TOXICOLOGICAL SCIENCES **84**, 167–185 (2005)
doi:10.1093/toxsci/kfi004
Advance Access publication October 13, 2004

A Statistical Evaluation of Toxicity Study Designs for the Estimation of the Benchmark Dose in Continuous Endpoints

Wout Slob,*[,1] Mirjam Moerbeek,† Eija Rauniomaa* and Aldert H. Piersma*

*National Institute for Public Health and the Environment (RIVM), 3720 BA Bilthoven, The Netherlands and
†Department of Methodology and Statistics, Utrecht University, 3508 TC Utrecht, The Netherlands

Received August 16, 2004; accepted September 26, 2004

The benchmark approach is gaining attention as an alternative to the No-Observed-Adverse-Effect-Level (NOAEL) approach. However, current guidelines for the design of toxicity tests are based on assessing a NOAEL. It has been suggested that the current study design may not be optimal for assessing a Benchmark Dose (BMD). To further investigate this we performed three simulation studies in which a large number of designs were compared, focusing on continuous endpoints. Four fictitious endpoints were considered, their underlying dose-response curves having a linear, sublinear, supralinear, or sigmoidal shape. In each simulation run the BMD was derived from a model fitted to the generated data, where the selection of the model was based on that particular data set (according to a formal likelihood ratio test procedure). Thus, the model used for deriving the BMD in a single generated data set may not be the same as the one used for generating the data. In this way, model uncertainty is taken into account as well. The results show that the performance of a design is, first of all, determined by the total number of animals used. Distributing them over more dose groups does not result in a poorer performance of the study, despite the smaller number of animals per dose group. Dose placement is another crucial factor, and to minimize the risk of inadequate dose placement, the use of multiple dose studies is favorable. As a concomitant advantage, the use of multiple doses mitigates the disturbing effect of potential systematic errors in single dose groups. However, for endpoints with large residual variation (CV $\geq$ 18%) there is a substantial probability of not detecting the overall dose-response, and this probability increases in designs with increasing number of dose groups. In such situations, six dose groups may be used as a compromise. Designs with high dose levels (i.e., associated with relatively high effects) are helpful in estimating doses with smaller effects (such as the benchmark dose), and it appears bad practice to omit higher dose groups to improve the fit at lower doses. The typical 28-day study design of four dose groups with five animals (per sex) may not be adequate to assess endpoints with large residual variation (CV $\geq$ 18%), both in assessing a benchmark dose and in assessing a NOAEL.

Key Words: **dose-response models; benchmark dose; critical effect dose; study design; dose levels; dose placement.**

The standardized toxicity tests as described in the OECD guidelines generally consist of one control group and three dose groups. Data obtained from such tests are used to determine a No-Observed-Adverse-Effect-Level (NOAEL) by statistically testing each dose group against the control. Several objections against this approach have been raised, and the benchmark approach has been introduced as an alternative to the NOAEL approach by Crump (1984). The benchmark approach estimates a dose-response curve on the basis of all available data, and from that curve it calculates a benchmark dose to be used as a point of departure (PoD) for deriving human exposure limits.

It has been suggested that the OECD design, being based on the NOAEL approach, may not be optimal for assessing a benchmark dose because of its limited number of dose groups. The benchmark dose is derived from the complete dose-response curve, which is influenced by both the number of dose groups and the number of animals per dose group. Woutersen *et al.* (2001) studied the effect of the number of dose groups by comparing the results of different designs in a real toxicity study. The use of real data appears favorable for obvious reasons. However, the problem is that the conclusions drawn from the results only hold for the particular situation in that particular study, while the number of possible situations (i.e., combinations of designs, shapes of dose-response relationships, and residual variance) is without bounds. Computer simulations offer the advantage that many more situations can be explored compared to the use of real toxicity data. Besides that, in computer simulations the 'true' dose-response relationship is known by definition, providing a reference for the results based on the (generated) data.

Various studies performed simulation studies that focused on quantal endpoints (e.g., Kavlock *et al.*, 1996; Kelly, 2001; Weller *et al.*, 1995). In this paper we focus on the efficacy of study designs in estimating the benchmark dose for continuous data. To cover various shapes of potential dose-response relationships, we include a linear, a convex, a concave, and a sigmoidal curve as representing the true underlying dose response.

[1] To whom correspondence should be addressed at National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA Bilthoven, The Netherlands. Fax: 31-30-2744475. E-mail: wout.slob@rivm.nl.

This paper describes three simulation studies. The first simulation study aims to gain further insight in the efficacy of various study designs by placing doses at the true effect sizes (which is a hypothetical situation, since in practice, the dose-response relationship is unknown in advance). The second simulation study compares study designs that aim to cover all four types of dose-response relationships, assuming they are not known in advance. Finally, the third simulation study quantifies the relationships between the number of dose groups/number of animals per dose group and the precision of the estimated benchmark dose, for various values of the residual variation.

MATERIALS AND METHODS

Definitions and terminology. We use the terminology introduced by Slob and Pieters (1998) for continuous endpoints. The Critical Effect Size (CES) is defined as a prespecified small percent change in the magnitude of (average) response that may be considered as a nonadverse effect size. Although the magnitude of this percent change may be chosen differently for different endpoints (based on biological arguments), we will use a value of 5% as a default. The associated dose is referred to as Critical Effect Dose (CED), and its lower (one-sided) 95%-confidence limit will be denoted by CED-L05 (or briefly, CEDL). Thus, the CEDL is analogous to the BMDL for quantal endpoints.

Models. We use a nested family of dose-response models $y = f(x)$ as introduced by Slob (2002):

- model 1: $y = a$; with $a > 0$
- model 2: $y = a \exp(x/b)$; with $a > 0$
- model 3: $y = a \exp(\pm(x/b)\, d)$; with $a > 0, b > 0, d \geq 1$
- model 4 $y = a\,[c - (c - 1)\exp(-x/b)]$; with $a > 0, b > 0, c > 0$
- model 5 $y = a\,[c - (c - 1)\exp(-(x/b)d)]$; with $a > 0, b > 0, c > 0, d \geq 1$

where y is the response and x is the dose. The parameter a is the background response (e.g., average body weight in the controls), and the parameter b is a scaling factor on dose ('absolute' potency factor). The parameter c determines the response at high doses for dose-response curves that level off. The parameter d can make the function more curved, to describe dose-responses that hardly change at low doses (threshold-like dose-response relationships). Models 2, 3, and 4 can be obtained from model 5 by setting $c = 0$ and/or $d = 1$. For models 3 and 5 we restrict the parameter d to be larger than one, since otherwise the slope at zero dose would be infinity, which might be implausible from a biological point of view. A likelihood ratio test may be used to test whether a more complicated model (i.e., an increased number of parameters to be estimated) gives a significantly better fit to the data. Once the model has been selected, the fitted model is used to calculate the CED for a given CES from

$$CES = \frac{f(CED) - f(0)}{f(0)}.$$

In this paper we use the likelihood ratio method (Crump, 1984; Moerbeek *et al*., 2004) to calculate the lower 95% confidence limit of the CED, the CED- L05.

Simulations

Curves used for generating the data. In each simulation study, four different endpoints are considered. These four endpoints show distinct dose-response relationships, which are indicated by numbers 2 to 5, corresponding to the model from which the response data were generated (see previous section). Parameter values were chosen such that the response at dose = 0 is equal to 1, and the response at the top dose (dose = 1) is equal to 1.3, (i.e., the effect size at the top dose is 30%). Furthermore, the parameters where chosen such that the four endpoints represent a linear, sublinear, supralinear, and sigmoidal shaped dose-response curve, respectively. They are shown in Figure 2 (left panels); for the parameter values of each endpoint, see Table 1. Note that the CEDs corresponding to a critical effect size of 5% are different, since the shapes of the curves differ. The same four curves (endpoints) were used in all simulation studies, with one exception: in simulation study 3 the curve for endpoint 5 was made less steep to make it more realistic.

TABLE 1
Parameter Values Used for Generating the Data

	Regression parameter				
Endpoint	a	b	c	d	CED
2 (linear)	1	3.82			0.186
3 (sublinear)	1	1.31		5	0.716
4 (supralinear)	1	0.167	1.3		0.030
5 (sigmoidal)	1	0.52	1.3	14	0.460
5b (sigmoidal)	1	0.52	1.3	4	0.340

Note. Endpoint numbers relate to the model numbers described in the methods section. Endpoint (model) 5 was used in simulation studies 1 and 2, while endpoint 5b was used in simulation study 3. The other endpoints were used in all simulation studies.

Generation of data and number of simulation runs. In the first two simulation studies, for each design and endpoint, 200 data sets were generated by random sampling from the true underlying dose-response curve. In particular, the data were randomly sampled from a lognormal distribution with median equal to the model value at that dose. In the third simulation study the number of runs was 500.

Model fitting to generated data. For each simulation run all models 1–5 were fitted to the generated data set, and the 'best' model was selected on the basis of the achieved maximum likelihoods for each model, taking the number of parameters in the model into account; i.e., a model with additional parameters was only accepted when it resulted in a significantly ($\alpha = 0.05$) better fit (likelihood ratio test). In this way, the model selected in each individual run depends on the generated data only, ignoring the 'true' model by which the data were generated. This mimics the real life situation, where the underlying dose-response 'model' is unknown, and the observed data are all we have. Hence, the results of our simulations include not only random sampling error, but also (to some extent) model uncertainty. Since the data were generated by assuming a lognormal distribution around the true dose-response curve, the log-transformed model is fitted to the log-transformed data. While in our view the lognormal rather than the normal distribution should be the default in analyzing biological (continuous) data, this is not a generally accepted strategy. One of the advantages of assuming a lognormal distribution for the residual variation is that it covers both skewed and nonskewed data as encountered in practice; in real data skewness is typically a function of the variation (more variable data being more skewed), which is also a property of the lognormal distribution. Dose-response data resulting from toxicity studies often show moderate scatter, and in that case, they will not be very skewed. Obviously, data with small scatter can be equally well described by the lognormal or the normal distribution. But skewed data cannot be described by the normal distribution, so the lognormal has a wider range of applicability. In addition, the log-transformation has the advantage that it tends to correct for heterogeneity of variances as often seen in biological data. Anyway, the simulation results that are presented here may still be useful for those who insist on assuming a normal distribution when the data are not clearly skewed.

Box-and-whisker plots. From the selected model the CED associated with a CES of 5% was calculated, and the CED-L05 was determined using the profile-likelihood (likelihood ratio) method (Crump, 1984; Moerbeek *et al*., 2004).

