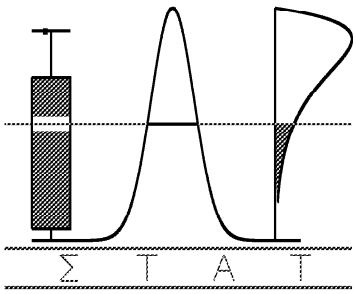
T E C H N I C A L

R E P O R T

0223

On the Use of Fractional Polynomial Predictors for Quantitative Risk Assessment in Developmental Toxicity Studies

C. Faes, H. Geys, M. Aerts, and G. Molenberghs



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

On the Use of Fractional Polynomial Predictors for Quantitative Risk Assessment in Developmental Toxicity Studies

C. Faes, H. Geys, M. Aerts and G. Molenberghs
Center for Statistics, Biostatistics, Limburgs Universitair Centrum, Diepenbeek, Belgium

Abstract

Developmental toxicity studies are designed to assess the potential adverse effects of an exposure on developing foetuses. Safe dose levels can be determined using dose-response modelling. To this end, it is important to investigate the effect of misspecifying the dose-response model on the safe dose. Since classical polynomial predictors are often of poor quality, there is a clear need for alternative specifications of the predictors, such as fractional polynomials. By means of simulations, we will show how fractional polynomial predictors may resolve possible model misspecifications and may thus yield more thrustworthy estimates of the benchmark doses.

Keywords: benchmark dose, beta-binomial model, conditional model, developmental toxicity, dose-response, fractional polynomials

1 Introduction

Lately, society has been increasingly concerned about public health problems. Especially problems related to fertility and pregnancy, birth defects, and developmental abnormalities are of major concern. Regulatory agencies, such as the U.S. Environmental Protection Agency (EPA) and the Food and Drug Administration (FDA) therefore stimulate reproductive and developmental toxicity research. One of the goals is to understand the causes of the problems. Further,

one wants to better protect people from exposures with an increased risk, such as drugs, harmful chemicals and other environmental hazards. A zero-exposure of all possible toxic agents would be the ideal case, although this is not realizable in modern society.

There are several strategies to investigate the relationship between possible harmful exposures and developmental toxicity. For example, epidemiology studies on humans can be used. However, reliable epidemiological information is often limited or unavailable (Budtz-Jørgensen, Keiding and Grandjean 2001). As an alternative, controlled experiments in laboratory animals can be conducted in advance of human exposure. Drugs and other possible toxic agents are specifically tested on pregnant animals to safeguard against possible teratogenic effects (such as malformations and low birth weight) on the human foetus. For ethical reasons, animal studies afford a greater level of control than epidemiological studies. However, methods for extrapolating the results to humans are still being developed and refined. In this work we will focus on a typical developmental toxicity study with a Segment II design. This involves exposing pregnant dams (mice, rats and occasionally rabbits) during the period of major organogenesis and structural development to a compound of interest. Dose levels for this design consist of a control group and three or four exposed groups, each with 20 to 30 pregnant animals. The dams are sacrificed just prior to normal delivery, at which time the uterus is removed and the contents are thoroughly examined for the occurence of defects. The primary outcomes of interest are thus typically dichotomous.

An important issue in developmental toxicity is the risk assessment. Risk assessment can be defined as "the use of available information to evaluate and estimate exposure to a substance and its consequent adverse health effects" (Roberts and Abernathy 1996), and thus deals with safety issues and regulation of exposures with potential adverse effects. One goal of interest in the area of risk assessment is the examination of the dose-response relationship, i.e., the dependence of a particular outcome (e.g. the number of dead foetuses, the risk of a malformed foetus, ...) on the dose which is administered to the dam. Statistical analysis must account for the structure of the data typical for developmental toxicity research. For

instance, rodents have multiple births, so clustering of offsprings within litters will complicate the analysis. Two different probability models describing such data are briefly introduced in Section 3.1. Another important goal in the risk assessment process is to determine a safe level of exposure, i.e., quantitative risk assessment (QRA). Different approaches to estimate a safe dose exist. Quantitative risk assessment can be performed via the "No Observable Adverse Effect Level" (NOAEL) approach. The NOAEL, however, has been criticized for its poor statistical properties (Leisenring and Ryan 1992). The estimation of the NOAEL depends on the design of the experiment, on the sample size and on the number of dose groups, and it does not allow calculating a measure of variability of the estimation. Alternatively, QRA can be based on the fitted dose-response models (Crump 1984). This has a number of important advantages. It allows adding a measure of variability to the point estimation of a safe dose, it can incorporate special features of the structure of developmental toxicity studies, ... (Williams and Ryan 1996). Because of the disadvantages of the NOAEL approach, and because of the benefits of basing quantitative risk assessment on dose-response modelling, the latter approach will be considered here. Of course, to get trustworthy results, models should fit the data well in all respects. Since classical polynomial predictors are often of poor quality, especially when low dose extrapolation is envisaged, there is a clear need for alternative specification of the predictors describing main effects and associations. Due to the small number of dose groups in developmental toxicity studies, penalized splines and other non-parametric methods (Simonoff 1996) are less suitable. Further, the use of non-linear predictor functions invokes non-trivial statistical problems, such as the lack of identifiability of the null hypothesis of no dose effect (Davidian and Giltinan 1995). Fractional polynomial predictors (Royston and Altman 1994) provide a more elegant approach, and still fall within the realm of (generalized) linear models. They are the topic of interest in Section 3.2. We will investigate the behaviour of fractional polynomials in the context of quantitative risk assessment through extensive simulations. The fractional polynomials are much more flexible to attain the correct benchmark dose than conventional polynomials, as will be seen in Section 4 and 5. In addition they can correct for possible misspecification of the probability models.

