FUNDAMENTAL AND APPLIED TOXICOLOGY 23, 496–509 (1994)

Dose–Response Assessment for Developmental Toxicity

III. Statistical Models[1]

BRUCE C. ALLEN,* ROBERT J. KAVLOCK,† CAROLE A. KIMMEL,‡ AND ELAINE M. FAUSTMAN§

*K. S. Crump Division, ICF Kaiser, Ruston, Louisiana 71270; †Developmental Toxicology Division, Health Effects Research Laboratory, Office of Health Research, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711; ‡Reproductive and Developmental Toxicology Branch, Human Health Assessment Group, Office of Health and Environmental Assessment, U.S. Environmental Protection Agency, Washington, DC 20460; and §Department of Environmental Health, University of Washington, and Affiliate of the Child Development and Mental Retardation Center, Seattle, Washington 98195

Received May 28, 1993; accepted May 16, 1994

Dose–Response Assessment for Developmental Toxicity. III. Statistical Models. ALLEN, B. C., KAVLOCK, R. J., KIMMEL, C. A., AND FAUSTMAN, E. M. (1994). *Fundam. Appl. Toxicol.* 23, 496–509.

Although quantitative modeling has been central to cancer risk assessment for years, the concept of dose–response modeling for developmental effects is relatively new. The benchmark dose (BMD) approach has been proposed for use with developmental (as well as other noncancer) endpoints for determining reference doses and reference concentrations. Statistical models appropriate for representing the unique features of developmental toxicity testing have been developed and applied (K. Rai and J. Van Ryzin, 1985, *Biometrics* 41, 1–9; L. Kupper, C. Portier, M. Hogan, and E. Yamamoto, 1986, *Biometrics* 42, 85–98; R. Kodell, R. Howe, J. Chen, and D. Gaylor, 1991, *Risk Anal.* 11, 583–590). Generalizations of those models (designated the RVR, LOG, and NCTR models, respectively) account for the correlations among observations in individual fetuses or implant within litters; the potential for variables other than dose, such as litter size, to affect the probability of adverse outcome; and the possibility of a threshold dose below which background response rates are unaltered. The generalized models were applied to a database of 607 endpoints with significant dose-related increases in response rate. It was determined that the models were generally capable of fitting the observed dose–response patterns, with the LOG model appearing to be superior with respect to fit. A significant contributor to the ability of the LOG model to fit the data was its flexibility with respect to the representation of the dependence of response probability on litter size, a trait not shared by the other two models. Litter size appeared to be a significant covariable for predicting response rates, even when intralitter correlation was accounted for by assuming a β-binomial distribution for the observations among individual fetuses. In contrast, a threshold dose parameter did not appear to be necessary to adequately describe the observed dose–response patterns. BMD estimates (corresponding to 5% additional risk) from all three models were similar to one another and to BMDs estimated from other, generic dose–response models (not specifically designed for developmental toxicity testing) that modeled average proportion of fetuses affected. The BMDs at the 5% level of risk were similar to no observed adverse effect levels determined by statistical tests of trend. Greater emphasis on and further examination of dose–response modeling for developmental toxicity testing are needed; biologically based approaches that consider the continuum of developmental effects induced in such tests should be encouraged.

Quantitative dose–response modeling has been central to cancer risk assessment for many years (Anderson *et al.*, 1983). In the context of cancer risk assessment, dose–response models are used to estimate risks to animal and human populations associated with chemical exposures, typically for exposure levels that are very low relative to levels used in toxicological studies. This use of dose–response modeling usually entails extrapolation below the experimental dose and response range used to estimate parameters of the models, i.e., it involves estimation of the probabilities of response based on the predictions of the dose–response models.

The concept of dose–response modeling for noncancer endpoints, including those associated with developmental toxicity, is relatively new. Traditionally, the risk assessment approach for noncancer effects has involved determination of no observed adverse effect levels (NOAELs) and the estimation of reference doses or concentrations via the applica-

[1] Although the research described in this article has been supported by the U.S. Environmental Protection Agency (through assistance agreement CR-816253-01-0 to the University of Washington), it has not been subjected to Agency review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

0272-0590/94 $6.00

tion of uncertainty and modifying factors to the NOAEL for the critical effect (U.S. EPA, 1991).

The benchmark dose (BMD) approach has been proposed as an alternative to the NOAEL approach (Crump, 1984; Kimmel and Gaylor, 1988). The BMD approach includes dose–response modeling as an essential component. However, the modeling proposed for the BMD approach is not intended to provide estimates of risk at levels of exposure far below the experimental dose range. Rather, the models are proposed as means of estimating a statistical lower bound on dose associated with a predefined level of risk, that level of risk typically being in the range of 1 to 10% (Crump, 1984; Kimmel and Gaylor, 1988). The lower bound on dose estimated in that manner is referred to as BMD. The advantages of the BMD approach over the NOAEL approach have been discussed and acknowledged (Crump, 1984; U.S. EPA, 1991).

The determination of NOAELs for developmental toxicity has been examined and described for a large database of developmental toxicity experiments by Faustman *et al.* (1994). Allen *et al.* (1994) compared "generic" dose–response models for estimation of BMDs for developmental toxicity experiments from the same database. The models used for those comparisons were "generic" in the sense that they were not developed specifically for features of developmental toxicity experiments. They do not consider the correlated nature of the observations from typical developmental toxicity experiments (the so-called intralitter correlation) nor do they include covariables (such as litter size) that might affect outcome. Finally, they do not consider multiple outcomes in litters and fetuses.

Recognition of the correlation problem and the effect it has on the variability of the observations have been at the core of efforts to develop models more appropriate for developmental toxicity (see Williams, 1975; Kupper and Haseman, 1978; Haseman and Kupper, 1979; Gladen, 1979; Crump *et al.*, 1990 for discussion, especially in the context of the related issue of statistical testing). The effect of correlated responses in developmental toxicity tests is to produce extrabinomial variation, variation greater than would be expected in a binomial (uncorrelated) setting, thus affecting the calculation of confidence limits around maximum likelihood estimates and the significance of dose-group differences. The present investigation focuses on three models (Rai and Van Ryzin, 1985; Kupper *et al.*, 1986; Kodell *et al.*, 1991) designed to account for extrabinomial variation observed in tests of developmental toxicity and which have been expanded or generalized to include other features of interest in dose–response modeling of developmental toxicity, specifically incorporation of litter size as a covariable and of a "threshold dose" parameter.

Rai and Van Ryzin (1985) proposed a model that considered the probability of response for an individual fetus to be a function of both dose and litter size. The Rai and Van Ryzin model has two factors, one being interpreted as the probability of effect on the litter environment (which depends only on dose level) and the second interpreted as the probability of effect on a fetus conditional on there being an effect on the litter environment. The second conditional probability factor is a function of both dose level and litter size. Their original model attempted to account for all observed extrabinomial variation by incorporation of the litter size variable: for a fixed litter size and dose level, responses were assumed to be binomial. Rai and Van Ryzin also assumed that litter sizes could be described by Poisson distributions with means that were nonlinear functions of dose level. Faustman *et al.* (1989) applied the model of Rai and Van Ryzin to several developmental toxicity endpoints from the NTP database and suggested that this model was able to represent the observed dose–response relationships for prenatal death and malformation.

Kupper *et al.* (1986) proposed the use of a log-logistic type model and evaluated it in a simulation study. In such a model, the logit of the expected probability of response at dose d_i, logit(μ_i), was assumed to be a straight-line function of $\ln(d_i)$. The extrabinomial variation of responses observed in developmental toxicity studies was accounted for by assuming that the observations have a β-binomial distribution. Such a distribution arises by assuming that the responses among fetuses within a litter are conditionally independent, given the underlying probability of response, but that the underlying probability of response varies from dam to dam according to a β distribution. The resulting β-binomial distribution yields dose group-specific intralitter correlations and displays extrabinomial variation. The model of Kupper *et al.* (1986) did not consider litter size as a covariable for probability of effect, nor did it allow for nonzero background rates.

Kodell *et al.* (1991) developed a model of developmental toxicity that also assumed β-binomial variation of responses. Their model treated litter size as a possible covariable for estimating response probability, although the manner in which litter size was introduced was different from that proposed by Rai and Van Ryzin (1985). In the model of Kodell *et al.* (1991) litter size acted as a modifier of the parameters describing background response rate and dose-response slope. The model of Kodell *et al.* (1991) allowed for nonzero background response rates and introduced a "threshold dose" parameter. For modeling purposes, the threshold dose parameter estimated the dose below which response rates were not distinguishable from the background response rates.

Our investigation of generalized versions of these models examined their behavior over a large database of developmental toxicity studies. All three models were expanded to include litter size as a covariable, a threshold dose parameter, Weibull (or in the case of the log-logistic model, Weibull-like) dose–response, and dose group-specific intralitter

correlations to incorporate β-binomial variability. The fits of the models and the importance of the parameters of the models were determined for endpoints in the database. BMD predictions from these models were compared to those of the generic benchmark dose models (Allen *et al.*, 1994) and to statistically determined NOAELs (Faustman *et al.*, 1994). Conclusions and recommendations with respect to the use of these models for BMD calculations are provided.