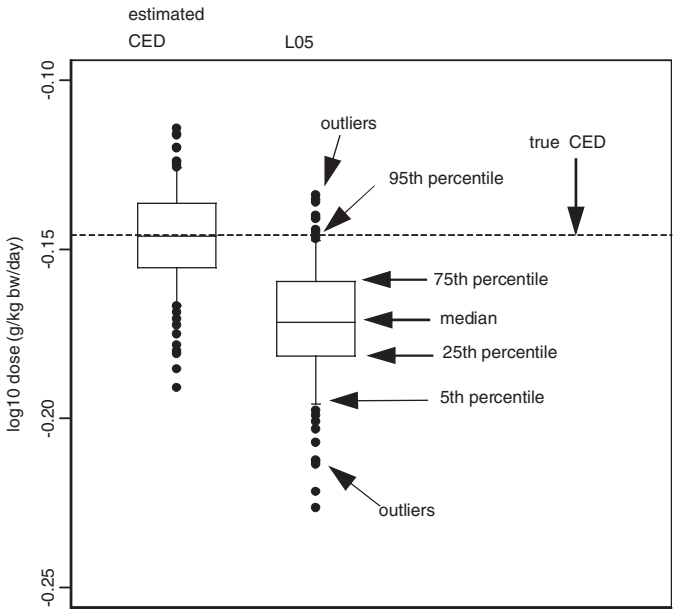

FIG. 1. Interpretation of the Box-Whisker plots used for expressing the results in Figures 2, 3, and 4. Note that this plot represents the ideal situation where the median of the estimated CEDs and the 95th percentile of the L05 coincide with the true CED, indicated by the dashed line.

Next, the resulting CEDs and CED-L05s from all runs were taken together and plotted in box-and-whisker plots. Figure 1 shows how these box-and-whisker plots should be read.

Further specifics. In all simulation studies we assumed that, on log-scale, the individual observations show homogenous variances around the dose-response function, resulting in a single value for the residual variance. This implies that on the original scale the (residual) Coefficient of Variation (CV) is homogenous (independent from the predicted response). In the first two simulation studies we assumed a fairly small residual variance (var = 0.001 for response on natural log-scale, equivalent to CV = 3%), fixing the total number of animals at 80. In the third simulation study the total size of the design was varied between 20 and 500 animals in total. The latter study also addressed larger residual variances, *viz.* CV = 10% and 18%. These values were chosen based on a collection of available results from dose-response analyses relating to all sorts of endpoints typically measured in toxicity studies.

Both the maximum response (30% change compared to background response) and the value of CES (5% change compared to background) were kept the same in all simulations. Regarding the former, the value of 30% was based on available dose-response modeling results as being a realistic value for the lower end of the range of maximum changes in response as encountered in various endpoints in toxicity studies. The fixed value of 5% for CES in all simulations may seem to be restrictive at first sight; for some endpoints one may want to choose higher values of CES (e.g., for serum enzyme levels). However, the overall study design is intended to pick up the smaller changes that are considered toxicologically relevant. A 5% change in response seems to be a reasonable choice for representing a low value for CES. This point is further discussed at the end of the discussion section.

SIMULATION STUDY 1

Theoretically, dose levels should preferably be chosen such that they result in different, gradually increasing (or decreasing) response levels. Hence, the first simulation study evaluates designs where dose levels were chosen in relation to effect sizes (for their values, see Table 2). Although, in practice, dose levels cannot be chosen in this way (the dose-response curve being unknown in the designing phase), this approach gives theoretical information on the influence of number of dose groups and dose placement with respect to the estimation of the CED. We studied nine designs, with the number of dose groups ranging from four to six (see Table 2 and Fig. 2). All designs include a control group. In discussing the results, doses associated with effect sizes smaller than 15% will be called low, and high when they are in the range 15–30%. Designs 1–3 correspond to both high and low doses, designs 4–6 correspond to low doses only, and designs 7–9 correspond to high doses only (see left panels of Fig. 2). Within each of these triplets the first design has four dose groups, the second five, and the third six (including the controls). The total number of animals is kept (approximately) the same in all designs, while all dose groups within a design are of equal size. This results in a total of 78 animals for the designs with six dose groups (n = 13), and 80 animals for the designs with four (n = 20) or five (n = 16) dose groups.

TABLE 2
Effect Sizes Used for Dose Placement in Simulation Study 1

	Design	Effect sizes (%) at selected doses					
Low and high doses	1	0	7.5	15	22.5		
	2	0	7.5	12.5	17.5	22.5	
	3	0	5	10	15	20	25
Low doses only	4	0	5	10	15		
	5	0	4	8	12	15	
	6	0	3	6	9	12	15
High doses only	7	0	15	22	30		
	8	0	15	22	25	30	
	9	0	15	18	22	26	30

Results

The results of simulation study 1 are shown in the right panels of Figure 2. In the majority of the resulting Box-and-Whisker plots the median estimated CED and the 95th percentile of the L05 are close to the true CED. The second triplet of designs (low doses) performs less well in the supralinear endpoint (4), and to a lesser extent in the linear endpoint (2), while the third triplet (high doses) performs best in both these endpoints. On the other hand, the third triplet of designs (high doses) perform less well for the sigmoidal endpoint (5); here, the designs with both low and high doses perform best. However, the high-dose designs in the sigmoidal endpoint (5) tend to underestimate the CED, while the low-dose designs in the supralinear endpoint (4) tend to overestimate it, which is a more serious error from the perspective of risk assessment. These results show that doses associated with higher effect sizes are informative and useful for the purpose of estimating a CED associated with lower effect sizes. This

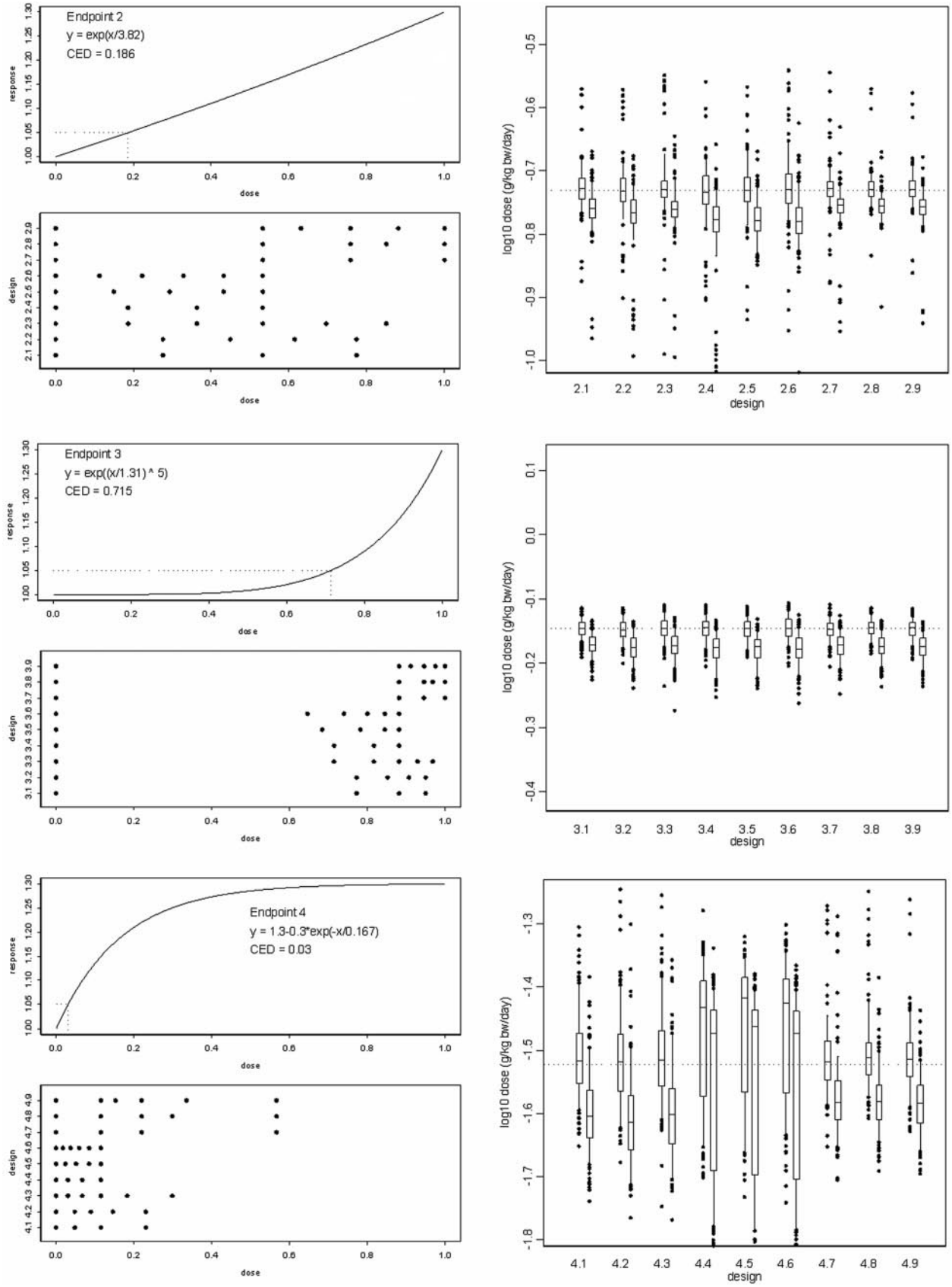

FIG. 2. Results of simulation study 1. Doses based on effect size. For explanation see text.

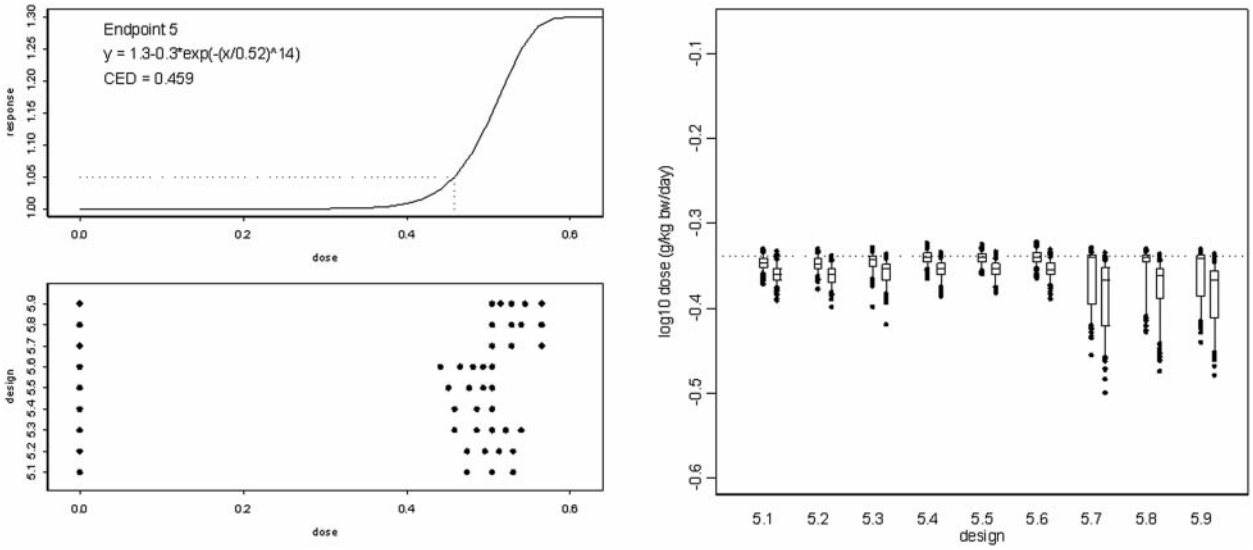

FIG. 2. Continued

indicates that choosing (or relying on) dose levels in the region of interest (i.e., lower doses) only may not be the best strategy.

