.

Ryan, L. 1992. Quantitative risk assessment for developmental toxicity. Biometrics 48:163-174.

Sawyer, C.; Peto, R.; Bernstein, L.; Pike, M.C. (1984) Calculation of carcinogenic potency from long-term animal carcinogenesis experiments. Biometrics 40:27–40.

Seber, G. A. F. and Wild, C. J. (1989) *Nonlinear Regression*. Wiley. New York.

Setzer, R.W.; Rogers, J.M. (1991) Assessing developmental hazard: the reliability of the A/D ratio. Teratology 44:653-665.

Simpson, D. G., Carroll, R. J., Zhou, H., Guth, D. J. (1996a) Interval censoring and marginal analysis in ordinal regression. J. Agr. Biol. Environ. Statistics 1: 354-376.

Simpson, D. G., Carroll, R. J., Zhou, H., Guth, D. J. (1996b) Weighted logistic regression and robust analysis of diverse toxicology data. Commun. In Statist.–Meth. 25: 2615–2632,

Stone, M. (1998) Akaike's Criteria. in *Encyclopedia of Biostatistics*, Armitage, P. and Colton, T., eds. Wiley, New York.

U.S. Environmental Protection Agency (EPA) (1986a) Guidelines for carcinogen risk assessment. Federal Register 51(185):33992–34003.

U.S. Environmental Protection Agency (EPA) (1986b) Science Advisory Board Comments

U.S. Environmental Protection Agency (EPA) (1987) Hazardous substances; reportable quantity adjustments; proposed rules. Federal Register 52(50):8140–8186.

U.S. Environmental Protection Agency (EPA) (1988a) Science Advisory Board Comments.

U.S. Environmental Protection Agency (EPA) (1988b) Science Advisory Board Comments.

U.S. Environmental Protection Agency (EPA) (1988c) Methodology for evaluating potential carcinogenicity in support of reportable quantity adjustments pursuant to CERCLA section 102. Washington: report no. EPA/600/8–89/053.

U.S. Environmental Protection Agency (EPA) (1989a) Reportable quantity adjustments; delisting of ammonium thiosulfate; final rules. Federal Register 54(155):33418–33484.

U.S. Environmental Protection Agency (EPA) (1989b) Technical background document to support rulemaking pursuant to CERCLA section 102, volume 3., Washington: Office of Solid Waste and Emergency Response.

U.S. Environmental Protection Agency (EPA) (1989c) Science Advisory Board Comments.

U.S. Environmental Protection Agency (EPA) (1991a) Guidelines for developmental toxicity risk assessment; notice. Fed Regist, 56:63798-63826.

US Environmental Protection Agency (EPA) (1991b) Regulatory impact analysis of proposed national primary drinking water regulation for lead and copper. Prepared by Wade Miller Associates, Inc. April.

U.S. Environmental Protection Agency (EPA) (1992) Draft report: a cross-species scaling factor for carcinogen risk assessment based on equivalence of $mg/kg^{3/4}/day$; notice. Federal Register 57(109):24152–24173.

U.S. Environmental Protection Agency (1994a) Ranking of pollutants with respect to hazard to human health; proposed rule. Federal Register 59.

U.S. Environmental Protection Agency (1994b) Technical background document to support rulemaking pursuant to the Clean Air Act—section 112(g): ranking of pollutants with respect to hazard to human health. Research Triangle Park, NC: report no. EPA–450/3–92–010.

U.S. Environmental Protection Agency (1994c) Methods for derivation of inhalation reference dose concentrations and application of inhalation dosimetry. Office of Health and Environmental Assessment, Environmental Criteria and Assessment Office, Research Triangle Park, MC. EPA/600/8-90/066F.

U.S. Environmental Protection Agency (1995a) Reportable quantity adjustments; final rule. Federal Register 60(112):30926–30962.

U.S. Environmental Protection Agency (1995b) Technical background document to support

rulemaking pursuant to CERCLA section 102, vol. 7. Washington: Office of Solid Waste and Emergency Response.

U.S. Environmental Protection Agency (1995c) The use of the benchmark dose approach in health risk assessment. Office of Research and Development, Washington, DC: EPA/630/R-94/007, February.

U.S. Environmental Protection Agency (1995d) Health assessment document for diesel emissions. Washington, EPA/600/8–90/057Bb.

U.S. Environmental Protection Agency (1995e) Benchmark dose concentration analysis for carbon disulfide. Internal Report.

U.S. Environmental Protection Agency (EPA) (1995f) Proposed guidelines for neurotoxicity risk assessment; notice. Fed Regist, 60:52032-52056

U.S. Environmental Protection Agency (EPA) (1995g) Manganese document

U.S. Environmental Protection Agency (1996a) Proposed guidelines for carcinogen risk assessment. Federal Register 61(79):17960-18011.

U.S. Environmental Protection Agency (1996b) Guidelines for Reproductive Toxicity Risk Assessment; notice. Federal Register (draft).

U.S. Environmental Protection Agency (EPA) (1996c) Integrated Risk Information System (IRIS). Online. National Center for Environmental Assessment, Washington, DC.

Van Ryzin, J. (1980) Quantitative risk assessment. J. Occup. Med. 22(5):321–326.

Venzon, D. J. and Moolgavkar, S. H. (1988) A method for computing profile-likelihood-based confidence intervals. Appl. Statist. 37: 87-94.

Wartenberg, D.; Gallo, M.A. (1990) The fallacy of ranking possible carcinogen hazards using the TD_{50}. Risk Analysis 10(4):609–613.

West, R. W. and Kodell, R. L. (1999). A comparison of methods of benchmark-dose estimation for continuous response data. Risk Analysis 19: 453 – 459.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**, 949-952.

Zeger, S. L.; Liang, K. Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42: 121-130.

Zhu, Y.; Krewski, D.; Ross, W.H. (1994) Dose-response models for correlated multinomial data from developmental toxicity studies. Applied Statistics 43:583-598.

EXAMPLES

1. Introduction

The following examples were selected to illustrate some important aspects of computing BMDs and BMDLs for single data sets and single endpoints. Of course, other decisions, not illustrated here with examples, need to be made before a POD is determined, in particular which endpoints and data sets to model, and how to select a POD from among several BMDLs.

2. Quantal Data: Selecting a Model

This example illustrates computing a benchmark dose for a simple quantal data set, using the dose-response models available in BMDS. The main point is to illustrate selecting a benchmark dose, given that the critical data set and benchmark response level have already been selected. In addition, it provides some background into why, in four commonly used models for quantal data, available in EPA's BMDS package (Weibull, log-logistic, log-probit, and gamma), a parameter ("power" or "slope") is often constrained to be no less than 1.0.

Consider the following dose-response data:

Dose	Number Affected	Fraction Affected	Number of Animals
0	1	0.02	50
21	15	0.31	49
60	20	0.44	45

We want to compute a benchmark dose and BMDL for an extra risk of 0.10 (as suggested by this document), using a one-sided 95% confidence interval. If we define the BMD to correspond to an extra risk of 0.10 (= *BMR*), then, if *P(BMD)* is the proportion of affected animals at the BMD,

and *P(0)* is the proportion in the control group, *BMR* is defined to be $BMR = \frac{P(BMD) - P(0)}{1 - P(0)}$. This can be rearranged to yield $P(BMD) = P(0) + [1 - P(0)]BMR$).Since we are looking for a BMR of 0.10, that will correspond to a response of 0.02 + (0.98 * 0.1) = 0.118. Notice that 31% of the tested animals were affected in the lowest non-control dose.

Thus the expected response at the BMD is substantially lower than the lowest observed response. We need to be aware that model choice will have some effect on the BMD calculation.

First, we fit a number of models to the data.

Results of fitting the models, sorted in order of increasing AIC [= -2 × (LL - p), where LL is the log-likelihood at the maximum likelihood estimates, and p is the degrees of freedom of the model; generally everything else being equal, lower AIC values are preferred]:

Model	χ^2	P-value	AIC	BMD	BMDL
log-logistic (slope ≥ 1)	0.93	0.34	136.907	7.21	4.93
log-probit (unconstrained)	0	NA[1]	137.995	2.75	NA
Weibull (unconstrained)	0	NA	137.995	1.71	NA
log-logistic (unconstrained)	0	NA	137.995	2.25	NA
gamma (unconstrained)	0	NA	137.995	1.33	~0
Multistage (degree=2)	2.27	0.13	138.17	9.29	6.92
gamma (power ≥ 1)	2.27	0.13	138.17	9.29	6.92
Weibull (power ≥1)	2.27	0.13	138.17	9.29	6.92
log-probit (slope ≥ 1)	6.05	0.0139	141.692	14.82	11.53
probit	7.83	0.0051	144.448	19.50	15.71
logistic	8.30	0.004	145.179	20.95	16.78

[1] Degrees of freedom are 0, since there are three dose groups and three estimated parameters.