METHODS

Database. The database used for this investigation has been described by Faustman *et al.* (1994). For each experiment, up to nine endpoints were defined, based on adverse outcomes including fetuses with three major types of malformations and dead implants. Endpoints representing the combination of the malformation categories (defining all malformed fetuses) and combinations of malformed fetuses with dead implants were included when possible. A total of 1825 endpoints from 246 developmental toxicity experiments were defined.

The developmental toxicity models under investigation (see below) were applied to the endpoints for which there were statistically significant dose-related increases in response rates. Mantel–Haenszel trend tests were employed to determine whether or not statistically significant increased response rates existed for either the percentage of litters affected (an affected litter having one or more affected fetuses or implants) or average proportion of fetuses or implants affected. If one or both of the trend tests for a particular endpoint indicated a significant dose-related trend ($\alpha = 0.05$), then that endpoint was included in the analysis.

NOAELs and generic BMDs. Statistically defined NOAELs for this database were described by Faustman *et al.* (1994). A NOAEL was defined to be the largest dose in an experiment for which the trend of increasing response rates with increasing dose was not statistically significant. These statistical NOAELs were specific to each endpoint; endpoints from the same experiment (representing different measures of adverse outcome in that experiment) might have different NOAELs. Two NOAELs were defined for each endpoint; one based on the quantal measure, percentage of litters affected (QNOAEL), and the other based on the continuous measure of response, average proportion of fetuses affected (CNOAEL).

BMDs derived from generic models were described by Allen *et al.* (1994) for the same database. Two sets of BMDs for each endpoint were calculated, analogous to the NOAELs. For percentage of litters affected, the BMDs were labeled $QBMD_{05}$, for example, indicating the use of the quantal metric and an additional risk level of 5% for defining the BMD (see Crump, 1984). Similarly, a BMD for average proportion of fetuses affected and a 5% increased risk level was referred to as $CBMD_{05}$. BMDs corresponding to 1 and 10% additional risk were also calculated. The generic BMDs and the statistically defined NOAELs were compared (Allen *et al.*, 1994): the $CBMD_{05}$s were similar on average to the NOAELs, so attention was restricted here to the 5% risk level.

Developmental toxicity models. Three developmental toxicity (DT) models were used to describe the dose–response relationships for the data on malformed fetuses and dead implants. The first was a generalization of the model presented by Rai and Van Ryzin (1985) and was defined as

$$P(d,s) = [1 - \exp\{-(\alpha + \beta(d - d_0)^\gamma)\}] * \exp\{-s(\theta_1 + \theta_2(d - d_0))\}, \quad (1)$$

where d was dose, s was litter size, and d_0 was the threshold dose parameter. The threshold dose parameter as well as the parameters α, β, γ, θ_1, and θ_2 were estimated for each endpoint by methods of maximum likelihood. This model was referred to as the RVR model and yielded BMDs referred to as $RBMD_{05}$ (corresponding to 5% additional risk).

The second model, referred to as the LOG model, generalized the log-logistic type model discussed by Kupper *et al.* (1986). The LOG model was defined by the following equation:

$$P(d,s) = \alpha + \theta_1 s + [1 - \alpha - \theta_1 s]/[1 + \exp\{\beta + \theta_2 s - \gamma \log(d - d_0)\}], \quad (2)$$

where parameters d_0, α, β, γ, θ_1, and θ_2 were estimated for each endpoint by methods of maximum likelihood. A BMD from this model corresponding to 5% additional risk was referred to as $LBMD_{05}$.

The model proposed by Kodell *et al.* (1991) was defined as

$$P(d,s) = 1 - \exp\{-[(\alpha + \theta_1 s) + (\beta + \theta_2 s)(d - d_0)^\gamma]\} \quad (3)$$

Maximum likelihood methods were used to estimate the parameters d_0, α, β, γ, θ_1, and θ_2. This model was referred to as the NCTR (National Center for Toxicological Research) model and the BMDs corresponding to 5% additional risk from this model were labeled $NBMD_{05}$.

Note that each of the parameters designated by a Greek letter has a different interpretation from one model to another. The parameter α, for example, does not represent the same thing in the LOG, NCTR, or RVR models.

In addition to the parameters displayed in Eqs. (1) through (3), intralitter correlation parameters were estimated for each model, one for each dose group. The maximum likelihood fits of each model to an endpoint included estimation of intralitter correlation parameters (separately for each model). The correlation parameters arise as a result of the β-binomial distribution that was assumed for all three models.

All BMDs from the models were 95% lower bounds on doses corresponding to fixed levels of additional risk. Attention was focused here on 5% additional risk.

DT model implementation. The RVR, LOG, and NCTR models were run using the software programs TERAVAN, TERALOG, and TERAMOD, respectively (Howe *et al.*, 1992). These programs allow implementation of the models in the full parametric form shown in Eqs. (1) through (3).

The full parametric form was utilized in most cases. However, about 20% of the endpoints included in this investigation lacked a satisfactory value for litter size. This occurred when live fetuses from each litter were separated into two groups, one of which underwent skeletal examinations and the other of which received visceral examinations. In those cases, the number of fetuses examined for skeletal malformations or for visceral malformations did not represent all live births. The number of fetuses examined, though related to the total litter size, was not felt to be representative of litter size in the manner in which the DT models were designed to account for litter size. Thus, for the endpoints representing fetuses with skeletal malformations or visceral malformations and the combination of those with dead implants, from experiments that divided live births into two groups, the models were run by assuming that the parameters θ_1 and θ_2 were zero, i.e., by removing litter size as a covariable in the equations defining probability of response.

For all endpoints, the models were also run fixing the threshold dose parameter at zero. For the endpoints considered to have satisfactory values for litter size, the models were also run with litter size removed as a covariable in the models. A representation of all the model runs is provided in Fig. 1. Because the RVR and NCTR models are identical when litter size is removed as a covariable, the RVR model was not run when litter size was ignored.

Model fit and comparison of DT models. χ^2 statistics were used to assess the fit of the models to the observed dose–response data. Observed and expected numbers of affected fetuses were determined for each litter size in each dose group, as were variance estimates. The variance terms

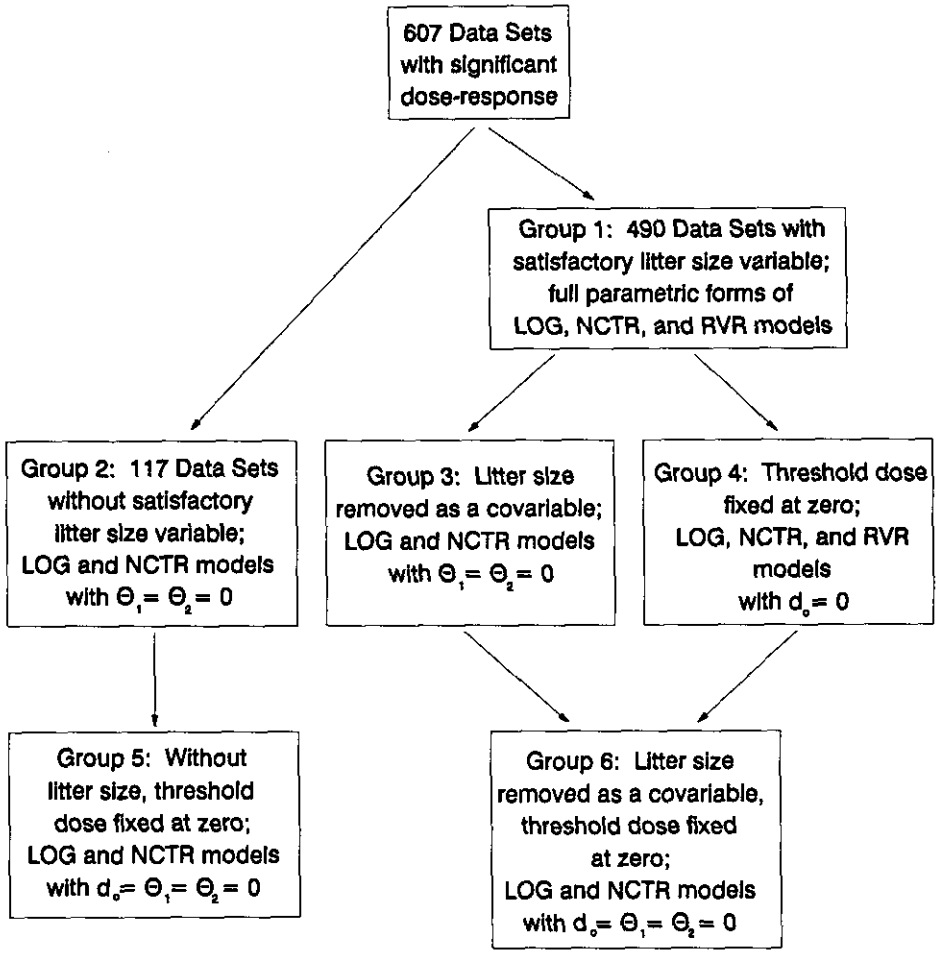

FIG. 1. Developmental toxicity modeling of database. Schematic showing the scope of the investigation. Models were run for each of the endpoints, which fell in either Group 1 or Group 2. Additional runs with less than the full parametric form of the models (setting some parameters to zero as shown) were completed in Groups 3–6.

used to define the χ^2 test statistic accounted for the correlated (β-binomial) nature of the observations. For the assessments of fit, a degree of freedom was allowed for each distinct dose/litter size pair. Degrees of freedom were reduced if necessary by combining neighboring distinct pairs (within dose group) to eliminate expected numbers less than 1; when such collapsing was required, observed numbers for the combined distinct pairs were summed, as were the expected numbers and variances.