The three designs within each triplet perform equally well. This implies that, given the same total number of animals, the number of dose groups had no particular impact on the performance in these simulations where dose placement was 'normalized' with respect to effect sizes.

Comparing the four endpoints, the linear endpoint (2) is somewhat exceptional in showing both low and high outliers. This can be explained by the fact that the more or less linear curve for this endpoint is in fact an intermediate between a sublinear and a supralinear curve. Hence, in simulation runs for which the dose-response data happened to be relatively sublinear or supralinear, it may happen that model 3 (sublinear) or 4 (supralinear) fitted significantly better, respectively. Since the decision to use a model with one more parameter was based on a likelihood ratio test with $\alpha = 0.05$, such a decision may be expected in about 5% of the simulation runs. Indeed, the number of outliers resulting from the simulations was very close to 5%, equally divided into low and high outliers, corresponding to selection of the sublinear model 3 and the supralinear model 4 in these cases, respectively. The best results were obtained for the sublinear and sigmoidal endpoints 3 and 5 (note that the vertical scales in the right panels of Fig. 2 are of equal length). Since these two endpoints have threshold-like dose response curves, this observation is not unfortunate, if the usual assumption that most (noncarcinogenic) compounds should result in threshold-like dose-response relationships is realistic. The supralinear endpoint (4) gives the poorest (least precise) results.

The reason that in the supralinear endpoint (4) the low-dose designs tend to overestimate the true CED is as follows. In these designs the dose-response is almost linear within the range of observation, and hence model 2 was often selected as the 'best' model. Model 2 having a sublinear rather than a supralinear shape will tend to overestimate the CED. The (high) outliers for the other two triplets of designs (4.1–4.3 and 4.7–4.9) correspond to simulation runs where model 5 (i.e., a sigmoidal curve) was selected as the 'best' model, resulting in too high CED estimates.

The fact that in (sigmoidal) endpoint 5 the third triplet of designs (high doses) performs poorest can be explained by the fact that the CED is here estimated by extrapolation: the CED is lower than the lowest nonzero dose level in these designs.

In (sigmoidal) endpoint 5 the best results were obtained with the second triplet of designs (low doses). Although the data were generated by (sigmoidal) model 5, (sublinear) model 3 was always selected as the best fitting model for these designs. This illustrates that selecting the 'right' model is not critical. What matters is that the selected model sufficiently mimics the true shape of the dose-response relationship (in the relevant range).

Comparing the results for the four endpoints shows that there is no single design that performs best for each endpoint. In general, designs with doses around the CED only (which might be intuitively the best thing to do) actually do not perform best; inclusion of higher doses improves the performance. Designs with both low and high doses appear to be the best compromise to cover different endpoints.

SIMULATION STUDY 2

When the four endpoints considered are imagined to relate to a single experiment, the study design should cover all four of them. The obvious compromise is to place the dose levels evenly distributed on the dose scale. This section discusses the efficiency of such designs with ten, six, or four dose levels, grouped as the first, second, and third triplet of designs,

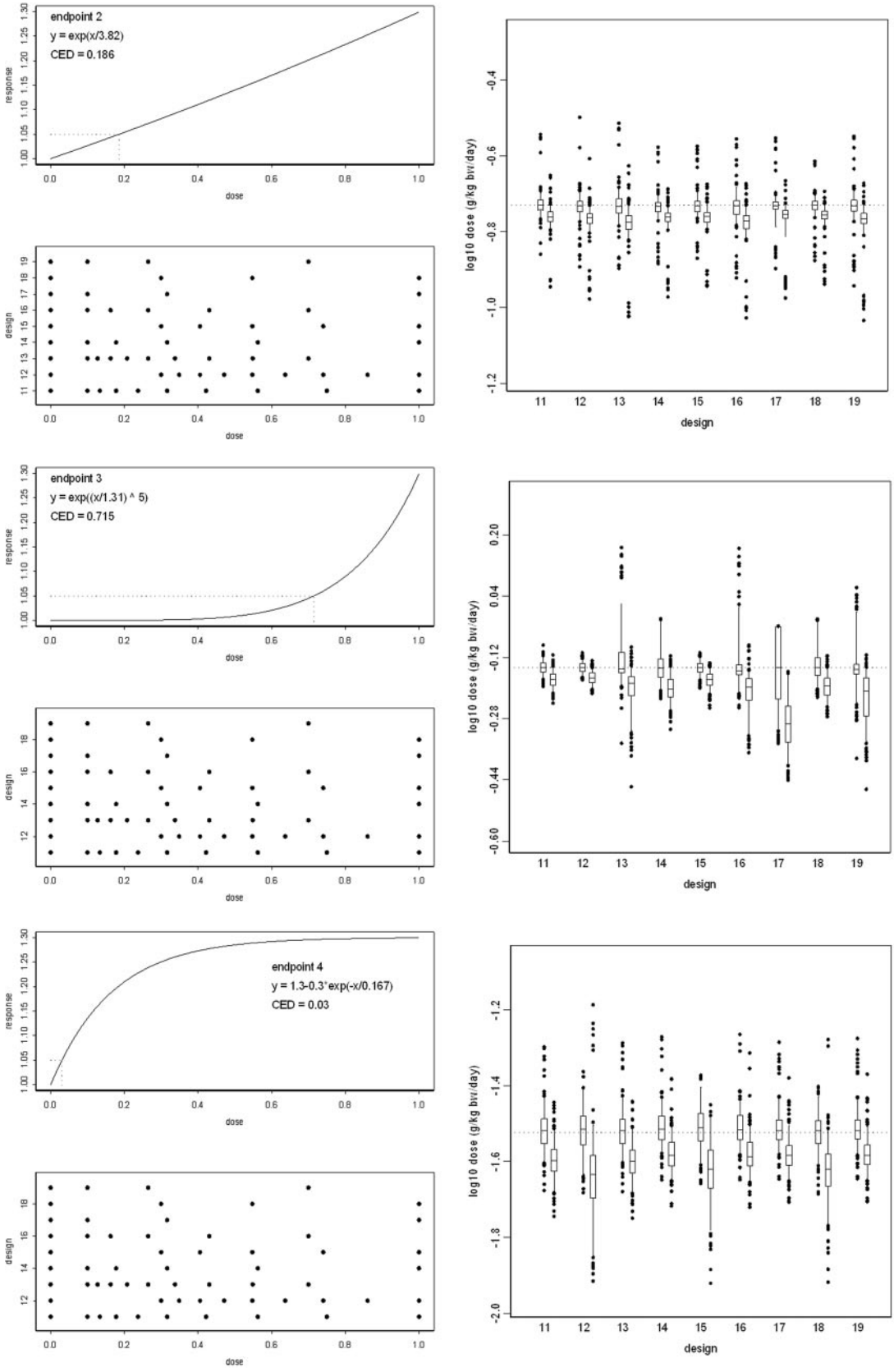

FIG. 3. Results of simulation study 2. Doses equidistant on log-dose scale.

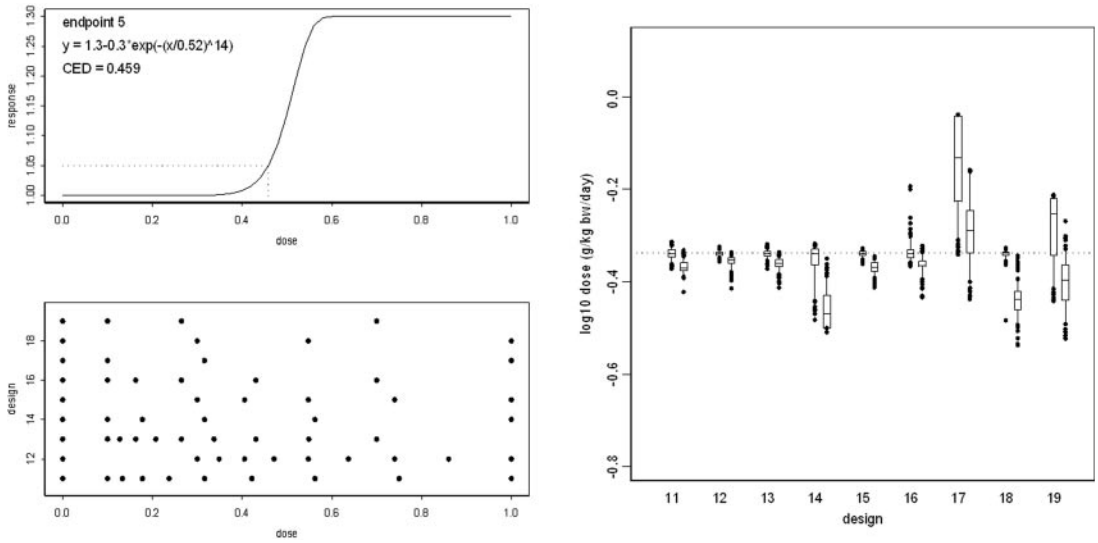

FIG. 3. Continued

respectively. In the first design of each triplet, the lowest dose = 0.1 and the highest dose = 1. In the second design of each triplet, the lowest (nonzero) dose is increased to 0.3, and in the third design of each triplet the highest dose is decreased to 0.7. Within each design the dose groups are of equal size, with a total of 78 animals for designs of six dose groups, and a total of 80 animals for the other designs. Each design contains a control group receiving dose zero. The other doses are equidistantly spaced between the specified highest and lowest dose, either on the log-dose or on the dose scale. The designs with doses equidistantly spaced at the log-dose scale are numbered 11–19, those on dose scale 21–29. The designs and results are given in Figures 3 and 4 for both dose scales, respectively. Note that the scales of the *y*-axis in the right panels in Figures 3 and 4 are different from those in Figure 2.

Results

For (linear) endpoint 2 the results are similar to those in simulation study 1. Again, the nine designs examined here perform almost equally well, while log-dose scaling or dose scaling does not make much difference. Apparently, for an approximately linear dose response, the design is not very critical (given a total number of animals). The outliers are due to simulation runs for which the data were generated such that model 3 (high outliers) or model 4 (low outliers) was selected as best fitting model (as was also found in simulation study 1).

For (sublinear) endpoint 3 the designs do give different results. For instance, the performance is influenced by the highest dose. There are more outliers in the estimated CED (and L05) if the highest dose is set at 0.7 (the third design of each triplet), in particular when doses are equidistant on dose-scale. In all these 'low-dose' designs the highest dose is near the CED, so that the highest expected effect is only just above 5%. Hence, the signal-noise ratio is unfavorable in these ('low-dose') designs, and indeed, in a small fraction of simulation runs model 1 (no response) was selected. This confirms the observation already made in simulation study 1 that doses associated with relatively high responses are helpful in estimating a dose associated with a relatively low response. Another pattern that shows up is that the performance tends to increase with increasing number of dose groups.

In (supralinear) endpoint 4 the number of dose groups and the placement of dose levels hardly influence the results. Some high outliers were obtained for designs due to model 5 being selected as the appropriate model. Designs with lowest dose = 0.3 resulted in some low CED-L05 values. In these designs there is, in terms of effect size, a large gap between the response in the controls and that in the lowest dose group.

