Toxicology and Applied Pharmacology 254 (2011) 181–191

Contents lists available at ScienceDirect

Toxicology and Applied Pharmacology

journal homepage: www.elsevier.com/locate/ytaap

Introduction to benchmark dose methods and U.S. EPA's benchmark dose software (BMDS) version 2.1.1

J. Allen Davis [a,*], Jeffrey S. Gift [a], Q. Jay Zhao [b]

[a] U.S. Environmental Protection Agency, National Center for Environmental Assessment, Research Triangle Park, NC 27711, USA
[b] U.S. Environmental Protection Agency, National Center for Environmental Assessment, Cincinnati, OH 45268, USA

ARTICLE INFO

Article history:
Received 23 September 2009
Revised 28 January 2010
Accepted 24 October 2010
Available online 27 October 2010

Keywords:
Benchmark dose
Risk assessment
Benchmark dose software
NOAEL
Dose–response
Point of departure
Dichotomous data
Continuous data
Nested data

ABSTRACT

Traditionally, the No-Observed-Adverse-Effect-Level (NOAEL) approach has been used to determine the point of departure (POD) from animal toxicology data for use in human health risk assessments. However, this approach is subject to substantial limitations that have been well defined, such as strict dependence on the dose selection, dose spacing, and sample size of the study from which the critical effect has been identified. Also, the NOAEL approach fails to take into consideration the shape of the dose–response curve and other related information. The benchmark dose (BMD) method, originally proposed as an alternative to the NOAEL methodology in the 1980s, addresses many of the limitations of the NOAEL method. It is less dependent on dose selection and spacing, and it takes into account the shape of the dose–response curve. In addition, the estimation of a BMD 95% lower bound confidence limit (BMDL) results in a POD that appropriately accounts for study quality (i.e., sample size). With the recent advent of user-friendly BMD software programs, including the U.S. Environmental Protection Agency's (U.S. EPA) Benchmark Dose Software (BMDS), BMD has become the method of choice for many health organizations world-wide. This paper discusses the BMD methods and corresponding software (i.e., BMDS version 2.1.1) that have been developed by the U.S. EPA, and includes a comparison with recently released European Food Safety Authority (EFSA) BMD guidance.

Published by Elsevier Inc.

Contents

[*] Corresponding author. U.S. EPA, NCEA, B243-01, Research Triangle Park, NC 27711, USA. Fax: +1 919 541 0245.
E-mail address: davis.allen@epa.gov (J.A. Davis).

0041-008X/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.taap.2010.10.016

Introduction and background

The U.S. Environmental Protection Agency (U.S. EPA) and other domestic and international health agencies routinely conduct risk assessments for human health effects that may arise due to exposure to chemicals in the environment. The current risk assessment process used by U.S. EPA is based on the National Research Council model (NRC, 1983) and includes hazard identification, dose–response assessment, exposure assessment, and risk characterization. Dose–response assessment, a key step in the process with respect to U.S. EPA's risk assessment mission, attempts to characterize the relationship between exposures to environmental pollutants or toxicants and observed adverse health outcomes. Traditionally, this analysis has been carried out differently for cancer and non-cancer endpoints due to assumed differences in the modes of actions for these disease processes. As the mechanistic understanding of disease processes has increased over time, however, the distinction between cancer and non-cancer endpoints has become less pronounced, allowing for a more harmonized approach to dose–response assessment (U.S. EPA, 2000).

The NOAEL/LOAEL approach

The main difference in how cancer and non-cancer risks have been characterized previously relates to the assumption that cancer risks are linear in the low-dose range and can arise from exposure to any concentration of pollutant, whereas non-cancer effects were assumed to exhibit threshold characteristics such that exposure below a certain concentration is associated with no appreciable risk of an adverse outcome. Quantitative methods used to analyze cancer risks involved fitting mathematical models to the observed tumor incidence data and extrapolating risk in a linear fashion to lower dose levels (U.S. EPA 1986a, 2005). Non-cancer risks, however, were analyzed by calculating a reference concentration (RfC) or reference dose (RfD) based on the identification of a No-Observed-Adverse-Effect-Level (NOAEL) as an estimate of the threshold, or a Lowest-Observed-Adverse-Effect-Level (LOAEL) as an alternative to the NOAEL. A NOAEL is the highest dose in a study at which no biologically or statistically significant effect is observed. A LOAEL is the lowest dose at which a significant adverse effect is detected. The NOAEL, or LOAEL in the absence of a NOAEL, then served as the point of departure (POD) for the calculation of the RfC or RfD through the application of uncertainty factors (UFs) designed to account for uncertainty in the data (e.g., lack of information about the differences in toxicokinetics and toxicodynamics between the test species and humans, differences among human individuals, or lack of knowledge in the database about other potential hazards for a particular chemical). After application of the UFs, the RfC or RfD is presumed to represent a daily dose over a lifetime with no appreciable risk of adverse non-cancer health outcomes.

However, the NOAEL/LOAEL approach for determining PODs is associated with limitations (see Table 1) that have been well documented (U.S. EPA, 2000). First and foremost is that an estimate of a point of departure based on a NOAEL is limited to the doses utilized in the study and is highly dependent on dose spacing. If a study does not utilize appropriate dose levels or spacing, then the ability of the NOAEL approach to determine a POD that approximates a low response is compromised. This dependency on dose spacing also makes it difficult to compare NOAELs across chemicals or studies. Additionally, the NOAEL approach is highly dependent on sample size. As the identification of the NOAEL primarily relies on the ability of a study to statistically detect a change in response in a dose group relative to the concurrent control group, studies with smaller sample sizes per dose group or a high background (i.e., control) response rate will have lower statistical power to detect small changes. As a result, these types of studies typically generate higher NOAELs. Thus, the NOAEL approach can result in a higher (i.e., less health protective) POD for studies with insufficient sample size and poor study design. Furthermore, the NOAEL method does not appropriately account for variability and uncertainty in the data that arise due to random errors (e.g., errors in animal dosing and response measurements). Because the NOAEL approach relies only on identification of a no-effect dose, it does not consider the shape of the dose–response curve. When a NOAEL cannot be identified in a study (e.g., there are statistically significant responses in all dose groups), a LOAEL cannot then be used to reliably extrapolate to a NOAEL. Lastly, even though a NOAEL has historically been thought of as a no-effect level, depending on the study design, it can represent effect levels as high as 10% (Allen et al., 1994) and does not imply that lower levels of exposure are without risk.

Table 1
Advantages and limitations of the NOAEL and BMD methods.

BMD advantages	**NOAEL limitations**
• Not limited to experimental doses	• Highly dependent on dose selection
• Less dependent on dose spacing	• Highly dependent on sample size
• Appropriately accounts for variability and uncertainty resulting from study quality	• Does not account for variability and uncertainty in the experimental results (e.g., does not account for study quality appropriately)
• Takes into account the shape of the dose–response curve and other related information	• Dose–response information (e.g., shape of dose–response curve) not taken into account
• Corresponds to consistent response level and can be used to compare results across chemicals and studies	• Does not correspond to consistent response levels for comparisons across studies
• Flexibility in determining biologically significant rates	• A LOAEL cannot be used to derive a NOAEL
BMD limitations	**NOAEL advantages**
• Ability to estimate BMD may be limited by the format of data presented	• Can be used when data is not amenable for BMD modeling
• Time consuming	• Easy to derive
• More complicated decision making process	• Has been the standard method for deriving a POD for decades (e.g., is familiar to most risk assessors)

The benchmark dose approach

Cognizant of the limitations of the NOAEL/LOAEL approach and in an attempt to address those limitations, the scientific community began proposing and using the benchmark dose (BMD) methodology as an alternative method for determining PODs. Crump (1984) detailed the use of the BMD approach in the determination of allowable daily intakes, the U.S. EPA Risk Assessment Forum published a report (U.S. EPA, 1995a) on the use of the BMD method in health risk assessments, and the U.S. EPA's Integrated Risk Information System (IRIS) first used the BMD approach for determining a POD for the calculation of a RfD in the toxicological review of methylmercury (U.S. EPA, 1995b).

