Available online at www.sciencedirect.com
SCIENCE DIRECT®
Regulatory Toxicology and Pharmacology 43 (2005) 280–291
Regulatory Toxicology and Pharmacology
www.elsevier.com/locate/yrtph

The role of the benchmark dose in a regulatory context

Kim Z. Travis *, Ian Pate, Zoe K. Welsh

Syngenta CTL, Alderley Park, Macclesfield, Cheshire, SK10 4TJ, UK

Received 31 January 2005
Available online 6 September 2005

Abstract

The use of no observed adverse effect levels (NOAELs) as a way of interpreting toxicology studies carries a number of problems, and the benchmark dose (BMD), or its lower confidence limit have been proposed as potential replacements. In practice, the theoretical advantages of the BMD approach are often outweighed by the practical disadvantages posed in a regulatory context. Attempts to seek consensus for the routine use of BMD methodology tend to involve diluting its potential advantages as much as they address the disadvantages, resulting in a relatively complex interpolation tool that delivers little more than the NOAEL. It is time to recognise that the BMD will never entirely replace the NOAEL. The two methods can have complementary roles. The NOAEL is well suited as a routine simple summary of effects in toxicology studies, whilst the BMD can be a higher tier approach for the interpretation of the most critical studies in a regulatory data package.

Keywords: Benchmark dose; NOAEL; Chemical regulation; Risk assessment

1. Introduction

The majority of chemical risk assessments for human are based on the determination of a no observed adverse effect level (NOAEL) from toxicology studies. The NOAEL is determined as the lowest dose tested which results in a response that is not significantly different from the control value, when judged by a statistical test and expert judgement, and which is also considered to be adverse. The NOAEL approach is technically flawed in several ways, and dissatisfaction with this approach led to the development of the benchmark dose (BMD) (Crump, 1984). The BMD is the dose that results in a predetermined level of adverse response, i.e., the critical effect size. The lower confidence limit (BMDL) of the BMD is often taken as the starting point for determining allowable exposure levels.

It seemed that it was simply a matter of time before the BMD would replace the NOAEL as the regulatory tool of choice. Yet nearly 20 years later, the NOAEL is still the dominant approach in routine use. Why is this? Many technical problems have been identified with the BMD methodology, but these tend to be viewed as temporary obstacles that will be overcome with more effort. In contrast, few have advocated the retention of the traditional NOAEL approach, so perhaps the current situation simply reflects resistance to change amongst regulatory toxicologists? This paper presents an alternative view. In reality, all techniques have advantages and disadvantages, and a new technique will be adopted if it offers a favourable balance of these to the user community. In practice, the theoretical advantages of the BMD approach are often outweighed by the practical disadvantages it poses in a regulatory context. This paper presents an overview of these advantages and disadvan-

* Corresponding author. Fax: +44 0 1625 510762.
E-mail address: Kim.Travis@Syngenta.com (K.Z. Travis).

0273-2300/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.yrtph.2005.07.003

tages, and considers the implications for the role of the BMD.

2. Good and bad aspects of the NOAEL approach

Plus points about the NOAEL approach include the following.

2.1. Ease of understanding

People feel they know what it means and feel confident in expressing this to others.

2.2. Ease of checking

In most cases you can look at a study variable summarised in a simple table or graph and see how the reported NOAEL relates to the data in front of you. If an error has been made it will probably be apparent by visual inspection.

2.3. Intuitive appeal

Any variety of "no effect" level, e.g., NOAEL, conveys the impression that there is no effect. This is intuitively appealing, reassuring, and politically acceptable, even though it a perception not based on science.

2.4. Familiarity

Having been used for so long, the NOAEL is ingrained in the regulatory process. However, this is only an explanation for resistance to change, not truly a plus point of the approach per se.

Set against these plus points are a number of problems that have been claimed to apply to the NOAEL approach.

2.5. NOAELs can only be set at dose levels studied

Undeniable, with the consequence that two identical chemicals could end up with different NOAELs because different dose levels were studied. If studies had more dose levels then this would be less problematic. Designing a new study builds on the results of studies that have gone before, and dose levels are chosen based partly on this prior information. This Bayesian approach should tend to reduce this undesirable limitation of the NOAEL approach to some unknowable extent, by seeking to ensure that there is a dose placed in the region of the anticipated NOAEL. An unfortunate side-effect of NOAELs only being set at dose levels, is that ratios of NOAEL between different chemicals, species, or study durations, can be misleading and are more variable than the true ratios of toxicity (Brand et al., 1999).

2.6. NOAELs use little information about the shape of the dose–response curve

True and unfortunate given that there must be information in the dose–response curve of value in the regulatory process.

2.7. NOAELs are not compatible with the use of PBPK models

A PBPK model will tend to predict a smooth continuous response to increasing dose. If you rigidly stick to the notion that you must regulate based on a supposed zero response, then it is hard to interpret such model results in a regulatory context. It is this rigidity, rather than the use of the NOAEL approach, which might limit the regulatory application of a PBPK model. PBPK modelling is a powerful way of integrating biological understanding into decision-making, in a way that any approach based on a single variable in a single study cannot do (whether NOAEL or BMD). So the NOAEL is fully compatible with PBPK modelling, with the NOAEL as Tier 1 and the PBPK as some much higher tier.

2.8. What if there is no NOAEL in a study?

It is the responsibility of the experimenter to place doses in the areas of interest, including at the bottom end of the dose–response curve. If an NOAEL is not achieved in a study, then the lowest dose is a lowest observed adverse effect level (LOAEL), and this is generally used in risk assessment with an additional safety factor. In quantitative terms, this is somewhat unsatisfactory.