In (sigmoidal) endpoint 5 the designs with four dose groups give poor results. The reason is that in these designs the placement of doses is more likely to be unfortunate. For instance, in designs 17 and 19, three doses fall in the first horizontal part of the curve, with only the fourth dose being associated with a substantial effect. Clearly, such data do not give any clue to a sigmoidal shape, and in all runs the model selected was model 3 with a strong curvature (i.e., starting to increase only near the top dose), leading to an overestimation of the CED.

Overall, equal dose spacing on the dose versus the log-dose scale did not make much difference. Only in (sublinear) endpoint 3, especially for the third triplet of designs, performance differed clearly. However, this may be explained by the fact that, in the designs with log-dose scaling, three out of the four doses happened to fall in the region where the effect size was very small. Furthermore, the results of this second simulation study confirm that a design with fewer dose levels is more likely to

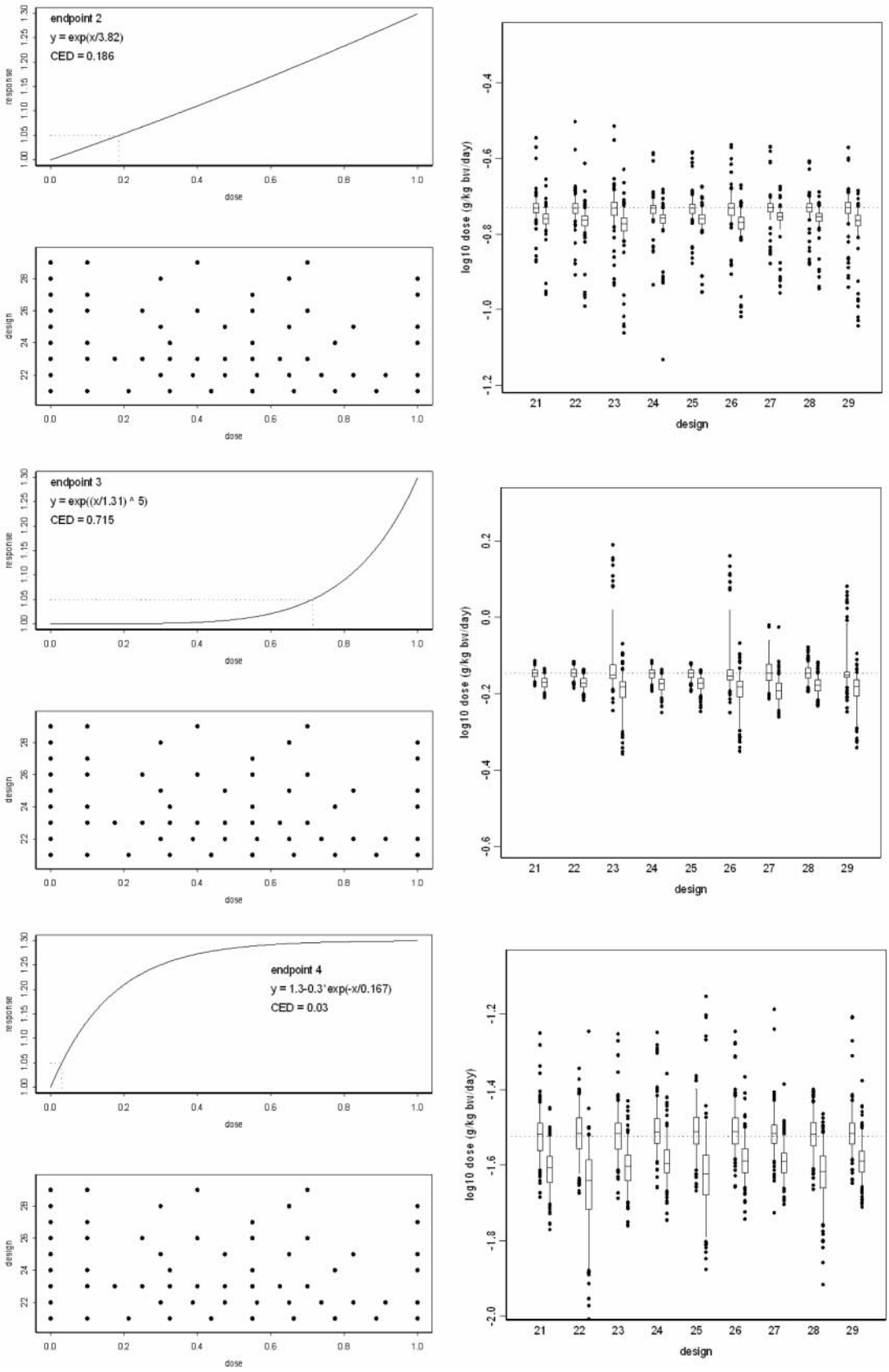

FIG. 4. Results of simulation study 2. Doses equidistant on dose scale.

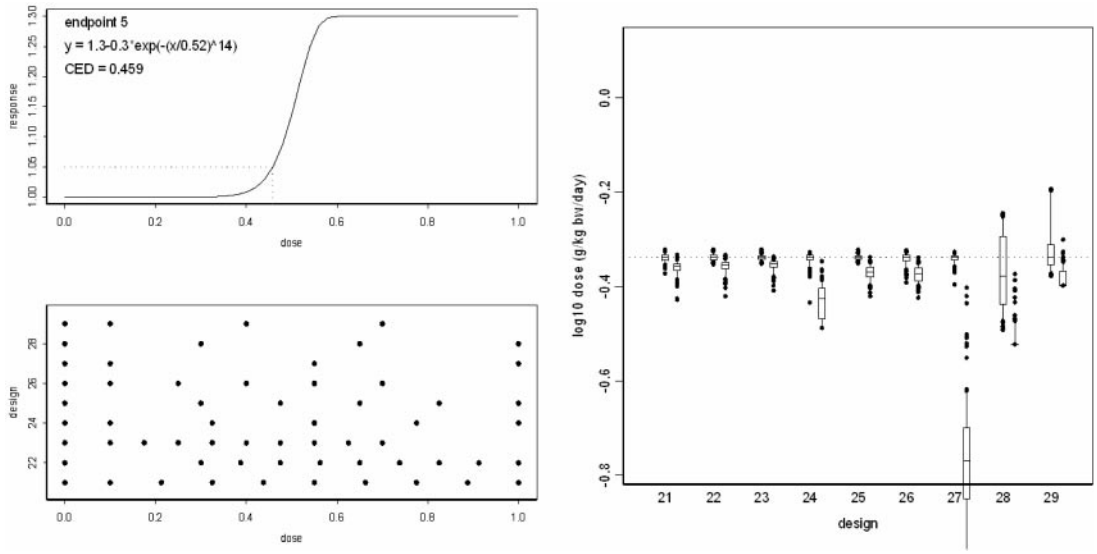

FIG. 4. Continued

give inadequate results, because it is more likely that the applied doses are poorly placed. Finally, it is found that the lowest or the highest dose level may influence the efficacy of a design. A good design is one where the highest dose is associated with a sufficiently large effect size, while the lowest dose is associated with an effect size that is not too large (preferably smaller than the CES for that endpoint). This implies that the dose range should not be too narrow, when differently shaped dose-response curves are to be covered in the same study.

SIMULATION STUDY 3

The aim of the third simulation study was to further quantify the precision of the estimated CED in relation to the number of dose groups and number of animals per dose group. While the first three endpoints (i.e., dose-response curves for generating the data) were left the same, the last one (endpoint 5) was modified and chosen to be less steep to make it more realistic (see Table 1).

A total of 17 designs were examined, each with a control group, a lowest nonzero dose of 0.1, and a top dose of 1. The designs had four, six, eight, or ten dose groups, placed equidistantly on log-scale between dose 0.1 and dose 1 (see Fig. 5), each with 5, 10, 20, or 50 animals per dose group, comprising a total of 16 designs. The 17th design consisted of six dose groups with 4, 3, 3, 3, 3, 4 animals, consecutively. This design was added as a counterpart of the 4×5 design (i.e., four dose groups of five animals) in having more dose groups but the same total number of animals. The counterparts of the 4×10, 4×20, and 4×50 designs are the 8×5, 8×10, and 10×20 designs. In this way each of the four-dose-groups designs has a counterpart with more dose groups.

These designs were examined for three different values for the residual variance (on natural log-scale): 0.001, 0.01, and 0.032. These values are equivalent to Coefficients of Variation of around 3%, 10%, and 18%. The choice of these values was based on a preliminary set of available results from toxicity studies analyzed by the benchmark approach, showing a median CV of around 10%. In this database, some endpoints (mostly biochemical measurements) showed a much higher residual variance, but these endpoints also tended to show much larger effect sizes than 30% (which is used as the maximum effect size in the underlying true dose-response relationships in our simulations).

For each combination of endpoint and design, 500 simulation runs were performed. The efficacy of the designs was quantified as follows. For each run the distance between the CED-L05 and the true CED was expressed as a factor, resulting in a set of 500 efficacy factors. We use the word efficacy here since the ratio of L05 to true CED includes both precision and accuracy (i.e., random error and bias). Figures 6–8 summarize these results (discussed below). In these plots the medians of the 500 efficacy factors resulting for each design is plotted against the total number of animals in the design. If possible, the 5th and 95th percentiles are plotted as well (as in Fig. 6) by vertical lines reflecting the range comprising 90% of the efficacy factors. Ideally, the 5th percentile should coincide with the horizontal line where the factor equals unity (i.e., the vertical lines should just touch the horizontal line). Note that the counterparts of the four-dose-groups designs (i.e., designs with more than four dose groups but the same total number of animals) are plotted after a small shift to the right, for reasons of visibility.

Not all simulation runs resulted in a quantitative value for the CED-L05. On the one hand, the generated data set could be such that model 1 (i.e., no response) was not significantly improved by

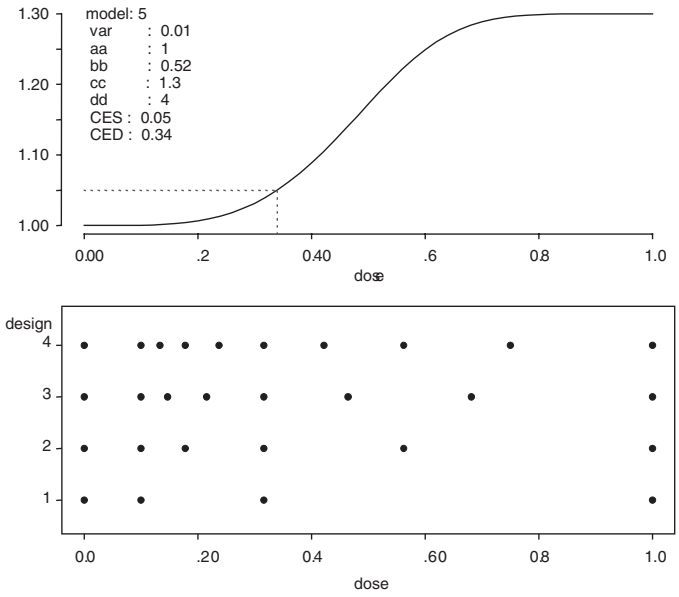

FIG. 5. Dose placement used in simulation study 3 (lower panel), together with the dose-response relationship for endpoint 5b (upper panel).

one of the other models. In that case the CED (as well as the L05) is in fact estimated to be infinite. On the other hand, the generated data could be such that the associated log-likelihood profile was extremely flat and/or did not reach the lower 95% confidence level (as indicated by the Chi-square criterion). In that case the CED-L05 is estimated to be extremely (or infinitely) small, i.e., (close to) zero. Figures 9–11 show the frequencies of runs resulting in CED-L05 being infinite or 'zero.' These figures are discussed below.