Eight of the models have chi-squared values that exceed the recommended cutoff P-value of 0.1 (this includes four models with perfect fits, even though their P-values are undefined because

there are no degrees of freedom left to test the chi-square statistic). The model with the best AIC is the log-logistic model with slope parameter constrained to be no less than 1. For this model, the standardized residuals [*i.e.*,(observed value - expected value)/standard error] are all small:

Dose	Est._Prob.	Expected	Observed	Size	Residual
0.0000	0.0218	1.091	1	50	-0.0881
21.0000	0.2609	12.784	15	49	0.7208
60.0000	0.4917	22.125	20	45	-0.6335

and a visual examination seems OK, since the predicted curve comes well within the confidence limits for each data point:

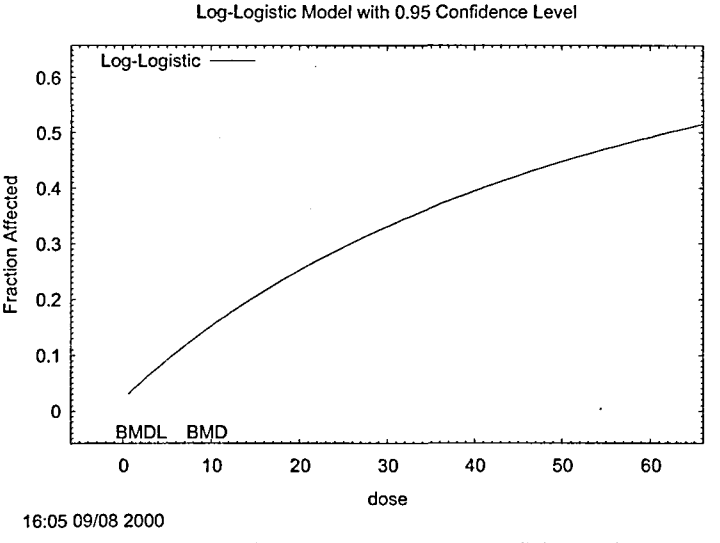

Figure A-2.1 Example data with 95% confidence limits, and constrained log-logistic model fit.

Four other models have only slightly greater AIC values and perfectly fit the data, the models with unconstrained slope or power parameters. Their AIC values are greater than that for the constrained log logistic only because an extra parameter counts against them: BMDS does not assign a model degree of freedom to parameters that end up on a constraint, so that model has only 2 degrees of freedom, while the models with unconstrained parameters have 3. The BMDs computed from the unconstrained models differ slightly among themselves, but are all quite a bit smaller than that computed from the constrained log-logistic, and, finally, there seems to be a problem with computing a BMDL for those models. Nevertheless, these models also describe

the data quite well, as the following graph of the unconstrained log-logistic model fit attests:

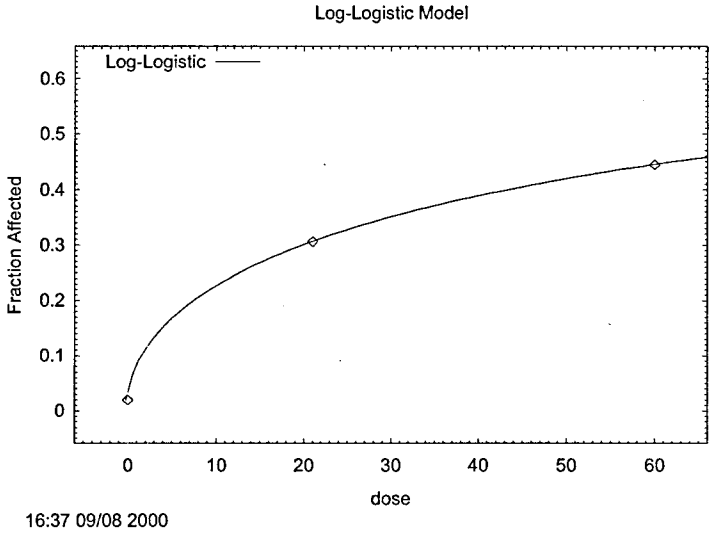

Figure A-2.2 Example data, 95% confidence limits, and unconstrained log-logistic model fit.

The main difference between the two log-logistic curves is in the region between the control and the lowest dose, where the unconstrained model curves upward more sharply than does the constrained model, which accounts for the lower BMDs from these models.

Finally, three models, the second-degree multistage, constrained Weibull, and constrained gamma, all give exactly the same fit and BMD prediction: in fact, for these data, they are really the same model. While the P-value for the fit is approaching the recommended cutoff, the AIC is only slightly worse than that for the unconstrained models. The predicted values and residuals are summarized in the table below:

Dose	Est._Prob.	Expected	Observed	Size	Scaled Residual
0.0000	0.0251	1.257	1	50	-0.232
21.0000	0.2318	11.356	15	49	1.234
60.0000	0.5064	22.787	20	45	-0.831

The fit at the lower two doses is a little worse than it was for the constrained log-logistic,

and this is apparent with close inspection of the graph, shown in Figure A-2.3.

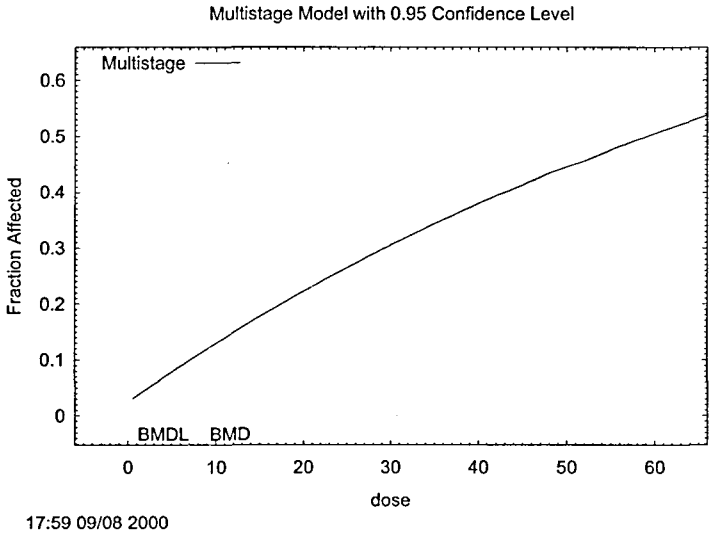

Figure A-2.3 Example data with 95% confidence limits, and second degree multistage model fit.

The primary question to be addressed here is, "Which model should I use to compute the BMD and BMDL". In this case, since the AIC of the constrained log-logistic model is slightly below those for the other models, the constrained log-logistic model can be considered preferable to them. However, the three lowest AIC values in the table above are so similar that it might be tempting to consider the models with the perfect fit to the data, even though this guidance recommends against using models such as the Weibull or log-logistic without constraining the power parameter or the slope parameter to be no less than 1.0. The rest of the narrative of this example is devoted to showing why allowing the slope parameter to be less than 1.0 might not be such a good idea.

The answer centers on the interpretation of BMDLs and how they are computed. When BMDLs are computed using the profile likelihood approach, this is particularly easy to visualize. In this case, conceptually at least, the BMD is treated as a parameter in the dose-response model, and, for each in a range of BMD values, the other parameters in the model are adjusted to maximize the log-likelihood while keeping the BMD constant at the selected value. The resulting curve, plotting the log-likelihood as a function of BMD value, is called a profile likelihood. This curve has a maximum at the BMD that corresponds to the maximum likelihood estimates for all the parameters, and drops off for values above and below that point. The BMDL

for a $(1 - \alpha) \times 100\%$ confidence interval is the BMD value where the log-likelihood is reduced from the maximum value by $(\chi^2_{1df, 2\alpha})/2$. Since we compute one-sided confidence intervals, we need only consider the shape of the curve below the maximum likelihood estimate for the BMD. The upper left hand panel of Figure A-2.4 shows this half of the profile likelihood for the BMD for the constrained log-logistic model fitted to the example data. The horizontal line indicates the critical value of the log-likelihood for determining a 95% confidence limit.

Since each BMD value on the x-axis of the figure has corresponding model parameter estimates, we can examine the plausibility of the dose-response curves we are claiming are consistent with the data. The upper right panel of Figure A-2.4 shows this for some BMD values including the maximum likelihood estimate (lowest dose-response curve) and the lower confidence limit (highest dose-response curve). Although clearly the range of curves does not exhaust the set of plausible dose-response curves one might consider for these data, they are certainly all plausible shapes. So, not only does the maximum likelihood fit of the model to the data represent a plausible dose-response shape, so do all the models between that and the model implied by the lower confidence bound on the BMD.

The story is different for the unconstrained log-logistic model, illustrated in the lower two panels of Figure A-2.4. First of all, the profile likelihood is substantially flatter for this model. It never even achieves the necessary drop in log-likelihood for there to be a lower 95% confidence limit, indicated by the horizontal line (the lowest, left-most point on the curve is the limiting value as BMD approaches 0). This explains why there is no BMDL for this model in the table: the confidence limit includes 0! The lower right panel shows the dose-response curves that correspond to the BMD values indicated by X's in the lower left panel. While the maximum likelihood fit may be a plausible fit to the data, the curves become increasing implausible as BMD drops, with the curve shooting up more and more rapidly from the control response as the BMD for the model is reduced. The log-likelihood is never reduced very much, because there is little evidence for trend in the responses at the two non-control doses. Indeed, at the limiting value for the BMD, 0, the curve is discontinuous: the control is fit perfectly, and the non-control responses are fit by a horizontal line, and the log-likelihood is not reduced sufficiently to reject this model as a plausible fit to the data! This situation often occurs when models such as the log-

logistic (also log-probit) are fit without constraining the slope parameter, and the Weibull, gamma, Hill, or power models are fit without constraining the power parameter. The implausibility of the curves that sometimes result when such models are fit to data is why this document recommends that such models not be used with unconstrained power or slope parameters, or only with great care.

In conclusion, for the reasons stated above, the log-logistic model, with the slope constrained to be greater than one, is selected as the preferred model for these data. This gives a BMD of 7.2 and BMDL of 4.9 for an extra risk of 10% for this dataset.