2.9. NOAELs favour bad experiments

The argument here is that poor experimental conduct and statistical design will be "favoured" by the NOAEL method (in as much as a higher NOAEL is to be viewed as favourable). This is frequently suggested as the fatal flaw for the NOAEL approach, but the phenomenon has been overestimated in a regulatory context. For example, having fewer animals will always tend to result in a higher NOAEL, but in a regulatory context the number of animals per dose level is tightly controlled by standard protocols and for animal welfare reasons (Crump, 1984). So experiments with too few animals would tend to be rejected for regulatory use, other than perhaps as supplementary information. Other sources of study variability will, if increased, tend to result in higher NOAELs, though if a regulator sees results that are unreasonably variable, he/she always retains the right to reject the study or at least to not use it as the basis for decision-making.

3. The BMD approach measured against the good and bad aspects of the NOAEL approach

The starting point for gauging the BMD approach is to measure it against the good and bad aspects of the NOAEL, which is done in this section. The following section will then consider the *new* good and bad aspects introduced by the BMD.

3.1. Ease of understanding

The overall concept of the BMD is probably no harder to grasp than the NOAEL. However, so long as there are not agreed standards for its definition and calculation, this will complicate explanations and hinder understanding. Arguably, if methodology can be standardised, a decision-maker would no more need to understand the BMD calculations than he/she needs to understand Student's *t* test to regulate based on NOAELs. A specific aspect of the BMD, which can cause confusion, is that it is not constrained to lie between the two doses that bound the critical effect size. For example, consider two doses that result in 8 and 22% effects. With a critical effect size of 10% it is entirely possible that the BMD is less than the dose at which an 8% effect was seen. Similarly, if the critical effect size was 20% then the BMD could be greater than the dose at which a 22% effect was seen. This is scientifically reasonable, but may make the BMD harder to explain.

3.2. Ease of checking

Visual inspection would easily indicate if a calculated BMD is incongruent with the data, but the whole calculation would need to be repeated in order to check any quoted BMD.

3.3. Intuitive appeal

By corresponding to a specific level of "harm," the BMD lacks the intuitive appeal of a "no effect" level. Scientists would see this as an advantage for the BMD, but some non-scientists would tend to see this as a powerful disadvantage. Is the customer (typically non-scientific) always right?

3.4. Familiarity

BMD is not familiar, though this should not be thought of as an obstacle or nothing would ever change.

3.5. NOAELs can only be set at dose levels studied

A clear advantage of the BMD is that this does not apply. Unlike ratios of NOAELs, ratios of BMDs should be less variable and closer to the true ratios in toxicity, though it has recently been shown that ratios of BMDs are not free from statistical artefacts (Brand et al., 2001). The advantage over NOAELs is greatly reduced if lower confidence limits are used (BMDLs), as is typically proposed in a regulatory context. Because these depend strongly on animal numbers and study variability (Murrell et al., 1998), ratios of BMDLs will relate very poorly to true ratios of toxicity.

3.6. NOAELs use little information about the shape of the dose–response curve

By fitting a dose–response curve, the BMD method certainly requires information about the full dose–response curve. But in practice the value of the BMD is in most cases almost entirely dependant on data from the two dose levels that bracket the BMD (Gephart et al., 2001). Therefore, the BMD only really uses information about the nature of the dose–response curve in the region of interest, rather than information about the shape of the whole dose–response.

3.7. NOAELs are not compatible with the use of PBPK models

The BMD philosophy is closer to that of a PBPK model (e.g., Schlosser et al., 2003).

3.8. What if there is no NOAEL in a study?

A BMD could be estimated in this situation. However, it is likely that the lowest dose would be above the BMD or BMDL, in which case this relies on a potentially dangerous extrapolation that will depend highly on the specific dose–response model used. This would present all the difficulties of low-dose extrapolation of carcinogens (e.g., Edler and Kopp-Schneider, 1998), though in a milder form.

3.9. NOAELs favour bad experiments

It is generally supposed that the BMD is neutral with respect to study quality (variability and numbers of animals). It is clear that BMDLs favour "good" experiments (e.g., more animals, better experimental conduct, and statistical design). But every argument has a reverse side—is it really a good thing for a method to reward the use of unnecessarily large numbers of animals in experiments? Fortunately, animal welfare regulations should curb any pressure in this direction.

4. Technical difficulties with the BMD approach

4.1. How to measure effect size—quantal data?

To use the BMD approach we first need to decide on what units to use for the response variable, i.e., the

y-axis on the dose–response curve. For quantal data, the main options are additional risk and extra risk. If the proportions responding in the control and treated groups are $P(0)$ and $P(d)$ then additional risk is

$$P(d) - P(0)$$

and extra risk, is calculated from

$$\frac{P(d) - P(0)}{1 - P(0)}.$$

Familiarity with extra risk from carcinogen assessment in the USA tends to favour the adoption of this scale for BMD calculation, though traditionally toxicology uses additional risk, and epidemiology uses extra risk. Slob and Pieters (1998) argue that neither of these measures is meaningful for quantal data in the context of risk assessment. Variability determines the slope of the dose–response for quantal data (without variability the dose–response would be a step function), and much of this is inter-animal variability amongst an inbred laboratory strain, which is not relevant to human risk. They argue that only the average animal response is of interest (and the uncertainty in this average response), since inter-human variability is accounted for elsewhere in the calculation of a reference dose. From this line of argument, only the dose at which half of the animals respond is meaningful. This argument has merit, but goes against most current practice in regulatory toxicology, and presents practical difficulties if less than half of animals respond at the top dose. It also fails to recognise that there can be useful toxicological information in the slope of the dose–response for quantal effects. For example, acute directly acting toxins tend to result in steep dose–responses, whereas toxins acting through complex and long chains of events tend to result in flatter dose–responses. This remains a controversial area, with no obvious "best answer."

4.2. How to measure effect size—continuous data?

For continuous data, with the control and treated group means $M(0)$ and $M(d)$ respectively, we could use the absolute difference

$$M(d) - M(0)$$

the percentage effect

$$\frac{M(d) - M(0)}{M(0)} \cdot 100$$

the difference expressed in control standard deviations (Crump, 1984), against which strong arguments have been made (Murrell et al., 1998)

$$\frac{M(d) - M(0)}{\sigma(0)}$$

or the difference expressed in relation to the full range of effects (Murrell et al., 1998), against which powerful arguments have also been made (Crump, 2002)

$$\frac{M(d) - M(0)}{M(\max) - M(0)}.$$

No clear consensus exists as to which is preferable. A further quite different option has also been proposed, which is based on modelling the probability that a treated animal's response is greater or less than a control animal's response (Bosch et al., 1996).