Results

The results will be discussed for the situation of small (var = 0.001), intermediate (var = 0.01), and large (var = 0.032) residual variation, consecutively.

Small residual variation (var = 0.001). For the case of small residual variation, hardly any runs resulted in infinite or zero CED-L05 values, and for all cases the 5th and 95th percentiles of the 500 efficacy factors could be assessed. As Figure 6 shows, design efficacy is dominated by the total number of animals in the study. Apart from that, the number of dose groups plays a role. In (sublinear and sigmoidal) endpoints 3 and 5, designs with fewer dose groups perform poorer, in particular designs with four dose groups. In (linear and supralinear) endpoints 2 and 4, designs with fewer dose groups perform better, but only slightly so. In some cases (smaller designs, i.e., with respect to total n) the vertical lines cross the horizontal line, implying that the nominal coverage of the CED-L05 is not reached. In (sublinear) endpoint 3, the four-dose-groups designs result in a higher than nominal coverage (i.e., the estimated CED-L05 is unnecessarily low).

For the situation of small residual variation, even the smallest designs (total number of animals = 20) perform quite well, the median efficacy factor being less than a factor of two for all four endpoints, and the 95th percentile being less than a factor of three. Considering all four endpoints together, the efficacy of toxicity studies would be improved by using designs with more than four dose groups, given a small residual variation.

Intermediate residual variation (var = 0.01). With intermediate residual variation the fraction of nonnumerical CED-L05s was quite small (2% or less) in all cases, except for (supralinear) endpoint 4, where the fraction of runs resulting in CED-L05 = 'zero' reached 20% of the runs for the designs with total $n = 20$. As Figure 9 shows, there appears to be some relationship with the number of dose groups in the study, smaller numbers of dose groups being slightly favorable (for a given total study size).

Because of the high fraction of nonnumerical CED-L05s for (supralinear) endpoint 4 in the smaller designs, the 5th and 95th percentiles could not be calculated, and for these designs only the median efficacy factors are plotted in Figure 7. Apart from resulting in higher factors (for obvious reasons), the results for (linear) endpoint 2 are similar to those in Figure 6. In (sublinear) endpoint 3, the median factors for the four-dose-groups designs are again higher than designs with more dose groups. However, this effect is not seen for the 95th percentiles of the factors. This is due to the fact that, especially in the 4×5 and 4×10 designs, model 2 was selected in a large fraction of runs, resulting in a relatively low estimate of the CED. A similar phenomenon occurred in (sigmoidal) endpoint 5, where the pattern found in Figure 6 (small residual variation) was even more disturbed, due to the fact that model 2 was selected in an even larger fraction of runs. This disturbance caused by selecting model 2 for many generated data sets is also reflected by various of the 90%-intervals being skewed to the left here (in the figure downward), while they are normally skewed to the right.

Large residual variation (var = 0.032). With large residual variation it was found that, for all four endpoints, quite a large fraction of runs did not result in a significant response at all. As Figure 10 shows, these fractions reached around 30% for the smallest study size (total $n = 20$). There is a consistent pattern of designs with larger number of dose groups being more likely to find no significant response. Further, an increase of total study size from 20 to 40 gives a substantial improvement, for a given number of dose groups. The fraction of runs resulting in CED-L05 = 'zero' was again small, except for (supralinear) endpoint 4 (see Fig. 11), reaching levels of around 30%. This fraction decreased only slowly with the total study size, with a rather irregular pattern.

Because of the large fraction of nonnumerical efficacy factors for all four endpoints, Figure 8 only shows the median factors. The results for (linear) endpoint 2 are again comparable to those

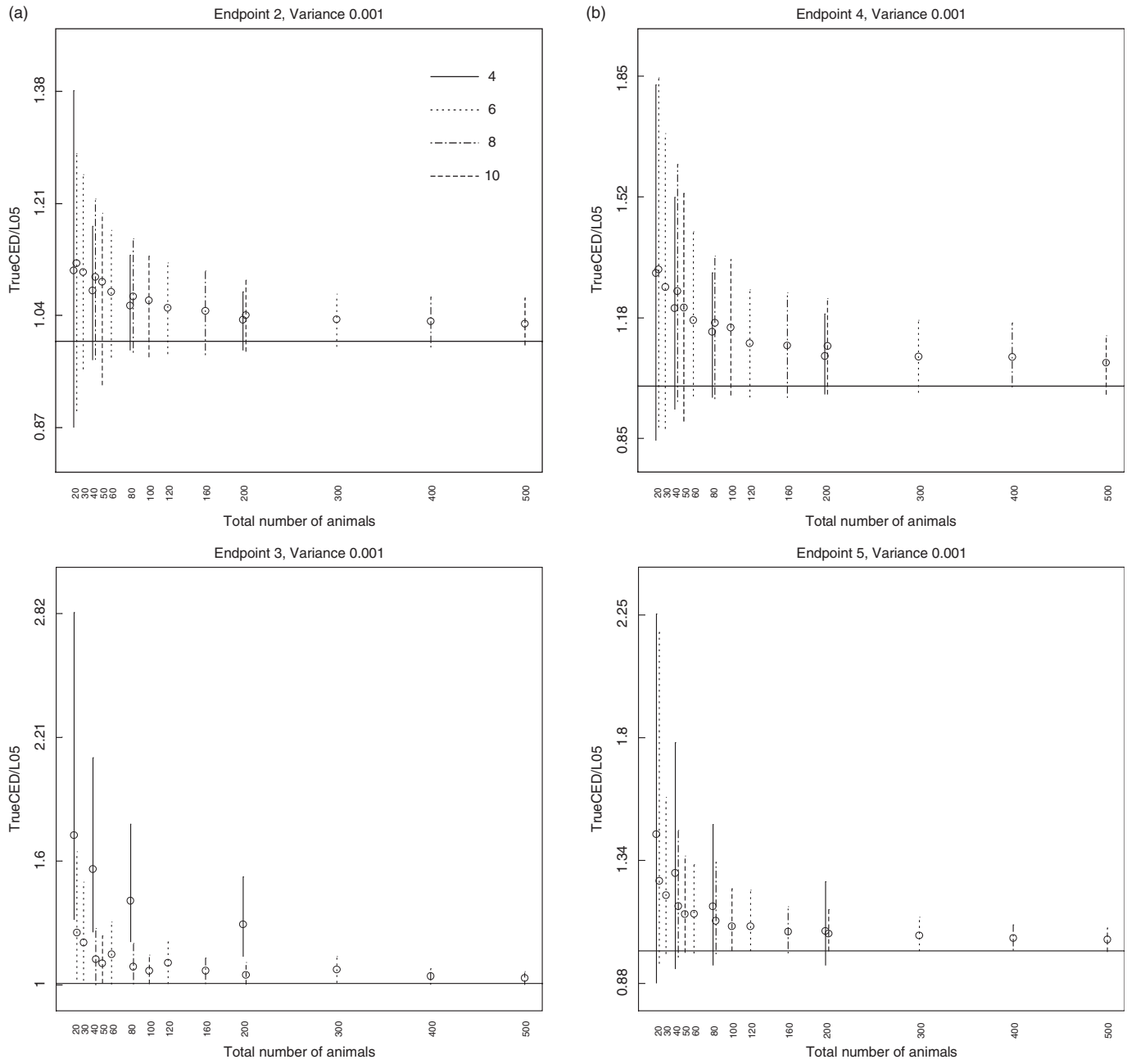

FIG. 6. Results of simulation study 3 for small residual variation (CV = 3%). Marks indicate the true CED divided by the median CED-L05; the vertical lines demarcate the 5th and 95th percentiles of these values.

found for the smaller variances. For (sublinear) endpoint 3, the multiple dose group designs show lower median efficacy factors, but it should be noted that these designs have a greater probability of not finding a response at all (see Fig. 10). In (supralinear) endpoint 4, the usual pattern is completely disturbed. In the smaller designs the majority of runs resulted in a CED-L05 that is *greater* than the true CED, obviously a quite undesirable situation. For the designs with total $n = 40$ and 50, it may be noted that the eight- and ten-dose-group designs are below the horizontal line, as opposed to the four- and six-dose-group designs of the same total size. In (sigmoidal) endpoint 5, the disturbing effect of selecting model 2 in a large fraction of runs is even greater than in Figure 7 (i.e., intermediate residual variation).

Benchmark Versus NOAEL Approach

To put the results shown in Figures 9, 10, and 11 into perspective, they were compared to the performance of the NOAEL approach for the same (four-dose-groups) designs and the same