Maximum values of the log-likelihoods (MLLs) were also recorded. These values were used to compare the three models with one another with respect to fit. Since all three models run on a particular endpoint had the same number of parameters, the model with the greatest MLL was determined to be the best fitting model for that endpoint.

Parameter significance. The contributions of the parameters d_0 and θ_1 and θ_2 to the performance of the models were assessed by examining the fits of the models to the data subsets both with and without those parameters in the models. Likelihood ratio statistics, based on differences in the MLLs, were used to assess the significance of differences between model fits with and without the threshold and litter size parameters. In relation to Fig. 1, the difference in MLLs for any particular endpoint from Group 1 and Group 4, Group 2 and Group 5, or Group 3 and Group 6 provided information relevant to the importance of the threshold dose parameter. Similarly, differences in MLLs for corresponding runs in Group 1 and Group 3 or Group 4 and Group 6 were relevant to the assessment of litter size as a covariable.

The likelihood ratio statistic for testing the significance of d_0 was distributed approximately as a χ^2 random variable with one degree of freedom. Thus, the threshold parameter d_0 was determined to significantly improve the fit of a model if the likelihood ratio statistic had a value greater than 3.84, which would happen only 5% of the time if a null hypothesis of no threshold ($d_0 = 0$) were true.

Similar likelihood ratio statistics were computed to assess the significance of the parameters θ_1 and θ_2. Because both of those parameters were set to zero simultaneously when litter size was removed as a covariable, the likelihood ratio statistic in this case was distributed approximately as a χ^2 random variable with two degrees of freedom. The parameters θ_1 and θ_2 were determined to be significant (litter size was determined to contribute significantly to the fit of a model to a data subset) if the likelihood ratio statistic was greater than 5.99, the 95th percentile of the two degree of freedom χ^2 distribution.

Comparisons of BMDs and NOAELs. The BMDs calculated for each model (e.g., $LBMD_{05}$, $NBMD_{05}$, or $RBMD_{05}$) were compared to one another, to the statistically defined NOAELs (QNOAEL and CNOAEL), and to $QBMD_{05}$ and $CBMD_{05}$ from the generic BMD models. Histograms of the ratios of $LBMD_{05}$, $NBMD_{05}$, and $RBMD_{05}$ to QNOAEL, CNOAEL,

$QBMD_{05}$, and $CBMD_{05}$ were used to examine the differences among the various estimates.

RESULTS

The 607 endpoints (from 141 studies) with significant dose-related increases in response rate are summarized in Table 1. A total of 93% of those endpoints would have been selected (i.e., determined to have significantly increased response rates) if the average proportion of fetuses affected was the only metric of response examined. Only 44 (7%) of the endpoints had a significant trend for percentage of litters affected but not for average proportion of fetuses affected.

TABLE 1
Summary of Endpoints with Significant Dose-Related Trends

	Number of endpoints having		
	Litter size covariable	No litter size covariable	Total
Total	490 (133)[a]	117 (46)	607 (141)
By source			
NTP/EPA	259	29	288
Wil Labs	141	46	187
Mobil	18	26	44
Argus	11	2	13
IRDC	61	14	75
By species			
Hamster	7	0	7
Mouse	96	4	100
Rabbit	146	16	162
Rat	241	97	338
By endpoint[b]			
DR	81	0	81
DM	51	0	51
DG	86	0	86
DS	52	36	88
DV	47	34	81
MO	40	0	40
GO	73	0	73
SO	34	25	59
VO	26	22	48
Statistically significant trend for:			
Only percentage of litters affected	38	6	44
Only average proportion of fetuses affected	155	22	177
Both	297	89	386

[a] Number of experiments from which these endpoints were obtained.

[b] DR, dead/resorbed (nonlive); DM, DR combined with malformed; DG, DR combined with gross (external) malformed; DS, DR combined with skeletal malformed; DV, DR combined with visceral malformed; MO, all malformed; GO, gross (external) malformed; SO, skeletal malformed; VO, visceral malformed.

TABLE 2
Summary of Model Fits to Endpoints and Assessment of Parameter Significance

	Model		
	LOG	NCTR	RVR
Number of endpoints with poor fits by group[a]			
Group 1 (N = 490)	39	50	56
Group 2 (N = 117)	30	33	—
Group 3 (N = 490)	43	41	—
Group 4 (N = 490)	29	42	35
Group 5 (N = 117)	33	31	—
Group 6 (N = 490)	46	40	—
Number of endpoints with significant:[b]			
Threshold dose parameter (d_0)			
Group 1 vs Group 4	7 (490)[c]	8 (489)	13 (469)
Group 2 vs Group 5	0 (113)	0 (117)	—
Group 3 vs Group 6	3 (482)	8 (490)	—
Litter size as a covariable			
Group 1 vs Group 3	118 (490)	60 (490)	42 (469)
Group 4 vs Group 6	118 (482)	64 (488)	42 (489)

[a] Counts of poor fits include endpoints for which model failed to converge as well as endpoints for which a χ^2 test of fit was rejected at the 0.01 level of significance. See Fig. 1 for definition of groups.

[b] Significance of d_0 or the litter size covariable indicates that d_0 or litter size, respectively, contributed significantly (as assessed by likelihood ratio tests) to the model fit.

[c] Number of endpoints for which the particular model converged in both of the comparison groups.

DT Model Fits and Significance of Parameters

Table 2 summarizes the results for tests of the goodness-of-fit of the models and of the importance of the threshold dose parameter and litter size covariable in the models.

In general, the models could adequately describe the dose–response pattern in about the same number of endpoints, although the models differed somewhat with respect to the endpoints that could be fit adequately (see below). When litter size was included as a covariable (Groups 1 and 4; see Fig. 1), 29 to 56 (6 to 11%) of the endpoints were not "fittable" by one model or another. When litter size was not a covariable (Groups 2, 3, 5, and 6), as many as 28% of the endpoints could not be fit by the LOG or NCTR models.

Likelihood ratio tests of the significance of the threshold parameter, d_0, revealed that d_0 contributed significantly to the fit of the model in very few cases. Among the data subsets where litter size was a covariable, the threshold parameter contributed significantly in only seven or eight (1.4%) of the cases for the LOG and NCTR models, and in about 4% of the cases for the RVR model. There were no instances in which the threshold parameter was significant for either the LOG or NCTR model when litter size was not included as a covariable. All of these rates are less than the rate that would be expected by chance alone (5%).

In no case was the addition of a threshold dose parameter sufficient to improve a poor model fit (observed vs expected, tested by χ^2 statistics). That is, for each endpoint in

Groups 4, 5, or 6 (run without the threshold dose parameter) that was poorly fit by a model, the same model also provided a poor fit to that endpoint even when the threshold dose parameter was allowed to be estimated (in Groups 1, 2, and 3, respectively).

There was substantial indication that the litter size covariable was an important contributor to the ability of the models to fit the data. For the LOG model, 118 (24%) of the likelihood ratio tests indicated that litter size was a significant factor. The importance of litter size was less evident for the NCTR model (12–13% of the likelihood ratio tests were significant) and for the RVR model (about 9% of the tests were significant).

In contrast to the case with the threshold dose parameter, the addition of litter size as a covariable was often sufficient to improve a poor model fit to acceptable levels (χ^2 p value > 0.01). For example, of the 43 endpoints in Group 3 for which the LOG model provided a poor fit without the litter size covariable, 23 of them were adequately fit by the LOG model with the litter size covariable (in Group 1).

The contribution of litter size to the ability of the models to fit the data did appear to be somewhat dependent on the type of endpoint. For the DS and DV endpoints (dead implants combined with fetuses having skeletal malformations or visceral malformations, respectively), litter size was a significant contributor more often than it was for the other endpoint types (Group 1 MLLs compared to Group 3 MLLs). Unfortunately, meaningful litter size information is often not available for these endpoints because the fetuses within litters are sometimes divided for skeletal or visceral examinations.

As mentioned above, the endpoints that the three models could not adequately fit differed somewhat from model to model. Figure 2 displays the relationships among the models, with respect to goodness-of-fit for Groups 4 and 5, the groups that do not include the threshold dose parameter but do, when possible (i.e., for Group 4), include a litter size covariable. In Group 4, 16 endpoints could not be fit by any model. In Group 5, there was considerable overlap for endpoints not adequately fit by the LOG model or the NCTR model.