No doubt this is not an end to the list of possible approaches. This diversity is not easily reconciled with the needs of regulatory processes, which lean so heavily on convention and precedent.

4.3. Difference in approach between quantal and continuous data

There has been a concern that the use of different measures of effect size for quantal and continuous data will lead to BMDs based on one being systematically lower than those based on the other. For this reason it might be suggested that continuous data be degraded to a quantal scale to ensure uniformity. However, this discards valuable information in the dataset and has undesirable consequences (Gaylor, 1996). Crump (1995) showed that there is a certain relationship between expressing continuous data as a difference relative to control variation in the one hand, and the extra risk approach used for quantal data on the other. This concept was developed further with the "hybrid approach," which could be thought of as converting continuous data to a scale of extra risk, without having to degrade the data (Crump, 2002; Gaylor and Slikker, 1990; Sand et al., 2003).

4.4. Choice of dose–response model

A huge range of dose–response models is available, and it is not the purpose of this document to review these (see Filipsson et al., 2003). Simplistically, as long as the model fits the data well it should not much matter which model is chosen. But, if the BMD or BMDL is below the lowest dose or higher than the highest dose, then the choice of model will be critical, and the result may well depend more on the choice of model than on the data. BMDLs will frequently be lower than the lowest dose, especially if a low critical effect size is chosen.

There seems to be a consensus that it would be advantageous to restrict model choice, and several suites of models have been proposed for continuous data (e.g., Crump, 2002; Slob, 2002). These suites of models offer a range of flexibility of curve shape, may be constrained in biologically meaningful ways (e.g., monotonic responses, preventing negative values), and may vary in the

number of fitted parameters, so that the most parsimonious model can be selected. In particular, nested models simplify model selection (Slob, 2002). For quantal data, the typical models used are not nested and many have the same number of parameters. One solution is to combine BMD estimates from different models, weighted according to the quality of fit of each model to the data (Garcia and Setzer, 2003). Use of a defined suite of models is helpful in a regulatory context, as it should prevent the use of models that are too flexible for the datasets they are being used on—this can lead to unreasonably extreme BMD estimates.

Biologists tend to argue for the use of biologically meaningful models (e.g., Health Council of the Netherlands, 2003). But as long as there is no extrapolation outside the range of tested doses, then an empirical curve that smoothes the data and interpolates between the doses, without overfitting the data, is all that is needed. The data are by definition biologically meaningful, so any curve that fits the data well is also biologically meaningful.

4.5. Estimation method for BMDL

Having chosen a measure of effect size, and chosen a dose–response model, there are several ways to proceed to estimate the BMDL, even for a simple randomised study design (e.g., see Gephart et al., 2001). The estimation of limits for parameters is a science in itself. A software package for regulatory use would settle on a single approach for a given situation.

4.6. Study design factors and missing values

Current benchmark dose software requires the input of the arithmetic mean and standard deviation for each treatment group, and may also require the number of animals in each group. Where the study design contains factors other than treatment, e.g., replicates, these inputs may not be appropriate summaries of the data. As an example, Table 1A contains data for plasma potassium in males at termination in a 28-day rat study. The study consisted of three treatment groups and one control group each with five males per group. Statistically significant reductions were seen at 100 (9%) and 1000 (21%) mg/kg/day using a two-sided Student's t test. The experimental design consisted of five replicates or blocks each containing one male per group. Fig. 1A shows the individual values by replicate with the number indicating the treatment group ($1 =$ control, $2 = 10$ mg/kg/day, $3 = 100$ mg/kg/day, and $4 = 1000$ mg/kg/day). Clear effects are seen at 1000 mg/kg/day with the 100 mg/kg/day animals also being slightly lower in each replicate.

The use of the replicate structure balances known external sources of variation, e.g., post-mortem personnel, order of bleed across treatment groups. Animals in each replicate may be processed by the same individual and/or procedures carried out in replicate order. This example has no replicate effect, but let us assume that the order of bleed has an effect on potassium levels and that this effect will lead to increased potassium levels over time. Animals are bled in replicate order. Treatment group differences are still clearly visible (Fig. 1B). The means in each group have increased but the difference between each group has stayed the same. However, the variability within each group has increased and the standard deviations for each group are significantly higher (Table 1B). The t test now fails to identify any statistically significant differences. The standard deviation contains not only the inherent variability in the data but also a component associated with the experimental design. For the purposes of statistical analysis, the data need to be analysed using analysis of variance rather than a two-group t test. The analysis of variance calculates variability associated with each experimental design factor and uses only the residual variability to compare between treatments. Similarly for dose–response modelling, use of the standard deviation can grossly overestimate the study variability and give misleading estimates of BMD and BMDL.

Missing values are a common feature of toxicity studies, e.g., if an animal does not survive to the end of the experiment. The effect of missing values, particularly in small studies, can be significant. Table 1C is the same data as in Fig. 1B, except the control animal in replicate 5 died before the end of the study. The loss of the replicate 5 control animal, where all values were higher due to the experimental design, has led to a reduction in the control group arithmetic mean. This has reduced the treatment effects at 100 and 1000 mg/kg/day from 9 and 21 to 2 and 12%, respectively. This is not due to any change in the values collected but reflects one missing control group value in the presence of an important experimental design factor.

