NATIONAL INSTITUTE OF PUBLIC HEALTH AND THE ENVIRONMENT
BILTHOVEN, THE NETHERLANDS

Report no. 620110 006

Few large, or many small dose groups? An evaluation of toxicological study designs using computer simulations

W. Slob, M.N. Pieters

August 1997

This study was conducted by order of the Directorate-General for Environmental Protection, Directorate for Chemicals, Safety and Radiation Protection, within the framework of project number 620110, Human Effect Assessment I.

National Institute of Public Health and the Environment, P.O. Box 1, 3720 BA Bilthoven, The Netherlands, tel: 030-2749111, fax: 030-2742971

MAILING LIST

1	Directie SVS/S
2	Plv. Directeur-Generaal Milieubeheer, Dr Ir B.C.J. Zoeteman
3	Dr D.W.G. Jung, DGM, SVS/S
4	Dr Ir P.C. Bragt, VWS, HIGB
5	Dr H. Roelfzema, VWS, Gezondheidsbeleid/Consumentenveiligheid en omgevingsrisico's
6	Depot Nederlandse Publikaties en Nederlandse Bibliografie
7	Dr P.E.T. Douben, Royal Commission on Environmental Pollution, Room 648, Church House, Great Smith Street, London SW1P 3BZ
8	Dr R. Kodell, FDA/NCTR, 3900 NCTR Road, Jefferson AR 72079
9	Prof. M. Razzaghi, Bloomsburg University, Dep. of Mathematics, Bloomsburg, PA 17815
10	Dr J.H.M. Schobben, RWS-RIKZ, PO Box 20907, 2500 EX Den Haag
11	Directie RIVM
12	SBD/Voorlichting & Public Relations
13	Dr Ir G. de Mik, Directeur Sector 6
14	Dr W.H. Könemann, Hoofd CSR
15	Dr C.J. van Leeuwen, CSR
16	Dr J. de Bruijn, CSR
17	Dr G.J.A. Speijers, CSR
18	Ing P.J.C.M. Janssen, CSR
19	Drs T. Vermeire, CSR
20	Dr ir H.J.G.M. Derks, Hoofd LBO
21	Dr A. Opperhuizen, Hoofd LEO
22	Dr Ir M.J. Zeilmaker, LEO
23	Dr Ir E.H.J.M. Jansen, LEO
24	Dr R.B. Beems, LPI
25	Dr J. Garssen, LPI
26	Dr P.F.M. Teunis, MGB
27-28	Auteurs
29	Bureau Rapportenregistratie
30	SBD/Voorlichting & Public Relations
31	Bibliotheek RIVM
32-52	Bureau Rapportenbeheer

CONTENTS

ABSTRACT

The benchmark approach of analyzing toxicological data does not require many subjects per dose group. In a number of simulation studies we theoretically examined the consequences of increasing the number of dose levels in a toxicological study at the expense of the number of subjects per group. The main conclusion of these simulation studies is that a multiple dose experiment is more likely to give an accurate point estimate of the "Critical Effect Dose", or benchmark dose, without loss of statistical precision. Therefore, study designs with more dose groups may be recommendable.

SAMENVATTING

De "benchmark dosis" als alternatieve karakterisering van het "no-adverse-effect niveau" in toxicologische dierstudies is de laatste jaren sterk in opgang. In deze alternatieve benadering komt de vaststelling van de NOAEL met behulp van significantie toetsing niet meer voor. In plaats daarvan worden de dosis-respons gegevens geanalyseerd met behulp van regressie-analyse. De gefitte curve wordt dan gebruikt om de "critical effect dose" (CED) te schatten behorende bij een bepaalde gepostuleerde effectgrootte.

Deze alternatieve analyse-methode heeft als consequentie dat het niet meer nodig is om een minimum aantal dieren per dosisgroep te vereisen, waardoor het mogelijk wordt om bij gelijk aantal gebruikte dieren meer dosisgroepen in te zetten. Met enkele computer-simulatiestudies onderzochten we de theoretische consequenties van het gebruik van meer dosisgroepen, door drie proefopzetten met vier, tien en veertig dosisgroepen, maar gelijkblijvend totaal aantal dieren, onderling te vergelijken.