2 Quantitative Risk Assessment

The standard approach to quantitative risk assessment based on dose-response modelling requires the specification of an adverse event, along with its risk expressed as a function of dose. For developmental toxicity studies where offsprings are clustered within litters, there are several ways to define the concept of an adverse effect. From a biological perspective one might argue that it is important to take into account the health of the entire litter when modelling risk as a function of dose. Therefore, we will focus on the probability that at least one foetus of the litter has the adverse effect under consideration. Thus risk assessment will be litter-based, with the risk function $r(d)$ representing the probability of observing a malformation at dose level d for at least one foetus within the litter (Declerck, Molenberghs, Aerts and Ryan 2000).

Based on this probability, a common measure for the excess risk over background is given by

$$r^*(d) = \frac{r(d) - r(0)}{1 - r(0)},$$

where greater weights are given to outcomes with larger background risk. Assuming that the chemical results in more adverse effects at non-zero dose d compared to dose level 0, the excess risk ranges from 0 to 1. This definition of the excess risk measures the relative increase in risk above background.

The benchmark dose is then defined as the dose corresponding to a very small increase in risk over background. More formally, the benchmark dose (BMD_q) is defined as the dose satisfying $r^*(d) = q$, where q corresponds to a pre-specified level of increased response and is typically specified as $0.01, 1, 5$ or 10% (Crump 1984). Of course, the use of dose-response models to set a safe limit of exposure is far more complicated than determining a NOAEL, but it offers a number of important advantages. We can account for special features of the data, we can incorporate other covariates of interest, we obtain a measure of the degree of variability, etc. (Williams and Ryan 1996).

Because the dose-response curve is estimated from the data and has inherent variability,

the benchmark dose itself is an estimate of the true dose that would result in the corresponding level of excess risk. This sampling uncertainty for the model on which the benchmark dose is based can be acknowledged, by replacing the benchmark dose by a lower confidence limit. Several approaches exist (Williams and Ryan 1996, Kimmel and Gaylor 1988, Crump and Howe 1983,...). A well known approach is the use of the lower effective dose, where an upper limit for the risk function is used to determine a safe dose level. The lower effective dose (LED_q) is thus defined as the solution of

$$\hat{r}^*(d) + 1.645\sqrt{\widehat{\mathrm{Var}}\big(\hat{r}^*(d)\big)} = q,$$

where q corresponds with the pre-specified level of increased response, and the variance of the estimated increased risk function $\hat{r}^*(d)$ is estimated as

$$\widehat{\mathrm{Var}}\big(\hat{r}^*(d)\big) = \Big(\frac{\partial r^*(d)}{\partial \boldsymbol{\beta}}\Big)^T \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})\Big(\frac{\partial r^*(d)}{\partial \boldsymbol{\beta}}\Big)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}},$$

with $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

3 Dose-Response Models

When performing risk assessment based on the fitted dose-response model, the model should fit the data well. This has implications for both the model family chosen (the probability model), as well as for the form of the predictors. While the probability model can take special features of the data into account, the predictor model must take care of the flexibility of the model.

3.1 Probability Model

A dose-response model describing developmental toxicity data must take the structure of the data into account. Interest goes to the risk of observing a malformation (skeletal, visceral or external), binary coded as absent/present. Further, we must account for the litter effect induced by the clustering of offsprings within litters. Different types of probability models are available,

namely conditional, marginal and cluster-specific models. The answer to the question of which model family is to be preferred depends principally on the research question(s) to be answered. In conditionally specified models the probability of a positive response for one member of the cluster is modeled conditionally on other outcomes for the same cluster, while marginal models relate the covariates directly to the marginal probabilities. Cluster-specific models differ from the previous models by the inclusion of parameters that are specific to the cluster. Several other issues are involved: should analysis be based on a multivariate outcome rather than on a collapsed version (any malformation), should the litter size be incorporated into the model, etc. (Chen and Kodell 1989). In this paper, we restrict attention to a selection of likelihood-based dose-response models for univariate clustered binary data: the beta-binomial model (Williams 1975) and the conditional exponential family model of Molenberghs and Ryan (1999). Due to the popularity of marginal and random-effects models for correlated binary data, the conditional models have received little attention, especially in the context of multivariate clustered data. The conditional approach has been criticized because the interpretation of the dose effect on the risk of one outcome is conditional on the responses of other outcomes for the same individual, outcomes of other individuals and the litter size (Diggle, Liang and Zeger 1994). Molenberghs, Declerck and Aerts (1998) and Aerts, Declerck and Molenberghs (1997) have compared marginal, conditional and random-effects models. Their results are encouraging for the conditional models, since they are competitive for the dose effect testing and for benchmark dose estimation, and because they are computationally fast and stable.