With the development of software packages such as the U.S. EPA's benchmark dose software (BMDS), the BMD method has become the preferred dose–response assessment method within the U.S. EPA, as it effectively addresses many of the limitations of the NOAEL/LOAEL approach (see Table 1). The primary goal of BMD modeling is to define a POD that is largely independent of study design. It involves the fitting of various mathematical models to the observed data and estimating the BMD, which is the central estimate of the dose or concentration that produces a predetermined change in the response rate of an adverse effect. This predetermined change in response is called the benchmark response (BMR). For example, a BMD for a specific response could be the dose estimated to cause a BMR defined as a 10% increase in the number of animals developing a particular histopathological lesion (for a dichotomous endpoint), or the dose that results in a BMR defined as a change in body weight equal to one standard deviation of the control mean (for a continuous endpoint). Because a dose–response curve is generated across the entire dose range, the estimated BMD is not limited to the experimental doses like the NOAEL approach, and it is affected to a lesser degree by dose spacing (although inadequate dose spacing can preclude BMD modeling). Another benefit of BMD modeling compared to the NOAEL approach is that it can appropriately account for variability and uncertainty in the experimental results. The BMD method adopted by the U.S. EPA does this by calculating the BMDL, or the 95% lower-bound confidence limit on the BMD. Unlike the NOAEL approach, where studies with lower statistical power (e.g., small sample sizes or high background rates) return higher NOAELs, the BMD method calculates lower BMDLs for less sensitive studies (see Fig. 1 and Table 2).

Additional benefits of the BMD method are that it takes into account the shape of the entire dose–response curve and it can be used to compare results across chemicals and studies as the BMD reflects a constant level of response (i.e., the BMR) allowing for flexibility in determining biologically significant rates. Some potential hindrances to the application of the BMD method is that it is more time consuming and involves a more complicated decision making process than the NOAEL method. However, these are difficulties that can be compensated for with user-friendly software and do not represent fundamental inadequacies in the method's ability to appropriately analyze data and estimate PODs. BMD modeling also relies on availability of more detailed data reporting than the NOAEL approach, which may preclude its use in some situations. The BMDL currently serves as the POD for most non-cancer (RfC or RfD) and cancer risk estimates derived by the U.S. EPA.

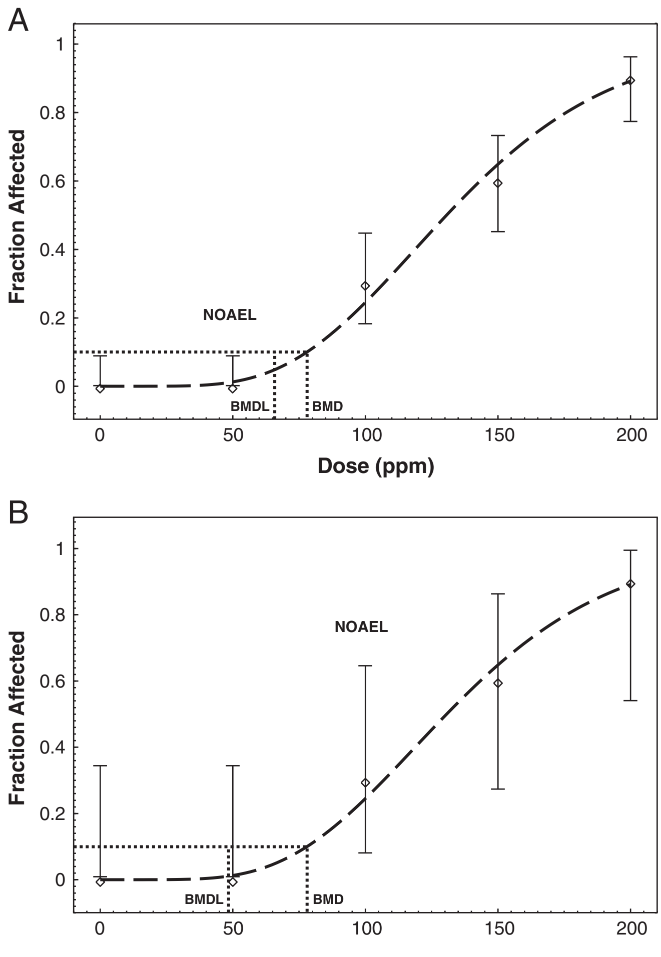

Fig. 1. Illustration of the effect sample size has on NOAEL and BMD and BMDL estimation. The incidence of effect at each dose group is represented by diamonds. The error bars represent the 95% confidence limits for the incidence. The fitted model is represented by the dashed line and the BMD and BMDL estimates are represented by the dotted line. Graph A represents a dataset with 50 animals per dose group; the NOAEL has been determined to be 50 ppm as it is the highest dose to fail to reach statistical significance (see Table 2); the BMD (corresponding to a BMR of 10% extra risk) is determined to be 78 ppm and the BMDL is 66 ppm (see Table 2). Graph B represents a dataset with only 10 animals per dose group, but with the same percentage of animals responding at each dose as Graph A. In this case, due to the decreased statistical ability to detect a treatment effect, the NOAEL has been determined to be 100 ppm. The BMDL in this situation has decreased to 48 ppm because the 95% lower-bound confidence limit is larger, reflecting the increased variability and uncertainty in the data due to the smaller sample size. Therefore, as sample size decreases, resulting in decreased power to detect treatment effects, the NOAEL method returns higher POD estimates whereas the BMD approach returns lower, more health protective, PODs. Note the increase in the size of the confidence limits around the observed incidence between Graphs A and B.

Table 2
Determination of NOAEL and BMDL dependent on dose group sample size.

Animals per dose group	Dose (ppm)	Incidence	Fisher's exact *p*-value	NOAEL[a]	BMD	BMDL
50	0	0	1.00	50	78.05	65.85
	50	0	1.00			
	100	15	<0.001			
	150	30	<0.001			
	200	45	<0.001			
10	0	0	1.00	100	78.05	48.40
	50	0	1.00			
	100	3	0.105			
	150	6	0.005			
	200	9	<0.001			

[a] NOAEL determined based on highest dose with Fisher's exact *p*-value of ≤0.1.

In the following sections, this paper will more thoroughly describe the BMD methodology and detail its application and use in the risk assessment process (for a more in-depth discussion of BMD methods, please see the BMDS website; http://www.epa.gov/ncea/bmds/). In addition, U.S. EPA's Benchmark Dose Software (BMDS) version 2.1.1 will be discussed with a particular focus on improvements recently incorporated into the latest version of the software. Additional tools for use in the analysis of certain toxicological data will also be presented. Lastly U.S. EPA's guidance on the use of BMDS and how it differs from other agency's guidance will be covered.

U.S. EPA BMD methodology

Data evaluation and study and endpoint selection

As outlined by the NRC (1983), the first step in the risk assessment process is hazard identification, which is the identification of health effects observed due to exposure to a particular chemical and determination of the critical effect on which to base NOAELs or BMDs and BMDLs. The initial process of hazard identification is fundamentally the same for both the NOAEL and BMD approach, and the U.S. EPA has published relevant guidance for endpoints specific to carcinogenicity, developmental toxicity, neurotoxicity, and mutagenicity (U.S. EPA 1986b, 1991, 1998, 2005). Data evaluations that are general to both the NOAEL and BMD approach include common considerations of the overall quality of a particular study. For example, were sufficiently large sample sizes used in order to adequately detect treatment responses? Were adequate exposure durations and relevant routes of exposure used? Did the study measure endpoints of concern? Did the study follow a sound quality control procedure such as good laboratory practice (GLP)? In addition to these general data quality considerations, there are additional BMD-specific issues to consider in the identification of datasets that are amenable to BMD modeling.

When determining which datasets are available and appropriate for BMD modeling, all relevant studies identified by the user should be evaluated carefully. Once potential critical effects have been selected, their adequacy for BMD modeling should be determined according to the following minimal data requirements.

First, has the data been reported in the study in such a manner that will allow it to be modeled? For dichotomous data, in addition to the sample size in each dose group, responses reported as incidence (i.e., 20/50 animals displaying an effect) or as a percentage (i.e., 40% animals displaying an effect) would be appropriate for modeling. Dichotomous data qualitatively reported without mention of specific information about incidence or percentage of response would fail to meet the minimal requirements for BMD modeling. Continuous data should be reported as a mean measure of biological effect (e.g., body weight or organ weight), along with a measure of variability (e.g., standard deviation or error) and the number of animals at each dose level, in order to be considered appropriate for modeling. Alternatively, individual animal data can be used if available.