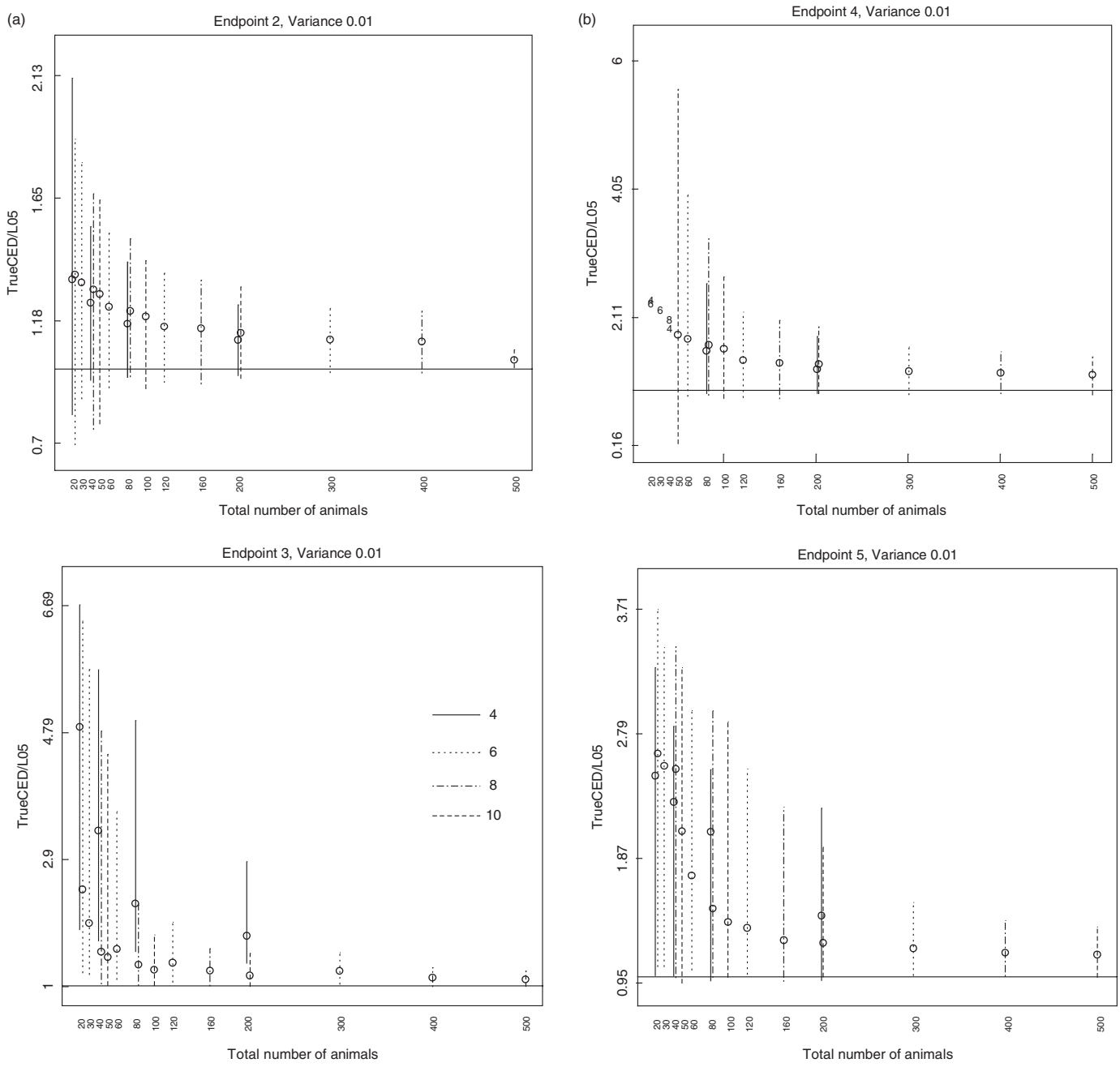

FIG. 7. Results of simulation study 3 for intermediate residual variation (CV = 10%). Marks indicate the true CED divided by the median CED-L05; the vertical lines demarcate the 5th and 95th percentiles of these values. For some situations, the percentiles could not be calculated; here only the medians are plotted, the number indicating the number of dose groups in the design.

endpoints. To that end, we performed simulation studies where the generated data were analyzed according to the NOAEL approach. Table 3 summarizes the results for the designs 4×5, 4×10, 4×20, and 4×50, where the residual variance was 0.032 (large variation).

The outcome 'no significant dose-response' in the benchmark approach may be considered analogous to the outcome 'none of the doses significant' in the NOAEL approach. Except for (supralinear) endpoint 4, the fractions of these outcomes are found to be comparable for the various situations examined, so that the NOAEL approach and the BMD approach perform similarly in this respect. However, it may be noted that, in the NOAEL approach, a fraction of runs resulted in an outcome where a significant dose is followed by a nonsignificant dose (observed response not monotone). If such happens in a real-life analysis, the significant dose will often be considered as a coincidental result, with the ensuing conclusion of no overall response. When this is taken

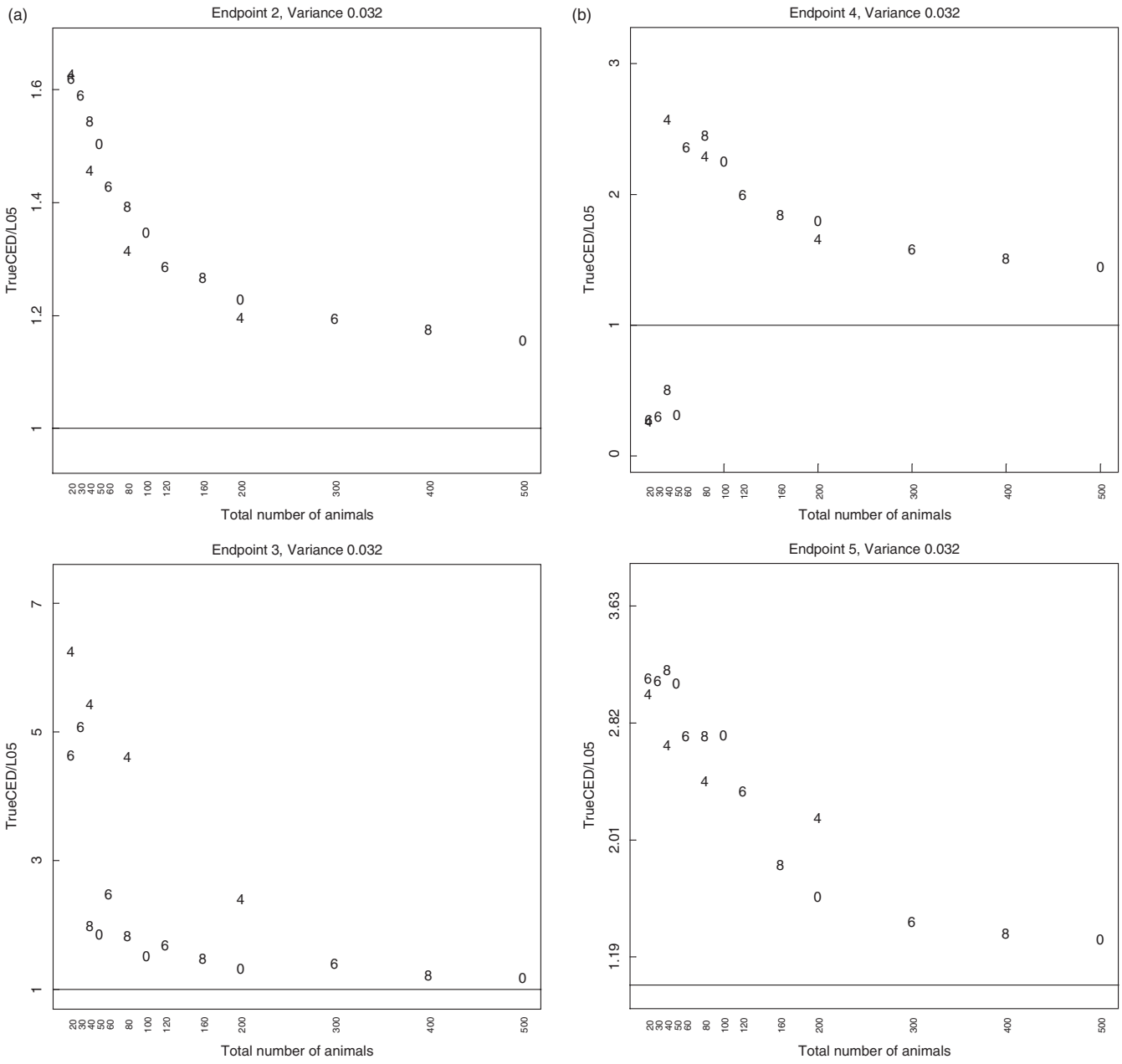

FIG. 8. Results of simulation study 3 for large residual variation (CV = 18%). The true CED divided by the median CED-L05 is plotted for each situation, the number indicating the number of dose groups in the design (0 = ten dose groups).

into account, the NOAEL approach behaves in fact poorer with respect to detecting an effect that does in fact exist. In (supralinear) endpoint 4, there is the additional circumstance that a certain fraction of runs resulted in a NOAEL where the associated effect size was in reality 25%. Clearly, treating such large effect sizes as though there were no (biologically significant) effect would be quite anti-conservative for many toxicological endpoints that are considered in practice. Overall, it may thus be concluded that the NOAEL approach more often fails to detect an existing response than the benchmark approach.

The outcome of an extremely low CED-L05 may be considered analogous to the result 'LOAEL only' in the NOAEL approach: both situations do not allow any conclusion on an upper bound of a 'no-adverse-effect' level. Here, a fundamental advantage of the BMD approach over the NOAEL approach becomes apparent. In the BMD approach the probability of this undesirable finding decreases by improving the study (by increasing sample size and/or decreasing residual variation). In the NOAEL approach, however, this probability increases in better studies. Indeed, the probability of the result 'LOAEL only' gets to 100% for sufficiently large sample size,

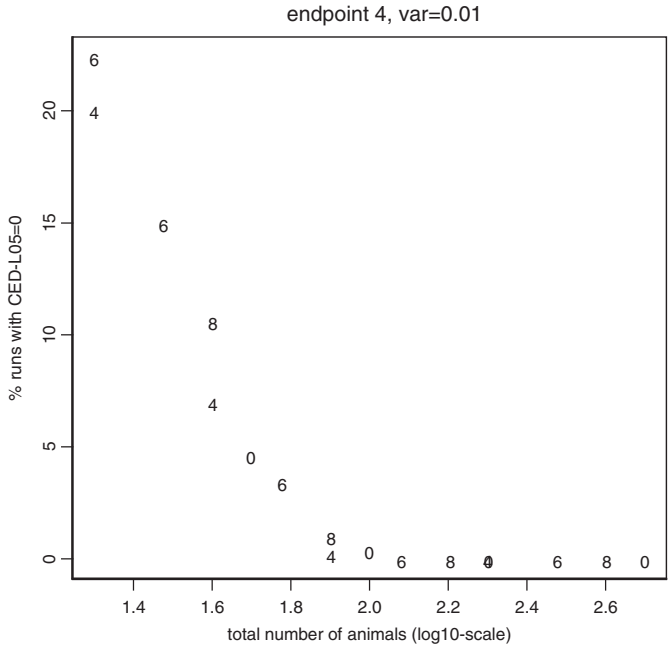

FIG. 9. Fractions of runs resulting in CED-L05 = 'zero' for (supralinear) endpoint 4 and intermediate residual variation (CV = 10%). Plotted numbers indicate the number of dose groups in the design (0 = ten dose groups).

and/or sufficiently small variance, as illustrated in Table 3 for (supralinear) endpoint 4. As a second advantage, in situations where the BMD approach does not result in a CED-L05, it does yield a point estimate of the CED. Hence, at least some reliable information on a 'no-adverse-effect dose' is provided by the benchmark approach (as long as it is not obtained by extrapolation). The result 'LOAEL only,' on the other hand, does not provide any information whatsoever on the potential NOAEL.

DISCUSSION

Number of Dose Groups

The influence of the number of dose groups on performance of the design is complicated. First of all, in comparing study designs with different number of dose levels, dose placement is usually different as well. For instance, the poorer performance of the four-dose-groups design for (sublinear) endpoint 3 in Figure 6 is probably due to the coincidental unfavorable dose placement (only one dose level associated with a measurable effect). This interpretation is supported by two observations. First, in simulation study 1, where doses where 'normalized' with respect to the associated effect sizes, the number of dose groups did not make much difference. And second, without such normalization (simulation study 2) dose placement showed a substantial effect when comparing designs with the same number of dose groups (Figs. 3–4).

It may be concluded that, given a certain total number of animals, designs with more dose groups are better, because they decrease the probability of unfavorable dose placement. The ensuing smaller number of animals per dose group was not found to have a negative impact on the efficacy factors in any of the simulations performed. There is, however, one complicating factor. In the situation of large residual variation (VC = 18%) the probability of missing an existing dose-response altogether increases with more dose groups while maintaining the same total n (see Fig. 10). This appears to be a quite systematic effect, as further illustrated in Figure 12, where regression curves are fitted to the results for (supralinear) endpoint 4. Therefore, the advantage of more dose groups is counteracted by this disadvantage of not detecting any response at all, in the case of large residual variation. Therefore, when toxicological endpoints of interest are known (from historical data) to have CVs close to 18% or more, it may be wise to apply a compromise and use six dose groups, in particular when the total number of animals is small.