Uit deze simulaties blijkt dat vergroten van het aantal dosisgroepen ten koste van het aantal dieren per groep, niet leidt tot een vermindering van de statistische precisie. De kracht van proefopzetten met meerdere dosisgroepen is gelegen in het feit dat zij beter in staat zijn te discrimineren tussen verschillende regressiemodellen volgens een "goodness of fit" criterium. Daardoor zal een proefopzet met meerdere dosisgroepen meer kans hebben dat de CED adequaat geschat wordt dan een proefopzet met slechts enkele dosisgroepen, zoals nu voorgeschreven door de internationale richtlijnen (bijv. OECD, EPA).

SUMMARY

The benchmark dose as an alternative way of characterizing the "no-adverse-effect level" in toxicological studies is gaining attention in the field of risk assessment. In this alternative approach the assessment of the NOAEL by significance testing has been dropped. Instead, dose-response data are analysed by regression analysis. The fitted curve is used to estimate the "critical effect dose"(CED) associated with a postulated effect size.

As an implication of this alternative approach of analysis, the requirement of having a minimum number of animals in each dose groups can be omitted, so that the use of more dose groups becomes feasible. We discuss a number of computer simulation studies to theoretically compare three study designs with four, ten and fourty dose groups, but with an equal total number of animals.

These simulation studies show that increasing the number of dose levels at the expense of the number of animals per group does not result in a loss of statistical precision. The strength of multiple dose experiments is that they are better able to discriminate between various regression models by a given goodness of fit criterion. As a result, multiple dose experiments are more likely to give an accurate estimate of the CED than designs with few dose groups.

1. INTRODUCTION

In the standard approach for deriving a reference dose (RfD, or ADI, TDI) for chemicals the NOAEL plays a key role. The NOAEL is considered to be the highest dose that does not evoke adverse effects in laboratory animals. Its value results from comparing the various dose groups with the control group in a particular study by applying tests of significance. A typical property of a significance test is that its power decreases with sample size. As a result, the NOAEL will be assessed at a higher value the less animals are used. International guidelines (e.g, OECD, USEPA) therefore prescribe a minimum number of animals per dose group, e.g., 20 animals of both sexes in chronic toxicological studies, and 50 animals of both sexes for chronic carcinogenic (or combined) studies. Because of this minimum sample size requirement, toxicological studies typically have a limited number of dose groups, to keep the total number of animals used within certain bounds. This study design with few dose groups is a logical consequence of the type of statistical analysis used, i.e. significance testing.

In a previous report (Slob and Pieters 1997) we discussed a probabilistic approach as an alternative to the classical NOAEL / uncertainty factor approach. In this alternative approach the assessment of the NOAEL by significance testing has been replaced by a benchmark type of approach. In this approach a regression model is fitted to the data, reflecting the dose-response relationship for a particular endpoint. From this fitted model, the Critical Effect Dose (CED) is estimated, i.e. the dose at which the response has changed to an extent that can be considered as being adverse. An important advantage of this alternative approach of analyzing toxicological studies is that, in contrast to the NOAEL, the precision of the estimated CED can be quantified. As a consequence, the required number of animals in a toxicological study can be related to the desired precision of the outcomes of the study.

As a second implication of this alternative approach of analysis, the requirement of having a minimum number of animals in each dose groups has become unnecessary. As a matter of fact, when the aim is to estimate the dose-response relationship as a whole, it appears intuitively better to have more dose groups, at the expense of the number of animals per dose group. This raises the question if a study design with more dose groups but less animals per group loses statistical precision. A second question to be addressed is, how designs with various numbers of dose groups compare, and if a optimum number of dose groups can be appointed (from a statistical point of view). Before carrying out a practical toxicological study with more dose groups, it is worthwhile to first get more insight in this alternative study design by performing a number of computer simulation studies. This report discusses the main results of these simulation studies.

2. METHODS

In three consecutive simulation studies we compared three study designs, having different numbers of dose levels, but an equal total number (80) of animals:

Design A. 4 dose groups, 20 animals per dose group
Design B. 10 dose groups, 8 animals per dose group
Design C. 40 dose groups, 2 animals per dose group

The data (type: continuous) were generated by calculating the expected response given by a particular regression model at the relevant dose level, and by multiplying that value by an error term that was randomly drawn from a lognormal distribution with median unity and variance on (natural) log-scale equal to 0.2 (i.e. a coefficient of variation of 47% for the response on the original scale).