Second, is there a biologically or statistically significant trend in response? Generally, a graded monotonic response that reaches statistical significance is required in order for a dataset to be considered appropriate for BMD modeling. However, in the case of particularly adverse or rare endpoints, dose–responses that display a monotonic dose–response, but fail to reach statistical significance may be considered to be biologically important and can be modeled given sufficient justification.

Third, are there sufficient dose groups to support BMD modeling? BMD modeling is essentially a curve-fitting exercise. This curve fitting relies on a sufficient amount of information being provided in the data set which informs the shape of the curve, especially at low dose. The fewer dose groups available in a dataset, the less information is provided about the dose–response curve. Also, the fewer the dose groups, the fewer models that will be able to be fit to the data, as the number of dose groups must be equal to or larger than the number of estimated parameters in the model. Ideally, datasets should contain at least three dose groups in addition to the control group. Datasets with only one non-control dose group are generally not suitable for BMD modeling since they contain minimal information about the shape of the dose–response curve. Datasets with two non-control dose groups may be fit by some models, but they may be insufficient to adequately describe the dose–response curve, and therefore, may ultimately be unsuitable as well.

Finally, are the observed dose–response relationships appropriate for BMD modeling? Generally, datasets with two doses with responses in excess of the control and with responses that define the low end of the dose–response curve (especially near the BMR) are preferred. Datasets in which response is only observed at the high dose are usually not suitable for BMD modeling (see Fig. 2a); however, if the response is in the range of the BMR and the study utilizes large sample sizes, there may be increased certainty in the calculated BMDs and BMDLs (Kavlock et al., 1996). Datasets in which all non-control doses exhibit similar levels of response are also not suitable for BMD modeling (see Fig. 2b); in this case minimal information is known regarding the shape of the dose–response curve below the lowest experimental dose. The “true” BMD may be just below the lowest dose, or very close to the control dose; therefore, uncertainty regarding the BMD is too great in datasets such as this. In the case of Fig. 2c, where there is a clear dose–response, but the lowest dose displays a response much greater than the desired BMR (e.g., 10%), limited information is known about the dose–response below the lowest dose. Again, the “true” BMD may lie anywhere below the lowest dose and uncertainty remains high. Datasets such as this are usually determined to be unsuitable for BMD analysis. It is important to note that in datasets such as these, dose–response models are often able to fit the data suitably from a mathematical standpoint. However, it is essential that the user exercise appropriate scientific judgment when determining what datasets are appropriate for BMD modeling from a risk assessment or biological basis.

After all available studies and potential critical effects have been evaluated for suitability, datasets meeting the minimal suitability criteria, as described above, can be analyzed using the BMD methodology. To maintain consistency and reproducibility, the U.S. EPA has developed a six-step process for BMD analysis (see Fig. 3). The six steps involved in the BMD analysis are (1) choice of a BMR, (2) selecting a set of models, (3) assessing model fit, (4) model selection when BMDLs are divergent, (5) model selection when BMDLs are not divergent, and (6) data reporting.

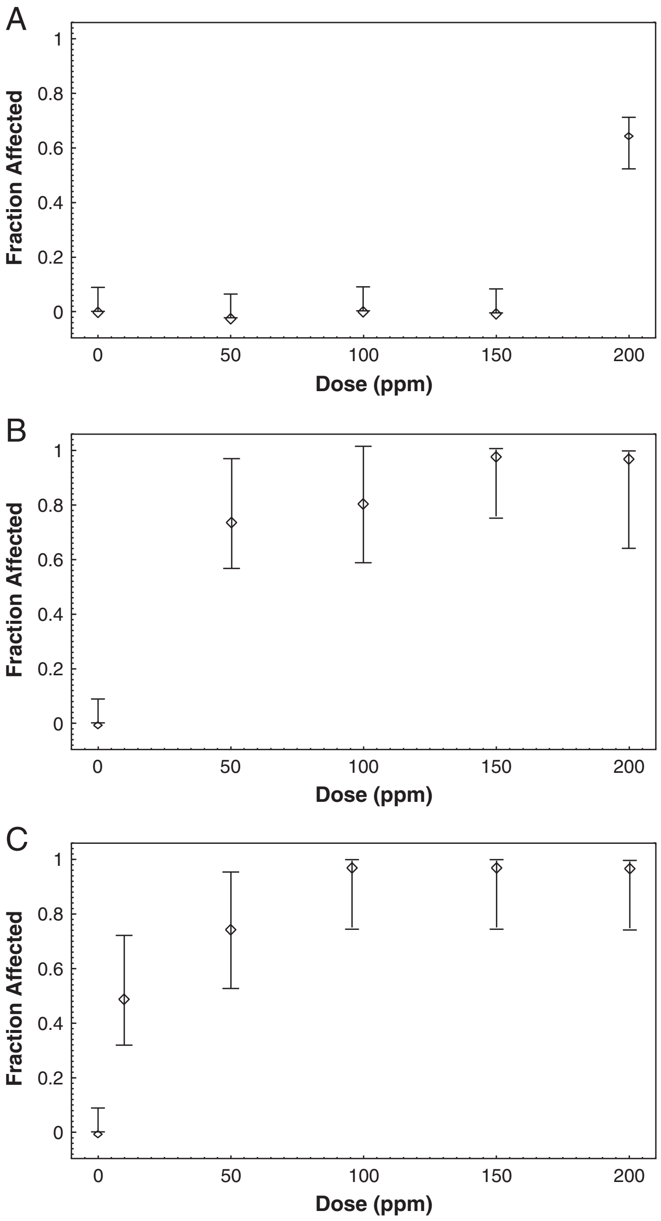

Fig. 2. Examples of datasets that are suboptimal for BMD modeling. Graph A shows a dataset where only the high dose shows response over control. Datasets of this nature are usually not amenable to BMD modeling but may return appropriate BMD and BMDL estimates when the response level is near the chosen BMR. Graph B shows a dataset where the level of response is comparable in all non-control doses. In this case, minimal information is known about where the “true” BMD may lay due to a lack of information about the shape of the dose–response curve below the lowest experimental dose. Graph C shows a dataset with a clear dose–response but with a high level of response at the lowest dose. Again, information about the dose–response is limited in this case, making this dataset of minimal utility in BMD modeling.

BMD analysis step 1: Choice of BMR

The BMR is selected based on scientific judgments regarding the dataset being analyzed and ideally should be near the low end of the range of increased risks that can be detected by a bioassay. However, a

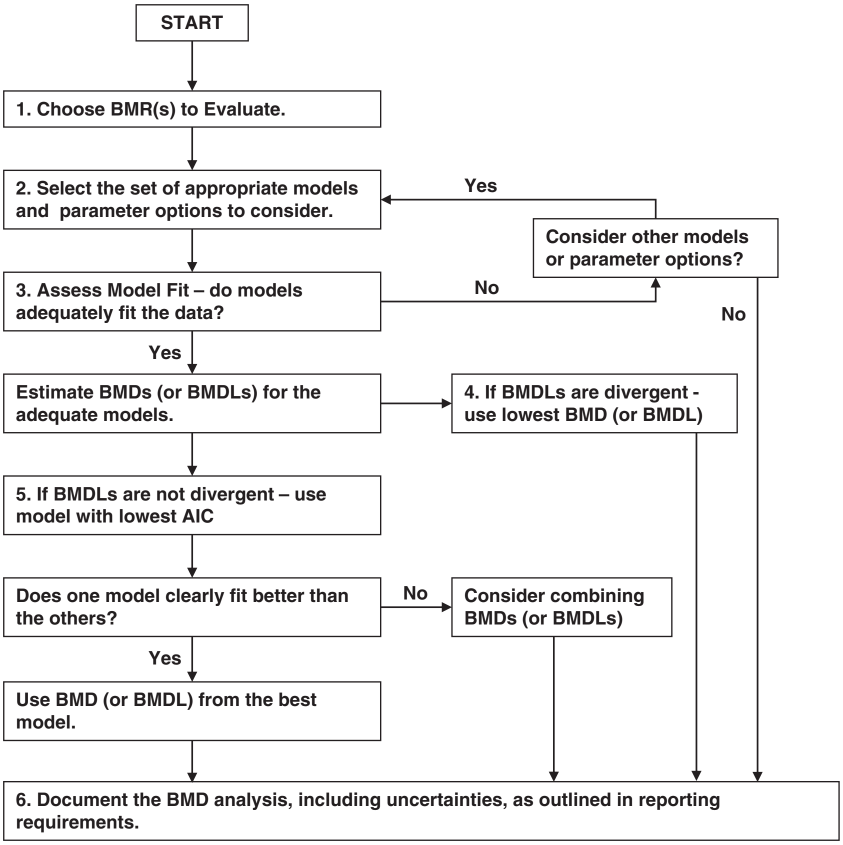
START
1. Choose BMR(s) to Evaluate.
2. Select the set of appropriate models and parameter options to consider.
Yes
Consider other models or parameter options?
3. Assess Model Fit – do models adequately fit the data?
No
No
Yes
Estimate BMDs (or BMDLs) for the adequate models.
4. If BMDLs are divergent - use lowest BMD (or BMDL)
5. If BMDLs are not divergent – use model with lowest AIC
Does one model clearly fit better than the others?
No
Consider combining BMDs (or BMDLs)
Yes
Use BMD (or BMDL) from the best model.
6. Document the BMD analysis, including uncertainties, as outlined in reporting requirements.