Complications of Real Data

As discussed in the introduction, evaluating designs based on computer simulations rather than on real data has the advantage that many more different situations can be explored. However, data generated by the computer differ from real data in being based on perfectly performed (i.e., perfectly randomized) experiments. In reality this is not the case: it is very hard to perform an experiment where all experimental factors and all treatments (other than the intended treatment) are perfectly randomized. For instance, in diet studies the order of feeding is usually randomized at the level of dose group, but not on the level of the individual animal. Another obvious example is that animals in the same dose group may be housed in the same cage. Since the experimental units are live animals that may respond to many factors other than the intended experimental treatment (in toxicity studies: the dose), systematic errors in response at the dose group level can easily arise in any real experiment. In the computer simulations only random errors were taken into account, while systematic errors among treatment groups caused by other experimental factors than the dose were assumed not to exist. In the statistical analysis of real toxicity data potential systematic errors usually cannot be distinguished from the random error, and they will be pooled with the residual (random) variation. The ensuing increase in residual variation will automatically be reflected in larger confidence intervals of the CED, and as such it is relatively harmless. However, systematic errors in dose groups may also lead to a bias in the estimate of the CED, by causing a bias in the perceived dose-response relationship, in particular when such errors occur in the controls or in the top dose. In that case, the estimated parameters of the fitted model may be biased (e.g., dose-response is estimated to be far too steep), or a less appropriate model may be fitted (e.g., model 3 is fitted, while the real dose-response behaves like model 2).

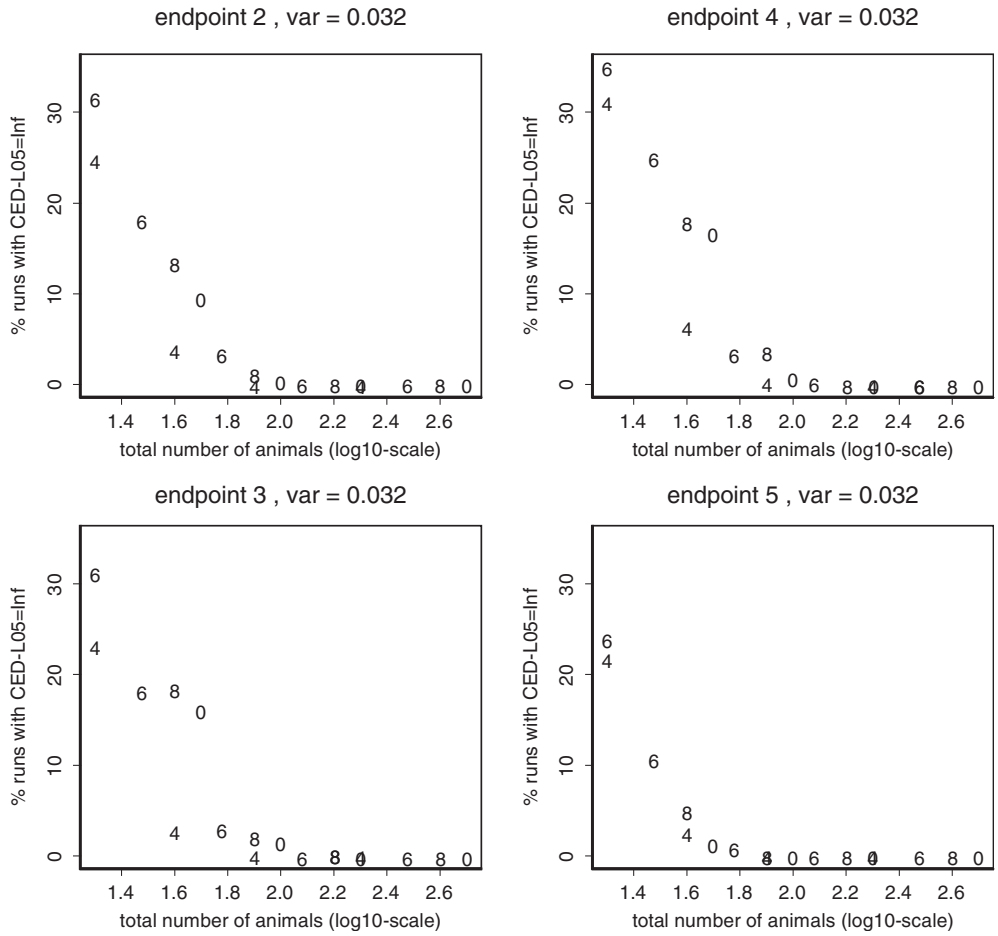

FIG. 10. Fractions of runs resulting in CED-L05 = 'infinite' for large residual variation (CV = 18%). Plotted numbers indicate the number of dose groups in the design (0 = ten dose groups).

Systematic deviations can usually not be recognized from the dose-response data, at least not unequivocally. What can be done is to mitigate their potential impact by increasing the number of dose groups in the study design (or, by applying a design in which each dose group has replicated subgroups that are treated as such with respect to randomization in the protocol). In this light, a multiple-dose study may be regarded as a way to have more treatment groups, and this may be regarded as an additional advantage of multiple dose studies (i.e., to minimize the effect of potential systematic errors in single treatment groups).

Parenthetically, it may be worthwhile to further explore the other option mentioned, that of using replicated subgroups at the same dose, and compare such designs with equivalent multiple-dose studies without replicated subgroups. The simulation studies required should then include systematic errors in generating the data.

High Doses

Our simulation studies showed that designs including relatively high doses usually perform better in estimating the CED than designs with (low) doses around the CED only. Indeed, it does not seem to be a wise strategy to omit higher dose levels when doing so results in a better fit of the model in the lower dose range. The usual argument here is that different biological mechanisms may play a role at higher doses as compared to lower doses. However, such does not imply that the overall dose-response relationship consists of two distinct parts that cannot be described by a single curve. After all, the fitted model does not reflect any biological mechanisms; it merely assumes that the underlying dose response follows a smooth curve over the whole dose range. If biological mechanisms change when going from lower to higher doses, this change will result in a (maybe rapidly, but nonetheless gradually) changing local slope of the dose-response. The purpose of fitting a dose-response model is exactly to describe this change, and there is no *a priori* reason to believe that such may not be possible when mechanisms change. As discussed above, toxicity studies are never perfectly randomized, and systematic errors in the data can easily arise. Obviously, systematic errors can just as well occur in dose groups close to the CED while the effect observed in the top dose is accurate, as the other way around. The fact that

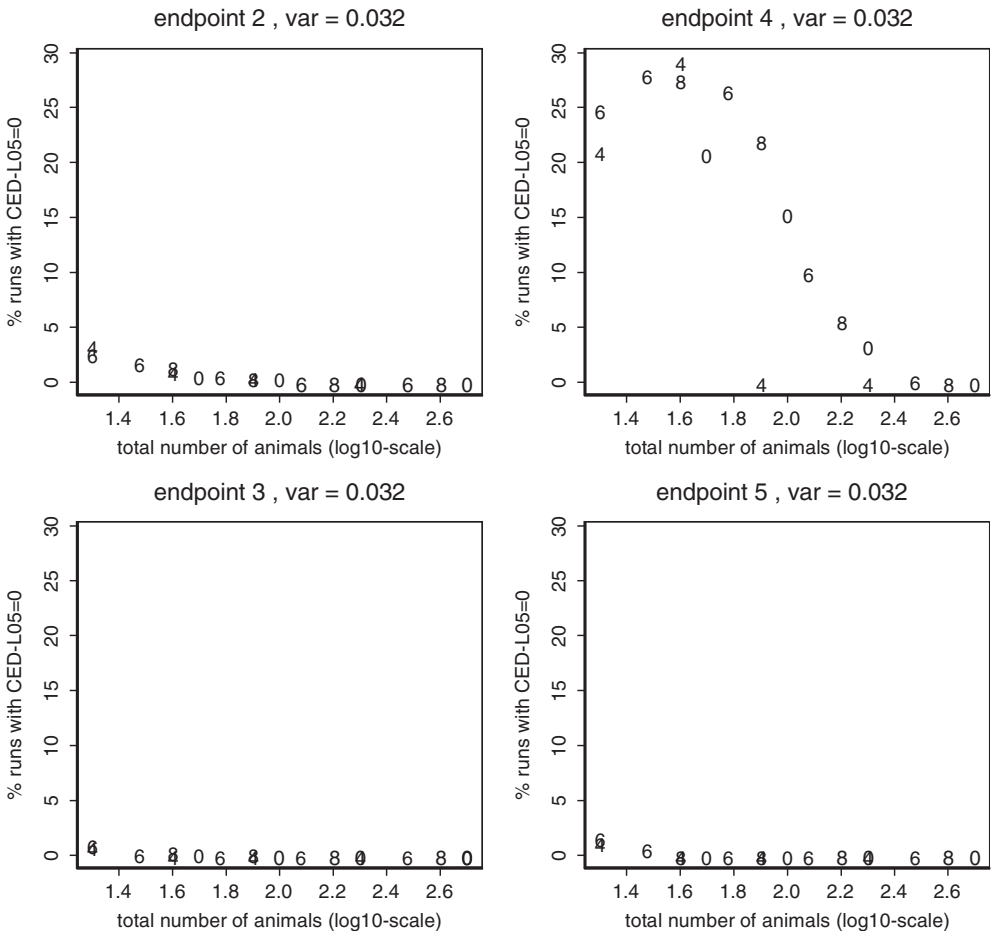

FIG. 11. Fractions of runs resulting in CED-L05 = 'zero' for large residual variation (CV = 18%). Plotted numbers indicate the number of dose groups in the design (0 = ten dose groups).

the goodness of fit is significantly improved by leaving out the top dose is therefore not a valid argument to do so. The strategy of leaving out higher dose groups may therefore lead to biased estimates of the CED (or BMD), and should generally be avoided.

Comparing Endpoints

Endpoint (supralinear) turned out to be a relatively difficult shape of a dose-response curve to be assessed. Already for the intermediate variance (var = 0.01) a fraction of runs did not result in a numerical value for the CED-L05 (see Fig. 9). In Figure 2, where the doses were 'normalized' with respect to true effect sizes, there is a striking difference in performance between (sublinear and sigmoidal) endpoints 3 and 4, even though the shapes of these two endpoints are each other's mirror image.

The 'Right' Model

In discussions on dose-response modeling it is often pointed out that biological arguments should support the choice of the model to be fitted. This point has very limited value. As an illustration, consider simulation study 1, (sigmoidal) endpoint 5. Here, the best results were obtained with the 'low dose' design, even though model 3 was usually selected. Imagine that biological knowledge were available saying that the dose response must level off (in fact we knew this in the simulations). If this were used as an argument to select model 5 instead of model 3 for these data, the result would be poorer (wider confidence intervals for the CED). This illustrates that the 'wrong' model can perform very well, and even better than the model that is biologically more plausible (or even 'true'). This insight can also be understood from the fact that in fitting dose-response models, it is only the data that contain information; the model is just a dummy that is chosen for its flexibility to follow the data. Therefore, the choice of design is far more important than the choice of the model.