For each generated data set a regression model was fitted (nonlinear regression analysis). From this fitted model the Critical Effect Dose (CED) was determined, i.e. the dose associated with the Critical Effect Size (CES).

3. RESULTS

3.1. Simulation study 1

The first simulation study addresses the comparative performances of the three study designs in estimating the CED, given the (hypothetical) situation that the "true" dose response relation is known. The data were generated from the power model

$$Y = a + b\ X^c\ ,$$

with Y and X denoting response and dose, respectively, and with a=0.2, b=0.075, c=2. The ("true") CED, associated with a CES=0.1, in this particular case is 0.52.

In each simulation run this power model was fitted to the generated data, for all three study designs (see Fig. 1 for an illustration). In each run, the CED and the lower- and upper 95%-confidence limits were calculated. For all three study designs it was found that the back-estimated CED was close to the real value, while the confidence intervals were similar among study designs (see Table 1).

Table 1. Results of simulation study 1.

study design	point estimate of CED[1)] (true value: 0.52)	length of 95%-conf. interval[1)]
A (4 doses)	0.515	0.278
B (10 doses)	0.509	0.283
C (40 doses)	0.516	0.306

[1)] Mean of 100 simulation runs.

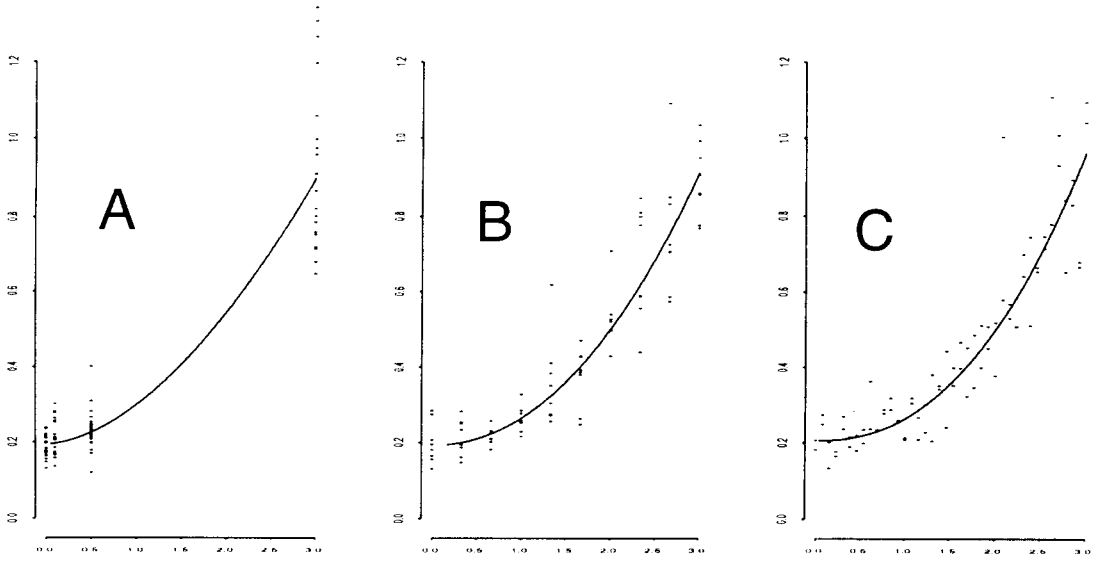

Fig.1. Three particular simulation runs for designs A, B and C. Abscissa: dose, ordinate: response. Dots: generated data; curves: fitted power models.

3.2. Simulation study 2

Obviously, for real data the "true" regression model is unknown. The only information useful for deciding which regression model to use for calculating the CED, is the model's goodness of fit to the data. In the second simulation study we address this aspect of data fitting for the three study designs.

To that end, data were generated data from the Weibull model

$$Y = a + d(1 - \exp(-(X / b)^c))$$

with a = 0.2, b = 2, c = 3, d = 0.7. Subsequently, not only the Weibull, but also the power model was fitted to each generated data set (see Fig. 2, for an illustration). The critical effect size was again chosen at 0.1, with an associated ("true") CED of 0.28.