Fig. 3. U.S. EPA's BMD analysis framework.

choice of a low BMR, far away from the lowest observed responses, will result in extrapolation outside the range of the observed data and can impart high model dependence on the results. In other words, choice of a low BMR would result in different models returning drastically different BMDs and BMDL estimates. In situations such as this, confidence in the model results is low and BMD modeling should not be used or a higher BMR should be chosen, if scientifically justified.

For dichotomous (quantal) endpoints, a BMR of 10% extra risk[1] has typically been used to calculate BMDs as this response level is at or near the limit of sensitivity in most cancer bioassays (usually with sample sizes of approximately 50 per dose group) and non-cancer bioassays of comparable size. Certain study types such as developmental toxicity and epidemiological studies often have much larger sample sizes, which provide greater statistical power to detect lower levels of response; accordingly, lower BMRs may justified in these situations. Biological considerations can also be used to justify lower BMRs as in the case of particularly adverse or frank effects. In situations where a BMR other than 10% is justified, the calculated BMD and BMDLs should be presented alongside those calculated using a BMR of 10% for comparison purposes.

For continuous endpoints, a greater variety of BMR types are available for consideration. Ideally, a user will have information about what level of response can be considered biologically significant. For example, a 10% reduction in body weight may be scientifically supported as being an adverse level of response. In the absence of data signifying what level of response is biologically adverse, a change in the mean response equal to one SD of the control mean can be used as the BMR. This BMR is based on the determination from Crump (1995), such that for a continuous endpoint in a normally distributed population, if 1% of the animals in the control group are assumed to have an "abnormal response," a change in the mean response of 1.1 SD will result in 10% of the remaining animals reaching the abnormal response level. This response, in terms of percent of animals considered to be adversely responding, is comparable to the 10% BMR commonly used for dichotomous data. As with dichotomous endpoints, a lower BMR can be used for certain datasets given appropriate statistical or biological justifications (e.g., BMR = 0.5 SD for severely adverse effects).

BMD analysis step 2: Select a set of appropriate models

Once a particular dataset is considered suitable for BMD modeling and an appropriate BMR has been chosen, a group of models (selected according to the nature of the data) is fit to the data. The selection of a group of models will be based on the nature of the endpoint being modeled (e.g., whether dichotomous or continuous) and the experimental design used to generate the data of interest (e.g., nested study designs and number of dose groups utilized). See *Available Models in BMDS version 2.1.1* for further information.

[1] It should be noted that the statistical concept of "extra risk" is different from "added risk". For example, for an endpoint with a background response rate of 5%, a BMR of 10% added risk would correspond to the dose that elicits a 15% response (i.e., $0.10+0.05=0.15$). For "extra risk" however, the risk of a particular level of response is limited to the animals that remain able to respond. In other words, for the same endpoint with a 5% background rate, only 95% of the animals in a given population remain able to exhibit the effect in response to the exposure. Therefore a 10% extra risk would correspond to the dose that elicits a 14.5% response (i.e., $[0.10\times0.95]+0.05=0.145$). Use of extra risk as the basis of a BMR will result in lower BMDs than for added risk, except when the background response is 0, in which case the BMDs will be equal. It is current U.S. EPA guidance that extra risk be used for modeling dichotomous data (U.S. EPA, 2000).

BMD analysis step 3: Assessing model Fit

In the case of BMDS, the method of maximum likelihood is used to estimate the parameter values for the various models so that the fitted model describes the observed data as well as possible. Once all selected models have been fit to the data, a series of scientific judgments must be made to ensure that the fitted models adequately describe the observed data. Linear and non-linear models do not necessarily have a biological interpretation, and often there will be no biological basis on which to base model selection. In the absence of any biological basis of model selection, criteria for model selection is solely based on whether various models describe the data mathematically. The criteria used by the U.S. EPA to make this determination are a global goodness-of-fit value (*p*-value), a measurement of local fit (χ^2 scaled residual values for each individual dose group), and a visual inspection of the model fit.

The global goodness-of-fit *p*-value provides a measure of how closely the dose–response model fits the observed data at each dose group across the entire range of doses. In BMDS, the goodness-of-fit *p*-value for dichotomous data is derived from a χ^2 statistic, whereas the *p*-value for continuous data is derived from a likelihood ratio test. In both cases, a small *p*-value indicates poor model fit (i.e., the probability of obtaining a test statistic at least this extreme is unlikely assuming the null hypothesis that the observed data is represented by the model). In most cases, the U.S. EPA measures *p*-values for model fit against a significance level of $\alpha = 0.1$ (i.e., *p*-values<0.1 indicate inadequate model fit). It is important to note that the global goodness-of-fit *p*-values should not be used to compare model fit between models because the *p*-values depend not only on the overall model fit to the data points, but also on the degrees of freedom of the model, which vary due to the number of parameters employed in each model.

In addition to the global goodness-of-fit *p*-value, BMDS contains two other tests of model fit (based on likelihood ratios) for continuous models. When modeling continuous data in BMDS, the distribution of continuous measures is assumed to be normal, with either a constant variance or a variance that changes as a power function of the mean. As the BMD and BMDL are estimated based on either an assumed constant or modeled variance, it is critical to be able to estimate the variance based on these assumptions. A decision regarding the assumed variance model must be made before running continuous models, and the additional two tests of model fit inform the user whether their assumptions about variance were acceptable. As with the goodness-of-fit statistic, small *p*-values indicate that either the variance is not constant (first additional test), or that the variance model used by BMDS is not adequate in describing the non-homogenous variance (second additional test).

Since the goal of BMD modeling is to fit a model to dose–response data that describes the data set, especially at the lower end of the dose range, models within BMDS report χ^2-scaled residual values[2] for individual dose groups. Evaluation of local fit is important, as the software may occasionally converge on a less than optimal solution where, in order to fit the high-dose region of the dose–response curve, model fit is compromised in the low-dose region. The scaled residual for each dose group provides a standardized distance between the model prediction and the observed response. A smaller scaled residual indicates a better local fit, and a scaled residual equal to zero at a particular dose group would indicate the model provides an exact fit at that dose group. When comparing two model fits in the low-dose region, the model with a smaller scaled residual at the response level closest to the BMD is preferred. Any scaled residual with an absolute value greater than 2 should be a cause for concern regarding model fit.

Finally, visual inspection of the plotted dose–response curve is also important. It can give an additional indication of how well the model fits the data that is often difficult to glean from the two previous statistical methods. For example, a compromise in the fitting in the low-dose range due to the fitting of the high-dose data might not be detected by the evaluation of goodness-of-fit *p*-values or scaled residuals. Similarly, visual inspection is also useful when a model provides an unrealistic curve fit (i.e., non-monotonic or "wavy" curves).

If, after assessment of model fit is performed, multiple appropriately fitting models are identified, a final selection of the "best" fitting model must be made. Ideally, this decision should be made based on biological considerations (i.e., mode of action), but in the absence of that information, two alternative methods may be used, as explained in steps 4 and 5.

BMD analysis step 4: Model selection when BMDLs are divergent

U.S. EPA deems that if the BMDLs from the adequately fitting models are not sufficiently close (based on the user's scientific judgment), this may serve as an indication of model dependence and selection of the lowest BMDL can be justified in order to derive a health protective estimate of the POD. This situation may arise when the BMR is much lower than the observed response at the lowest dose and illustrates that, as a model extrapolates further from the observed range of the data, the results become more dependent on the particular model than on the actual data.