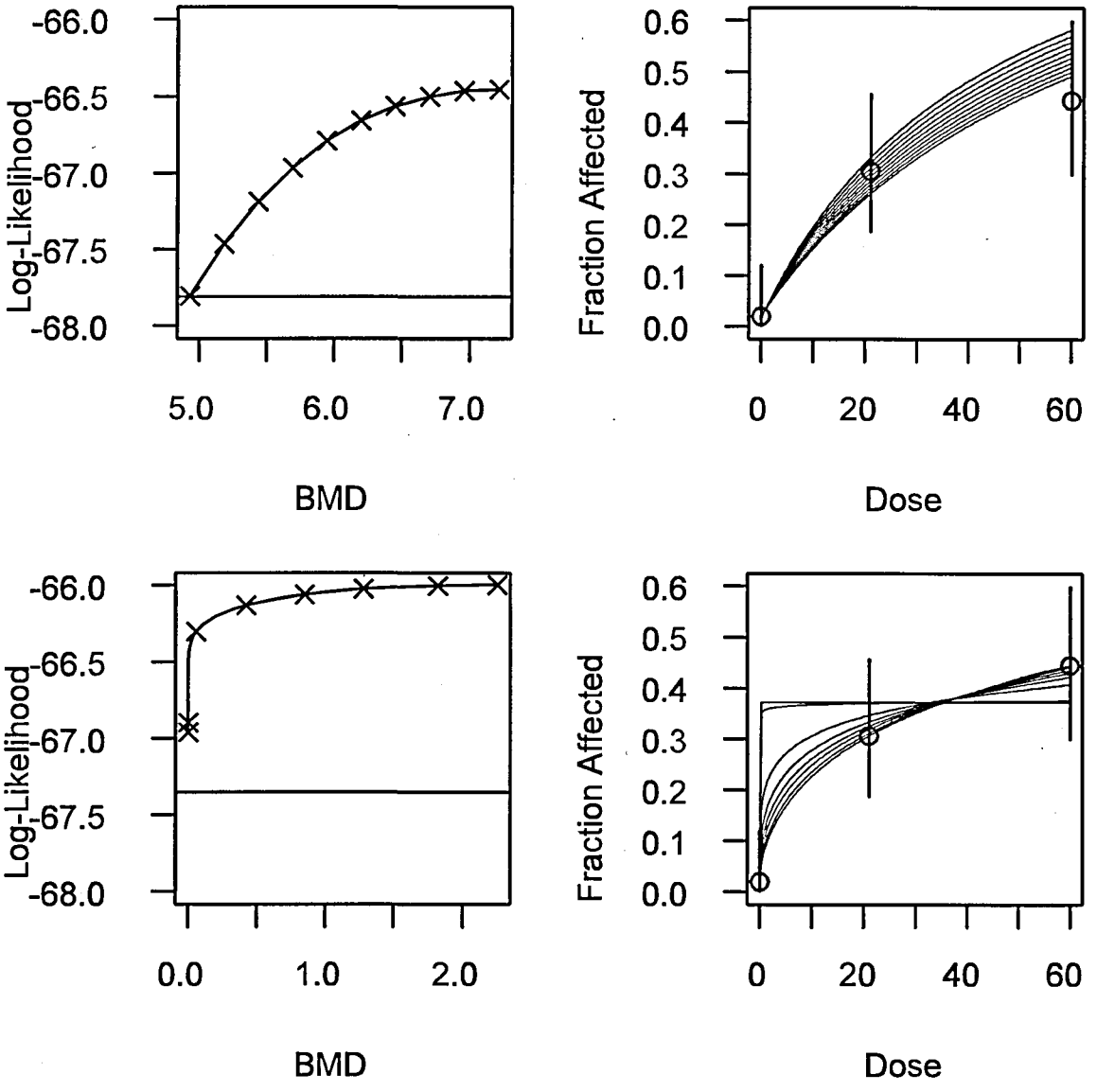

Figure A-2.4 Profile likelihoods (left) and corresponding dose-response curves (right) for log-logistic models fit to the example data set. The top two figures correspond to models with slope constrained to be no less than 1.0; the bottom two figures correspond to models with slope constrained to be positive. "Xs" on the profile likelihoods correspond t the plotted dose response curves. Vertical lines on the dose-response graphs indicate 95% confidence limits for the data means. The horizontal lines in the profile likelihood plots correspond to the likelihood value that defines the 95% confidence limit for the BMD.

3. Continuous Data: Getting a Good-Fitting Model

This example illustrates some of the care required when using non-linear modeling software and some of the data manipulation that may be required to get an adequate model fit for computing a BMD and BMDL. Several points are being made here: (1) convergence of a nonlinear model does not guarantee that maximum likelihood estimates have been achieved; sometimes some common sense and refitting is required to get MLEs; (2) even once maximum likelihood estimates have been achieved, the model may not fit well enough, and other actions may need to be taken to get a better fitting model; (3) one of the BMRs for this example is 5% of the dynamic range of the response (Murrell *et al.*, 1998, suggest that the fraction of the dynamic range of a continuous variable may often be a good quantification of the biological significance of the change); sometimes it may require some common sense and ingenuity to compute the BMD corresponding to such a BMR. NOTE: Some of the behavior of this example depends on the way the April 3, 2000 version of the Hill model from BMDS, selects its initial values. Other software, and even later versions of the Hill model from BMDS, may well behave differently on these data. This does not indicate "bugs" in the software, but rather that, for some datasets, there can be multiple "local maxima" for the likelihood function; software that uses purely local methods for optimization (as does BMDS) can get trapped at a local maximum, and may require experimenting with alternative initial parameter values to assure convergence to a true global maximum of the likelihood function. Software packages differ in the algorithm used to select the starting parameter values for optimization, so may end up in different local maxima.

For this example, consider the following data set:

Dose	subject/group	Mean	Std. Dev.
0.	8	100.	30.4
0.3	8	98.24	49.8
1.	8	111.34	59.9
3.	8	172.16	58.4
10.	8	357.48	167.5

Dose	subject/group	Mean	Std. Dev.
30.	8	1695.03	260.9
100.	8	1576.11	169.7
300.	8	1896.22	141.7

The data represent a biochemical response in rats after dosing. For this example, we will compute a BMD as the dose where the response has increased over background by an amount of 5% of the range between the background and the maximum response, per the suggestion of Murrell, *et al.* (1998), as well as the dose where the mean has been displaced by one control standard deviation, as this document suggests. As can be seen from Figure A-3.1, the dose-response is clearly sigmoidal.

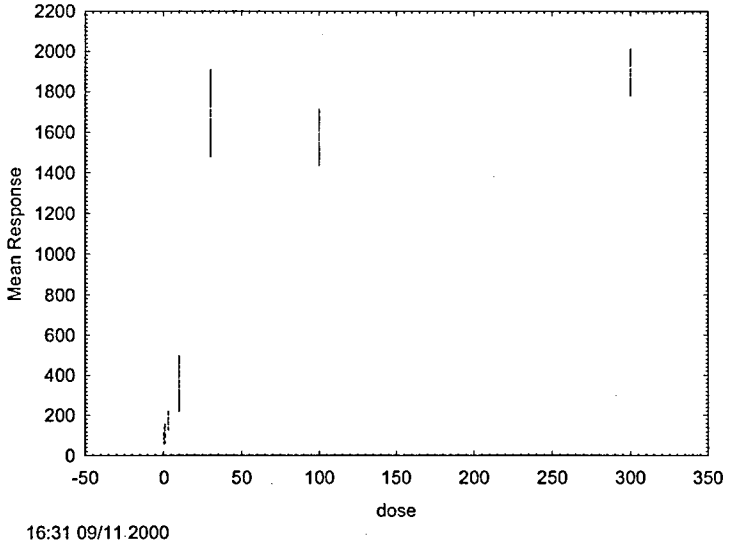

Figure A-3.1: Mean and 95% confidence intervals for example data.

It is natural in such data to fit a flexible model that allows a sigmoidal response; the Hill model is one such model, available in BMDS. Since it is usual in biochemical data for the variance to be proportional to the square of the mean (approximately), and since it looks as if the variance is larger in this data set for larger means, in general, for this example, we fit a model to the data in which the variance is modeled as being proportional to the power of the mean. That is, our model is:

$$\mu(d) = \lambda + \frac{V d^n}{k^n + d^n}$$
$$\sigma^2(d) = \alpha\left(\mu(d)\right)^\rho$$

where d represents dose, $\mu(d)$ represents the mean response, and $\sigma^2(d)$ represents the variance of the observations at dose d. Rough estimates of this model can be read off the graph of the data, and this provides a useful check of the fitting algorithm. When we fit a Hill model to the example data, we would expect λ (intercept) to be around 100, since that is about the background level of the response, V should be around 1600, since that is about the increment at the highest doses over the background level. k represents the dose where half the response has occurred, and should be in the range of $10 - 30$. Furthermore, based on experience, n should be relatively small, say between 1 and 10, and ρ ought to fall between 1 and 2, or so, since it is common for variances to be proportional to the square of means in such data.

If that model is fit to these data using the April 3, 2000 version of the Hill model from BMDS (the current version as of this writing), the fitting algorithm apparently converges on a solution. The parameter estimates from this solution are:

Variable	Estimate	Std. Err.
alpha	4381.57	2211.67
rho	0.266572	0.0668979
intercept	105.045	22.8759
v	1634.05	51.087
n	4.76591	1.62145
k	14.256	1.80324

Note that all the estimates are in their expected ranges except for the estimate of ρ (rho), which is 0.27, though we said we would have expected a value in the range 1-2.

The resulting predicted values are:

Dose	N	Obs Mean	Obs Std Dev	Est Mean	Est Std Dev	Chi^2 Res.
0	8	100	30.4	105	123	-0.115
0.3	8	98.2	49.8	105	123	-0.156
1	8	111	59.9	105	123	0.138
3	8	172	58.4	106	123	1.518
10	8	357	168	360	145	-0.059
30	8	1.7e+003	261	1.69e+003	178	0.028
100	8	1.58e+003	170	1.74e+003	179	-2.570
300	8	1.9e+003	142	1.74e+003	179	2.483

While the model predicts the mean values and the standard deviations at the higher doses pretty well, the standard deviations at the lower doses are overestimated by factors of 2 to 4. For future reference, the log-likelihood for this model fit is -345.786.