Consider an experiment involving N litters (pregnant dams), the ith of which contains n_i individuals, each of whom are examined for the presence or absence of a malformation. Suppose $Y_{ij} = 1$ indicates whether the jth individual in cluster i is abnormal, and 0 otherwise. Then, define $Z_i = \sum_{j=1}^{n_i} Y_{ij}$, the total number of malformations in cluster i. Covariates of interest are the treatment or dosing d_i given to cluster i. Further, we assume exchangeability within a litter, i.e., each foetus in the same litter has an identical malformation probability, and the association between each pair of foetuses within the same litter is equal.

Rather than modelling marginal functions directly, a popular approach is to assume a random effects model in which each litter has a random parameter. The beta-binomial model assumes that, conditional on litter size n_i and malformation probability of any foetus in litter i, the number of malformations Z_i in the ith cluster follows a binomial distribution. To account for the litter effect, i.e., the cluster effect, the underlying malformation probabilities are assumed to vary within a litter according to a beta distribution with mean π_i. This leads to the beta-binomial distribution of the number of malformations Z_i in cluster i, and is expressed by

$$f(z_i;\pi_i,\rho_i,n_i)=\binom{n_i}{z_i}\frac{B\big(\pi_i(\rho_i^{-1}-1)+z_i,(1-\pi_i)(\rho_i^{-1}-1)+n_i-z_i\big)}{B\big(\pi_i(\rho_i^{-1}-1),(1-\pi_i)(\rho_i^{-1}-1)\big)}, \tag{1}$$

where $B(.,.)$ denotes the beta function (Skellam 1948, Kleinman 1973). The association parameter ρ_i in this model indicates the correlation between two binary responses of litter i. Note that both the parameters π_i and ρ_i of the beta-binomial model have a marginal interpretation.

To model the marginal parameters π_i and ρ_i we use a composite link function. Since we have binary responses, we could use the logistic link function for the mean parameter π_i. However, other link functions, such as the probit link, the log-log link or the complementary log-log link, could be chosen too. An appropriate transformation for the association parameter ρ_i is Fisher's z-transformation. This gives us the following generalized linear regression relations

$$\begin{pmatrix}\ln\left(\frac{\pi_i}{1-\pi_i}\right)\\ \ln\left(\frac{1+\rho_i}{1-\rho_i}\right)\end{pmatrix}\equiv\eta_i=\boldsymbol{X_i\beta},$$

where X_i is a design matrix and β is a vector of unknown parameters. A frequently used model in literature is

$$\boldsymbol{X_i}=\begin{pmatrix}1 & d_i & 0\\ 0 & 0 & 1\end{pmatrix}\ \text{and}\ \boldsymbol{\beta}=\begin{pmatrix}\beta_0\\ \beta_d\\ \beta_a\end{pmatrix} \tag{2}$$

with a logit-linear dose trend for the mean parameter, and a constant association parameter ρ. Obviously, this model can be extended by adapting the design matrix and the vector of regression parameters, such that the logit of π_i depends on dose via e.g. a quadratic or higher

order polynomial function. Also, the association parameter ρ_i can be modeled as some function of dose.

Molenberghs and Ryan (1999) proposed a likelihood-based model for clustered binary data, based on a multivariate exponential family model (Cox 1972). The model describes the probability of an outcome given values for the other outcomes, and therefore is conditional in nature. Molenberghs and Ryan (MR) considered $Y_{ij} = 1$ if the jth foetus in cluster i exhibits the adverse event of interest, and -1 otherwise. This coding is preferred above the 1/0 coding, since it provides a parameterization that more naturally leads to desirable properties when the roles of success and failure are reversed (Cox and Wermuth 1994). They proposed the distribution of z_i, the number of individuals from cluster i with positive response, as

$$f(z_i; \theta_i, \delta_i, n_i) = \exp\Big\{\theta_i z_i - \delta_i z_i(n_i - z_i) - A(\theta_i, \delta_i)\Big\}, \tag{3}$$

with θ_i the main parameter, δ_i the association parameter describing the association between pairs of individuals within the ith cluster and $A(\theta_i, \delta_i)$ the normalizing constant. The parameters θ_i and δ_i can be modeled as $(\theta_i, \delta_i)' = X_i\beta$ with X_i and $\boldsymbol{\beta}$ as in (2). Note that, as with the beta-binomial model, this model reduces to the logistic regression model in the absence of clustering. More details about model properties and inference can be found in Molenberghs and Ryan (1999).

Subsequently we will show how the beta-binomial model and the conditional model of Molenberghs and Ryan (1999) can easily handle litter-based rates.