Table 1
Male plasma potassium levels from a 28-day rat study

	Dose level (mg/kg/day)			
	0	10	100	1000
(A) All data—no replicate effect				
Mean	3.84	3.90	3.50	3.02
SD	0.21	0.27	0.14	0.22
N	5	5	5	5
(B) All data with replicate effect				
Mean	4.92	4.98	4.58	4.10
SD	0.86	0.74	0.68	0.96
N	5	5	5	5
(C) Incomplete data with replicate effect				
Mean	4.68	4.98	4.58	4.10
SD	0.77	0.74	0.68	0.96
N	4	5	5	5

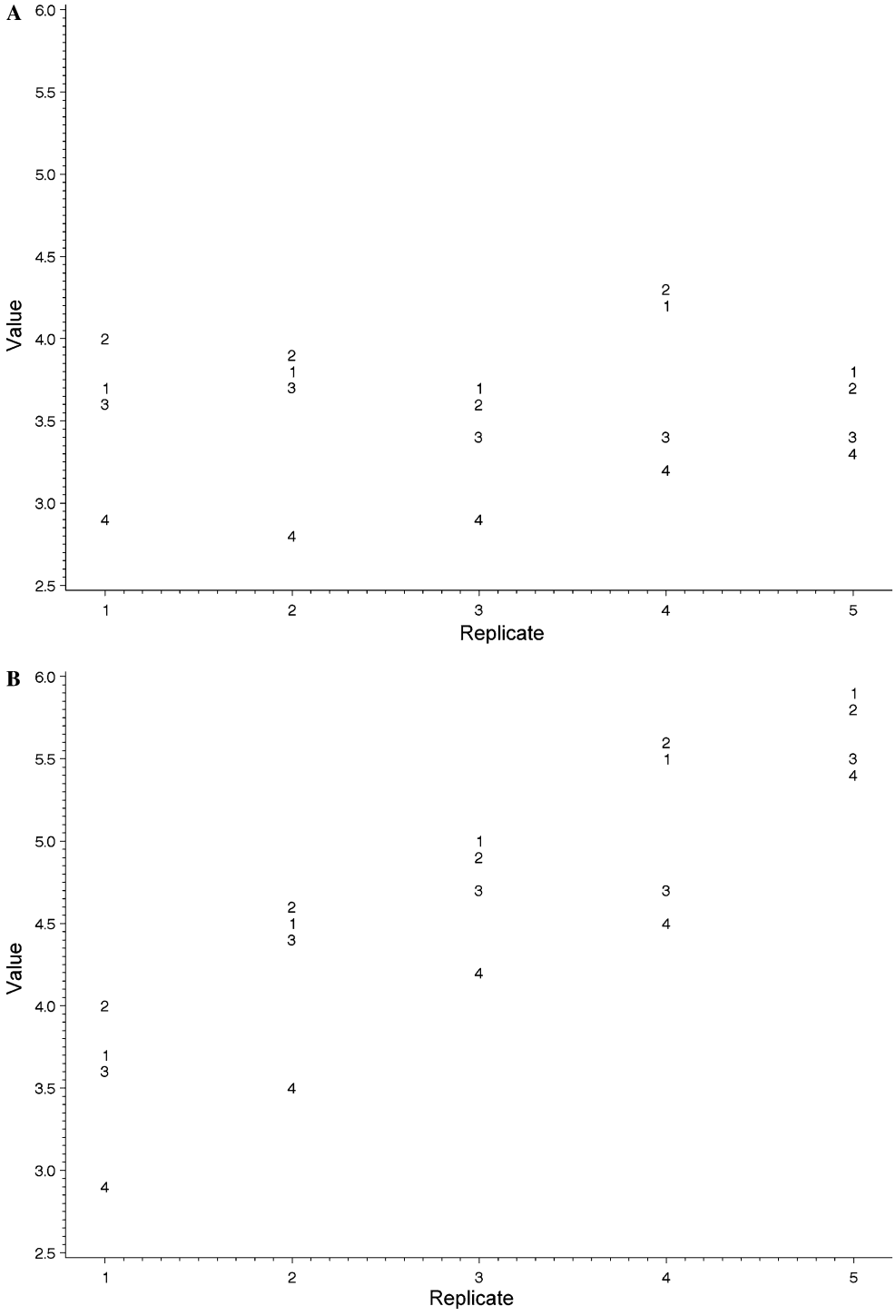

Fig. 1. (A) Individual values for male plasma potassium—no replicate effect. (B) Individual values for male plasma potassium—with replicate effect.

Gaylor et al. (1998), provide another example—the need to account for litter effects in estimating BMDs for reproduction studies. Methods to allow for all these study design and complicating factors exist, but cannot be performed with only summary statistics. Estimates of effect size based on simple summary statistics may be misleading.

4.7. Estimating a distribution of BMDs

It has been proposed that a reference dose (RfD) could be estimated using a distribution of BMDs, rather than a point estimate (Slob and Pieters, 1998). One approach to this is to use the bootstrap method (see e.g., Manly, 1997). Bootstrapping involves repeated sampling

Table 2
Female plasma phosphorous levels from a 90-day rat study

	Dietary concentration (ppm)			
	0	1	10	100
Mean	5.93	6.22	6.26	7.15
SD	0.57	0.98	0.77	0.94
N	12	12	12	12

from the observed data to produce simulated studies. For the example study in Table 2, the control group contained the values

5.1, 5.3, 5.4, 5.5, 5.9, 5.9, 6.0, 6.0, 6.1, 6.3, 6.4, 7.2.

A new simulated control group is generated by repeated random selection from these values, e.g.

5.3, 5.4, 5.4, 5.4, 5.5, 5.9, 6.0, 6.3, 6.4, 6.4, 6.4, 7.2.

A simulated sample is generated for each group to form a new simulated study, for which a BMD is estimated. This process is then repeated a large number of times to produce a distribution for the BMD. This procedure is trouble-free for some datasets, but potential problems can be illustrated by the test dataset in Table 2.

The inherent variability in the data can be seen from the results of 500 bootstrap samples (Fig. 2), where the distribution of the percentage difference from control is shown for each treatment group. The centre of each box represents the average effect size, i.e., 5, 5, and 20% in the 1, 10, and 100 ppm groups, respectively.