This may be the best this model can do, but it looks suspiciously like the fitting algorithm got caught in a local maximum of the likelihood surface, and that, perhaps, if we could get better initial values for some of the parameters, we could get a better set of estimates. Since the model for the mean seems to describe the data pretty well, we will restart the model, selecting the old estimates as initial values for the parameters of the model for the mean, and get new starting values for estimating the variance function parameters. These new estimates will come from regressing the log of the observed variance (that is, the square of the standard deviation), on the log of the observed mean (that is, $\log(var) = \log(\alpha) + \rho \log(mean)$). The parameter estimates from this regression are: ρ=1.0, log(α)=3.166, so the estimate of α is $e^{3.166} = 23.7$. Starting from these new values, the final estimates are:

Variable	Estimate	Std. Err.
alpha	24.8892	24.5755
rho	1.04671	0.162142
intercept	117.097	10.798
v	1629.2	64.9209
n	4.18855	1.33386
k	14.8385	1.86453

and the new predicted values:

Dose	N	Obs Mean	Obs Std Dev	Est Mean	Est Std Dev	Chi^2 Res.
0	8	100	30.4	117	60.3	-0.797
0.3	8	98.2	49.8	117	60.3	-0.882
1	8	111	59.9	117	60.3	-0.281
3	8	172	58.4	119	60.9	2.462
10	8	357	168	379	112	-0.556
30	8	1.7e+003	261	1.67e+003	242	0.351
100	8	1.58e+003	170	1.75e+003	248	-1.939
300	8	1.9e+003	142	1.75e+003	248	1.711

BMD = 7.3467

BMDL = 5.96733

The log-likelihood for this fit is -333.127, a substantial improvement over the previous fit. Furthermore, now not only do the estimated means accord with those observed, but the estimated standard deviations are a lot closer to those observed. Most likely, the current estimates are really the maximum likelihood estimates for this model and this dataset.

However, even though the fit is improved, neither the variance model (see the result of Test 3, below) nor the model for the mean (result of Test 4, below) fits the data, as the following excerpt from BMDS output for this example illustrates:

Likelihoods of Interest

Model	Log(likelihood)	DF	AIC
A1	-343.706	9	705.412
A2	-317.77	16	667.539
A3	-324.533	10	669.065
fitted	-333.127	6	678.253
R	-458.043	2	920.086

Explanation of Tests

```
Test 1:  Does response and/or variances differ among Dose levels?
         (A2 vs. R)
Test 2:  Are Variances Homogeneous? (A1 vs A2)
Test 3:  Are variances adequately modeled? (A2 vs. A3)
Test 4:  Does the Model for the Mean Fit? (A3 vs. fitted)

                    Tests of Interest

  Test    -2*log(Likelihood Ratio)  Test df        p-value

  Test 1              280.547          14          <.0001
  Test 2              51.8732           7          <.0001
  Test 3              13.5263           6           0.0354
  Test 4              17.1876           4        0.001777
```

What is going on? The table of fitted values, above (particularly the column labeled "chi^2 residuals") shows that the current model seriously underpredicts the response at a dose of 3, and misses the response at the two highest doses on either side. Furthermore, the model *over predicts* the standard deviation at the two highest doses (which is probably why the model for the variance is rejected). It is the under prediction at the lower doses that is most important, however, because that is in the region of the BMD, as far as this model can tell.

The three highest doses, at 30, 100 and 300, are quite far from the BMD; if we drop those doses, we will be eliminating doses whose responses the model cannot account for very well, and, since they are far from the BMD, we should not be eliminating much information about the actual location of the BMD. Furthermore, since the responses on the plateau have all been dropped, other monotonic dose-response models can be fit to the data. We consider three here: the Hill, a first degree polynomial (adding higher degree terms to the model did not add significantly to the ability of the model's ability to fit the data; the model used is $\mu(d) = \beta_0 + \beta_1 d$), and the power model ($\mu(d) = \beta_0 + \beta_1 d^{\gamma}$).

However, one of the BMRs we want to calculate is based on a change in the mean response equal to 5% of the range of the response (that is 5% of the maximum value minus the minimum value). In the Hill model, the BMD and BMDL corresponding to this change can be computed directly by the software in BMDS, but this is not so for the other models (since those

models to not allow for a horizontal asymptote). Furthermore, since this reduced data set really contains no information about the maximum response, even the Hill model's estimate of that is suspect (the estimate of the maximum value from the model reported in the above table is ridiculously large: 143289; with a huge standard error: 5.8×10^8, so it is clearly not useful for setting a BMR). The way around this is to calculate 5% of the observed dynamic range for this endpoint, and look for the dose that would result in an absolute change of this amount. The minimum value, based on the variance-weighted mean of the lower two dose groups, is 99.51, and the maximum value, based on the variance-weighted mean of the upper three dose groups, is 1758.3; 5% of the difference of the two is 82.9.

			5% Dyn. Range		1 SD Change	
Model	GOF P-value	AIC	BMD	BMDL	BMD	BMDL
polynomial	0.98	375.46	3.23	2.46	1.46	1.11
power	0.95	377.35	3.46	2.47	1.66	1.11
Hill	0.76	379.35	3.46	2.47	1.70	1.14

All three models fit the data well, according to both the summary results reported here and a more detailed examination of the graphs and residuals (not shown here), but the AIC for the polynomial model is somewhat better than that for the other two, so that is the model to choose to calculate the BMD and BMDL. That is, the BMD and BMDL based on 5% of the dynamic range of the response are 3.23 and 2.46; based on a one standard deviation change, 1.46 and 1.11.

This example illustrates three points, none of which is specific to modeling continuous data: (1) it is important to exercise some judgment when fitting models to data; no software package can guarantee that the parameters returned are actually maximum likelihood estimates, and the analyst may have to do some "tweaking" to get an acceptable answer; (2) we want models that describe the data well in the region of the BMR/BMD, which may involve some judicious narrowing of the dose range we attempt to model; (3) it may be necessary to exercise some creativity to compute BMDs for the BMR we want, and what scientific and risk analytic judgment dictate as desirable answer should not be subservient to what the software can do.

4. Cancer Bioassay Data: Modeling POD for Cancer Slope Factor

For cancer response modeling from standard cancer bioassay data, U.S. EPA is developing a specific algorithm which will be included in the BMDS package. The algorithm uses a multistage (polynomial) model with some constraints. As this model is under development at the time of this writing, the standard BMDS version of the multistage model will be used for the purposes of this example. Under EPA's proposed 1996 Guidelines for Carcinogen Risk Assessment, quantitative risk estimates from cancer bioassay data are typically calculated by modeling the data in the observed range to estimate a BMDL for a BMR of 10% extra risk, which is generally at the low end of the observable range for standard cancer bioassay data. This BMDL then serves as the "point of departure" for linear extrapolation or a nonlinear quantitative approach, as warranted by the mode of action of the carcinogen.

This example uses the dose-response data presented in EPA's 1988 Health and Environmental Effects Document for Dibromochloromethane for the quantitative estimate of carcinogenic risk from oral exposure. The rationale for study selection and endpoint selection, while important components of any comprehensive write-up of a BMD calculation, are beyond the scope of this quantitative example.

BMD Modeling for Dibromochloromethane

tumor type: hepatocellular adenoma or carcinoma
test animal: B6C3F1 mouse, female
route of exposure: gavage
study: NTP, 1985

DOSE-RESPONSE DATA

administered dose (mg/kg/day)	human equivalent dose (mg/kg/day)	tumor incidence
0	0	6/50
50	2.83	10/49
100	5.67	19/50

As discussed above, the multistage model was used because it is considered the default model for cancer bioassay data; although, in the future there will be a specific algorithm for modeling such cancer data. Similarly, a BMR of 10% extra risk was used, as is typical for standard cancer bioassay data.

BMR: 10%
model: multistage, extra risk

First, a second-degree (i.e., n-1) multistage model is fit to the data.

model form: background + (1-background) * [1-EXP(-beta1*dose^1-beta2*dose^2)]

parameter	estimate (MLEs)	std.error
background	0.12	0.132665
beta (1)	0.00930036	0.141898
beta (2)	0.00925286	0.0246904

AIC = 158.688
p-value = 1
$Chi^2 = 0$
residuals = 0

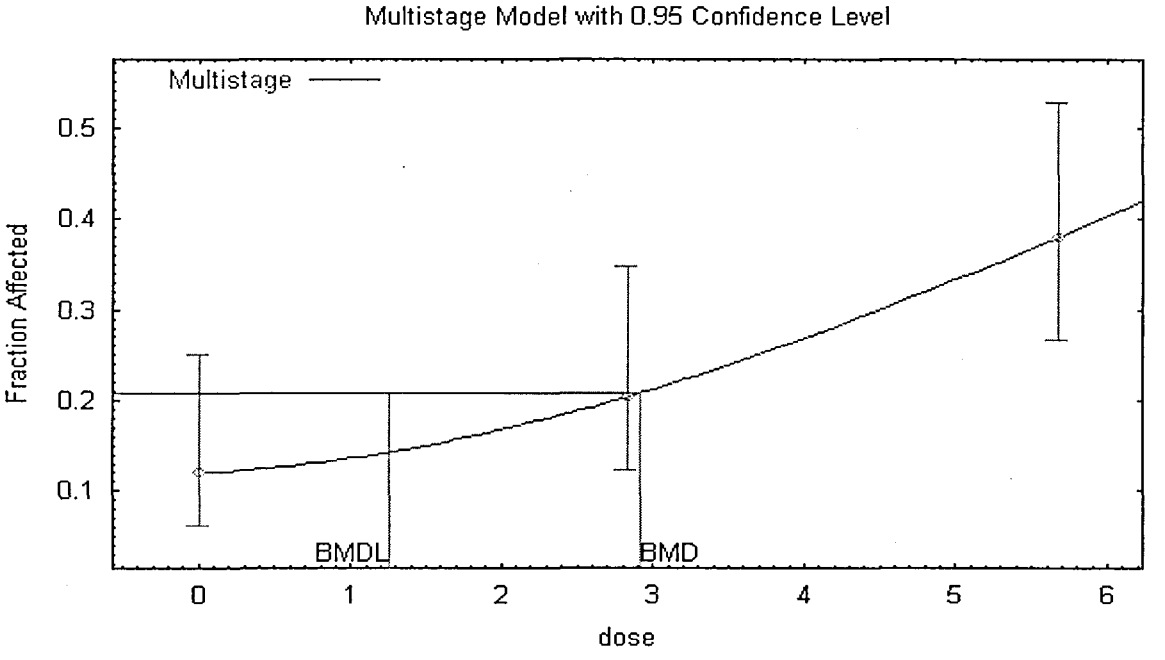

Figure A-4.1 Second-degree multistage model.