BMD analysis step 5: Model selection when BMDLs are not divergent

If the BMDL estimates for the appropriately fitting models are sufficiently close based on the users' best scientific judgment, there is no model dependence in the estimation of the BMDLs. Akaike's Information Criterion (AIC) can then be used to compare models from different families using a similar fitting method (e.g., maximum likelihood) (Akaike, 1973; $\mathrm{AIC} = -2\,L + 2p$, where L is the log-likelihood at the maximum likelihood estimates for the model parameters and p is the number of model estimated parameters). For models with a similar degree of fit, AIC penalizes a model for adding additional parameters when those added parameters do not significantly improve model fit. Therefore, the AIC rewards the less complex model. The model with the smallest AIC would be considered the model that most parsimoniously fits the data, and its BMDL would serve as the POD. The current technical guidance to use the smallest AIC, even when the differences are very small, is intended to prevent users from choosing models based on subjective and inconsistent criteria. When multiple models return the exact same AIC, their BMDLs can be averaged to obtain the POD.

BMD analysis step 6: Data reporting

It is important that after BMD modeling is conducted, it is reported in such a fashion that allows others to understand the rationale and justification for the scientific decisions and judgments made in the course of the analysis. The rationales for selection of the appropriate studies and endpoints, for the selection of the models used in the analysis, and choice of BMRs should be included. Modeling results, including model parameter estimates, goodness-of-fit statistics, scaled residuals, BMDs and BMDLs, as well as graphs of the dose–response curves, should be presented in such a way that allows for detailed comparison between models for each endpoint.

[2] BMDS calculates χ^{2-}scaled residuals for dichotomous data as follows: $\chi^2 = O_i - n_i p_i \,/\, \sqrt{n_i p_i (1 - p_i)}$ where n_i is the number in the *i*th group, O_i is the observed number of affected subject in the *i*th group, and p_i is the fitted proportion of affected subjects in the *i*th group. For continuous data, scaled residuals are calculated as: $\chi^2 = (O_i - E_i)/SE_i$where O_i is the observed value in the *i*th group, E_i is the estimated value in the i^{th} group, and SE_i is the estimated standard error in the *i*th group.

U.S. EPA's BMDS version 2.1.1

History

BMDS was developed by the U.S. EPA as a tool to facilitate the application of the BMD method in human health risk assessments. Development of BMDS began in 1995, when the National Center for Environmental Assessment (NCEA) started research into model development. The first prototype version of BMDS was released in 1997 and underwent extensive external and public reviews in 1998–1999 and quality assurance testing in 1999–2000. In September 2000, U.S. EPA released BMDS version 1.2 for public use. Subsequently released versions included: version 1.2.1 (October 2000, contained updated versions of the Polynomial and Hill models), version 1.3.1 (March 2001, contained updated, more compact and stable versions of the Polynomial, Power, and Hill models, as well as new Multistage, Weibull, and Gamma dichotomous models), version 1.3.2 (May 2003, contained revised Polynomial and Nested Logistic models), version 1.4.1 (February 2007, all models recompiled to improve speed and stability, a Multistage Cancer model added), and version 2.0 (July 2008, employed a completely new graphic user interface to facilitate batch processing of multiple models and contained a new set of background dose and background response models as well as the dichotomous Hill model for dose–responses that plateau at high doses). The current version of BMDS, version 2.1.1, contains interface improvements and several new models: a ten Berge model for concentration × time ($C \times t$) analyses, a toxicodiffusion model for neurotoxicology analyses and the new nested family of continuous exponential models.

Available models in BMDS version 2.1.1

In the current version of BMDS, version 2.1.1, there are 31 available models intended for the analysis of dichotomous, nested dichotomous, and continuous data as described below.

BMDS includes models for dichotomous endpoints in which the observations are statistically independent of each other. These models provide the probability that an animal will have an adverse response at a given dose. The actual number of animals that have an adverse response is assumed to be binomially distributed. Currently, BMDS contains 9 "traditional" models for dichotomous endpoints: Probit, Log-Probit, Logistic, Log-Logistic, Weibull, Quantal Linear, Gamma, Multistage, and Multistage Cancer. All dichotomous models can be written in the form:

$$\text{Prob}\{\text{response}\} = \gamma + (1\text{–}\gamma) \times F(\text{dose}, \alpha, \beta, \ldots.),$$

where $F(\text{dose}, \alpha, \beta, \ldots)$ is a cumulative distribution function specific to each model and γ, α, β, ... are parameters to be estimated using maximum likelihood methods. In addition to the options that are available to all dichotomous models (e.g., selection of BMR), there may be additional model-specific options to consider. Generally, these are options to restrict or constrain certain parameters so that the models do not return biologically questionable results. For example, a possible model-specific restriction is constraining either the slope parameter (Log-Logistic)[3] or power parameter (Weibull or Gamma) to ≥ 1 so that the slope of the dose–response curve does not become infinite in the region close to dose $= 0$. Another model-specific restriction that can be made is to constrain the β coefficients of the multistage model to ≥ 0 to prevent the curve from becoming non-monotonic, or "wavy." The multistage cancer model available in BMDS is the same model as the Multistage model except that its β coefficients are automatically constrained to ≥ 0, and it provides a cancer slope factor (i.e., BMR/BMDL) and a 95% upper bound on the BMD (i.e., BMDU) along with the estimated BMD and BMDL.

In addition to the "traditional" dichotomous models, BMDS version 2.1.1 incorporates new "alternative" dichotomous models, including a dichotomous Hill model designed for the evaluation of saturable responses and modified "traditional" models designed to allow for alternative approaches to evaluating background (see Table 3). The dichotomous Hill model contains a maximum probability of response (asymptote) parameter that allows the model to plateau at higher doses. Thus, this model is able to fit highly sigmoidal dose–response curves. Alternative Log-Probit, Log-Logistic, Gamma, and Weibull models incorporate a background dose parameter (η) in place of the background response parameter (γ). The incorporation of the background dose parameter provides low-dose linear alternatives to the traditional Log-Probit, Log-Logistic, Gamma, and Weibull models. In contrast to the other "traditional" models, the "traditional" Logistic and Probit models implicitly allow for a background dose effect. Therefore, the new versions of these two models incorporate an explicit background response parameter (γ). Consequently, for each type of dichotomous model in the current version of BMDS, there are now two model forms available, one with a background dose parameter and another with a background response parameter.[4]

BMDS also contains models for the analysis of "nested" dichotomous data. The most common application of these models will be in the analysis of developmental toxicity data in which the pregnant animal, or "dam," is individually exposed to the chemical of interest and the abnormal effects are assessed in their fetuses, embryos, or offspring (i.e., the "pups"). Statistically, the responses in the pups are not independent of one another; pups from one litter are more likely to respond similarly to each other than to pups from other litters. This is referred to as the "litter effect." The nested models in BMDS make available two separate parameters that can be estimated to account for this litter effect. The first, the litter-specific covariate, takes into account the condition of the dam *prior* to chemical treatment that may affect the development of the pups in the litter. The pregnant animal's body weight, the number of implantations, or live litter size could serve as indices of a dam's biological condition, and could be used as the litter-specific covariate. It is important to make sure that whatever biological index is used as the litter-specific covariate is not affected by experimental treatment. Otherwise, the covariate will dominate the true dose–response due to treatment. Therefore, the ideal covariate will often times be implantation sites as dosing in guideline developmental toxicity studies usually starts after implantation (i.e., gestational day (GD) 6). If treatment starts earlier than GD 6, care should be taken to ensure that implantation is not affected by dose if this index is to be used as the litter-specific covariate. The second parameter that attempts to control for the litter effect is intra-litter correlation. This parameter is explicitly estimated by the model and statistically describes the similarity in responses among pups of the same litter after the treatment has started. In order to determine whether or not the litter effect is affecting the model estimates of the BMD and BMDL, the nested models need to be run in an iterative fashion, alternatively estimating one, both, or neither of the litter effect parameters. Examination of magnitude of the model parameters values (i.e., are they non-zero?) and how their inclusion affects model fit (i.e., is the AIC lowered?) allows the user to determine whether they should be estimated in the BMD analysis. The three nested models currently available in BMDS version 2.1.1 are the Nested Logistic, NCTR, and Ran and Van Ryzin models.