For the beta-binomial model, the probability that at least one foetus in a litter of size n_i is abnormal, is

$$q(n_i; d) = 1 - \frac{B\big(\pi_i(\rho_i^{-1} - 1), (1 - \pi_i)(\rho_i^{-1} - 1) + n_i\big)}{B\big(\pi_i(\rho_i^{-1} - 1), (1 - \pi_i)(\rho_i^{-1} - 1)\big)}.$$

It can be shown that this expression equals

$$q(n_i; d) = 1 - \prod_{k=0}^{n_i - 1} \Big(1 - \pi_i + \frac{k\pi_i\rho_i}{1 + (k-1)\rho_i}\Big).$$

Now, consider all values of n_i, the number of implants, with non-zero probability $P(n_i)$. The litter-based risk, corresponding to some specified dose d, is given by

$$r(d) = \sum_{n_i} P(n_i) q(n_i; d),$$

which is an average of conditional probabilities $q(n_i; d)$ with weights $P(n_i)$. The excess risk can be computed as

$$r^*(d) = 1 - \frac{\sum_{n_i} P(n_i) \prod_{k=0}^{n_i-1} \big(1 - \pi_i(d) + k\pi_i(d)\rho_i/(1 + (k-1)\rho_i)\big)}{\sum_{n_i} P(n_i) \prod_{k=0}^{n_i-1} \big(1 - \pi_i(0) + k\pi_i(0)\rho_i/(1 + (k-1)\rho_i)\big)}.$$

The exponential model of Molenberghs and Ryan (1999) also allows easy calculation of quantities such as the probability that at least one littermate is affected. Given the number of viable foetuses n_i, the probability of observing at least one abnormal foetus in a cluster is

$$q(n_i; d) = 1 - \exp\big(-A_{n_i}(\boldsymbol{\Theta_i})\big).$$

Integrating over all possible values of n_i, we obtain the risk function

$$r(d) = \sum_{n_i=0}^{\infty} P(n_i)\big[1 - \exp\big(-A_{n_i}(\boldsymbol{\Theta_i})\big)\big],$$

where $P(n_i)$ is the probability of observing n_i viable foetuses in a pregnant dam. Using this equation, calculation of the excess risk $r^*(d)$ is straightforward.

3.2 Predictor Model

For risk assessment to be reliable, models should fit the data well in all aspects. Although classical polynomials are very customary, they are often inadequate, especially when low dose extrapolation is envisaged. Therefore, we need alternative specifications of the predictors describing main effects and associations. Apart from penalized spline methods and non-linear dose-response models (Davidian and Giltinan 1995), a very elegant alternative approach to classical polynomials, which falls within the realm of (generalized) linear methods, is given by

fractional polynomials (Royston and Altman 1994). They provide much more flexibly shaped curves than conventional polynomials, but in cases where the extension is not necessary, this family essentially reduces to conventional polynomials. Thus, their use is strongly recommended and considering a conventional and a fractional polynomial approach simultaneously, is certainly a worthwhile sensitivity analysis in an important public health matter such as the determination of safe limits for human exposure to potentially hazardous agents. Let us briefly describe the procedure.

For a given degree m and an argument $d > 0$ (e.g., dose), fractional polynomials are defined as

$$\beta_0 + \sum_{j=1}^{m} \beta_j d^{p_j},$$

where the β_j are regression parameters and $d^0 \equiv \ln(d)$ and the powers $p_1 < \ldots < p_m$ are positive or negative integers or fractions. Royston and Altman (1994) argue that polynomials with degree higher than 2 are rarely required in practice and further restrict the powers of dose to a small predefined set of possibly non-integer values: $\Pi = \{-2, -1, -1/2, 0, 1/2, 1, 2, \ldots, \max(3, m)\}$. For example, setting $m = 2$ generates 4 "quadratics" in powers of d (represented by $(1/d, 1/d^2)$, $(1/\sqrt{d}, 1/d)$, $(\sqrt{d}, d)$, (d, d^2)), a quadratic in $\ln(d)$ and other curves which have shapes different from those of conventional low degree polynomials.

The full definition includes possible "repeated powers" which involve powers of $\ln(d)$. For example, a fractional polynomial of degree $m = 3$ with powers $(-1, -1, 2)$ is of the form $\beta_0 + \beta_1 d^{-1} + \beta_2 d^{-1} \ln(d) + \beta_3 d^2$ (Royston and Altman 1994, Sauerbrei and Royston 1999).

4 Asymptotic Simulation Study

In this section we perform an asymptotic simulation study to investigate the effect of model misspecifications on quantitative risk assessment. In addition, we investigate to which extent the use of flexible predictor models based on fractional polynomials can correct for such misspecification.