BMD (ED_{10}) = 2.91 mg/kg/day

BMDL (LED_{10}; 95% confidence limit estimated by likelihood profile) = 1.25 mg/kg/day

The second-degree model provides a good fit. Next, a first-degree multistage model is fit to the data to see if a more parsimonious model can also provide an adequate fit.

model form: background + (1-background) * [1-EXP(-beta1*dose^1)]

parameter	estimate (MLEs)	std.error
background	0.111488	0.120556
beta (1)	0.0559807	0.0391492

AIC = 157.272

p-value = 0.4446

Goodness of Fit:

Dose	Est. Prob.	Expected	Observed	Size	Chi^2 Residuals
0.0000	0.1115	5.574	6	50	0.086
2.8300	0.2417	11.842	10	49	-0.205
5.6700	0.3531	17.657	19	50	0.118

Chi-square = 0.57 DF = 1 P-value = 0.4494

BMD (ED_{10}) = 1.88 mg/kg/day

BMDL (LED_{10}; 95% confidence limit estimated by likelihood profile) = 1.20 mg/kg/day

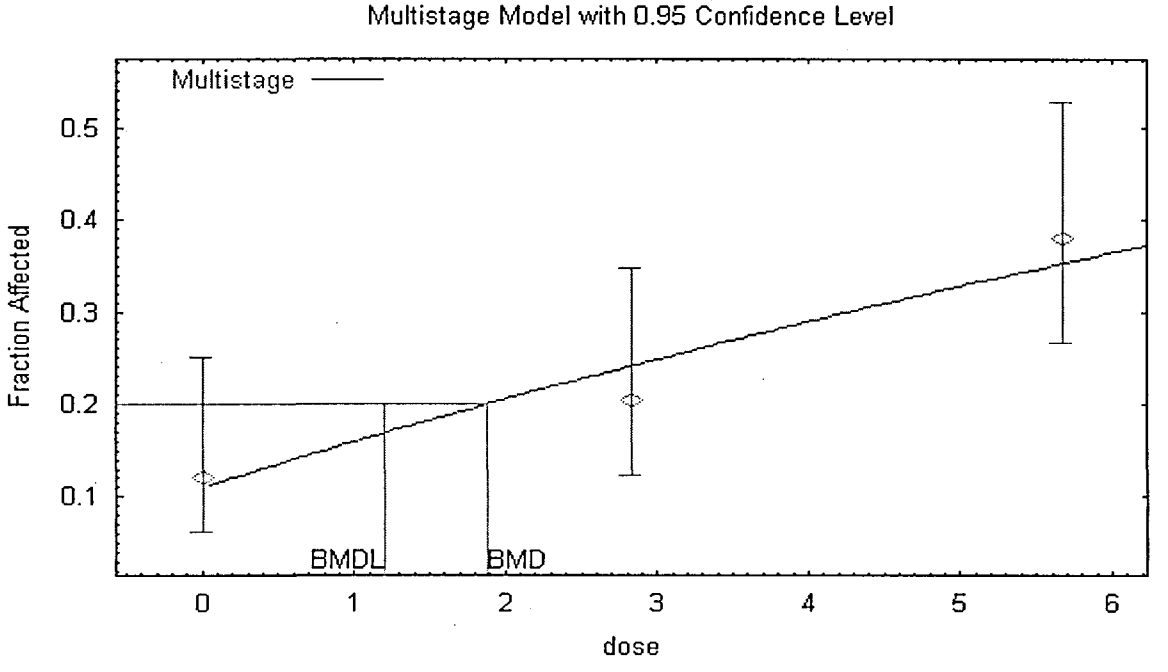

Figure A-4.2 First-degree multistage model.

The AIC is lower for the first-degree model suggesting that this is the preferred model; however, because the multistage model is really a family of k-degree models, a likelihood ratio test can be used to evaluate whether the improvement in fit afforded by estimating additional parameters is justified. In this case, the log likelihood for the second-degree model was -76.3439 and for the first-degree model was -76.6361. Thus twice the absolute difference in the log likelihoods is less than 3.84, i.e., a Chi-square with one degree of freedom (i.e., 2-1), suggesting that the first-degree multistage model is not significantly different from the second-degree model. Under the recommendations of the benchmark dose guidance, the more parsimonious first-degree model would be generally preferred. Final judgement on this may be subject to endpoint-specific guidance.

References

NTP (National Toxicology Program). 1985. Toxicology and carcinogenesis studies of chlorodibromomethane (CAS No. 124-48-1) in F344/N rats and B6C3F1 mice (gavage studies). NTP Tech. Report Series No. 282. NTIS PB 86-166675.

U.S. EPA. 1988. Health and Environmental Effects Document for Dibromochloromethane. Prepared by the Office of Health and Environmental Assessment, Environmental Criteria and Assessment Office, Cincinnati, OH. ECAO-CIN-GO40.

5. Developmental Toxicity Example

In general, data from developmental toxicity studies in rodents are best modeled using nested models. These models account for any intralitter correlation, or the tendency of littermates to respond similarly to one another relative to the other litters in a dose group. If this correlation (which may vary with dose) is not estimated, variance estimates, and hence the confidence limits on benchmark responses and doses, will generally be misspecified.

This example uses dose-response data reported by George et al. (1992), regarding the developmental toxicity of ethylene glycol diethyl ether administered orally to mice. As with the other examples in this guidance, this example illustrates fitting a model to one dose-response pattern. Note that the rationale for study selection and endpoint selection, while important components of any comprehensive BMD calculation write-up, are beyond the scope of this quantitative example.

The outcome modeled was prevalence of malformations, a quantal endpoint. The nested logistic model was considered for the purpose of illustrating fitting these quantal, nested data. Elements of the analysis addressing the reporting requirements in Section II.D. are documented in Table A-4.1, including a brief description of the experiment. The model input and model output data are summarized in Table A-4.2.

The nested logistic model demonstrated a reasonably good visual fit to the mean responses of the dose groups (not shown), but the goodness of fit p-value was 0.061, less than the value of 0.10 recommended in Section II.E. Since the coefficients which gauge the influence of litter size in predicting the response rate were fairly close to zero (0.0013 and -0.1507, respectively, not shown), suggesting that litter size was not important in this case, the model was re-fit without litter size. The resulting fit yielded a p-value of 0.184 ⑥, adequate for supporting BMD evaluation. Its AIC (at 450.6) was also slightly lower than the first fit (at 452.5).

Another variation on this model was also fitted, setting the intralitter correlations (the coefficients phi1 – phi5) to zero. This fit was not successful, with a goodness of fit p-value of 0 and an AIC of 570.4 (compare to 450.6, above). The intralitter correlations are therefore important for describing the observed variability in this data set.

The fitted model and the mean responses by dose group are shown in Figure A-5.1. The

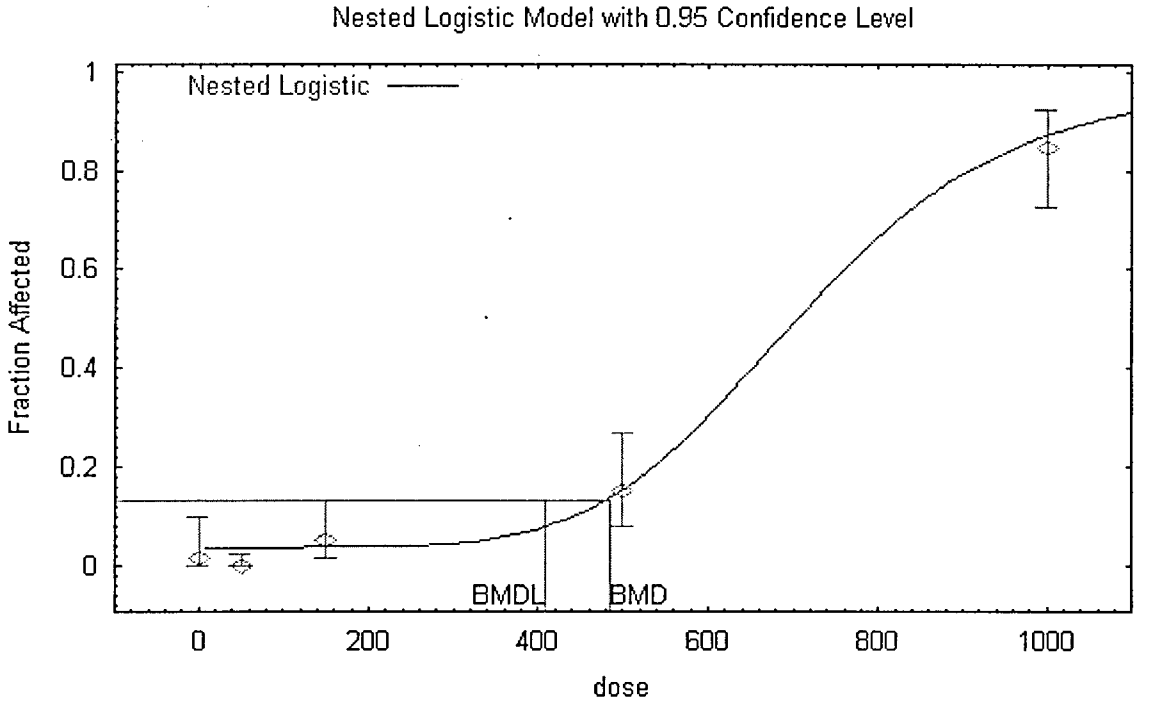

Figure A-5.1 Developmental Example Model Fit.

results for the selected nested logistic fit (the second fit described above) are provided in Table A-5.2.