Total Study Size

The results presented here indicate that the current OECD guidelines concerning the 28-day study design (i.e., four dose

TABLE 3
Results of 500 Simulation Runs Analyzed with the NOAEL Approach

		Design			
Endpoint	Outcome of analysis	4 × 5	4 × 10	4 × 20	4 × 50
2 linear	LOAEL = 0.1	5	6	9	14
	NOAEL = 0.1/ ES = 2.7%	13	22	34	61
	NOAEL = 0.3/ ES = 8.2%	48	62	55	23
	None significant/ ES = 30%	27	4	<1	
	No monotone response	8	6	2	2
3 sublinear	LOAEL = 0.1	2	2	2	2
	NOAEL = 0.1/ ES = 0.0003%	6	7	9	8
	NOAEL = 0.3/ ES = 0.06%	56	79	83	85
	None significant/ ES = 30%	23	3		
	No monotone response	12	8	6	5
4 supralinear	LOAEL = 0.1	22	47	74	97
	NOAEL = 0.1/ ES = 13%	28	38	25	3
	NOAEL = 0.3/ ES = 25%	18	9	1	
	None significant/ ES = 30%	18	3		
	No monotone response	14	3	1	
5 sigmoidal	LOAEL = 0.1	2	3	2	2
	NOAEL = 0.1/ ES = 0.04%	9	11	14	24
	NOAEL = 0.3/ ES = 3.1%	55	77	77	69
	None significant/ ES = 30%	24	3		
	No monotone response	10	7	6	5

Note. Based on designs with four dose groups, and with residual variance of 0.032 (CV = 18%). Percentages (out of 500 runs) are given that resulted in the associated outcome. ES = (true) effect size (in % change in response at that dose compared to response at dose zero, as defined by the underlying dose-response relationship, see Table 1).

groups with five animals [of both sexes]), may have a substantial probability of missing a response for an endpoint that has in reality changed by no less than 30% (at the highest dose). This holds for both the BMD approach and the NOAEL approach (the latter even more so). Therefore, one may question the adequacy of this design. Given the assumption that longer exposure duration would increase rather than decrease the size of an effect, it may be considered undesirable to miss effect sizes that are as large as 30% after only 28 days of exposure. When the data are analyzed by the benchmark approach, however, it is feasible to fit the dose-response model to the responses from males and females together (see Slob, 2002). Experience with simultaneous analyses of both sexes in many toxicity studies has shown that such is possible in most cases (e.g., Appel *et al*., 2001; Piersma *et al*., 2000; Woutersen *et al*., 2001). Hence, by using this feature of the benchmark approach, the OECD design for a 28-day study effectively performs as a design somewhere in between a 4 × 5 and a 4 × 10 design. It will be worthwhile to further assess the performance of designs when male and female response data are analyzed simultaneously in future simulation studies.

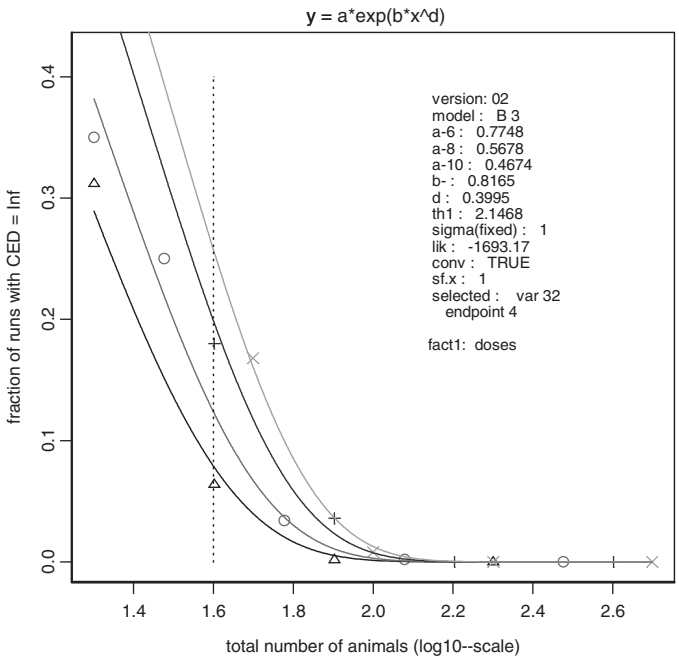

FIG. 12. Regression curves fitted to the fractions of runs not detecting a response (data from upper right panel in Fig. 10). Triangles, circles, plusses, and crosses indicate four, six, eight, and ten dose groups, respectively. The vertical dashed line indicates total $n = 40$.

Choice of Residual Variance

In designing a new toxicity study based on the results presented in Figures 6–8, a choice needs to be made for the residual variation. As discussed above, historical data showed that different endpoints may differ considerably in their associated residual variation. Since a toxicity study intends to evaluate a whole series of endpoints, one might argue that the design should be based on the higher end of the range of potentially occurring residual variances. Given our preliminary historical database of dose-response outputs, this would be a CV of 18%, or even more. However, the database also shows a correlation between the residual variation and the maximum response level: higher residual variations tend to occur in endpoints with higher maximum response levels. If one is willing to accept a higher value for CES in these endpoints, these endpoints are covered even if the design is based on simulation results based on a lower choice for the residual variation. To illustrate this point, consider an endpoint that reaches a maximum response level of a factor of 1.8 above background (as opposed to 1.3 in our simulations). For this

endpoint, exactly the same simulation results would be obtained when we scale the CES and the residual standard deviation (on log-scale) with the factor 1.8/1.3. This results in the following conditions:

1. The value of CES is changed from 1.05 to 1.11 (i.e., from a 5% to an 11% change relative to background). This value is obtained from $\ln(CES_{new}) : \ln(1.05) = \ln(1.8) : \ln(1.3)$. The idea of letting the value of CES depend on the maximum response appears intuitively reasonable, and has been discussed by Murrell *et al*. (1998), although their proposal is not exactly identical to the calculation given here.
2. The value of the residual variance on log-scale is changed by a factor of $[\ln(1.8)/\ln(1.3)]^2$.

As an example, an endpoint with maximum response of 1.8 above background with a residual variation of CV = 23% will have the same simulation results as in Figure 7 (i.e., for residual CV = 10%) regarding a CED associated with CES = 11%. Clearly, improving designs of toxicity studies in the future would be helped by better historical information on the values for residual variation and maximum response for the individual endpoints that typically result from toxicity studies.

MAIN CONCLUSIONS

- The main factor determining the efficacy of a design in assessing the CED-L05 (lower bound of CED) is the total number of animals used.
- Dose placement is another important factor for design efficiency. While an accurate dose placement depends on the shape of the underlying dose response, it is clearly found that high doses (i.e., doses associated with larger effects) are important in estimating a CED (i.e., a dose associated with smaller effects). Concentrating doses around the CED is not a good strategy, due to the fact that the signal-noise ratio in the observed response is unfavorable in this region.
- Distributing the total number of animals over more dose groups does not result in poorer performance as measured by the efficacy factor (i.e., distance from CED-L05 to true CED). In certain situations, designs with fewer (but larger) dose groups perform poorer, due to the fact that fewer dose groups are less likely to hit favorable dose levels. However, for endpoints with larger residual variation, the probability of not detecting the overall response increases with more dose groups. Yet, because in practice appropriate dose placement is hard to ensure in advance, the use of more dose groups is recommendable, for the sake of spreading the risk of inadequate dose placement. At the same time, more dose groups may mitigate the impact of potential systematic errors in single dose groups caused by unknown experimental factors and/or imperfect randomization in real life toxicity studies. When there is an interest in endpoints that are known to exhibit considerable residual variation (CV close to 18 % or even more), it may be wise to use six dose groups as a compromise.
- For endpoints with small residual variation (CV = 3%), all designs considered performed satisfactorily, the 95th percentile of the efficacy factor remaining within a factor of around three. For intermediate residual variation (CV = 12%), the efficacy factor remained within a factor of seven. However, for (supra-linear) endpoint 4 the 95th percentile of this factor could not be assessed in the case of designs with total $n < 50$, since more than 5% of the runs did not result in a numerical value for the CED-L05. For endpoints with large variation (CV = 18%) a substantial fraction (up to 30% for small total n) failed to detect any dose-response relationship.
- The adequacy of the OECD design for the 28-day study appears questionable, at least for endpoints with large residual variation (CV = 18% or more). For many endpoints the NOAEL approach can analyze these studies only for both sexes separately, as the response in the controls often differs between sexes. The advantage of the benchmark approach is that both sexes can often be analyzed simultaneously, resulting in an effective increase of the study size. This aspect should be further investigated in future simulation studies.
- In the case of large residual variation (CV = 18% or more) many of the commonly used designs (total n = 10, 20, or 40) appear quite ineffective. Reducing the residual variation (e.g., by making the experimental conditions as homogenous as possible) may improve this undesirable situation considerably.
- Selection of the 'right' model is not crucial. Some of our results showed that very good results can be obtained even though the fitted model was not the same model as the one that generated the data.
- Although the OECD design of four dose groups is based on the NOAEL approach, its efficacy in estimating a CEDL is quite reasonable, as long a dose placement is not highly unfavorable.

ACKNOWLEDGMENTS

This research was sponsored by the Netherlands Organization for Health Research and Development.

REFERENCES

Appel, M. J., Bouman, H. G. M., Pieters, M. N., and Slob, W. (2001). Evaluation of the applicability of the Benchmark approach to existing toxicological data. Bilthoven, RIVM/TNO Report No. 601930 001.

Crump, K. S. (1984). A new method for determining allowable daily intakes. *Fund. Appl. Toxicol.* **4,** 845–871.

Kavlock, R. J., Schmid, J. E., and Setzer, R. W., Jr. (1996). A simulation study on the influence of study design on the estimation of benchmark doses for developmental toxicity. *Risk Anal.* **16,** 399–410.

Kelly, G. E. (2001). The median lethal dose – design and estimation. *Statistician* **50,** 41–50.

Moerbeek, M., Piersma, A. H., and Slob, W. (2004). A comparison of three methods for calculating confidence intervals for the benchmark approach. *Risk Anal.* **24,** 31–40.

Murrell, J. A., Portier, C. J., and Morris, R. W. (1998). Characterizing dose response I: Critical Assessment of the Benchmark Dose Concept. *Risk Anal.* **18,** 13–26.

Piersma, A. H., Verhoef, A., Te Biesebeek, J. D., Pieters, M. N., and Slob, W. (2000). Developmental toxicity of butyl benzyl phtalate in the rat using a multiple dose study design. *Reprod. Toxicol.* **14,** 417–425.

Slob, W., and 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,** 787–798 (1998).

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

Weller, E. A., Catalano, P. J., and Williams, P. L. (1995). Implications of developmental toxicity study design for quantitative risk assessment. *Risk Anal.* **15,** 567–574.

Woutersen, R. A., Jonker, D., Stevenson, H., Te Biesebeek, J. D., and 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.