In order to get asymptotic information on the effect of model misspecification, we follow the particular recommendations of Rotnitzky and Wypij (1994). An artificial sample is constructed, where each possible realization is weighted according to its true probability under a given true model. In our case, we need to consider all realizations of the form (n_i, z_i, d_i), and have to specify: (1) $f(d_i)$, the relative frequenties of the dose groups, as prescribed by the design; (2) $f(n_i|d_i)$, the probability with which each cluster size can occur, possibly depending on the dose level (we will assume $f(n_i|d_i) = f(n_i)$), and (3) $f(z_i|n_i, d_i)$, the actual model probabilities. We assume that there are 4 dose groups, with one control group ($d_i = 0$) and three exposed groups ($d_i = 0.25, 0.5, 1.0$), and that each dose group has an equal probability (i.e., $f(d_i) = 1/4$). The number n_i of viable foetuses per cluster is assumed to follow a local linear smoothed version of the relative frequency distribution given in Table 1 of Kupper et al. (1996), which is considered representative of that encountered in actual experimental situations. Least squares cross-validation has been used to choose the bandwidth. The smoothed frequencies are presented in Aerts, Declerck and Molenberghs (1997).

Data are generated from the beta-binomial model with a given non-linear predictor for the mean parameter. Different dose trends on the mean parameter π of the true model can be considered. Here, we look at two different models. The first model (Model A) is defined as $\text{logit}(\pi) = \beta_0 + \beta_d \sinh^2(d)$, the second model (Model B) as $\text{logit}(\pi) = \beta_0 + \beta_d \cos(d)$. In both models, the association parameter ρ is kept constant. We use parameter settings that were encountered in real data sets (Price et al. 1985,1987). The parameter settings are summarized in Table 1. In the beta-binomial model the baseline risk is a function of both the intercept and association parameter. In model A, an intercept of -4 and association of 0.1, 0.3 and 0.5 corresponds to a background rate of respectively 16%, 11% and 9%. The Fisher-transformed correlation of 0.1, 0.3 and 0.5 corresponds to respectively a correlation of about 0.05, 0.15 and 0.24. The parameters of model B all correspond with the same baseline malformation rate of about 15%. A Fisher-transformed correlation of 0.1 is used (correlation of about 0.05). Figure 1 shows the dose-response models for the different parameter settings of model B considered here.

Table 1: Parameter Settings of the True (beta-binomial) Model.

parameter	Model A	Model B
	$\sinh^2(d)$	$\cos(d)$
intercept β_0	-4	$2k$
dose effect β_d	4, 6, 8	$-4-2k$
		$(k = 0, 1, \ldots, 7)$
association β_a	0.1, 0.3, 0.5	0.1

FIGURE 1 ABOUT HERE

In this research, the technique introduced by Rotnitzky and Wypij (1994) is tailored to compute "asymptotic" values of the estimated benchmark dose. The benchmark dose is determined for the artificial sample under three different models:

- the beta-binomial model, with a conventional linear predictor for the mean π and a constant association ρ (Model 1);
- the conditional model of Molenberghs and Ryan, with a conventional linear predictor for the main parameter θ and a constant association parameter ρ (Model 2);
- the conditional model of Molenberghs and Ryan, with the best fitting fractional polynomial predictor for the main parameter θ and a constant association parameter ρ (Model 3).

Different misspecifications occur in the above models. In the first model, the form of the predictor is misspecified. This often occurs in practice, when one uses a linear polynomial predictor. In the second model, the probability model is misspecified. Because the true dose-response model is unknown in general, this is a realistic misspecification. And also in the third model, the predictor model is misspecified, but here we try to correct for it using a fractional polynomial predictor.

In choosing the best fractional polynomial, we follow the ideas of Royston and Altman (1994). Polynomials with degree higher than two are not taken into account, and powers of dose are further resticted to the set of values $\Pi = \{-2, -1, -1/2, 0, 1/2, 1, 2, 3\}$. We consider as the best fractional polynomial model, the one producing the smallest value of Akaike's Information Criterium among the eight models with one regressor and 36 models with two regressors, and which shows a monotonic behaviour.

Results are summarized in Tables 2 (Model A) and 3 (Model B). The "true" benchmark dose is found by fitting the correct model (i.e., the model under which the data were generated) and by calculating the purely model based benchmark dose. The asymptotically estimated benchmark doses are determined under all three models (Models 1, 2 and 3).

Table 2: Asymptotic Estimation of Benchmark Dose under Model A

(β_0, β_d, ρ)	True Model	Model 1	Model 2	Model 3
(-4,4,0.1)	0.344	0.123	0.178	0.310
(-4,6,0.1)	0.283	0.082	0.169	0.257
(-4,8,0.1)	0.246	0.062	0.160	0.229
(-4,4,0.3)	0.390	0.160	0.186	0.354
(-4,6,0.3)	0.321	0.107	0.173	0.276
(-4,8,0.3)	0.279	0.080	0.160	0.242
(-4,4,0.5)	0.425	0.192	0.210	0.210
(-4,6,0.5)	0.350	0.128	0.185	0.305
(-4,8,0.5)	0.305	0.096	0.166	0.255