Lastly, BMDS contains models for the analysis of continuous endpoints. These types of effects are measured on a continuum, such as organ weight or enzyme levels. These models estimate the mean value of response, λ(dose), expected for any given dose. Unlike

[3] The Log-Probit model also contains a slope parameter, but restricting this parameter is not considered necessary because values < 1 do not cause the Log-Probit model curve to have infinite slope close to dose 0.

[4] Except for the background dose form of the Log-Logistic model, which is currently under development.

Table 3
Comparison of current BMDS quantal models and new models with alternative background dose or response parameter.

Model name[a]	Functional form of the model	Explicit background parameter	Low dose linearity?	Number of parameters
Multistage[b]	$\gamma+(1-\gamma)[1\text{-}\exp\{-\sum_{j=1}^{k}\beta_j X^j\}]$	γ	Yes if $\beta_1>0$ No if $\beta_1=0$	1 + k
multistage_bgd	$1\text{-}\exp\{-\sum_{j=1}^{k}\beta_j (X+\eta)^j\}$	η	Yes	1 + k
logistic	$[1+\exp\{-(\alpha+\beta X)\}]^{-1}$	None	Yes	2
logistic_bgr	$\gamma+(1-\gamma)[1+\exp\{-(\alpha+\beta X)\}]^{-1}$	γ	Yes	3
probit	$\Phi\{\alpha+\beta X\}$	None	Yes	2
probit_bgr	$\gamma+(1-\gamma)\Phi\{\alpha+\beta X\}$	γ	Yes	3
log-logistic	$\gamma+(1-\gamma)[1+\exp\{-(\alpha+\beta\log\{X\})\}]^{-1}$	γ	No[c]	3
log-logistic_bgd[d]	$[1+\exp\{-(\alpha+\beta\log\{X+\eta\})\}]^{-1}$	η	Yes	3
log-probit	$\gamma+(1-\gamma)\Phi\{\alpha+\beta\log\{X\}\}$	γ	No[c]	3
log-probit_bgd	$\Phi\{\alpha+\beta\log\{X+\eta\}\}$	η	Yes	3
gamma	$\gamma+(1-\gamma)[\int_0^{\beta x} t^{\alpha-1}e^t dt]/\Gamma(\alpha)$	γ	No[e]	3
gamma_bgd	$G(\beta(d+\eta), \alpha/\Gamma(\alpha)$, where $G=$ the incomplete gamma function	η	Yes	3
Weibull	$\gamma+(1-\gamma)[1-\exp\{-\beta X^{\alpha}\}]$	γ	No[e]	3
Weibull_bgd	$[1-\exp\{-\beta(X+\eta)^{\alpha}\}]$	η	Yes	3

[a] Names in regular type denote "traditional" models that currently exist in BMDS. Names in italics denote "alternative" models that are new to BMDS and represent alternative forms to "traditional" models. "bgr" indicates a model with an explicit background response parameter, "bgd" indicates a model with an explicit background dose term.
[b] The cancer model is identical to the multistage model except that $\beta>0$ is automatically enforced.
[c] If slope parameter is ≥1, dose–response slope approaches 0 as dose approaches 0, if slope parameter <1, dose–response slope approaches ∞ as dose approaches 0.
[d] Currently under development for inclusion in future version of BMDS.
[e] If power parameter is ≥1, dose–response slope approaches 0 as dose approaches 0, if power parameter <1, dose–response slope approaches ∞ as dose approaches 0.

dichotomous models, where adversity is reported as an increase in incidence of effect, adverse continuous effects can manifest as either a decrease or increase in mean effect compared to control. Because continuous responses can either be an increase or decrease in effect, BMDS allows the user to manually set the adverse direction as "up" or "down" for a specific endpoint, or lets the software determine the direction automatically. Another important distinction between dichotomous and continuous models is that the variance of continuous measures must be modeled (see *BMD analysis step 3: Assessing model fit*). BMDS assumes that the distribution of continuous measures is normally distributed and provides options for assuming that the variance is either constant with dose, or increases as a power function of dose:

$$\sigma_i^2 = \alpha[\mu(\text{dose}_i)]^{\rho}.$$

Currently there are five models available in BMDS version 2.1.1 for modeling continuous data: linear (i.e., first-degree polynomial), Polynomial, Power, Hill, and Exponential. As with dichotomous models, there are model-specific restrictions to consider. The first model-specific restriction that can be made is to constrain the power parameter (Power or Hill) to be ≥ 1 so that the slope of the dose–response curve does not become infinite in the region close to dose 0. The second model-specific restriction is to constrain the β coefficients of the Polynomial model to prevent the curve from becoming non-monotonic, or "wavy." In order to do so, the user must determine the adverse direction of response for a specific dataset and restrict the β coefficients to that direction. For example, if body weight decreases with increasing exposure, the β coefficients could be restricted to be "non-positive."

Unique among the continuous models is the newly incorporated exponential model in BMDS version 2.1.1. It is actually a nested family of four increasingly complex models that are fit simultaneously by BMDS. This approach to evaluating risk via a nested family of models was described by Slob (2002) and was originally implemented in the Netherland's National Institute for Public Health and the Environment (RIVM) PROAST software. The output of the Exponential model includes AIC values for the individual models, as well as likelihood ratio tests of the significance of the more complex models relative to the simpler models, allowing the user more than one means to determine which model is the most suitable (see *Differences in BMD guidance from other agencies*). The Exponential model is also the only model that currently allows the user to assume a lognormal distribution of the data. This assumption of lognormality reflects the distribution of the data, not how the modeling is performed. Being able to assume lognormality is important as many biological parameters are thought to be lognormally distributed. Another benefit of the more complex Exponential models is that they can model data that plateaus at the high dose. However, unlike the Hill model, which is symmetric at the low and high dose regions, the more complex Exponential models can display asymmetric curvature at the high and low doses.

New features and models in BMDS 2.1.1

The major improvement in recent versions of BMDS software (version 2.0 and later) is the availability of batch processing. With batch processing, multiple models can be run against one or more datasets simultaneously in what BMDS calls a "session." Batch processing greatly improves the efficiency and speed in which a user can analyze multiple endpoints in the course of a risk assessment. When batch processing is used, and multiple models are run simultaneously, BMDS automatically summarizes the results from the individual model outputs and tabulates them into one summary file for easy comparison of results between models.

Currently, BMDS contains the beta versions of two types of time-dependent models, for both continuous and dichotomous endpoints, that will be finalized in future versions of the software. The first is the "toxicodiffusion" model that allows for the modeling of time-course, or repeated response, data that are commonly generated from certain types of studies (i.e., neurotoxicity screening tests). In studies of this nature, it is common to expose an animal to a particular dose of the chemical agent and then record the response as a continuous endpoint over time in order to assess how both the exposure of interest and time impact the endpoint. Zhu (2005) developed the toxicodiffusion model to consider the time impact on possible toxic effects and provide predictions of responses as a function of both time and dose. Using the toxicodiffusion model, a user could not only determine the critical dose necessary to elicit a pre-defined level of response, but also how the time-course of exposure-related effects differs from the natural time-course in unexposed animals. The current toxicodiffusion model is run through the BMDS platform but requires that the user have downloaded and installed the latest version of the R programming software (R-project, 2009). The second time-dependent model type that will be finalized in BMDS is the ten

Berge $C \times T$ model (ten Berge, 1985; ten Berge and Zwart, 1989; ten Berge et al., 1986), which is used in the context of short-term exposures where both the concentration and duration of exposure are critical for predicting the risk of developing the endpoint of interest. Haber's law, which states that risk is dependent on $C \times T$, is the basis for the development of the ten Berge models. These models are intended to be applied to data that presents both dose and duration (again, usually acute or short term) of exposure as well as responses in dichotomous endpoints in order to construct a concentration-time-response relationship.

For a more in-depth discussion of the new features of BMDS version 2.1.1 and for extensive tutorials and on-line training materials, please refer to the BMDS website at http://www.epa.gov/ncea/bmds/ (U.S. EPA, 2008).

Future improvements to BMDS

The present version of BMDS, version 2.1.1, represents more than a decade's worth of model and software development. However, these research and development activities are ongoing and future versions of BMDS will build upon the progress already made and will include new software features increasing functionality of existing models as well as incorporating new models for the analysis of additional data types.