Reference

George JD, Price CJ, Marr MC, Kimmel CA, Schwetz BA, Morrissey RE. (1992). The Developmental Toxicity of Ethylene Glycol Diethyl Ether in Mice and Rabbits. *Fund. App. Tox.* 19:15-25.

Table A-5.1: Summary of benchmark dose estimate, and key to Table 4-2		
Study	Title/Identifier	Ethylene glycol diethyl ether, administered to mice by gavage (in mg/kg/day), days 6-15 of gestation (George et al., 1992)
	Rationale for study selection	Selected by developmental toxicologist as an adequate study
	Rationale for endpoints (effects)	Skeletal malformations - Developmental toxicologist selected as an important endpoint
	List dose response data used	See ④ in Table 4-2
Dose - Response Model	Form	Nested logistic model in BMDS package, see ① in Table 4-1
	Rationale	Fits a wide variety of dose-response shapes for nested data
	Estimation procedure	Maximum likelihood
	Estimates of model parameters with standard errors	See ②
	Goodness-of fit test statistics	See ③, ⑤, ⑥
	Standardized residuals	See ④, ⑤
Choice of BMR	Rationale	Quantal data, used default 10% extra risk level.
Benchmark Dose	Lower Confidence Limit Procedure	Likelihood profile
	BMD	485 mg/kg/day (⑦)
	BMDL	410 mg/kg/day (⑧)
Graphics	Data points	See mean response rates and confidence limits in Figure 4-1
	Fitted dose-response model	See Figure 4-1
	Confidence limits for fitted curve	Not provided

Table A-5.2: Output from Model Run (EPA BMDS NLogistic Model. Revision: 2.6, Date: 2000/03/03)

```
The probability function is:

Prob. = alpha + theta1*Rij + [1 - alpha - theta1*Rij]/ 1+exp(-beta-theta2*Rij-rho*log(Dose))], ①

        where Rij is the litter specific covariate.

Restrict Power rho >= 1.


Total number of observations = 105
Total number of records with missing values = 0
Total number of parameters in model = 10
Total number of specified parameters = 2


Maximum number of iterations = 250
Relative Function Convergence has been set to: 1e-008
Parameter Convergence has been set to: 1e-008

User specifies the following parameters:
        theta1 =          0
        theta2 =          0


                Default Initial Parameter Values
                       alpha =    0.0370596
                        beta =     -36.6368
                      theta1 =            0   Specified
                      theta2 =            0   Specified
                         rho =      5.56873
                        phi1 =      0.64095
                        phi2 =     0.999996
                        phi3 =     0.159461
                        phi4 =     0.284719
                        phi5 =     0.231641


                        Parameter Estimates                              ②

     Variable          Estimate          Std. Err.
        alpha         0.0370596          0.0142364
         beta          -36.6368           0.289861
          rho           5.56873            242.744
         phi1           0.64095           0.107174
         phi2          0.999996            0.13603
         phi3          0.159461           0.130185
         phi4          0.284719                  0
         phi5          0.231641                  0


        AIC:          450.56                                             ③


                            Litter Data                                  ④

          Lit.-Spec.              Litter                        chi-squared
  Dose       Cov.     Est._Prob.   Size    Expected   Observed   Residual
-------------------------------------------------------------------------
0.0000     6.0000       0.037        6       0.222          0     -0.2343
0.0000     8.0000       0.037        8       0.296          0     -0.2369
0.0000     8.0000       0.037        8       0.296          0     -0.2369
0.0000     9.0000       0.037        9       0.334          0     -0.2378
0.0000     9.0000       0.037        9       0.334          0     -0.2378
0.0000    10.0000       0.037       10       0.371          0     -0.2385
0.0000    10.0000       0.037       10       0.371          0     -0.2385
0.0000    11.0000       0.037       11       0.408          0     -0.2390
0.0000    11.0000       0.037       11       0.408          0     -0.2390
```

Table A-5.2: Output from Model Run (EPA BMDS NLogistic Model. Revision: 2.6, Date: 2000/03/03)

1	0.0000	11.0000	0.037	11	0.408	0	-0.2390
2	0.0000	11.0000	0.037	11	0.408	0	-0.2390
3	0.0000	11.0000	0.037	11	0.408	0	-0.2390
4	0.0000	11.0000	0.037	11	0.408	0	-0.2390
5	0.0000	11.0000	0.037	11	0.408	0	-0.2390
6	0.0000	11.0000	0.037	11	0.408	0	-0.2390
7	0.0000	12.0000	0.037	12	0.445	0	-0.2395
8	0.0000	14.0000	0.037	14	0.519	0	-0.2403
9	0.0000	14.0000	0.037	14	0.519	0	-0.2403
10	0.0000	14.0000	0.037	14	0.519	4	1.6122
11	0.0000	15.0000	0.037	15	0.556	0	-0.2406
12	0.0000	15.0000	0.037	15	0.556	0	-0.2406
13	0.0000	15.0000	0.037	15	0.556	0	-0.2406
14							
15	50.0000	2.0000	0.037	2	0.074	0	-0.1962
16	50.0000	5.0000	0.037	5	0.185	0	-0.1962
17	50.0000	9.0000	0.037	9	0.334	0	-0.1962
18	50.0000	9.0000	0.037	9	0.334	0	-0.1962
19	50.0000	9.0000	0.037	9	0.334	0	-0.1962
20	50.0000	10.0000	0.037	10	0.371	0	-0.1962
21	50.0000	10.0000	0.037	10	0.371	0	-0.1962
22	50.0000	11.0000	0.037	11	0.408	0	-0.1962
23	50.0000	12.0000	0.037	12	0.445	0	-0.1962
24	50.0000	12.0000	0.037	12	0.445	0	-0.1962
25	50.0000	12.0000	0.037	12	0.445	0	-0.1962
26	50.0000	12.0000	0.037	12	0.445	0	-0.1962
27	50.0000	13.0000	0.037	13	0.482	0	-0.1962
28	50.0000	13.0000	0.037	13	0.482	0	-0.1962
29	50.0000	13.0000	0.037	13	0.482	0	-0.1962
30	50.0000	13.0000	0.037	13	0.482	0	-0.1962
31	50.0000	13.0000	0.037	13	0.482	0	-0.1962
32	50.0000	14.0000	0.037	14	0.519	0	-0.1962
33	50.0000	15.0000	0.037	15	0.556	0	-0.1962
34							
35	150.0000	3.0000	0.037	3	0.112	0	-0.2965
36	150.0000	10.0000	0.037	10	0.372	1	0.6722
37	150.0000	10.0000	0.037	10	0.372	0	-0.3984
38	150.0000	11.0000	0.037	11	0.409	5	4.5396
39	150.0000	11.0000	0.037	11	0.409	4	3.5507
40	150.0000	11.0000	0.037	11	0.409	0	-0.4048
41	150.0000	12.0000	0.037	12	0.447	1	0.5086
42	150.0000	12.0000	0.037	12	0.447	0	-0.4104
43	150.0000	12.0000	0.037	12	0.447	0	-0.4104
44	150.0000	12.0000	0.037	12	0.447	0	-0.4104
45	150.0000	12.0000	0.037	12	0.447	0	-0.4104
46	150.0000	12.0000	0.037	12	0.447	0	-0.4104
47	150.0000	13.0000	0.037	13	0.484	1	0.4431
48	150.0000	13.0000	0.037	13	0.484	0	-0.4153
49	150.0000	13.0000	0.037	13	0.484	0	-0.4153
50	150.0000	13.0000	0.037	13	0.484	0	-0.4153
51	150.0000	13.0000	0.037	13	0.484	0	-0.4153
52	150.0000	14.0000	0.037	14	0.521	0	-0.4196
53	150.0000	14.0000	0.037	14	0.521	0	-0.4196
54	150.0000	15.0000	0.037	15	0.558	1	0.3352
55	150.0000	18.0000	0.037	18	0.670	0	-0.4330
56							
57	500.0000	6.0000	0.149	6	0.893	0	-0.6581
58	500.0000	8.0000	0.149	8	1.191	0	-0.6839
59	500.0000	9.0000	0.149	9	1.340	6	2.4099
60	500.0000	10.0000	0.149	10	1.489	2	0.2404
61	500.0000	10.0000	0.149	10	1.489	0	-0.7008
62	500.0000	10.0000	0.149	10	1.489	0	-0.7008
63	500.0000	11.0000	0.149	11	1.638	7	2.3153
64	500.0000	11.0000	0.149	11	1.638	4	1.0199
65	500.0000	11.0000	0.149	11	1.638	3	0.5881
66	500.0000	11.0000	0.149	11	1.638	2	0.1563
67	500.0000	11.0000	0.149	11	1.638	1	-0.2755
68	500.0000	11.0000	0.149	11	1.638	0	-0.7073
69	500.0000	11.0000	0.149	11	1.638	0	-0.7073
70	500.0000	12.0000	0.149	12	1.787	4	0.8828
71	500.0000	12.0000	0.149	12	1.787	1	-0.3139
72	500.0000	12.0000	0.149	12	1.787	0	-0.7128
73	500.0000	12.0000	0.149	12	1.787	0	-0.7128
74	500.0000	12.0000	0.149	12	1.787	0	-0.7128
75	500.0000	12.0000	0.149	12	1.787	1	-0.3139
76	500.0000	13.0000	0.149	13	1.936	6	1.5066