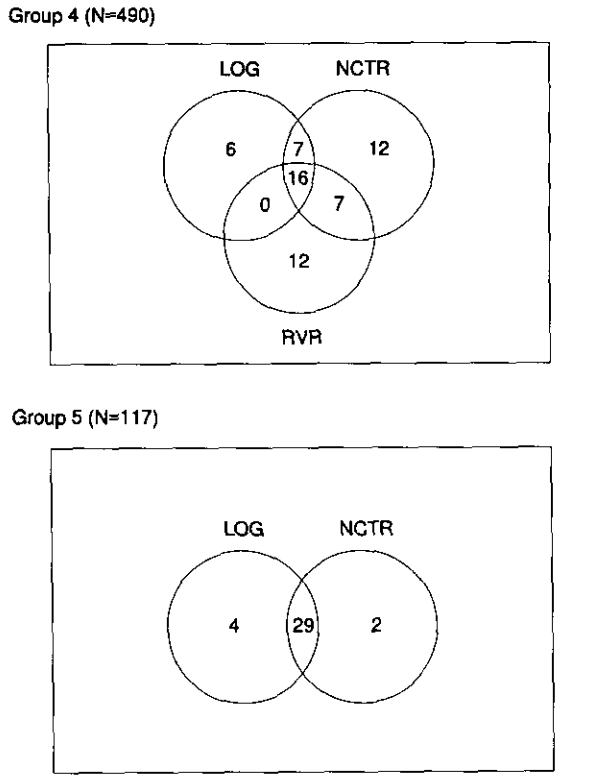

FIG. 2. Venn diagram of lack of model fits in Groups 4 and 5. Model runs in these groups did not include the threshold dose parameter, considered to be the "best" option. Lack of fit (either due to failure to converge or poor χ^2 fit to the observations) is indicated by the numbers inside the circles, where a circle corresponds to one of the models: LOG, NCTR, or RVR.

Comparison of DT Model Fits

Comparison of the MLLs across models provided an indication of the relative ability of the three models to describe the observed dose–response patterns and variations in litter responses within dose groups (Table 3). A larger MLL implied a better description of the data. The LOG model appeared to give a larger MLL than either the NCTR or RVR model in most of the cases examined.

Whenever litter size was included in the models (Groups 1 and 4), the LOG MLL was greater than the MLLs from the other two models a large majority of the time. The frequency of that occurrence was significantly different from the frequency expected if all models are equally likely to obtain the best fit (sign-test p values < 0.01). The comparison of the NCTR model and the RVR model revealed that the NCTR model was superior in a majority of cases (327 out of 469 and 352 out of 489, for Groups 1 and 4, respectively, where 469 and 489 are the number of cases for which both models converged and for which the two MLLs were different). Those frequencies, although not as large as the frequency with which the LOG was better than either the NCTR or RVR models, were also significantly greater than expected if the NCTR and RVR models were equally able to describe the results.

When litter size was not included, neither the LOG model nor the NCTR model was unequivocally better. In Groups 2 and 3 (which included the threshold dose parameter), the NCTR model was significantly better at describing the results. For the same endpoints, but with the LOG and NCTR models fit without the threshold dose parameter (Groups 5 and 6), the LOG model was better significantly more often.

The ability of the models to fit the observed results and

TABLE 3
Comparison of Maximum Likelihoods across Models

Group	Number of endpoints[a] with larger MLL from					
	LOG	vs NCTR	LOG	vs RVR	NCTR	vs RVR
1 (all parameters)	344*	146	397*	72	327*	142
2 and 3 (without litter size)	249	314*	—	—	—	—
4 (without threshold dose)	360*	129	415*	75	352*	137
5 and 6 (without litter size or threshold dose)	343*	243	—	—	—	—

[a] Includes endpoints fit by both of the models compared that had different MLLs.
* The difference is significant as assessed by a sign test.

the relative values of the likelihoods were not dependent on the type of endpoint. The models fit the dead implants endpoint, for example, with about the same frequency as they fit the total malformed or other endpoints. The MLL from the LOG model was greater than the MLL from the NCTR model 69–87% of the time for each type of endpoint (for Group 4); the average overall endpoint was 74% (Table 3).

BMD Comparisons

The relationships between the NOAELs or generic benchmarks and the benchmark dose estimates for the log-logistic model are displayed in Figs. 3 and 4, respectively. The results of analysis of Group 4 were selected for display because of the determination that litter size, but not threshold, was important in model performance. Risk estimates at the 5% added response level from the log-logistic model resembled both the QNOAEL and the CNOAEL (Fig. 3) and the continuous generic model results (Fig. 4). On the basis of standard deviations of the ratios, the results of the log-logistic model were more consistent with CBMD (a standard deviation for the ratio of 0.60) than with either NOAEL (standard deviations of 2.4 and 1.4 with the quantal and continuous NOAELs, respectively). As was observed when comparing the generic quantal model to the NOAELs (Allen *et al.*, 1994), the difference between the log-logistic model and the quantal generic model predictions was about three- to fivefold on average, with the latter usually (96% of the time) producing the more conservative result (Fig. 4, Table 4).

For Group 4, the three developmental-specific dose–response models yielded benchmark estimates at the 5% added risk level that were similar to one another (Fig. 5). Although overlap between the developmental model predictions was substantial, the RVR model tended to yield

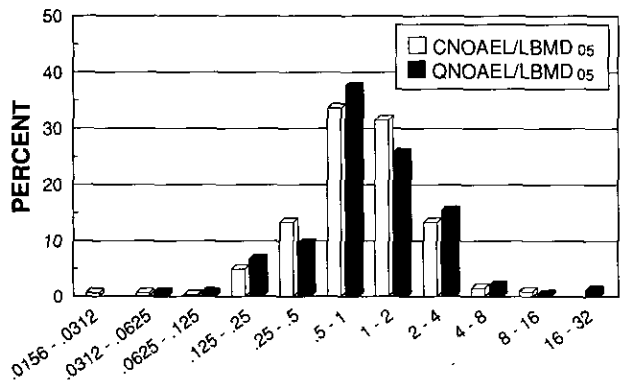

FIG. 3. Histograms of CNOAEL/LBMD$_{05}$ and QNOAEL/LBMD$_{05}$ ratios (Group 4). Ratios are shown for those endpoints with satisfactory fit by the LOG model, convergence to a LBMD$_{05}$ value in Group 4 and, for CNOAEL/LBMD$_{05}$, a significant dose-related increase in average proportion of fetuses affected and a non-zero CNOAEL (N = 324), or for QNOAEL/LBMD$_{05}$, a significant dose-related increase in number of litters with one or more affected fetuses and a non-zero QNOAEL (N = 253). The ratio intervals represent factors of 2. The mean (±SD) and median values for CNOAEL/LBMD$_{05}$ were 1.3 (±1.4) and 0.97, respectively; for QNOAEL/LBMD$_{05}$ they were 1.5 (±2.4) and 0.96.

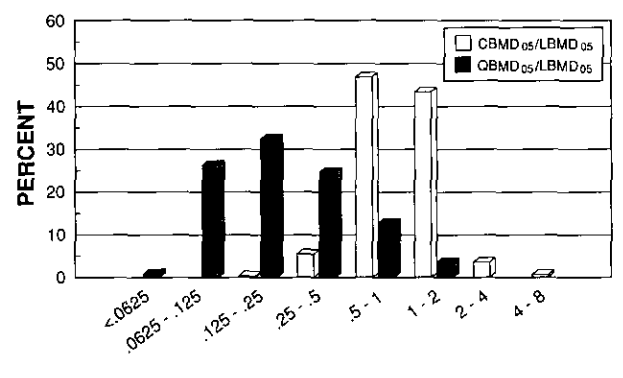

FIG. 4. Histograms of CBMD$_{05}$/LBMD$_{05}$ and QBMD$_{05}$/LBMD$_{05}$ ratios (Group 4). Ratios are shown for those endpoints with satisfactory fit by the LOG model, convergence to a LBMD$_{05}$ value in Group 4, and, for CBMD$_{05}$/LBMD$_{05}$, a significant dose-related increase in average proportion of fetuses affected and convergence of the continuous generic BMD model (N = 312), or for QBMD$_{05}$/LBMD$_{05}$, a significant dose-related increase in number of litters with one or more affected fetuses and convergence of the quantal generic BMD model (N = 260). The ratio intervals represent factors of 2. The mean (±SD) and median values for CBMD$_{05}$/LBMD$_{05}$ were 1.1 (±0.60) and 0.99, respectively; for QBMD$_{05}$/LBMD$_{05}$ they were 0.30 (±0.29) and 0.21.