Currently, a number of data analysis tools are under development for inclusion into BMDS, including methods relating to the underlying distribution of the data being analyzed and the identification of statistically significant trends in response. In the current version of BMDS, continuous endpoints are assumed to have a normal distribution around dose-specific means. This is the manner in which most data are reported in the scientific literature. However, it is known that many biological parameters can be lognormally distributed and assuming lognormal distribution may be more appropriate for certain datasets, especially ones in which the responses are constrained to only positive values (e.g., serum enzyme levels). The only model currently included in BMDS that allows for the assumption of a lognormal distribution is the exponential model. Research is presently underway to incorporate the ability to assume a lognormal distribution for any continuous model included in BMDS. Further, a normality test will be incorporated into BMDS to allow the user to explicitly determine the normality/lognormality of the data of interest rather than relying on a priori assumptions. Inclusion of the lognormality assumption for continuous models will be consistent with methods commonly used in Europe, where the PROAST dose–response software (RIVM, 2009) allows for such assumptions, thus harmonizing BMDS with European methods.

The current version of BMDS does not include a statistical trend test.[5] Visual inspection of the dose–response curve is currently the only method in which to determine whether a trend in response exists. Future versions of BMDS will incorporate explicit, statistical trend tests for all types of data in order to provide the user a quantitative measure of trends that exist in the dataset of interest. The specific tests that will be implemented will be the Cochran-Armitage test (Haseman, 1984) for dichotomous endpoints and the NOSTASOT test (Tukey et al., 1985) for continuous endpoints. Other analytical tools that are being considered for inclusion in future versions of BMDS include improvements in the initial estimation of model parameters, which will allow for better model convergence and performance, and the ability to perform model averaging in order to account for uncertainties in the model selection process (Wheeler and Bailer, 2007). Non-computational improvements in future versions of BMDS will include the capacity for user-defined summary table outputs and improved summary table export options.

In addition to added features intended to complement and enhance existing models, new methods and models are being developed for BMDS. One such method, the "hybrid" approach to modeling continuous endpoints, is intended to address the problems with explicitly dichotomizing continuous data and the difficulties surrounding interpreting and comparing the results of BMD analyses of dichotomous and continuous endpoints. The "hybrid" approach uses the distribution of the data being modeled to estimate the probability of an individual experiencing an adverse effect. Instead of explicitly dichotomizing the original data, the user defines a background rate of adversity which defines a cut-off in the tail of the distribution of the data. Based on the distribution cut-off value, the software can determine the dose that increases the probability of being in the tail which is equal to a user-defined risk level. The hybrid approach represents significant advantages over manually dichotomizing the original data. Often, the data necessary for dichotomizing the continuous endpoint is not available as continuous endpoint data are often reported as summary statistics (e.g., mean and standard deviation). Secondly, the hybrid approach uses all of the available information at hand and does not impose a loss of precision on the BMD and BMDL calculations. Inclusion of the hybrid approach into the BMDS platform will allow users to model continuous endpoints similarly to dichotomous data (e.g., express the BMR as a percent of animals affected) more easily and will allow for the consistent and accurate interpretation of modeling results obtained from both types of data.

Another model being developed for inclusion in BMDS is a multistage cancer model for the analysis of multiple tumors (i.e., "multi-tumor model") resulting from exposure to a chemical agent. The term "multiple tumors" does not refer to the incidence of multiple tumors in one animal, but the observation of two different tumor types (e.g., liver and lung tumors) with confirmed dose–response occurring in the same bioassay. If an assumption of independence is made in regards to the observed tumor types, application of the multi-tumor model estimates the risk of any combination of tumor types. In other words, the risk estimate will represent the risk of observing tumor A, tumor B, or tumor A and B in the same animal. The multi-tumor model has been developed for combination of two tumor types, but is currently undergoing expansion to handle more than two tumor types.

Lastly, another analytical tool being developed for BMDS is categorical regression software (i.e., CatReg). CatReg can be used to perform categorical regression analyses on toxicity data associated with up to two independent variables related to exposure (e.g., concentration and time) after the effects have been assigned to ordinal severity categories (e.g., no effect, mild, severe). In this way, the categorization of observed effects allows for the analysis of dichotomous, continuous, or descriptive data in the same manner. CatReg calculates the probabilities of the different severity categories over the continuum of the exposure-related variables. Another important aspect of CatReg analysis is that it allows for the meta-analysis of data from multiple toxicity studies simultaneously as long as the responses have been converted into ordinal data using the same category descriptions. CatReg is available as a stand-alone R package, but is currently being redesigned for inclusion in the BMDS platform.

Differences in BMD guidance from other agencies

The application of the BMD method represents significant improvement over the NOAEL method in the analysis of dose–response data for the determination of a POD; however, it is a more complicated analytical procedure that involves the application of scientific judgment to make certain statistical decisions. U.S. EPA has promulgated technical guidance representing what agency scientists have determined to be

[5] The closest approximation to a trend test is Test 1 for continuous models, which tests the hypothesis that mean responses or variances don't differ across dose levels. When the null hypothesis is rejected (i.e., $p<0.05$), it indicates that there may be differences across dose groups. However, Test 1 does not give an indication if the responses (or variances) are increasing or decreasing with dose.

Table 4
Comparison of BMDS and PROAST software.

	BMDS	PROAST
Environment	Can be run immediately as an executable in Windows	Splus or R software (free) is required
Time to get started	Short. Can be used immediately upon download	Steeper "learning curve." Requires basic knowledge of Splus or R
User Interface	Fully Windows based	Ordered process of answering multiple choice questions; Graphical User Interface available only for continuous data
Models:		
Continuous	Yes	Yes
Dichotomous	Yes	Yes
Nested continuous	No	Yes
Nested dichotomous	Yes	Yes
Categorical	No	Yes
Global goodness of fit test		
Continuous	Likelihood ratio test	Likelihood ratio test
Dichotomous	χ^2 *p*-value	Likelihood ratio test
Model selection criteria		
Model Dependence[a]	Lowest BMDL	Lowest BMDL
No Model Dependence[a]	Lowest AIC	Lowest BMDL
Confidence interval calculated using profile likelihood	Yes	Yes
Confidence interval calculated using bootstrapping	No	Yes
Covariates	No[b]	Yes
Automatic model fitting for nested models	Yes	Yes
Graphic output	Yes	Yes

[a]As determined by similarity between estimated BMDLs (see *Model selection*).
[b] Except for litter-specific covariate for nested dichotomous data.
Adapted from EFSA (2009).

the most appropriate methods to use when making decisions pertaining to model fit, model selection, and other modeling considerations (U.S. EPA, 2000). Recently, the European Food Safety Authority (EFSA) released guidance in regard to the use of the BMD method, including BMR selection issues and how model selection should be carried out (EFSA, 2009). EFSA also specifically discusses the use of BMDS and another software package, PROAST, in the analysis of dose–response data. PROAST was developed by the Netherland's National Institute for Public Health and the Environment (RIVM) for the analysis of toxicological dose–response data (RIVM, 2009). While very similar in scope to BMDS, PROAST is quite different in the execution of the software, as well as what statistical methods are utilized for determining how well models fit the data and the ultimate selection of the best model (see Table 4).

The guidance promulgated by EFSA differs from U.S. EPA guidance in a number of ways. The first is EFSA's recommendation for selection of the BMR for continuous endpoints (EFSA's and U.S EPA guidance are similar for dichotomous BMRs). EFSA recommends that a 5% response level is usually satisfactory for continuous data, as it is usually within the observed range of the data and should provide BMD and BMDL estimates that are not overly model dependent, based on the findings of Woutersen et al. (2001) and Sand et al. (2006). The U.S. EPA discourages using a percentage change as the basis for a BMR for continuous endpoints without a biological basis to do so; the same percent change can represent very different degrees of response for different endpoints. U.S. EPA's guidance instructs that a BMR of 1 control standard deviation is a more appropriate BMR for continuous endpoints because it takes into consideration the distribution of the data and is more comparable to the 10% extra risk BMR suggested for dichotomous endpoints.