Table A-5.2: Output from Model Run (EPA BMDS NLogistic Model. Revision: 2.6, Date: 2000/03/03)

500.0000	13.0000	0.149	13	1.936	0	-0.7176
500.0000	15.0000	0.149	15	2.234	0	-0.7255
1000.0000	3.0000	0.867	3	2.601	3	0.5609
1000.0000	3.0000	0.867	3	2.601	3	0.5609
1000.0000	3.0000	0.867	3	2.601	3	0.5609
1000.0000	3.0000	0.867	3	2.601	3	0.5609
1000.0000	3.0000	0.867	3	2.601	3	0.5609
1000.0000	9.0000	0.867	9	7.803	9	0.6958
1000.0000	9.0000	0.867	9	7.803	9	0.6958
1000.0000	9.0000	0.867	9	7.803	8	0.1147
1000.0000	10.0000	0.867	10	8.670	10	0.7053
1000.0000	10.0000	0.867	10	8.670	8	-0.3550
1000.0000	10.0000	0.867	10	8.670	7	-0.8851
1000.0000	10.0000	0.867	10	8.670	5	-1.9454
1000.0000	11.0000	0.867	11	9.536	11	0.7135
1000.0000	11.0000	0.867	11	9.536	11	0.7135
1000.0000	11.0000	0.867	11	9.536	5	-2.2115
1000.0000	12.0000	0.867	12	10.403	12	0.7204
1000.0000	12.0000	0.867	12	10.403	11	0.2692
1000.0000	12.0000	0.867	12	10.403	7	-1.5358
1000.0000	13.0000	0.867	13	11.270	13	0.7265
1000.0000	13.0000	0.867	13	11.270	8	-1.3737
1000.0000	14.0000	0.867	14	12.137	13	0.3389

Combine litters with adjacent levels of the litter-specific covariate
within dose groups until the expected count exceeds 3.0, to help improve
the fit of the X^2 statistic to chi-squared.

Grouped Data ⑤

Dose	Mean Lit.-Spec. Cov.	Expected	Observed	chi-squared Residual
0.0000	9.1111	3.039	0	-0.7043
0.0000	11.5000	3.409	0	-0.6744
0.0000	14.6000	2.705	4	0.2572
50.0000	8.9000	3.298	0	-0.5882
50.0000	12.7143	3.298	0	-0.5187
50.0000	14.5000	1.075	0	-0.2773
150.0000	10.2222	3.424	11	2.5976
150.0000	12.5714	3.275	1	-0.7591
150.0000	14.8000	2.754	1	-0.5991
500.0000	7.6667	3.425	6	0.8773
500.0000	10.0000	4.467	2	-0.6704
500.0000	11.0000	11.466	17	0.9031
500.0000	12.0000	10.722	6	-0.7689
500.0000	13.0000	3.872	6	0.5579
500.0000	15.0000	2.234	0	-0.7255
1000.0000	3.0000	13.004	15	1.2542
1000.0000	9.0000	23.408	26	0.8696
1000.0000	10.0000	34.678	30	-1.2400
1000.0000	11.0000	28.609	27	-0.4530
1000.0000	12.0000	31.210	30	-0.3153
1000.0000	13.0000	22.541	21	-0.4577
1000.0000	14.0000	12.137	13	0.3389

Chi-square = 17.35 DF = 13 P-value = 0.1837 ⑥

To calculate the BMD and BMDL, the litter specific covariate is fixed
at the mean litter specific covariate of control group: 11.227273

Benchmark Dose Computation

Specified effect = 0.1

Risk Type = Extra risk

Table A-5.2: Output from Model Run (EPA BMDS NLogistic Model. Revision: 2.6, Date: 2000/03/03)

```
Confidence level =          0.95

             BMD =       485.152                                    ⑦

            BMDL =       409.019                                    ⑧
```

6. Human Data

Opportunities for modeling human toxicological data are limited, and the human studies are less standardized than studies of experimental animals; thus modeling of human data is done on a case-specific basis. For some examples of benchmark dose modeling of human data, please refer to the following references; although it should be noted that these examples precede this benchmark dose modeling guidance and may not strictly adhere to the recommendations described herein. One example presented in EPA's IRIS database is for peripheral nervous system dysfunction induced by carbon disulfide in occupationally exposed workers (U.S. Environmental Protection Agency, 1995a). Another example in IRIS is for developmental neurologic abnormalities in human infants from exposure to methylmercury (U.S. Environmental Protection Agency, 1995b). More recent examples of benchmark dose modeling of methylmercury developmental neurologic effects from different databases are reported by Crump et al. (2000) and Budtz-Jorgensen et al. (2000).

References

IRIS (2000) ...

Budtz-Jorgensen E, Grandjean P, Keiding N, White RF, Weihe P (2000) Benchmark dose calculations of methylmercury-associated neurobehavioural deficits. Toxicology Letters 112-113:193-199.

Crump KS, Van Landingham C, Shamlaye C, Cox C, Davidson PW, et al. (2000) Benchmark concentrations for methylmercury obtained from the Seychelles Child Development Study. Environ Health Perspect 108:257-263.

U.S. Environmental Protection Agency (EPA). (1995a). Integrated Risk Information System (IRIS): Online substance file for carbon disulfide (http://www.epa.gov/ngispgm3/iris/index.html). National Center for Environmental Assessment, Washington, DC.

U.S. Environmental Protection Agency (EPA). (1995b). Integrated Risk Information System (IRIS): Online substance file for methylmercury (http://www.epa.gov/ngispgm3/iris/index.html). National Center for Environmental Assessment, Washington, DC.

GLOSSARY

Akaike Information Criteria (AIC) : A statistical procedure that provides a measure of the goodness-of-fit of a dose-response model to a set of data. AIC = -2 × (LL - p), where LL is the log-likelihood at the maximum likelihood fit, and p is the degrees of freedom of the model (usually, the number of parameters estimated).

Asymptotic Test : Statistical tests that approach known properties as sample sizes increase.

Benchmark Concentration (BMC): The concentration of a substance inhaled that is associated with a specified low incidence of risk, generally in the range of 1% to 10%, of a health effect; or the concentration associated with a specified measure or change of a biological effect.

Benchmark Dose (BMD) : An exposure due to a dose of a substance associated with a specified low incidence of risk, generally in the range of 1% to 10%, of a health effect; or the dose associated with a specified measure or change of a biological effect.

Benchmark Response (BMR): The response, generally expressed as in excess of background (see for example, Extra Risk), at which a benchmark dose or concentration is desired (see Benchmark Dose, Benchmark Concentration).

Beta-Binomial Distribution : A statistical distribution of clustered values, e.g., measures on offspring in a litter, where the average proportions of an event for clusters are described by a Beta distribution and the proportions of events in a cluster are described by a binomial distribution.

Binomial Distribution : The statistical distribution of the probabilities of observing 0,1,2,- - - ,n events in a sample of n independent trials each with the same individual probability that the event occurs.

BMCL: A lower one-sided confidence limit on the BMC.

BMDL: A lower one-sided confidence limit on the BMD.

Bootstrap : A statistical technique based on multiple resampling with replacement of the sample values or resampling of estimated distributions of the sample values that is used to calculate confidence limits or perform statistical tests for complex situations or where the distribution of an estimate or test statistic cannot be assumed.

Cancer Potency (Cancer Slope Factor) : A number that estimates the cancer risk (incidence) for a lifetime exposure to a substance per unit of dose. dose is generally expressed as mg / kg body wt / day.

Categorical Data : Results obtained where observations or measurements on individuals or samples are stratified according to degree or severity of an effect, e.g., none, mild, moderate,or

severe.

Chi-square Test : A statistical test used to examine the deviation of an observed number of events from an expected number of events.

Clustered Data : Measurements collected on some grouping of individuals, e.g., litters in reproductive and developmental studies.

Confidence Interval (Two-Sided) : An estimated interval from the lower to upper confidence limit of an estimate of a parameter. This interval is expected to include the true value of the parameter with a specified confidence percentage, e.g., 95% of such intervals are expected to include the true values of the estimated parameters.

Confidence Interval (One-Sided) : An interval below the estimated upper confidence limit, or interval above the estimated lower confidence limit, that is expected to include the true value of an estimated parameter with a specified confidence (percent of the time).

Confidence Limit : An estimated value below (or above) which the true value of an estimated parameter is expected to lie for a specified percentage of such estimated limits.

Constrained Dose-Response Model : Estimates of one or more parameters of the model are restricted to a specified range, e.g., equal to or greater than zero.

Continuous Data : Effects Measured on a continuum, e.g., organ weight or enzyme concentration, as opposed to quantal or categorical data where effects are classified by assignment to a class.

Convergence : Estimates of a parameter approach a single value with increasing sample size or increasing number of computer iterations.

Convex : Region of a dose-response relationship that curves upward, i.e., the slope becomes steeper with increasing dose.