TABLE 4
Summary Statistics for BMD and NOAEL Comparisons

Group	Model	Ratio	N^a	Mean (SD)	Median (1-Q range)^b	Percentage of endpoints with ratio <0.2	<0.5	<1	<2	<5
4	LOG	CNOAEL/$LBMD_{05}$	324	1.32 (1.37)	0.97 (0.91)	4.01	19.14	52.78	84.26	97.53
		QNOAEL/$LBMD_{05}$	253	1.52 (2.36)	0.96 (1.07)	5.53	17.79	55.34	81.03	96.84
		$CBMD_{05}$/$LBMD_{05}$	312	1.06 (0.60)	0.99 (0.33)	0.32	5.77	52.56	95.83	99.36
		$QBMD_{05}$/$LBMD_{05}$	260	0.30 (0.29)	0.21 (0.23)	47.69	83.85	96.54	100	100
	NCTR	CNOAEL/$NBMD_{05}$	331	1.32 (1.06)	1.05 (0.97)	3.32	13.90	46.83	82.78	99.09
		QNOAEL/$NBMD_{05}$	252	1.63 (2.96)	1.0 (1.06)	4.37	14.68	50.40	78.97	97.22
		$CBMD_{05}$/$NBMD_{05}$	325	1.08 (0.31)	1.03 (0.25)	0.0	0.92	40.0	97.85	100
		$QBMD_{05}$/$NBMD_{05}$	260	0.31 (0.28)	0.22 (0.23)	44.62	83.85	96.92	99.62	100
	RVR	CNOAEL/$RBMD_{05}$	266	1.32 (1.10)	1.06 (0.80)	4.89	14.29	45.11	84.96	98.12
		QNOAEL/$RBMD_{05}$	204	1.85 (3.46)	1.01 (1.00)	5.39	16.67	49.51	80.88	95.10
		$CBMD_{05}$/$RBMD_{05}$	256	1.04 (0.53)	0.94 (0.25)	0.0	1.56	59.38	96.09	100
		$QBMD_{05}$/$RBMD_{05}$	209	0.32 (0.36)	0.19 (0.22)	51.68	82.78	94.74	99.52	100
5	LOG	CNOAEL/$LBMD_{05}$	54	1.42 (1.02)	1.20 (1.16)	1.85	14.82	40.74	79.63	100
		QNOAEL/$LBMD_{05}$	49	2.00 (2.12)	1.49 (1.17)	2.04	14.29	28.57	73.47	91.84
		$CBMD_{05}$/$LBMD_{05}$	46	1.34 (0.94)	1.15 (0.61)	0.0	8.70	32.61	91.30	97.83
		$QBMD_{05}$/$LBMD_{05}$	53	0.32 (0.24)	0.25 (0.28)	33.96	83.02	96.23	100	100
	NCTR	CNOAEL/$NBMD_{05}$	62	1.40 (1.04)	1.13 (0.88)	1.61	9.68	41.94	80.65	98.39
		QNOAEL/$NBMD_{05}$	57	2.44 (5.78)	1.26 (1.47)	1.75	12.28	33.33	70.18	94.74
		$CBMD_{05}$/$NBMD_{05}$	52	1.33 (0.82)	1.13 (0.53)	0.0	1.92	30.77	92.31	98.08
		$QBMD_{05}$/$NBMD_{05}$	59	0.33 (0.24)	0.25 (0.24)	28.81	84.75	98.31	100	100

[a] The number of endpoints included in the comparison. Endpoints were excluded if the model(s) did not fit the data or if a valid BMD(s) could not be obtained (nonconvergence). For comparisons with CNOAEL and $CMBD_{05}$, only endpoints with a significant trend for proportion of fetuses affected were included. For comparisons with QNOAEL and $QBMD_{05}$, only endpoints with a significant trend for percentage of litters with one or more affected fetuses were included.

[b] Interquartile range.

results greater than those from the log-logistic model (p = 0.014 from a sign test) which, in turn, tended to be greater than results from the NCTR model ($p < 0.001$).

Because litter size was not always available from some study designs, we also examined results from Group 5 for comparison with results from Group 4. Relative to Group 4, similar findings were obtained from Group 5, although in this instance the log-logistic model tended to be slightly more conservative in relation to the QNOAEL and the generic continuous model (Figs. 6 and 7, Table 4). In fact, the results from the NCTR and log-logistic models, with two exceptions, did not differ by more than a factor of 2 (Fig. 7).

DISCUSSION

The BMD approach to defining reference doses and reference concentrations has many advantages over an approach that depends on the estimation of a NOAEL (Crump, 1984). The results of this investigation demonstrate that models exist that allow the estimation of BMDs and that are appropriate for commonly accepted protocols for conducting developmental toxicity tests. Such models take into consideration the correlated nature of the observations and allow factors other than dose (e.g., components of the litter environment such as litter size) to alter the probability of adverse outcome. The three DT models discussed

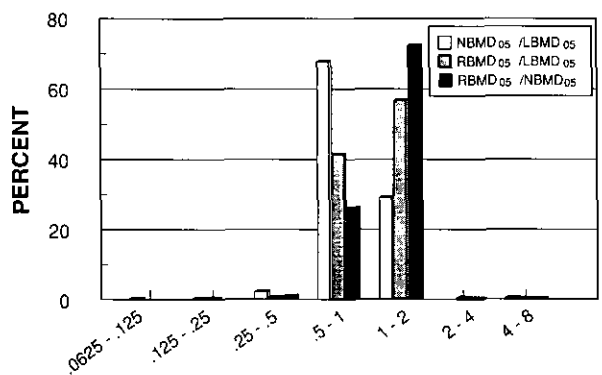

FIG. 5. Histograms of the ratios of BMD_{05}s from DT models (Group 4). Ratios are shown for all endpoints with satisfactory fit and convergence to lower bounds on dose in Group 4 for the two models being compared: the NCTR and LOG models (N = 346), the RVR and LOG models (N = 278), or the RVR and NCTR models (N = 283). The ratio intervals represent factors of 2. The mean (±SD) and median values for $NBMD_{05}$/$LBMD_{05}$ were 0.96 (±0.43) and 0.96, respectively; for $RBMD_{05}$/$LBMD_{05}$ they were 1.1 (±0.35) and 1.0; and for $RBMD_{05}$/$NBMD_{05}$ they were 1.1 (±0.32) and 1.1.

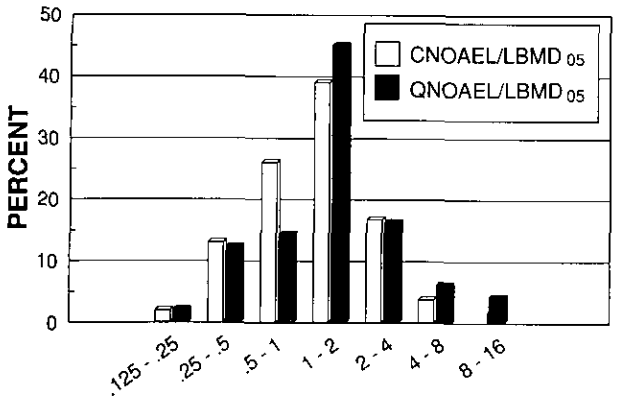

FIG. 6. Histograms of CNOAEL/LBMD$_{05}$ and QNOAEL/LBMD$_{05}$ ratios (Group 5). Ratios are shown for those endpoints with satisfactory fit by the LOG model, convergence to a LBMD$_{05}$ value in Group 5 and, for CNOAEL/LBMD$_{05}$, a significant dose-related increase in average proportion of fetuses affected, and a non-zero CNOAEL (N = 54), or for QNOAEL/LBMD$_{05}$, a significant dose-related increase in number of litters with one or more affected fetuses, and a non-zero QNOAEL (N = 49). The ratio intervals represent factors of 2. The mean (±SD) and median values for CNOAEL/LBMD$_{05}$ were 1.4 (±1.0) and 1.2, respectively; for QNOAEL/LBMD$_{05}$ they were 2.0 (±2.1) and 1.5.

here provide a powerful set of tools for estimating BMDs and ultimately the reference doses and reference concentrations that can form the basis for regulation and control of agents that cause developmental toxicity.

The three DT models investigated displayed an excellent capacity for fitting the dose–response patterns observed for over 600 endpoints from 141 actual developmental toxicity experiments. Although the endpoints from any single experiment were not all independent of one another (e.g., a total malformed endpoint would not be independent of an "all affected" endpoint, which includes the malformed fetuses as well), no two endpoints exactly duplicated one another. Inclusion of all the endpoints ensured that the most diverse set of dose–response patterns available was examined. The nonindependence of the endpoints may have resulted in a slight overrepresentation of endpoints that are poorly fit by the models (see example below), that have particular values for the ratios between NOAELs and BMDs, or that are better fit by one model rather than another. However, because there were no a priori reasons that any endpoint would display a particular pattern of model fit or NOAEL/BMD ratios, nonindependence of endpoints is unlikely to have affected the results of the investigation.