The second way in which EFSA guidance differs from U.S. EPA guidance is in how models are judged regarding fit. EFSA guidance for model fit involves two principles: deciding which model fits best within a "nested" family of increasingly complex models and then a determination of overall goodness-of-fit. Both principles are based on the likelihood ratio test. In the PROAST software there are three families of nested models: the Exponential models (also found in BMDS) and Hill models (only the most complex, four-parameter, form is available in BMDS) for continuous endpoints and the linearized Multistage models for dichotomous endpoints. For dichotomous endpoints, PROAST also contains all of the dichotomous models available in BMDS. In order to select the best model within a family of models, more complex models must be compared to the corresponding simpler models in order to determine whether the addition of extra parameters significantly improves the model fit. This is done in a step-wise fashion until the most "optimal" (parsimonious) model has been selected. Once the best model from a nested family has been chosen, that "fitted" model and any other singular models included in the analysis is then compared to the "full" model to determine goodness-of-fit. The full model is the model that perfectly fits the means (continuous data) or incidences (dichotomous data) at all dose levels. The U.S. EPA BMDS reports *p*-values derived from likelihood ratio test results (between "fitted" and "full" models)[6] and includes a nested analysis of Exponential models similar to that which is performed in PROAST. However, U.S. EPA recommends that each model fit be judged independently (before model comparisons among models of a nested family). In addition, BMD modeling is largely considered a curve-fitting exercise involving a suite of models, and U.S. EPA (2000) recommends that $\alpha = 0.1$ be used to compute the critical value for goodness-of-fit, instead of the more conventional value of 0.05 used by EFSA.[7] Finally, EFSA does not support the use of $\chi^{2\text{-}}$scaled residuals to assess local fit, a factor that the U.S. EPA considers important to ensure that the models of interest are providing good local fit, especially in the low-dose region.

Final model selection in PROAST using EFSA's guidance is solely dependent on the lowest BMDL. Unlike U.S. EPA guidance, no consideration is given to relative model fit or the divergence of BMDL results at this point. Pursuant to U.S. EPA's guidance, the "lowest BMDL" criterion is only used when BMDLs are considered to be sufficiently divergent indicating a high degree of model dependence (see *BMD analysis step 4*). When BMDLs are not sufficiently different and model dependence is unlikely, AIC values should be used

[6] BMDS also reports and U.S. EPA guidance refers to the χ^2 *p*-values for overall fit of dichotomous models, but the difference between the two p-value results is generally insignificant.

[7] An exception to this recommendation is when there is an a priori reason to prefer a specific model(s), such as the Multistage model in the case of cancer data, in which case the U.S. EPA allows that the more conventional values of $\alpha = 0.05$ or $\alpha = 0.01$ may be appropriate.

in order to determine which model most parsimoniously fits the data (see *BMD analysis step 5*).

Additional research and analysis is needed in order to determine how the differences between PROAST and BMDS guidance affect dose–response analyses. As the state-of-the-science evolves, so will the specific guidance promulgated by domestic and international health agencies, and some harmonization of methods can be reasonably anticipated. For example, currently within BMDS, the exponential models can be tested in a nested fashion, and the provided *p*-values, based on the likelihood ratio tests, can be used to make model selection.

Conclusions

This paper has provided a basic overview of BMD methodologies and how the BMD method represents a more scientific and quantitative method for the derivation of human health protective reference values than the NOAEL approach. The BMD method addresses many of the specific limitations of the NOAEL method and allows for a more detailed analysis of toxicological dose–response information. A general summary of the BMD method, including issues pertaining to BMR selection, model fit, and model selection, has been included. The focus of this paper has been U.S. EPA's Benchmark Modeling Software and its application in the current risk assessment paradigm. The current version of BMDS has been discussed in detail, including the suite of models available to users and recent improvements in the user interface that allow for more efficient analysis of multiple datasets from multiple studies. Planned improvements to the software have been discussed, as have the differences between U.S. EPA's and EFSA's guidance regarding the application of the BMD method.

The current version of BMDS is considered to represent the state-of-the-science in quantitative toxicological dose–response analysis. However, it can be reasonably expected that, as the BMD method grows in acceptance and usage, the needs of the risk assessment community respective to quantitative analyses will also grow. To that end, BMDS will continue to evolve in order to meet the changing needs and challenges faced by the scientific community in ensuring continued protection of public health.

Conflict of interest disclosure statement

The authors declare that they have no conflicts of interest.

References

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

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

ten Berge, W.F., 1985. The toxicity of methylisocyanate for rats. J. Hazard. Mater. 12, 309–311.

ten Berge, W.F., Zwart, A., Appelman, L.M., 1986. Concentration-time mortality response relationship of irritant and systemically acting vapors and gases. J. Hazard. Mater. 13, 301–309.

ten Berge, W.F., Zwart, A., 1989. More efficient use of animals in acute inhalation toxicity testing. J. Hazard. Mater. 21, 65–71.

Crump, K.S., 1984. A new method for determining allowable daily intakes. Fundam. Appl. Toxicol. 4, 854–871.

Crump, K.S., 1995. Calculation of benchmark doses from continuous data. Risk Anal. 15, 79–89.

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

Haseman, J., 1984. Statistical issues in the design, analysis, and interpretation of animal carcinogenicity studies. Environ. Health Perspect. 58, 385–392.

Kavlock, R.J., Schmid, J.E., Setzer Jr., R.W., 1996. A simulation study of the influence of study design on the estimation of benchmark doses for developmental toxicity. Risk Anal. 16, 391–403.

NRC (National Research Council), 1983. Risk assessment in the federal government: managing the process. National Academy Press, Washington, DC.

R-project (The R Project for Statistical Computing), 2009. http://www.r-project.org/.

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

Sand, S., von Rosen, D., Victorin, K., Falk Filipsson, A., 2006. Identification of a critical dose level for risk assessment: developments in benchmark dose analysis of continuous endpoints. Toxicol. Sci. 90, 241–251.

Slob, W., 2002. Dose–response modeling of continuous endpoints. Toxicol. Sci. 66, 298–312.

Tukey, J., Ciminera, J., Heyse, J., 1985. Testing the statistical certainty of a response to increasing doses of a drug. Biometrics 41, 295–301.

U.S. EPA (Environmental Protection Agency), 1986a. Guidelines for carcinogen risk assessment. Fed. Regist. 51 (185), 33992–34003.

U.S. EPA (Environmental Protection Agency), 1986b. Guidelines for mutagenicity risk assessment. Fed. Regist. 51 (185), 34006–34012 http://www.epa.gov/ncea/iris/backgr-d.htm.

U.S. EPA (Environmental Protection Agency), 1991. Guidelines for developmental toxicity risk assessment. Fed. Regist. 56 (234), 63798–63826 http://www.epa.gov/ncea/iris/backgr-d.htm.

U.S. EPA (Environmental Protection Agency), 1995a. Use of benchmark dose approach in health risk assessments. Risk Assessment Forum, Washington, DC; EPA/630/R-94/007. Available from the National Technical Information Service, Springfield, VA, PB95-213765.

U.S. EPA (Environmental Protection Agency), 1995b. Integrated risk information system (IRIS): online substance file for methylmercury. National Center for Environmental Assessment, Washington, DC.

U.S. EPA (Environmental Protection Agency), 1998. Guidelines for neurotoxicity risk assessment. Fed. Regist. 63 (93), 26926–26954 http://www.epa.gov/ncea/iris/backgr-d.htm.

U.S. EPA (Environmental Protection Agency), 2000. Benchmark dose technical guidance document (External Peer Review draft). Risk Assessment Forum, Washington, DC; EPA/630/R-00/001. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=20871

U.S. EPA (Environmental Protection Agency), 2005. Guidelines for carcinogen risk assessment. Fed. Regist. 70 (66), 17765–17817 http://www.epa.gov/ncea/iris/backgr-d.htm.

U.S. EPA (Environmental Protection Agency), 2008. Benchmark Dose Software (BMDS). http://www.epa.gov/NCEA/bmds/index.html2008.

Wheeler, M.W., Bailer, A.J., 2007. Properties of model-averaged BMDLs: a study of model averaging in dichotomous response risk estimation. Risk Anal. 27, 659–670.

Woutersen, R.A., Jonker, D., Stevenson, H., te Biesebeek, J.D., Slob, W., 2001. The BMD approach applied to a 28-day toxicity study with rhodorsil silane in rats: the impact of increasing the number of dose groups. Food Chem. Toxicol. 39, 697–707.

Zhu, Y., 2005. Dose-time-response modeling of longitudinal measurements for neurotoxicity risk assessment. Environmetrics 16, 603–617.