Let us first have a look at the results when data are generated under Model A (Table 2). None of the estimated benchmark doses are equal to the true benchmark dose. This of course is due to the misspecification of the model. The conventional polynomial results for both the beta-binomial and conditional model (respectively model 1 and 2) are very low compared with the true benchmark dose. There is a small decrease in asymptotic bias when the dose-parameter

Table 3: Asymptotic Estimation of Benchmark Dose under Model B

(β_0, β_d, ρ)	True Model	Model 1	Model 2	Model 3
(0,-4,0.1)	0.502	-0.014	0.261	0.484
(2,-6,0.1)	0.408	-0.016	0.204	0.401
(4,-8,0.1)	0.353	-0.013	0.179	0.351
(6,-10,0.1)	0.315	-0.011	0.165	0.315
(8,-12,0.1)	0.288	-0.009	0.155	0.283
(10,-14,0.1)	0.266	-0.008	0.147	0.256
(12,-16,0.1)	0.249	-0.007	0.143	0.234
(14,-18,0.1)	0.235	-0.006	0.144	0.219

increases, but the difference with the true benchmark dose stays much too large. While this seems cautious, Morgan (1992) warns that safe dose determination should be tempered by common sense. For example, blind use of an overly conservative procedure has been regarded as scientifically indefensible by the Scientific Committee of the British Food Safety Council (1980), since it may produce unrealistically low safe doses. The fractional polynomial results (Model 3) are much closer to the true benchmark dose. This seems to indicate that the fractional polynomials are much more flexible to attain the correct benchmark dose than the conventional linear polynomial. The only major discrepancy is seen for $(-4, 4, 0.3)$. Here, the best fitting fractional polynomial reduced to the conventional linear predictor.

When data are generated under Model B (Table 3), the conclusions are similar and even more encouraging for the fractional model. The estimated benchmark doses for the beta-binomial model with a linear predictor (Model 1) are very small, and even negative. This is due to the misspecification of the polynomial predictor, leading to unrealistically low doses. And also the estimated benchmark doses for the model of Molenberghs and Ryan with a conventional linear predictor (Model 2) take very small values. Again, the fractional polynomials seem to correct for the model misspecification. As can be seen, the estimated benchmark doses attained

from the model of Molenberghs and Ryan, using the best fitting fractional polynomial, are very close to the true benchmark dose.

In order to investigate whether these conclusions also hold for classical random samples, a small sample simulation study was performed.

5 Small Sample Simulations

The same models and parameter combinations as in the asymptotic study are investigated (Table 1). For each parameter setting, 1000 datasets of 30 observations per dose group were generated. The estimated benchmark doses were averaged at the end of the run, and mean squared errors (MSE) were calculated. Results are displayed in Tables 4 and 5.

Table 4: Small Sample Estimation of BMD (MSE) under Model A

(β_0, β_d, ρ)	True Model	Model 1	Model 2	Model 3
(-4,4,0.1)	0.344	0.126 (0.048)	0.186 (0.026)	0.295 (0.005)
(-4,6,0.1)	0.283	0.084 (0.040)	0.179 (0.012)	0.259 (0.002)
(-4,8,0.1)	0.246	0.064 (0.033)	0.167 (0.007)	0.235 (0.001)
(-4,4,0.3)	0.390	0.163 (0.052)	0.194 (0.040)	0.288 (0.019)
(-4,6,0.3)	0.321	0.052 (0.075)	0.170 (0.024)	0.230 (0.012)
(-4,8,0.3)	0.279	0.082 (0.039)	0.170 (0.013)	0.241 (0.003)
(-4,4,0.5)	0.425	0.196 (0.054)	0.214 (0.046)	0.288 (0.033)
(-4,6,0.5)	0.350	0.130 (0.049)	0.204 (0.023)	0.270 (0.012)
(-4,8,0.5)	0.305	0.099 (0.043)	0.180 (0.017)	0.250 (0.005)

The results of the asymptotic study and the small sample study are remarkably well in agreement. While the conventional polynomial results with both the beta-binomial and conditional model (respectively Model 1 and 2) are smaller than the true benchmark dose,

Table 5: Small Sample Estimation of BMD (MSE) under Model B

(β_0, β_d, ρ)	True Model	Model 1	Model 2	Model 3
(0,-4,0.1)	0.502	-0.016 (0.269)	0.274 (0.054)	0.400 (0.029)
(2,-6,0.1)	0.408	-0.016 (0.180)	0.209 (0.040)	0.337 (0.014)
(4,-8,0.1)	0.353	-0.013 (0.134)	0.185 (0.029)	0.308 (0.008)
(6,-10,0.1)	0.315	-0.011 (0.106)	0.170 (0.022)	0.269 (0.007)
(8,-12,0.1)	0.288	-0.009 (0.088)	0.161 (0.017)	0.245 (0.006)
(10,-14,0.1)	0.266	-0.007 (0.075)	0.153 (0.013)	0.222 (0.005)
(12,-16,0.1)	0.249	-0.007 (0.065)	0.151 (0.010)	0.210 (0.004)
(14,-18,0.1)	0.235	-0.006 (0.058)	0.150 (0.008)	0.201 (0.003)