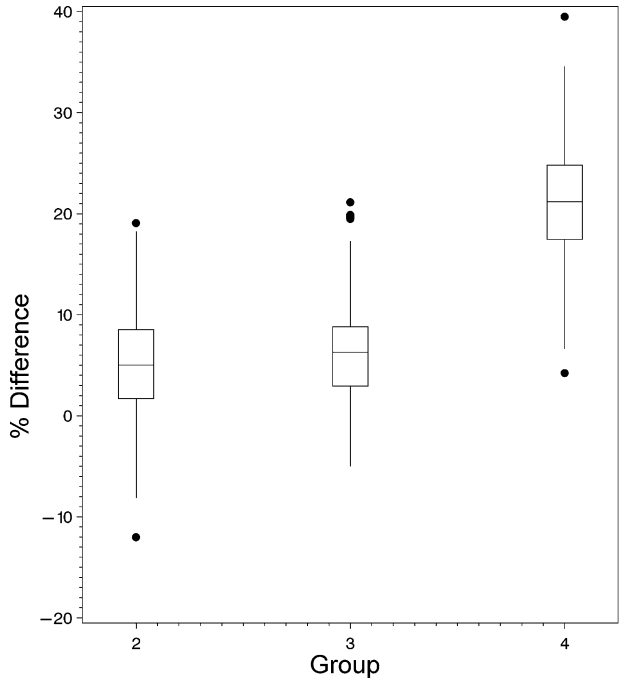

Fig. 2. Variability in effect size generated from 500 simulations (effects shown as % difference from the control).

The boxes contain 50% of the sample effect sizes and the whiskers contain 98%. The most extreme individual effects outside this range are shown as dots. For the 100 ppm group, the range of effect sizes seen varies from 4 to 40% and at the 1 ppm level effect sizes vary from a 12% reduction to a 20% increase.

This level of variability causes difficulties in estimating a distribution for the BMD, even when making a point estimate of the BMD is trouble-free. The Hill model, shown below, was chosen for the example data and it provided an adequate quality of fit (Fig. 3)

$$M(d) = \frac{\text{intercept} + v * \text{dose}^n}{\text{slope}^n + \text{dose}^n}.$$

Defining the critical effect size to be 10%, the BMD is estimated to be 29 ppm, whilst the BMDL is much lower at 0.9 ppm—these compare to a NOAEL of 10 ppm in this study. Moving now to estimating a distribution for the BMD, no estimate was possible from 15% of the simulated studies because a 10% effect was not achieved for these simulations. If the whole distribution of BMDs is taken forward to estimate a reference dose on a probabilistic basis, how should these 15% be taken into account, given that excluding them results in bias (Brand et al., 2001)? For those simulations where the BMD could be estimated, the distribution of BMDs is shown in Fig. 4. The distribution of BMDs is far broader than the neat examples typically shown in the literature, with values ranging from lower than the LOAEL to above the top dose studied. Although some of the spread in the BMD distribution may have been due to the choice of very flexible Hill model, use of a different model would not result in a fundamentally different result.

The use of bootstrapping is not straightforward. The Hill model provided an adequate fit to these study data, but could not be automatically fitted to the simulated study data, as it did not provide a good fit to all of the simulations. Consequently, the simulation process needed significant thought, and each simulation needed checking and possibly re-fitting using a different model. This makes it extremely time consuming to carry out routinely. Alternatively, should a single dose–response model be fitted to all simulations irrespective of the quality of fit?

These issues seem not to have been discussed in the literature. The common example studied has been cholinesterase inhibition, using datasets with a very sharp dose–response, and a massive range of observed effect sizes relative to study variability and to critical effect size. But study results showing small ranges of effect size are much more commonly seen in practice. Welfare concerns are also tending to limit effects at the top dose to being not much greater than that deemed critical for risk assessment (i.e., the critical effect size), as the concept of the maximum tolerated dose is increasingly challenged.

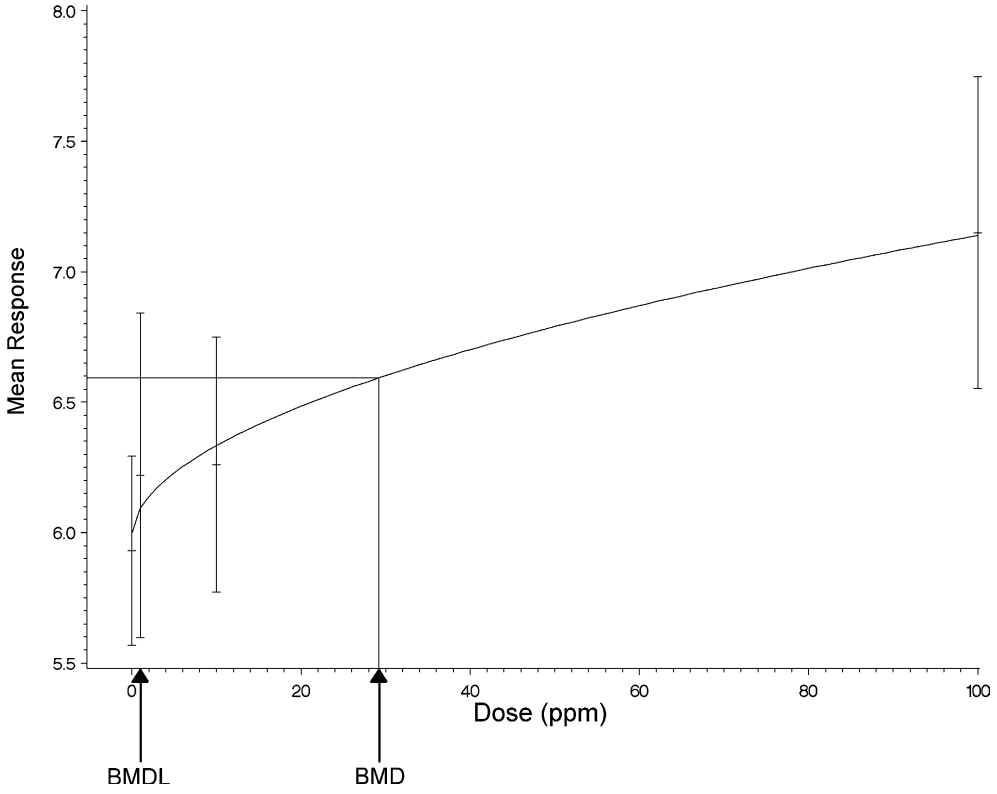

Fig. 3. Fit of the Hill model to data in Table 2.

