Environmental and Ecological Statistics **9,** 357–377, 2002

Robust benchmark dose determination based on profile score methods

GERDA CLAESKENS,[1] MARC AERTS,[2]
GEERT MOLENBERGHS[2] and LOUISE RYAN[3]

[1]*Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.*
E-mail: gerda@stat.tamu.edu
[2]*Biostatistics, Center for Statistics, Limburgs Universitair Centrum, B-3590 Diepenbeek, Belgium*
E-mail: marc.aerts@luc.ac.be; geert.molenberghs@luc.ac.be
[3]*Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, MA 02155, U.S.A.*
E-mail: ryan@jimmy.harvard.edu

Received September 2000; Revised December 2001

We investigate several methods commonly used to obtain a benchmark dose and show that those based on full likelihood or profile likelihood methods might have severe shortcomings. We propose two new profile likelihood-based approaches which overcome these problems. Another contribution is the extension of the benchmark dose determination to non full likelihood models, such as quasi-likelihood, generalized estimating equations, which are widely used in settings such as developmental toxicity where clustered data are encountered. This widening of the scope of application is possible by the use of (robust) score statistics. Benchmark dose methods are applied to a data set from a developmental toxicity study.

Keywords: clustered binary data, generalized estimating equations, likelihood ratio, profile likelihood, score statistic, toxicology

1. Introduction

This work is motivated by toxicological experiments designed to assess the potential adverse effects of drugs or other exposures on developing fetuses of pregnant rodents (usually mice or rats). In a typical study, pregnant females are assigned to different treatment groups, usually one control group and a small number of dosed groups. The exposure occurs early in gestation, the animals are sacrificed prior to term and the fetuses are examined for malformations. The quantity of interest is the proportion of malformed fetuses as a function of the dosage. Since littermates are likely to show similar behavior, this kind of experiments results in cluster correlated binary outcomes (malformation: yes or no). Emphasis can be placed on estimating a dose effect parameter, on testing the null hypothesis of no dose effect, or on determining a benchmark dose.

Limitations on the availability of epidemiological data in many settings means that risk assessment in toxicology is often based on experimental results in non-human species such

as mice and rats, which are exposed to some relatively high dose or exposure levels. Extrapolation of the results to humans, i.e., the prediction of the level of impact of the particular exposure on humans, is a complex issue that we do not discuss further. For more information and a discussion on the existence of a ''safe threshold'', refer to Gad (1999, Chapter 15).

Historically, the no-observed-adverse-effect-level, the highest dose that does not differ significantly from the control, has been used for risk assessment in non-cancer settings. Disadvantages of this method are well-known. A major drawback is that it is restricted to be one of the chosen experimental dose levels. Crump (1984) develops a more quantitative approach, based on dose response modeling techniques.

In Section 2 we will start from the full likelihood-based definition of benchmark dose determination to introduce the score-based version. Section 3 resumes the profile likelihood version and gives a score equivalent. Section 4 shows situations where these methods may have severe shortcomings. Possible solutions are given in Section 5. Further results from simulation studies are given in Section 6. These methods are applied to a dataset from the United States National Toxicology Program in Section 7. Some open problems and discussion are provided in Section 8.

1.1 *The benchmark dose*

A first step in estimating developmental risk involves fitting a dose-response model to the data. An example of a dose-response function is a linear model on the logit scale:

$$\pi(d;\boldsymbol{\beta}_\pi) = \frac{1}{[1+\exp\{-(\beta_0+\beta_d d)\}]}, \tag{1}$$

with $\pi(d;\boldsymbol{\beta}_\pi)$ the probability of the event of interest, e.g., malformation or death, for an individual fetus from a sacrificed pregnant female exposed to dose level d and with $\boldsymbol{\beta}_\pi = (\beta_0, \beta_d)$ the unknown model parameter. The probability $\pi(0;\boldsymbol{\beta})$ denotes the background response rate, corresponding to the control group.

Second, one defines an excess risk function. We will use the multiplicative or relative excess risk

$$r(d;\boldsymbol{\beta}) = \frac{\pi(d;\boldsymbol{\beta}) - \pi(0;\boldsymbol{\beta})}{1-\pi(0;\boldsymbol{\beta})}, \tag{2}$$

which measures the relative increase in response rate above background.