As an example, consider the fit of the three models to the entire database (Fig. 2). All three of the models were incapable of satisfactorily describing the dose–response pattern for only 45 of the 607 endpoints; 29 of those cases were from Group 5 (an appropriate litter size variable was not available) and 16 of the cases were from Group 4 (litter size variable available). The 16 endpoints from Group 4 are summarized in Table 5. Four of those endpoints were from a single experiment (the NTP study of 5-hydroxy tryptophan in mice); they all shared death/resorption as a component of their definitions and so they were not independent of one another. Note also that in all but 1 of the 16 cases, the average proportion of fetuses affected did not monotonically increase as dose increased. Although the lack of monotonicity did not always entail inability of the models to describe results, any dose–response model that assumes that the probability of response is monotone, as do the three DT models discussed here, will have difficulty fitting such patterns. It is possible that pharmacokinetic considerations might reveal that apparent differences in administered doses do not directly translate to differences in delivered doses and thus reduce some of the fit problems observed here. Note that such considerations modify the values of dose, not the models into which those dose values are input.

The difference between the rates of poor fit in Group 5 (29 out of 117; 25%) and Group 4 (16 out of 490; 3%) confirms the importance of litter size as a covariable for predicting rates of death/resorption and malformation. The importance of litter size as a covariable was also indicated by the comparison of the likelihoods in Groups 1 and 3 or Groups 4 and 6 (Table 3). For those comparisons, a substantial fraction of the endpoints modeled (9–24%, depending on the model) was significantly better described when litter size was included than when it was not included as a

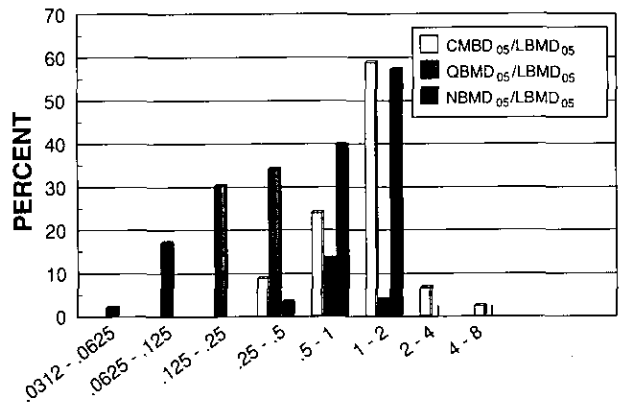

FIG. 7. Histograms of CBMD$_{05}$/LBMD$_{05}$, QBMD$_{05}$/LBMD$_{05}$, and NBMD$_{05}$/LBMD$_{05}$ ratios (Group 5). Ratios are shown for those endpoints with satisfactory fit by the LOG model, convergence to a LBMD$_{05}$ value in Group 5, and, for CBMD$_{05}$/LBMD$_{05}$, a significant dose-related increase in average proportion of fetuses affected and convergence of the continuous generic BMD model (N = 46), for QBMD$_{05}$/LBMD$_{05}$, a significant dose-related increase in number of litters with one or more affected fetuses and convergence of the quantal generic BMD model (N = 53), or for NBMD$_{05}$/LBMD$_{05}$, satisfactory fit by the NCTR model and convergence to a NBMD$_{05}$ value in Group 5 (N = 63). The ratio intervals represent factors of 2. The mean (±SD) and median values for CBMD$_{05}$/LBMD$_{05}$ were 1.3 (±0.94) and 1.2, respectively; for QBMD$_{05}$/LBMD$_{05}$ they were 0.32 (±0.24) and 0.25; and for NBMD$_{05}$/LBMD$_{05}$ they were 0.99 (±0.21) and 1.0.

TABLE 5
Dose–Response Patterns Not Fit by Any DT Model[a]

Number of dose groups	Dose–response pattern	Number of endpoints with that pattern	χ^2 p values[b]
4	Average proportion affected in control and low dose groups greater than that in third group	1	3E-5–2E-3
	Average proportion affected in controls greater than that in second and third groups	2	2E-6–5E-3
	Average proportion affected in controls greater than that in low dose group	1	3E-3–6E-3
5	None affected in first three groups	1	3E-3–8E-3
	Average proportion affected in controls greater than that in low dose group	1	3E-3–7E-3
	Average proportion affected in controls greater than that in next two dose grops	2	5E-4–2E-3
	Average proportion affected in controls greater than that in low dose; third greater than fourth	3	3E-5–6E-3
	Average proportion affected in third group greater than that in fourth group	2	3E-3–8E-3
	Monotone increasing average proportion affected; nonmonotone proportion of litters with one or more affected fetus	1	1E-6–8E-3
6	Fifth and second groups have greatest average proportions affected	2	2E-3–7E-3

[a] Results from the Group 4 runs, without the threshold dose parameter.
[b] Range of p values from the three models, assessing the fit of the predicted to the observed numbers affected. 2E-3 denotes 2×10^{-3}, for example.

covariable. Moreover, of the 43 cases for which the LOG model could not satisfactorily fit endpoints in the Group 3 runs (without litter size), 23 (53%) of those unsatisfactory fits were improved to satisfactory fits in the corresponding runs with litter size (in Group 1; see Table 2).

Williams (1987) and Carr and Portier (1991), who have discussed the role of litter size as a covariable in the context of the original model of Rai and Van Ryzin (1985), concluded that litter size alone was not sufficient to explain the extrabinomial variation typically observed in response rates from developmental toxicity studies. The results of the present investigation suggest that, while litter size may not completely account for that variation, it does increase the ability of the models to represent actual death/resorption and malformation results even when extrabinomial variation is accounted for by other means. That is, the β-binomial distributions assumed by these models account for extrabinomial variation, but litter size as a covariable still significantly increases model suitability in a substantial number of cases.

It is interesting that the contribution of litter size to the model is smallest for the generalized RVR model (Table 2). This fact and the tendency for the NCTR model and, especially, the LOG model to fit the data better than the RVR model (Table 3) are related to the constraints imposed on the litter size covariable in the three models.

In the RVR model, the function controlling the effect of litter size can be expressed as

$$f_{\mathrm{RVR}}(s) = \exp\{-s(\theta_1 + \theta_2 d)\},$$

where the term in parentheses is constrained to be positive. For a fixed dose level, $f(s)$, the probability of response decreases (or remains constant) as litter size increases. This is true for all dose levels, although the rate of decrease may differ from one dose level to another.

The NCTR model can be expressed as follows to emphasize the role of litter size:

$$P_{\mathrm{NCTR}}(d, s) = 1 - g(d)^{*}\exp\{-s(\theta_1 + \theta_2 d^{\gamma})\}$$

Although there are constraints on the parameters θ_1 and θ_2, $(\theta_1 + \theta_2 d^{\gamma})$ can assume both positive and negative values. Thus, the probability of response can be either an increasing or a decreasing function of litter size. Moreover, $(\theta_1 + \theta_2 d^{\gamma})$ may be positive for some dose levels and negative for other dose levels; the probability of response may increase as litter size increases for some dose levels in any particular experiment while probability of response decreases as litter size increases for other doses in that experiment. The NCTR model has more flexibility with respect to litter size than does the RVR model.

The LOG model is even more flexible with respect to litter size than either the RVR or the NCTR models. Whereas the NCTR and RVR models are restricted to monotone changes in probability of response as litter size changes, for a fixed dose level, the LOG model can represent relationships that are nonmonotonic, even when dose is constant. Specifically, because the LOG model involves a ratio with $(r - \theta_1 s)$ in the numerator and $(1 + h(d)^{*}\exp\{\theta_2 s\})$ in the denominator, it can describe "U-shaped" relationships between probability of response and litter size. Those relationships have probability of response decreasing for smaller litter sizes to a minimum value, at say s_{m}, and then increasing for litter sizes greater than s_{m}. The presence or absence of the U shape, the rate of decrease