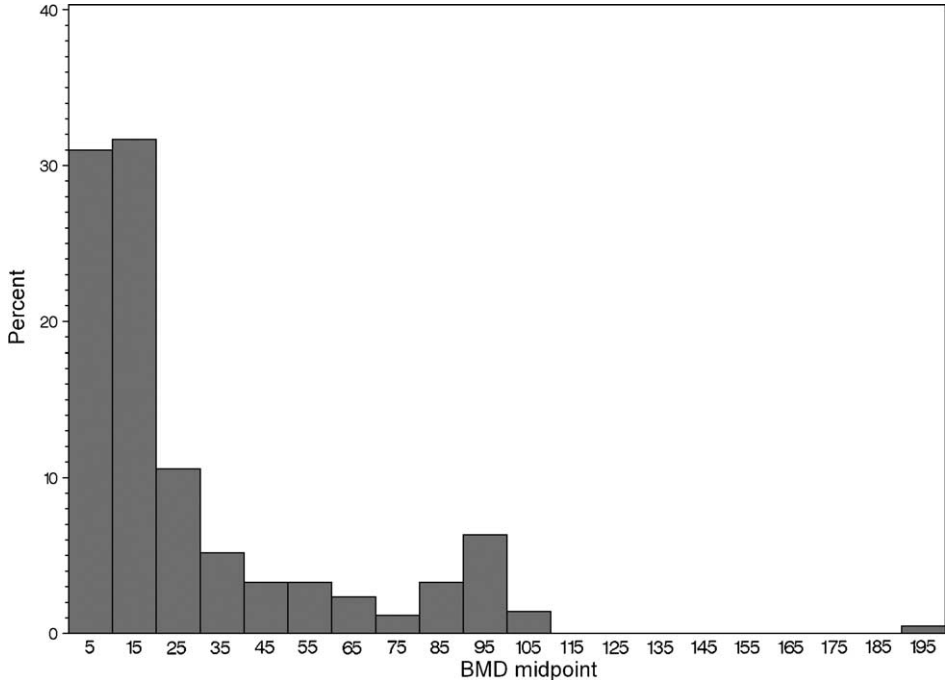

Fig. 4. Estimated distribution of BMD from 500 simulations (excludes 15% of simulations where BMD could not be estimated).

5. Regulatory difficulties with the BMD approach

As well as technical issues, there are a number of issues to do with the suitability of the BMD approach for use in regulation.

5.1. Choosing a critical effect size

It is proposed that the explicit setting of a critical effect size as the basis for a BMD is an advantage over the NOAEL approach (e.g., Slob and Pieters, 1998).

Technically this is true, but in a regulatory setting it can be problematic and a practical disadvantage. Given that there isn't international regulatory consensus on what *effects* are to be considered adverse, and what are not (Lewis et al., 2002), it is hard to imagine that international consensus could ever be arrived at in determining exactly what *magnitude* of each effect is to be considered adverse.

Without a predetermined regulatory critical effect size for each effect, the routine operation of the BMD method becomes very difficult in a regulatory context because risk management becomes mixed up with risk assessment. It is helpful in the regulatory process for the registrant to analyse study results as a part of reporting them, and for all the risk assessment to be done before results go to regulatory risk managers for their interpretation as a late step in the process. This division between risk assessment and risk management is familiar in US regulation, and has more recently become institutionalised in some European countries. For the risk manager to need to choose critical effect sizes for multiple variables in multiple studies, long before he/she is familiar with the data or issues, and long before ordinarily becoming involved in the process, is difficult. In practice, this means that BMDs will have to be recalculated, perhaps more than once, very late in the regulatory process. This will consume a lot of time, especially if summary statistics are not sufficient to perform the calculations.

5.2. Choosing a probability for the BMDL

Unless standardised, choosing a probability for the lower confidence limit of the BMD (to derive a BMDL) is also problematic. The need to choose a probability of exceeding the critical effect size is the conceptually equivalent problem in the case of the hybrid approach.

5.3. Credibility of the BMDL

In practice the BMDL can be below a study dose where no effect was observed. Although this can be rationalised in scientific terms, it stretches the credibility of the method. If regulatory action is taken on the basis of such a BMDL, the registrant is effectively being told "you didn't see anything at this dose, but if you had used more animals we think you *might* have seen an effect."

5.4. Barrier to harmonisation

Faustman (1996) retells an anecdote about the great variation in NOAELs determined by five OECD member countries in reviewing the same developmental and reproductive toxicity databases, with the NOAELs commonly varying by a factor of 20 or more. It is suggested that the use of BMD methodology would address much of this problem. However the need for international agreement on various aspects of BMD methodology, especially the choice of critical effect size, would seem to increase the barriers to international harmonisation. There are certainly many more ways to calculate a BMD than there are to calculate a NOAEL.

5.5. Different software gives different answers

In the hands of an expert, with a suitable dataset, differences in BMD derived from different software packages should be small (e.g., Filipsson and Victorin, 2003). But any time that calculations result in multiple answers to the same question this is not likely to be conducive to the smooth operation of the regulatory process.

5.6. Not always possible to get an answer

For a variety of reasons, it is not always possible to estimate a BMD from a given dataset (e.g., Clewell et al., 1997). If at all frequent, this would be undesirable in a regulatory context and from an animal welfare perspective.