the fractional polynomial results (Model 3) are much closer to the true benchmark dose. For instance, for the parameter setting $(-4, 6, 0.5)$, the estimated benchmark doses for models 1, 2 and 3 correspond with an increased risk of respectively 1%, 3% and 5%, while the true benchmark dose corresponds with an increase of 10%. In Figure 2, we present the 1000 benchmark doses for the different datasets generated from Model A with parameters (-4,6,0.5) in a scatterplot matrix. Benchmark doses of the three different models are compared, and also the true benchmark dose is marked (by a "T") on the figures. It is clear that Model 3 is the most flexible model in attaining the correct benchmark dose. Although use of a more flexible model yields higher standard errors, the increase in variability is small compared to the increase in bias, as summarized by the mean squared error.

Figure 2 About Here

To acknowledge the sampling uncertainty for the model on which the benchmark dose is based, we replace the BMD by the LED. Table 6 summarizes the LED estimations for model B. Results for Model A are similar. For both conventional and fractional polynomial predictors in the conditional model of Molenberghs and Ryan (Model 2 and 3), we show the (mean) estimated

LED, the percentage of the LED's smaller than the true BMD, the mean difference of the LED's smaller than the true BMD and the mean distance of the LED's larger than the true BMD.

Table 6: Lower Effective Dose when True Model has $\cos(d)$ Trend(Model B)

True Model	Model 2			Model 3			
parameters	LED	perc <	dist <	LED	perc <	dist <	dist >
(0,-4,0.1)	0.229	100	0.273	0.357	97.4	0.150	0.018
(2,-6,0.1)	0.174	100	0.235	0.302	96.4	0.111	0.013
(4,-8,0.1)	0.153	100	0.200	0.275	94.2	0.084	0.013
(6,-10,0.1)	0.140	100	0.175	0.238	94.6	0.082	0.013
(8,-12,0.1)	0.134	100	0.154	0.215	94.8	0.078	0.016
(10,-14,0.1)	0.153	100	0.139	0.195	98.2	0.073	0.010
(12,-16,0.1)	0.127	100	0.122	0.184	98.7	0.066	0.009
(14,-18,0.1)	0.126	100	0.109	0.176	99.2	0.059	0.005

The estimated LED's using fractional polynomial predictors, seem to behave quite well for realistic datasets. Around 95% of the lower effective doses of the 1000 generated samples are smaller than the true BMD, while the difference with the true benchmark dose stays small. Also the small percentage of the estimated doses which are larger than the true benchmark dose are very close to the true BMD. When using Model 2, all estimated doses are smaller than the true benchmark dose, moreover the distance with the true safe dose is large. This confirms the conclusion that the estimated dose is too small when using the conventional linear predictor. In contrast, the fractional polynomials provide satisfactory results.

This indeed shows that, in order to determine a safe limit of exposure, models should fit the data well. This has implications for both the model family chosen as well as for the form of the predictors. Even when the probability model is known, unreliable and unrealistically safe doses can be found. This demonstrates the importance of the shape of the predictors. In practice

however, the true dose-response model is not known. Moreover, the choice between different dose-response models is often subjective and can affect the quantitative risk assessment. Using a flexible polynomial predictor, such as a fractional polynomial, can partly solve the effects of model misspecification on QRA.

6 Concluding Remarks

Developmental toxicity studies are complicated by the hierarchical, clustered and multivariate nature of the data. As a consequence, a multitude of modelling strategies have been proposed in literature. Such choices are often subjective and can affect the quantitative risk assessment based on the fitted models. A study of the possible effects of misspecifying the dose-response model on QRA is therefore an important issue.

Blind use of conventional linear predictors in the dose-response model can yield unrealistically low or unreliable safe doses, even when the probability model is well specified. Flexible parametric models cannot only correct for a misspecification of the predictor model, they can even correct for possible misspecification of the probability model. Therefore, the fractional polynomial approach is important when searching for safe limits for human exposure to hazardous agents.

One concern, often raised for developmental toxicity studies is the danger of potential overfitting. Indeed, a standard teratology study typically involves no more than 4 or 5 different dose levels. Therefore, we have restricted ourselves to a (small) discrete set of fractional polynomials, with degree one or two. In general however, more design points would be desirable, but, from a practical point of view, such experiments are hard to manage in the developmental toxicity context. Other flexible parametric models could be considered too, such as models based on non-linear predictors, penalized splines, ... In contrast with the fractional polynomials, which are easy to handle with, these methods pose non-trivial methodological challenges.

When defining the LED as a lower confidence limit for the BMD, we considered a two-

sided confidence interval. Future research will focus on one-sided confidence intervals, not necessarily based on asymptotic normality. Another topic of current research is the determination of the BMD based on several adverse effects including continuous outcomes like weight.

Acknowledgment

The first two authors gratefully acknowledge support from the Institute for the Promotion of Innovation by Science and Technology (IWT) in Flanders, Belgium. Research supported by a PAI program P5/24 of the Belgian Federal Government (Federal Office for Scientific, Technical, and Cultural Affairs).