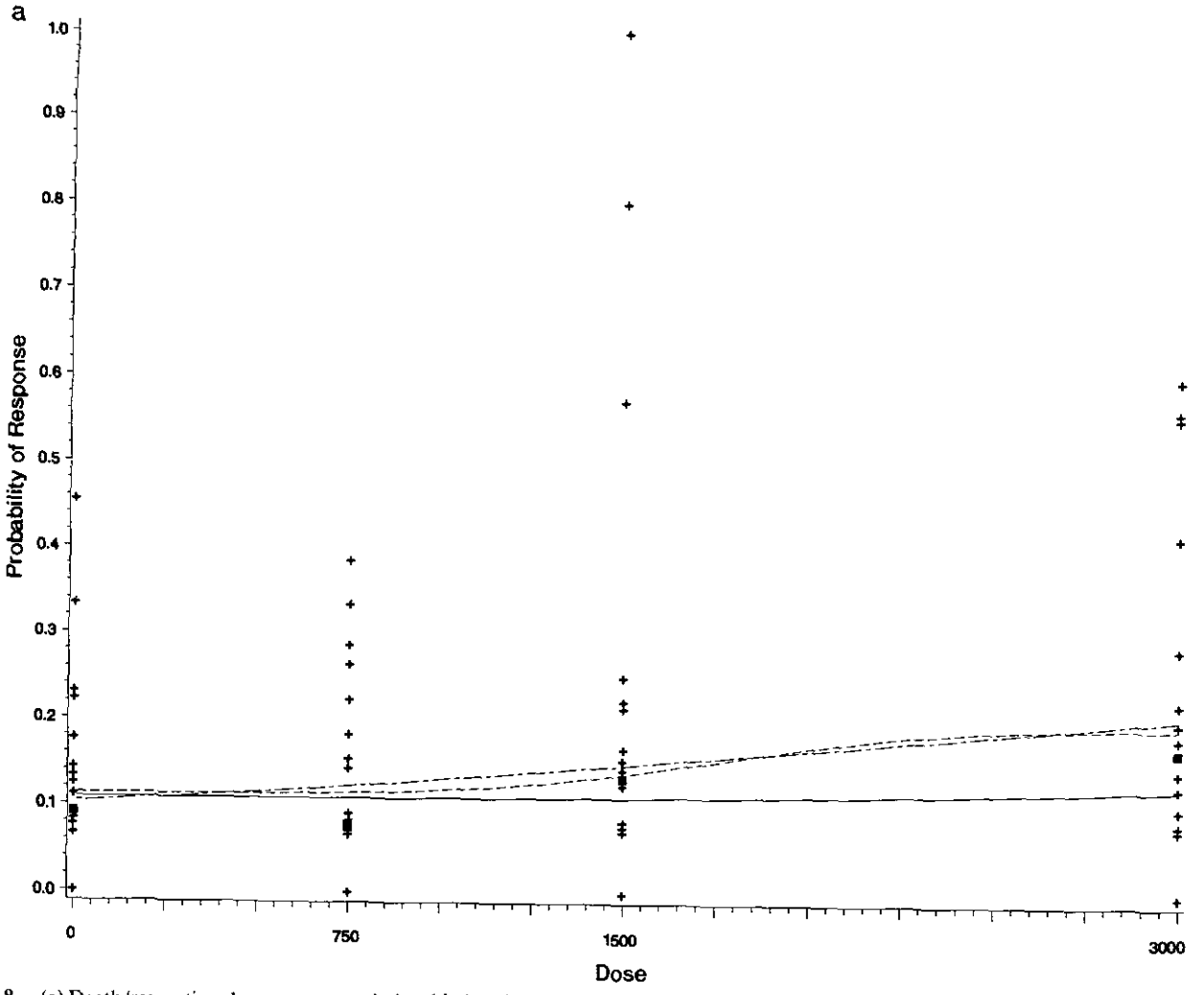

FIG. 8. (a) Death/resorption dose–response relationship in mice exposed to ethylene glycol. Individual response rates (+) and median response rates (■) for each dose group are compared to the predictions from the LOG (——), NCTR (– · – –), and RVR (– – –) models for a fixed litter size. The litter size was fixed at the average size observed over all groups (12.547 implants per dam). (b) Observed and predicted death/resorption response rates by dose and litter size in mice exposed to ethylene glycol. Observed values (means shown as circles) are compared to the predictions from the LOG model (——) and the RVR model (- - -). The vertical bars represent ranges of observed values, not confidence intervals. Note the nonlinearity of the LOG model predictions as a function of litter size, reflecting the nonmonotonic pattern of the observed means.

and increase for the probability of response, and the value of s_m can differ from one dose level to another.

An example of the advantage of the LOG model with respect to litter size is provided by the NTP study of ethylene glycol exposure in mice. The LOG model provided a satisfactory fit to the data (χ^2 p value of 0.45), the RVR model fit marginally (p value of 0.06), and the NCTR model fit poorly (p value of 0.008). The dose–response predictions of the three models for a fixed litter size (Fig. 8a) do not indicate overwhelming differences among the models in comparison to the observed response rates. A better indication is provided by examining the litter size-specific predictions (Fig. 8b). Although the number of dams within a dose group with any particular litter size (number of implants) is small, it appears that the rate of death/resorption within the dose group is not a monotonic function of litter size. The RVR model (as well as the NCTR model, not shown) predicted decreasing response rates for increasing litter size. The LOG model, on the other hand, predicted U-shaped litter size–response curves for all but the control group (for which it predicted increasing response rates with increasing litter size). The litter size at which the minimum value of the LOG-predicted response rate was attained increased from 9 to 12 to 13 as the dose increased from 750 to 1500 to 3000 mg/kg.

The relative freedom of the LOG model to describe litter size–response relationships allows it to fit observed results more accurately than the other two models. This accounts, at least in part, for the tendency of that model to provide the largest of the likelihoods among the three models (Table 3).

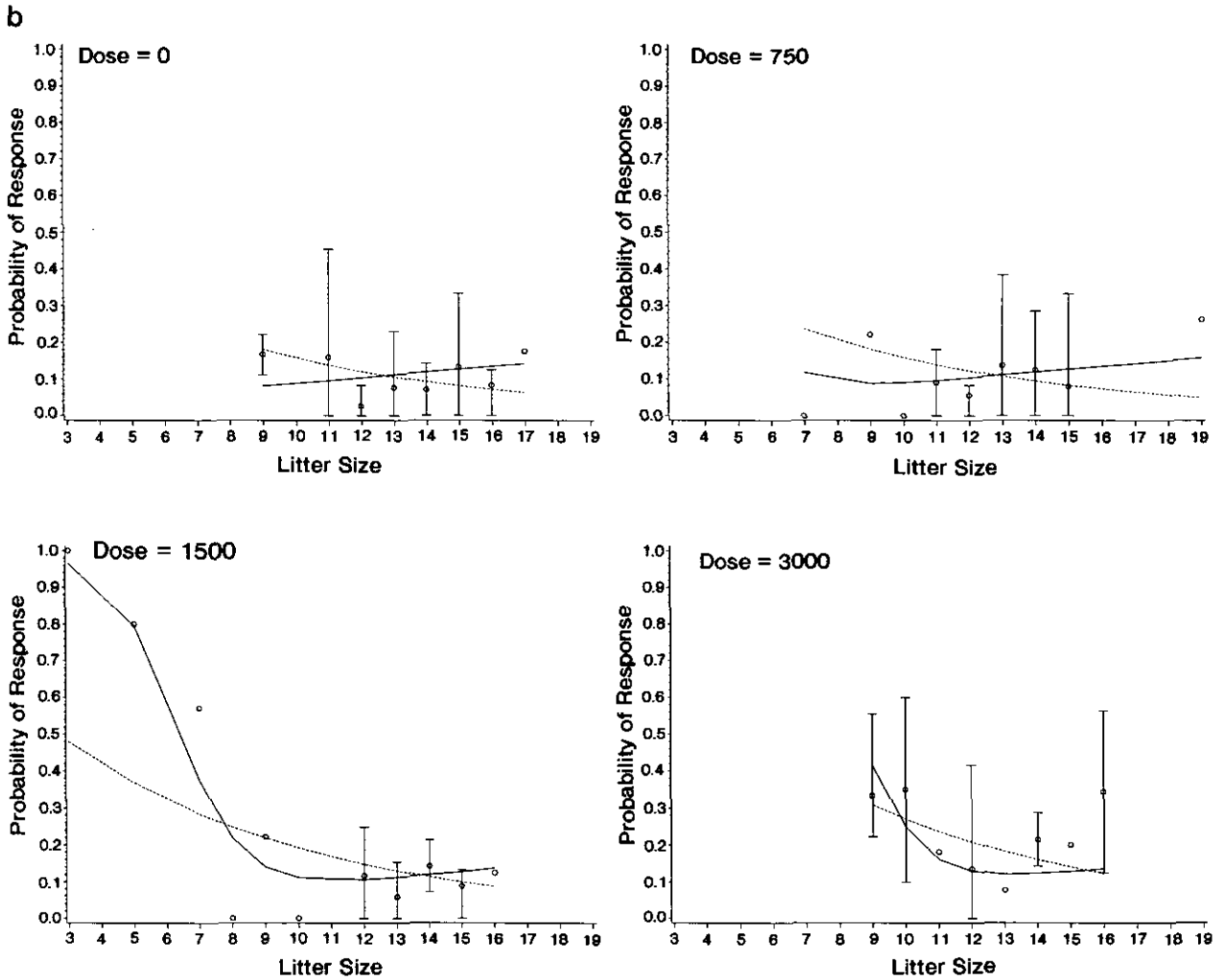

FIG. 8—*Continued*

It also accounts for the greater importance of litter size in the LOG model compared to the other two models (Table 2): if the other two models cannot adequately represent the observed litter size–response relationship, then taking litter size out of those models will not significantly change the fits of those models to the data.

The significance of litter size as a covariable, especially in a model that is free to fit a wide range of relationships between litter size and response, is, as stated earlier, present even when extrabinomial variation is built into these models by the β-binomial assumption. The importance of litter size as an explanatory variable, especially for the DS and DV endpoints, suggests that developmental toxicity experiments be designed so that an appropriate litter size variable is available for modeling all endpoints. This was not the case for some of the experiments in our database, namely those that separated the live births into two groups, one of which received visceral examinations while the other received skeletal examinations. Alternatively, extensions of DT models could be considered that account for the "real" litter size (number of implants or number of live births) as well as the number of fetuses examined.