5.7. Need for additional expertise

Even standardisation of BMD software packages does not remove the need for expertise, both toxicological and statistical, to arrive at sensible answers (Filipsson and Victorin, 2003). This expertise requirement is certainly greater than for the NOAEL approach, which for simple datasets can be packaged into a hand-turning technique. The need for additional statistical expertise in institutions conducting studies, and in regulatory authorities, is a disadvantage of the BMD approach (Barnes et al., 1995), but not an insurmountable obstacle if the benefits are sufficient.

6. Overall benefits evaluation

Having analysed all factors, the main ones remaining to distinguish the BMD and NOAEL approach are summarised in Table 3. It can be seen that there is a long list of disadvantages of introducing the BMD approach that could be remedied with further progress, especially with agreement on standard approaches. But some of the current points of disagreement are so contentious, that there is likely to be a price to pay for achieving consensus. The effects of this can already be seen. For example, the difficulty of agreeing a different critical effect size for each effect has resulted in suggestions to set a universal critical effect size for all effects (Barnes et al., 1995), which eliminates a potential advantage of the method, i.e., that the severity and nature of the effect can be

Table 3
Summary of advantages and disadvantages of moving to the BMD approach

Advantages
Set on a continuous scale
Choice of critical effect size can reflect the severity and nature of each effect
Ratios of BMD more informative than ratios of NOAELs (does not apply to BMDLs)
Does not favour "bad" experiments
Disadvantages (fixable)
Unfamiliarity and ease of understanding
Lack of consensus regarding
•Effect size definition
•Choice of critical effect size
•BMDL probability
•BMDL estimation method
Difficulty of analysis for non-trivial study designs and missing values
Software standards
Barrier to international harmonisation
Disadvantages (not fixable)
Less intuitive appeal than NOAEL concept
Mixes risk assessment and risk management

reflected in the choice of critical effect size (Crump, 2002). Similarly, a criterion for the choice of critical effect size and probability for the BMDL seems to be that when tested against a range of datasets, the resulting BMDLs should centre around the NOAELs. This is only human nature, despite best intentions not to consider the NOAEL to be a "gold standard" (Barnes et al., 1995). Indeed, much of the reason why BMDL is favoured by many over BMD seems to be that it gives values closer to the NOAEL. Using BMDLs that are close to NOAELs enables mixed data packages, i.e., collections of studies including ones for which the BMDL method can be applied and others for which it cannot, to be handled in a pragmatic way using both techniques. Yet ratios of BMDLs are far less consistent than ratios of BMDs, so another key advantage of moving away from NOAELs—more consistent decision making—would be largely dispensed with if BMDs are abandoned in favour of BMDLs.

Gephart et al. (2001) have shown that the BMD can in practice become little more than a complex way of interpolating between two points. The watering down of the method necessary to achieve consensus degrades the benefits still further. The remaining advantages are the possibility that the NOAEL might encourage "bad" experiments, which should already be under control in a regulatory context, and the fact that BMDs are set on a continuous scale. This latter advantage is normally intangible, but becomes important if a probabilistic BMD is determined and used to set a probabilistic reference dose (e.g., Slob and Pieters, 1998).

7. Concluding remarks

Five important criteria for evaluating the BMD or BMDL as a possible replacement for the NOAEL in regulatory risk assessment are as follows.

- Does it give the risk manager more certain information on which to base decisions (Barnes et al., 1995)?
- Does it result in more consistent decision-making?
- Does it result in public health benefits (Barnes et al., 1995)?
- Does it help the user?
- Does it work for most datasets?

It is hard to answer yes unequivocally to any of these questions. Overall the balance of advantages and disadvantages in a regulatory context is not favourable, which might explain the slow uptake of the method in this community. Advocates of the BMD would point out that the playing field is not level—the BMD is being compared with the NOAEL in the context of study designs are not suited to the BMD. This is certainly true. If the NOAEL method were abandoned, studies might be able use more treatment groups and far fewer animals per group, defining the shape of the dose–response much better than now yet with few animals in total which would be a valuable welfare benefit. But the power of t tests on such datasets would be very low, so we would be committed to the BMD with no hope of reverting to a normal NOAEL interpretation if necessary—a high-risk strategy. Perhaps this is why the study design modifications so far evaluated to fit the BMD method have been relatively minor modifications of conventional designs (e.g., Woutersen et al., 2001). In practice, current regulatory datasets are a mixed collection of studies reasonably well suited to the BMD method, and those that are completely unsuited. Examples of the latter are many studies on dogs and other non-rodent species, where the number of animals per group is typically very low, but the studies are highly valued by regulators. With mixed datasets of this kind, the dose–response assessment may need both NOAEL and BMD methodologies to be used.

Degrading the BMD technique to gain consensus for its wholesale use is not in the end helpful, as is dilutes the advantages as well as the disadvantages. Perhaps it is time to recognise that the BMD will never entirely replace the NOAEL in general use. It is has been pointed out that the BMD method must be applied to all variables in a study, because you cannot assume that the most sensitive endpoint is the same when gauged by the NOAEL and BMD approaches (Gephart et al., 2001). Presumably on the same basis, every study in a regulatory package ought to be evaluated with the BMD methodology. If the BMDL were used for regulation, this means the study with least informational value

in the whole regulatory package (e.g., fewest animals) is quite likely to be the one with the lowest BMDL (Crump, 2002). Is it appropriate for the reference doses of chemicals to be determined according to the weakness of their weakest study?

Reducing the interpretation of a regulatory data package to an automated process is neither necessary nor helpful. Expert judgment (and NOAELs) can guide an evaluator to the most critical study and critical endpoint for a chemical, and at this point the BMD approach could be invoked. Using the BMD as a higher tier approach for the critical endpoint in the package circumvents many of the disadvantages of the BMD method, whilst retaining the advantages. A critical effect size only needs to be decided for one variable, for example. Limited resources can be focussed on a considered evaluation of the dose–response and statistical issues for this one dataset, resulting either in a point estimate, or in a distribution for use in probabilistic risk assessment. An example is the boron risk assessment proposed by Murray (1996), in which a pivotal toxicity study was identified (a rat developmental toxicity study), and this was then analysed in greater detail, including the use of a BMD approach. The US-EPA used a similar approach for the cumulative risk assessment for organophosphorous pesticides (EPA, 2001), though because of the number of chemicals involved the amount of BMD analysis needed was still prodigious.