For each fitted regression model the dose (CED) associated with the CES of 0.1 was derived, together with its 95%-confidence interval. As Table 2 shows, the true CED is badly estimated when fitting the power model in all three study designs. When the Weibull model is fitted, all three study designs perform equally well in estimating the CED. This result is in concordance with simulation study 1. The point is, however, that in study design A, it is impossible to discriminate between the two regression models: they both fit the data equally well, as indicated by comparable values for the residual variances. For study designs B and C, on the other hand, the fit of the Weibull model was found to be better, as indicated by a lower average value of the residual variance (also in the individual data sets the fit was always better for the Weibull). Therefore, designs B and C allow to distinguish between the Weibull and the power model. This means that, in analyzing real data, these designs have a higher probability of giving an

accurate estimate for the CED, by choosing the model that gives the best goodness of fit.

Table 2. Results of simulation study 2.

study design		point estimate of CED[1)] (true value: 0.28)	residual variance
power model fitted			
A	(4 doses)	0.14	0.038
B	(10 doses)	0.095	0.043
C	(40 doses)	0.098	0.042
Weibulll model fitted			
A	(4 doses)	0.29	0.038
B	(10 doses)	0.28	0.037
C	(40 doses)	0.29	0.037

[1)] Mean of 50 simulation runs.