The apparent significance and importance of the litter size covariable were not shared by the threshold dose parameter. No matter which model was considered or whether litter size was included as a covariable or not, the threshold dose parameter added little to the ability of the DT models to fit the data (Table 2). In no case did inclusion of the threshold dose parameter change a model fit from unacceptable (χ^2 p value < 0.01) to acceptable. In the few cases in which the likelihood ratio tests indicated the significance of the threshold parameter, the χ^2 fit to the data was acceptable even without the threshold.

There are practical and theoretical advantages to dropping a threshold dose parameter from the models. Operationally, the programs implementing the models, especially the RVR model, execute more rapidly and converge more often when the threshold dose parameter is eliminated.

Moreover, elimination of that parameter gives an extra degree of freedom for the χ^2 goodness-of-fit statistic, which can be important for assessing model fit when the expected (predicted) numbers of affected fetuses are small. In that case, litters of similar size are combined (within dose group), which reduces the degrees of freedom but yields expected numbers large enough for a good approximation in the χ^2 test.

The theoretical advantage of ignoring a threshold dose parameter is that it avoids confusion concerning the existence of a "biological" threshold for the endpoint being modeled. When included, the estimate of the threshold dose parameter in the models is based solely on the observed response rates at the doses included in a particular study. That estimate would not reflect other biological considerations related to the existence of a biological threshold (e.g., repair, compensation, redundancy, etc.) nor would it necessarily provide the best estimate of the value for such a threshold. Given the complexity of the biological processes leading to observable developmental toxicity, it is not clear how the model-based threshold dose parameter would relate to a biological threshold, if one exists. The model-based threshold dose is simply a parameter that gives the model an additional degree of flexibility to describe the dose–response pattern observed; it estimates the dose below which the best-fitting model predicts no increase in response rates for a particular experiment. Based on the results of this investigation, this added flexibility is not needed to describe observed dose–response patterns.

Although the LOG model has theoretical advantages over the NCTR and RVR models with respect to the handling of litter size as a covariable, an advantage that was empirically apparent for the set of endpoints analyzed in this investigation (Table 3), it was not always the case that the LOG model provided the best fit. The RVR and NCTR models gave a greater maximized likelihood than the LOG model for about 15 and 26%, respectively, of the endpoints analyzed in Group 4 (with litter size but without a threshold dose parameter). The three models, considered as a battery of available dose–response functions, allow some flexibility to describe the variety of developmental toxicity results represented in the database. Judicious choice of a model, based in part on examination and consideration of the role of litter size and other determinants of model fit, will facilitate estimation of BMDs that will be appropriate for developmental toxicants. Because the BMD_{05}s estimated by the three DT models tended to be very similar to one another (Fig. 5), the basis for model selection need not hinge on the relative magnitude of the BMD estimates themselves. In this sense, the results of this investigation support the argument (Crump, 1984) that model choice is not a critical factor for BMD estimation (because model-based extrapolation to low doses is not required).

The BMD approach for assessing and regulating risks of developmental toxicity is proposed as an improvement over the NOAEL approach (Crump, 1984; Kimmel and Gaylor, 1988). Part of that improvement includes proper accounting for sample size through use of statistically appropriate lower confidence limits on dose. It is important, then, to have statistical models that represent the underlying features of the data; a danger is that unrealistic models will underestimate the variability and overestimate BMDs. The DT models discussed here do estimate the dose– and litter size–response relationships while accounting for extra-binomial variation and therefore provide a suitable basis for developmental toxicity BMD estimation.

Ryan (1992) discusses alternative approaches to risk assessment for developmental toxicity, emphasizing the generalized estimating equations (GEE) technique (Williams, 1982; Liang and Zeger, 1986; Zeger and Liang, 1986; McCullagh and Nelder, 1989). The GEE approach does not require a specific distributional assumption about the nature of the correlation among responses. The results of this investigation suggest, however, that the correlation structure entailed by the β-binomial assumption represents very well the actual patterns observed in developmental toxicity tests. The DT models presented here, incorporating both a β-binomial assumption and litter size as a covariate, are more flexible than those compared by Ryan to the GEE approach.

The statistical models presented here serve the purposes of BMD estimation. We also hope that consideration and application of such models will spur interest in refinement and extension of the models, for example, as suggested above, in relation to litter size as distinct from the number of fetuses examined. Another example of a modeling extension includes consideration of multiple outcomes simultaneously (Chen *et al.*, 1991; Ryan, 1992; Catalano *et al.*, 1993). The ultimate extension, requiring the interaction of developmental toxicologists, pharmacokineticists, and modelers, will be the development of more biologically based models that will move us beyond the BMD or NOAEL paradigm and allow us to consider the continuum of developmental effects ranging from functional deficits to growth alteration to malformation and death (Kimmel and Gaylor, 1988).

ACKNOWLEDGMENTS

The investigators thank the following investigators and laboratories for so generously allowing us to use their experimental data: Dr. Bern Schwetz, NTP/NIEHS: Dr. Joseph Holson, Mark Nemec, and Stan Kopp, Wil Labs; Dr. Maureen Feuston, Environmental and Health Sciences Laboratory, Mobil Oil Corp.; Drs. Mildred Christian and Allen Hoberman, Argus Labs; Dr. Kate Smith, EPA Labs; Mr. Mike Narotsky, ManTech; and Mr. James Schardein, IRDC Labs. Special thanks go to Lynn Williams for tireless manuscript preparation and revision; to Zamyat Kirby for technical assistance in preparing the data; and to Cynthia Van Landingham and

S. Eric Brooks for expert programming, data management, and graphics support.

REFERENCES

Allen, B. C., Kavlock, R. J., Kimmel, C. A., and Faustman, E. M. (1994). Dose-response assessment for developmental toxicity. II. Comparison of generic benchmark dose estimates with no observed adverse effect levels. *Fundam. Appl. Toxicol.* **23,** 487–495.

Anderson, E., and Carcinogen Assessment Group of the U.S. Environmental Protection Agency (1983). Quantitative approaches in use to assess cancer risk. *Risk Anal.* **3,** 277–295.

Carr, G., and Portier, C. (1991). An evaluation of the Rai and Van Ryzin dose-response model in teratology. *Risk Anal.* **11,** 111–120.

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

Chen, J., Kodell, R., Howe, R. B., *et al.* (1991). Analysis of trinomial responses from reproductive and developmental toxicology experiments. *Biometrics* **47,** 1049–1058.

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

Crump, K., Howe, R., and Kodell, R. (1990). Permutation tests for detecting teratogenic effects. In *Statistics in Toxicology* (D. Krewski and C. Franklin, Eds.), pp. 347–375. Gordon and Breach Science Publishers, New York.

Faustman, E. M., Allen, B. C., Kavlock, R. J., and Kimmel, C. A. (1994). Dose-response assessment for developmental toxicity. I. Characterization of database and determination of no observed adverse effect levels. *Fundam. Appl. Toxicol.* **23,** 478–486.

Faustman, E. M., Wellington, D. G., Smith, W. P., and Kimmel, C. A. (1989). Characterization of a developmental toxicity dose-response model. *Environ. Health Perspect.* **78,** 229–241.

Gladen, B. (1979). The use of the jackknife to estimate proportions from toxicological data in the presence of litter effects. *J. Amer. Stat. Assoc.* **746,** 278–283.

Haseman, J. K., and Kupper, L. L. (1979). Analysis of dichotomous response data from certain toxicological experiments. *Biometrics* **35,** 281–293.

Howe, R. B., Van Landingham, C., and Allen, B. C. (1992). TERAMOD, TERALOG, and TERAVAN; software developed by K. S. Crump Division of Clement International. TERAMOD marketed and maintained under a cooperative research and development agreement with National Center for Toxicological Research, Jefferson, AR.

Kimmel, C., and Gaylor, G. (1988). Issues in qualitative and quantitative risk analysis for developmental toxicology. *Risk Anal.* **8,** 15–21.

Kodell, R., Howe, R., Chen, J., and Gaylor, D. (1991). Mathematical modelling of reproductive and developmental toxic effects for quantitative risk assessment. *Risk Anal.* **11,** 583–590.

Kupper, L., and Haseman, J. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **34,** 69–76.

Kupper, L., Portier, C., Hogan, M., and Yamamoto, E. (1986). The impact of litter effects on dose-response modeling in teratology. *Biometrics* **42,** 85–98.

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

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

Rai, K., and Van Ryzin, J. (1985). A dose-response model for teratological experiments involving quantal response. *Biometrics* **41,** 1–9.

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

U.S. Environmental Protection Agency (EPA) (1991). Guidelines for developmental toxicity risk assessment. *Fed. Reg.* **56,** 63798–63826.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31,** 941–952.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Appl. Stat.* **31,** 144–148.

Williams, D. A. (1987). Dose-response models for teratological experiments. *Biometrics* **43,** 1013–1016.

Zeger, S., and Liang, K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.