Perhaps the ultimate weakness of the BMD method is that it is not a sufficiently radical departure from conventional practice. It appears to consider the whole dose–response, but does not actually do so. It purports to be a better basis for risk assessment, yet it does not help transform conventional safety assessment into a true risk assessment that addresses the magnitude and frequency of effects. Liberating the BMD or any other new methodology from the requirement to be able to completely replace the use of the NOAEL could lead to further fruitful technical developments.

It is time to recognise the conventional NOAEL approach as a valuable tool that provides a simple, repeatable, and checkable summary of effects in toxicology studies. Each study is just conducted in one laboratory, with one animal strain, under one set of conditions, so the danger of over-interpretation is real. More labour- and discussion-intensive methods such as the BMD can be used to best advantage as a higher tier approach, once the critical studies and endpoints have been identified across the entire data package.

Acknowledgments

Financial support for this work from the UK Department of Food and Rural Affairs (DEFRA) is acknowledged. The authors are grateful to John Doe for helpful discussions concerning the work reported here.

References

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. Regul. Toxicol. Pharmacol. 21, 296–306.

Bosch, R.J., Wypij, D., Ryan, L.M., 1996. A semiparametric approach to risk assessment for quantitative outcomes. Risk Anal. 16, 657–665.

Brand, K.P., Catalano, P.J., Hammitt, J.K., Rhomberg, L., Evans, J.S., 2001. Limitations to empirical extrapolation studies: the base of BMD ratios. Risk Anal. 21 (4), 625–640.

Brand, K.P., Rhomberg, L., Evans, J.S., 1999. Estimating noncancer uncertainty factors: are ratios of NOAELs uninformative?. Risk Anal. 19 (2) 295–308.

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

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.

Crump, K.S., 2002. Critical issues in benchmark calculations from continuous data. Crit. Rev. Toxicol. 32 (3), 133–153.

Edler, L., Kopp-Schneider, A., 1998. Statistical methods for low dose exposure. Mutation Res. 405, 227–236.

Faustman, E.M., 1996. Review of noncancer risk assessment: application of the benchmark dose methods. Prepared for the Commission on risk assessment and risk management. Available from: http://www.riskworld.com/Nreports/1996/risk_rpt/pdf/faustman.pdf.

Filipsson, A.F., Sand, S., Nilsson, J., Victorin, K., 2003. The benchmark dose method—review of available models, and recommendations for application in health risk assessment. Crit. Rev. Toxicol. 33 (5), 505–542.

Filipsson, A.F., Victorin, K., 2003. Comparison of available benchmark dose softwares and models using trichloroethylene as a model substance. Regul. Toxicol. Pharmacol. 37, 343–355.

Garcia, R.I., Setzer, R.W., 2003. A multimodel approach for calculating benchmark dose. Poster Abstract, Society for Risk Analysis 23rd Annual Meeting, Baltimore, December 2003, p. 69.

Gaylor, D.W., 1996. Quantalization of continuous data for benchmark dose estimation. Regul. Toxicol. Pharmacol. 24, 246–250.

Gaylor, D., Ryan, L., Krewski, D., Zhu, Y., 1998. Procedures for calculating benchmark doses for health risk assessment. Regul. Toxicol. Pharmacol. 28, 150–164.

Gaylor, D.W., Slikker, W., 1990. Risk assessment for neurotoxic effects. Neurotoxicology 11, 211–218.

Gephart, L.A., Salminen, W.F., Nicolich, M.J., Pelekis, M., 2001. Evaluation of subchronic toxicity data using the benchmark dose approach. Regul. Toxicol. Pharmacol. 33, 37–59.

Health Council of the Netherlands, 2003. Benchmark dose method: derivation of health-based recommended exposure limits in new perspective. The Hague: Health Council of the Netherlands publication no. 2003/06E.

Lewis, R.W., Billington, R., Debryune, E., Gamer, A., Lang, B., Carpanini, F., 2002. Recognition of adverse and nonadverse effects in toxicity studies. Toxicol. Pathol. 30 (1), 66–74.

Manly, B.F.J., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology, second edition. Chapman & Hall, London.

Murray, F.J., 1996. A human health risk assessment of boron (boric acid and borax) in drinking water. Regul. Toxicol. Pharmacol. 22 (3), 221–230.

Murrell, J.A., Portier, C.J., Morris, R.W., 1998. Characterizing dose–response: I: critical assessment of the benchmark dose concept. Risk Anal. 18 (1), 13–26.

Sand, S.J., von Rosen, D., Filipsson, A.F., 2003. Benchmark calculations in risk assessment using continuous dose–response information: the influence of variance and the determination of a cut-off value. Risk Anal. 23 (5), 1059–1068.

Schlosser, P.M., Lilly, P.D., Conolly, R.B., Janszen, D.B., Kimball, J.S., 2003. Benchmark dose risk assessment for formaldehyde using airflow modelling and a single-compartment, DNA–protein cross-link dosimetry model to estimate human equivalent doses. Risk Anal. 23 (3), 473–487.

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

Slob, W., Pieters, M.N., 1998. A probabilistic approach for deriving acceptable human intake limits and human health risks from toxicological studies: general framework. Risk Anal. 18 (6), 787–798.

U.S. Environmental Protection Agency, 2001. Preliminary cumulative hazard and dose–response assessment for organophosphorous pesticides: determination of relative potency and points of departure for cholinesterase inhibition. US.EPA, Washington DC. Available from: http://www.epa.gov/oppsrrd1/cumulative/.

Woutersen, R.A., Jonker, D., Stevenson, H., te Biesebeek, J.D., Slob, W., 2001. The benchmark 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.