Correlated Binomial Distribution : Clustered data where the individual values in a cluster ,e.g., a litter, each have the same probability of expressing an effect.

Covariate : An independent variable other than dose that may influence the outcome of an effect, e.g., age, body weight, or polymorphism.

Coverage : See confidence intervals or confidence limits.

Cubic : An effect is a function of a measure raised to the third power.

Degrees of Freedom : For dose-response model fitting, the number of data points minus the number of model parameters estimated from the data.

Delta Method : Variance of a function of random variables approximated from the derivatives of the function with respect to the random variables and the variances of the random variables.

Dichotomous Data : Quantal data where an effect for an individual may be classified by one of two possibilities, e.g., dead or alive, with or without a specific type of tumor.

Dispersion : Variation (differences) from a central (mean or median) value.

Dose-Response Model : A mathematical relationship (function) that relates (predicts) a measure of an effect to a dose.

Dose-Response Trend : Relationship between incidence or severity of a biological effect and a function of dose. Simply the slope for a linear dose-response.

EC_x : Effective exposure concentration associated with a biological effect in x% of the individuals. Often used for inhalation exposures based on the airborne concentration.

ED_x : Effective dose associated with a biological effect in x% of the individuals. Dose may be the external exposure often expressed in mg per day of the substance per kg body weight raised to a power (generally 1, 3/4, or 2/3) or area under the curve (AUC) in blood or target tissue where the substance remains in the body over a period of time.

Estimate : An empirical value derived from data for a parameter.

Excess Risk : Proportion of individuals or animals observed or estimated to possess an effect in addition to the spontaneous background risk.

Extra Risk: $[P(d)-P(0)]/[1 - P(0)]$, where $P(d)$ is the risk at a dose = d and $P(0)$ is the background risk at zero dose.

Gamma Distribution : A unimodal statistical distribution (relative proportion of responders as a function of some measure) that is restricted to effects greater than or equal to zero that can describe a wide variety of shapes, e.g., flat, peaked, asymmetrical.

Gaussian (Normal) Distribution : A unimodal symmetrical (bell-shaped) distribution where the most prevalent value is the mean (average) and the spread is measured by the standard deviation. Mathematically, the distribution varies from minus infinity with zero probability to plus infinity with zero probability.

Generalized Estimating Equation (GEE) : A statistical technique used for estimating parameters that requires only specification of the first two moments of the distribution of the estimator as opposed to a complete specification of the distribution.

Goodness-of -Fit : A statistic that measures the dispersion of data about a dose-response curve in order to provide a test for rejection of a model due to lack of an adequate fit, e.g., a P-value < 0.1.

Hazard Identification : Detection of an adverse biological effect, or precursor to an adverse effect, as a result of exposure to a substance.

Hill Equation : A dose-response curve, frequently used for enzyme kinetics, that monotonically approaches an asymptote (maximum value) as a function of dose raised to a power.

Hybrid Model : For continuous data establishes abnormal values based on the extremes in controls (unexposed individuals or animals) and estimates the risk of abnormal levels as a function of dose.

Incidence : Proportion or probability of individuals or animals exhibiting an effect, that varies from zero to one, sometimes expressed as a percent from 0% to 100%.

Independence : The result in one animal or individual does not influence the result in another animal or individual.

Intercept Term : The estimated value at zero dose or the dose corresponding to a zero effect.

Least Squares : A statistical procedure that estimates the values of dose-response parameters such that the sum of squares of deviations of data points from their estimated values is minimized, i.e., minimizes the estimated variance.

Likelihood Ratio : Ratio of the probability that the observed data arise from a set of model parameters relative to the maximum probability that arises from the set of maximum likelihood estimates.

Linear Dose-Response Model : The amount of change in a response is proportional to the amount of change in some function of dose.

Linearized Multistage Model : Dose-response model based on the multistage model of carcinogenesis that is restricted to a form that is approximately linear at low doses.

Local Maximum : Mathematical solution that maximizes a function in a region that may not be the overall global maximum.

Likelihood Function : Relative probabilities that various values of population parameters would arise from the sample observations.

Logistic Model : A sigmoid (S-shaped) function that relates the proportion of individuals with a specified characteristic to an independent variable, e.g., dose.

Log Transformation : Logarithm of raw data.

Maximum Likelihood Estimate (MLE) : Estimate of a population parameter most likely to have produced the sample observations.

Michaelis-Menten Equation : A dose-response curve, frequently used for enzyme kinetics, with maximum slope at zero dose that approaches a maximum asymptote at increasing dose.

Margin of Exposure (MOE) : Ratio of a dose that produces a specified effect, e.g., a benchmark dose, to an expected human dose.

Moment Estimates : A statistical estimation procedure that equates population moments to sample moments.

Monotonic Dose-Response : A dose-response that never decreases as dose increases. A monotonic function may be flat (constant) up to a threshold dose or may be flat at high doses if a biological limit, e.g., saturation, is attained.

Multinomial : Animals or individuals may be classified by more than two (binomial) categories, e.g., in a reproductive study fetuses may be : dead, alive normal, or alive abnormal.

Nonlinear Dose-Response Model : Mathematical relationship that cannot be expressed simply as the change in response being proportional to the amount of change of some function of dose.

Objective Function : Choice of function that is optimized for maximum likelihood estimation.

Ordinal Data : Integers designating the rank, order, or counts.

P-Value : In testing a hypothesis, the probability of a type I error (false positive) . The probability that the sample (experimental) results are compatible with a specific hypothesis.

Parameter : A value used to numerically describe a population of values, e.g., the mean and standard deviation; or a value used to describe a dose-response curve, e.g., the intercept and the slope of a linear dose-response.

Point of Departure (POD) : The point on a dose-response curve established from experimental data, e.g., the benchmark dose, generally corresponding to an estimated low effect level (e.g., 1% to 10% incidence of an effect). Depending on the mode of action and available data, some form of extrapolation below the POD may be employed for low-dose risk assessment or the POD may be divided by a series of uncertainty factors to arrive at a reference dose.

Polynomial : A mathematical function of the sum of a constant, linear term, quadratic term, cubic term, etc.

Probability : The proportion (on a scale of 0 to 1) of cases for which a particular event occurs. Zero indicates the event never occurs and one indicates the event always occurs.

Probability Distribution : A mathematical description of the relative probabilities of all possible outcomes of a measurement.

Probit Function : Assumes that the relative probabilities of effects as a function of dose are described by a Normal distribution. The cumulative probability as a function of dose has a sigmoid shape.

Profile Likelihood : A plot of the likelihood function versus the estimated value of a parameter.

Quadratic Term : A quantity in a mathematical formula that is raised to the second power (squared).

Quantal Data : Dichotomous (Binomial) classification where an individual or animal is placed in one of two categories, e.g., dead or alive, with or without a particular type of tumor, normal or abnormal level of a hormone.

Quantile : Percentile (cumulative probability) of a distribution that ranges from zero to the 100th percentile.

Quasi-Likelihood : Likelihood function that is not totally defined and generally based on only an expression including the mean and variance.

Rectangular Hyperbola : A mathematical function of the form y squared equals x squared plus c squared, where x and y are variables and c is a constant.

Regression Analysis : A statistical process that produces a mathematical function (regression equation) that relates a dependent variable (biological effect) to independent variable, e.g., dose rate, duration of exposure, age.

Repeated Measures : A biological endpoint is measured for the same individual or animal at different times (ages).

Residual Variance : The variance in experimental measurements remaining after accounting for the variance due to the independent variables, e.g., dose rate, duration of exposure, age. Typically referred to as the inherent unaccountable experimental variation.

Residuals : The numerical differences between observed and estimated effects.

Reference Concentration (RfC) : An estimate of the concentration of daily exposure to a substance (with uncertainty spanning perhaps an order of magnitude) for a human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime.

Reference Dose (RfD) : Replace " concentration " by " dose " in the above definition.

Risk : Probability that an animal or individual exhibits a particular adverse effect for a specified exposure, expressed on a probability scale of 0 to 1. May be expressed as the proportion of a population effected and often converted to the percent effected.

Risk Characterization : The process of combining dose-response information with exposure information in order to estimate risk.

S-Plus : Computer software for performing statistical analyses.

SAS : Computer software for performing statistical analyses.

Second Degree : A mathematical function that contains a quadratic term.

Shape Parameter : The exponent on dose in a dose-response function that dictates the curvature of the function.

Significance (Statistical Significance) : See P-value.

Threshold Dose : Dose below which a specified biological effect does not occur, generally for a particular population. Hence, the threshold dose is for the most sensitive individual in a population.

Uncertainty : The unknown effects of parameters, variables, or relationships that cannot or have not been verified or estimated by measurement or experimentation.

Uncertainty Factor : The value (often a default value of 10) used as a divisor of a NOAEL, LOAEL, or benchmark dose to calculate a RfC or RfD. Uncertainty factors are applied as needed for extrapolation of results in experimental animals to humans, interindividual variability including sensitive subgroups, extrapolation from a LOAEL to a NOAEL, extrapolation of results from subchronic exposures to chronic exposures, and database inadequacies.

Unconstrained Dose-Response Model : No restrictions imposed on the estimates of parameters.

Upper-Tail Probability : Probability that a variable exceeds a specified value.

Variability: Observable diversity in biological sensitivity or response, and in exposure parameters (such as breathing rates, food consumption, etc.) These differences can be better understood, but generally not reduced by further research.

Variance : Measure of variability , standard deviation squared.

Weibull : Form of a dose-response curve characterized by a relatively shallow slope at low doses that increases sharply as dose increases before leveling off at high doses.

Weighted Least Squares Estimate : Parameter estimate obtained by minimizing the sum of squares of observed and estimated values weighted by a function, frequently the reciprocal of the variance of an observation.