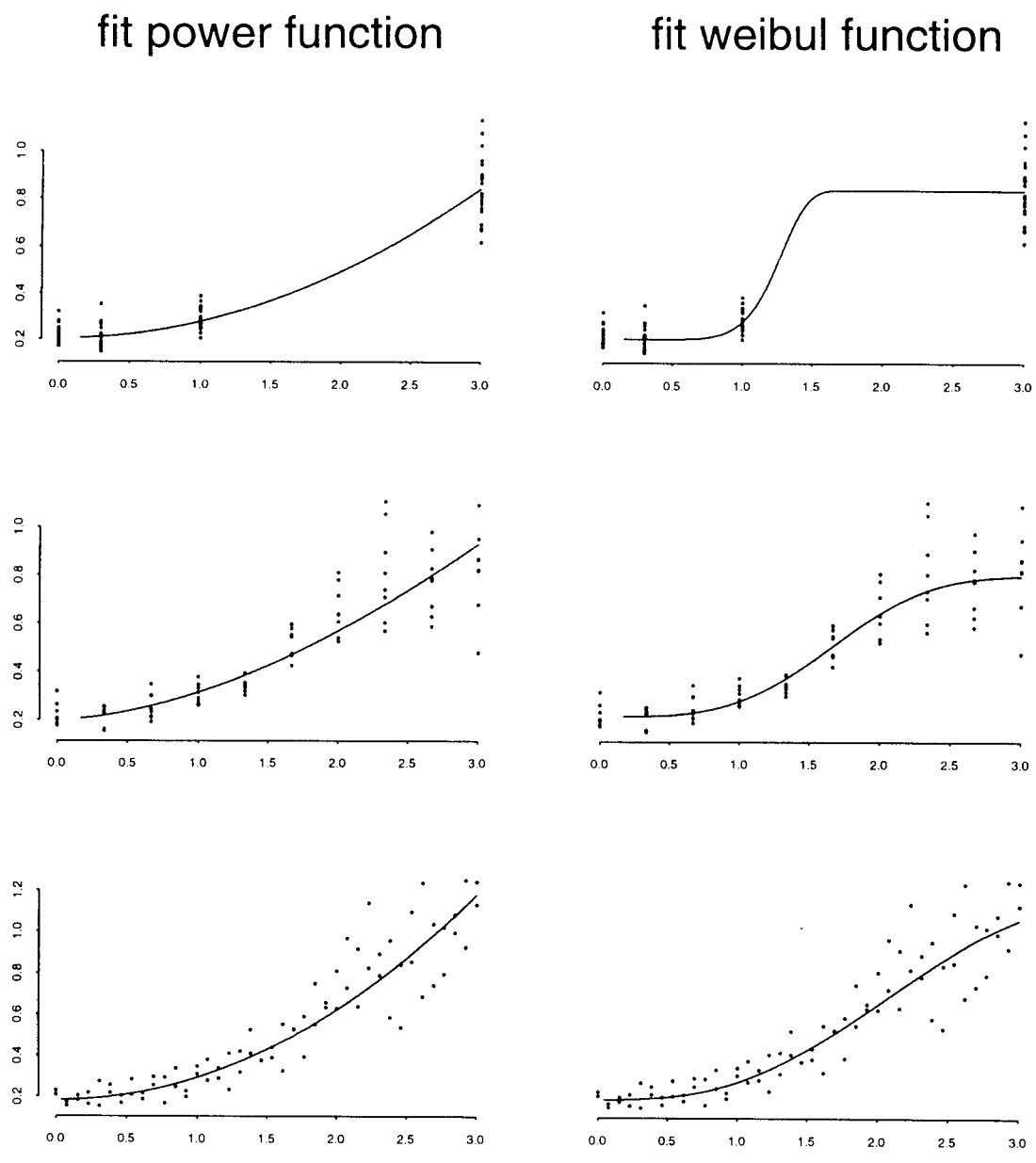

Fig. 2. Six particular simulation runs for designs A (upper panels), B (middle panels) and C (lower panels). In the left hand panels a power model is fitted, in the right hand panels a Weibull model.

3.3. Simulation study 3

The third simulation study differs from the second only with respect to the critical effect size: here it was set at 10^{-6}, to study the performance of the three designs in low-dose extrapolations. Table 3 shows that, when the wrong regression model is fitted, the CED is again poorly estimated, and even to a much greater extent as compared to the previous simulation study. This is in accordance with the well-known observation in carcinogenicity studies, that low-dose extrapolation is extremely sensitive for the model chosen. Yet, when the right model is fitted, the low-dose extrapolation is fairly accurate for study designs B and C. Again, in design A one has no clue for deciding

accurate for study designs B and C. Again, in design A one has no clue for deciding what model to use for deriving the CED, while in designs B and C the residual variance is clearly lower for the right model.

Table 3. Results of simulation study 3.

study design	point estimate of CED[1)] (true value: 13.2)	residual variance
power model fitted		
A (4 doses)	1.21	0.038
B (10 doses)	0.543	0.043
C (40 doses)	0.569	0.042
Weibull model fitted		
A (4 doses)	9.12	0.038
B (10 doses)	12.6	0.038
C (40 doses)	11.3	0.038

Dose values were multiplied by 1000, compared to simulation study 2.
[1)] Mean of 50 simulation runs.

4. CONCLUSIONS

The results of the simulation studies indicate that distributing the same number of animals to more dose groups does not result in a loss of statistical precision. For a given total number of experimental animals, the number of dose levels does not appear to have much impact on the precision of the estimated critical effect dose (CED); the precision of that estimate only depends on the total number of animals used in the experiment, and not on the way that animals are allocated to dose groups. In this respect study designs with low or high numbers of dose groups perform similarly.

The strength of a multiple dose experiment, however, is given by its potential to discriminate between various candidate response models. Our results clearly showed, that fitting the "wrong" model, results in inaccurate estimates of the CED. The weakness of prevailing study designs, i.e. designs with only three or four dose groups, is there inability to discriminate between dose-response models. Therefore, these designs have a larger probability of giving inaccurate estimates of the CED than designs with more dose groups.

The main conclusion of our simulation studies is that a multiple dose experiment is more likely to give an accurate point estimate of the CED, without loss of statistical precision. Therefore, study designs with more dose groups are recommendable. Although from a statistical point of view there is no objection to have as many dose levels as experimental subjects (i.e., a single animal per dose level), it is usually more practical to have fewer dose groups. Since our simulation results indicated that an extremely high number of dose levels does not perform better than a more moderate number of dose levels, an intermediate number of dose levels appears the best option.

Since these theoretical results show that a multiple dose experiment is better suited for estimating the critical-effect dose, this alternative study design will be tested in practice by performing a 28 days toxicological study with multiple doses, and compare the results with a simultaneously performed study according to the OECD guidelines, i.e. four dose groups including the control.

REFERENCES

Slob, W. and Pieters, M.N. (1997).
A probabilistic approach for deriving human RfDs and human health risks from toxicological studies: general framework.
RIVM, Report no. 620110 005.