References

Aerts, M., Declerck, L., Molenberghs, G.(1997). "Likelihood Misspecification and Safe Dose Determination for Clustered Binary Data," *Environmetrics*, **8**, 613–627.

Budtz-Jørgensen, E., Keiding, N., Grandjean, P.(2001). "Benchmark Dose Calculation from Epidemiological Data," *Biometrics*, **57**, 698–706.

Chen, J.J., Kodell, R.L. (1989). "Quantitative Risk Assessment for Teratologic Effects," *Journal of the American Statistical Association*, **84**, 966–971.

Cox, D.R. (1972). "The Analysis of Multivariate Binary Data," *Applied Statistics*, **21**, 113–120.

Cox, D.R., Wermuth, N. (1994). "A Note on the Quadratic Exponential Binary Distribution," *Biometrika*, **81**, 403–408.

Crump, K.S., Howe, R.B. (1983). "A review of methods for calculating statistical confidence limits in low dose extrapolation," in Clayson, D.B., Krewki, D. and Mundro, I.

(eds.),*TOxicological Risk Assessment. Volume I: Biological and Statistical Criteria*, Boca Raton: CRC Press, pp. 187–203.

Davidian, M., Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*, London: Chapman and Hall.

Declerck, L., Molenberghs, G., Aerts, M., Ryan, L.(2000). "Litter-based methods in developmental toxicity risk assessment," *Environmental and Ecological Statistics*, **7**, 57–76.

Diggle, P.J., Liang, K.Y., Zeger, S.L. (1994). *Analysis of Longitudinal data*, Oxford: Oxford University Press.

Kimmel, G.L., Gaylor, D.W.(1988). "Issues in Qualitative and Quantitative Risk Analysis for Developmental Toxicology," *Risk Analysis*, **8**, 15–20.

Kleinman, J.C. (1973). "Properties with extraneous variance: single and independent samples," *Journal of the American Statistical Association*, **68**, 46–54.

Leisenring, W., Ryan, L. (1992). "Statistical Properties of the NOAEL," *Regulatory Toxicology and Pharmacology*, **15**, 161.

Molenberghs, G., Declerck, L., Aerts, M (1998). "Misspecifying the Likelihood for Clustered Binary Data," *Computational Statistics and Data Analysis*, **26**, 327–350.

Molenberghs, G., Ryan, L.M. (1999). "An Exponential Family Model for Clustered Multivariate Binary Data," *Environmetrics*, **10**, 279–300.

Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*, Chapman and Hall, London.

Price, C.J., Kimmel, C.A., Tyl, R.W., and Marr, M.C. (1985). "The Developmental Toxicity of Ethylene Glycol in Rats and Mice," *Toxicology and Applied Pharmacology*, **81**, 113–127.

Price, C.J., Kimmel, C.A., George, J.D., and Marr, M.C. (1987). "The Developmental Toxicity of Diethylene Glycol Dimethyl Ether in Mice," *Fundamental and Applied Toxicology*, **8**, 115–126.

Royston, P., Altman, D.G. (1994). "Regression using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling," *Applied Statistics*, **43**, 429–467.

Roberts, W.C., Abernathy, C.O. (1996). "Risk assessment: principles and methodologies," in A. Fan and L.W. Chang (eds.), *Toxicology and Risk Assessment, Principles, Methods and Applications*, New York: Marcel Dekker Inc., pp. 245–270.

Rotnitzky, A., Wypij, D. (1994). "A Note on the Bias of Estimators with Missing Data," *Biometrics*, **50**, pp. 1163–1170.

Sauerbrei, W., Royston, P. (1999). "Building Multivariate Prognostic and Diagnostic Models: Transformation of the Predictors by Using Fractional Polynomials," *Journal of the Royal Statistical Society, Series A*, **162**, 71–94.

Scientific Committee of the Food Safety Council(1980). "Proposed system for food safety assessment," *Food and Cosmetic Toxicology*, **16**, Supplement 2, 1–136 (1978). Revised report published June 1980 by the Food Safety Council, Washington, DC.

Simonoff, J.S. (1996). *Smoothing methods in statistics*, New York: Springer.

Skellam, J.G. (1948). "A probability Distribution derived from the binomial distribution by regarding the probability of success as variable betwees the sets of trials," *Journal of the Royal Statistical Society, Series B*, **10**, 257–261.

U.S. Environmental Protection Agency(1991). "Guidelines for Developmental Toxicity Risk Assessment," *Federal Register*, **56**, 63798 63826.

Williams, D.A. (1975). "The analysis of binary responses from toxicology experiments involving reproduction and teratogenicity," *Biometrics*, **38**, 150.

Williams, P.L., Ryan, L.M. (1996). "Dose-Response Models for Developmental Toxicology", in R.D. Hood (ed.), *Handbook of Developmental Toxicology*, New York: CRC Press, pp. 635–666.