The benchmark dose (BD) is defined by Crump (1984) as a statistical lower confidence limit for the dose corresponding to a specific increase in excess risk.

For practical application, this means that once we have specified the level of excess risk, say q, we need to solve the excess risk function for the dose d; i.e., find d such that $r(d;\boldsymbol{\beta}) = q$. This dose level d is called the effective dose (ED). Popular values of q are 0.05, 0.01 or 10^{-4} depending on the specific area of application. A point estimator for ED can be defined as the solution to $r(d;\widehat{\boldsymbol{\beta}}) = q$ where $\widehat{\boldsymbol{\beta}}$ is a consistent estimator for $\boldsymbol{\beta}$. A more useful estimator is one which accounts for the sampling variability (like an interval estimator). Since we are only interested in a ''safe'' estimator (low dose), the BD is defined as the lower confidence limit for ED.

Another but related method of taking the extra amount of variability into account, which arises from the estimated dose-response curve, is to construct a pointwise upper confidence limit on the entire excess risk function, and next, find the dosage value corresponding to the point on this curve where the excess risk equals the specified value. This dose value is referred to as the lower effective dose (Kimmel and Gaylor, 1988). We will not consider this method, but focus attention to BD determination instead.

It is important to note that the definition of a BD depends on the particular method used for the construction of a confidence interval. Different methods will lead to different BDs. For an overview of some of these, see Williams and Ryan (1996).

1.2 *Dose-response models for clustered binary data*

Here, we briefly describe the models and estimation methods used in this paper. Since the definition of a BD is based on marginal parameters, we focus on a selection of marginal models. Next, we restrict attention to a univariate binary malformation index. Let $\boldsymbol{y} = (y_1, \ldots, y_m)$ denote the vector of binary outcomes in a cluster of size m. In the present context, y_j denotes whether the jth fetus of that litter is malformed or not. Then $z = \sum_{j=1}^{m} y_j$ is the total number of malformed fetuses of this litter.

1.2.1 The beta-binomial model

As a full likelihood model, we used the well-established beta-binomial (BB) model (Skellam, 1948). The BB model assumes a random malformation probability P in a cluster to come from a beta distribution with mean π (see, for example, Kleinman, 1973). Its probability density function is given by

$$f(z;(\pi,\rho)) = \binom{m}{z} \frac{B(\pi(\rho^{-1}-1)+z,(1-\pi)(\rho^{-1}-1)+m-z)}{B(\pi(\rho^{-1}-1),(1-\pi)(\rho^{-1}-1))}, \tag{3}$$

where ρ is the intraclass correlation (assuming exchangeability) and $B(.,.)$ denotes the beta function. It can be shown that the corresponding log-likelihood is

$$\ln\binom{m}{z} + \sum_{r=0}^{z-1} \ln\left(\pi + \frac{r\rho}{1-\rho}\right) + \sum_{r=0}^{m-z-1} \ln\left(1-\pi+\frac{r\rho}{1-\rho}\right) - \sum_{r=0}^{m-1} \ln\left(1+\frac{r\rho}{1-\rho}\right).$$

The marginal parameters π and ρ can be modeled as a function of dose d via the logistic link function for π and Fisher's z-transform for ρ. As an example, consider a linear dose effect on logit(π) and a constant intra-litter correlation:

$$\begin{pmatrix} \ln(\frac{\pi}{1-\pi}) \\ \ln(\frac{1+\rho}{1-\rho}) \end{pmatrix} = \boldsymbol{X\beta} = \begin{pmatrix} 1 & d & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_d \\ \beta_a \end{pmatrix}. \tag{4}$$

Note that in case of no intra-litter association ($\beta_a = 0$ and hence $\rho = 0$), the model reduces to ordinary logistic regression.

1.2.2 The Bahadur model

The second full likelihood model considered here is the Bahadur (1961) model. Again assuming exchangeability and putting all three- and higher-order correlations equal to zero, the probability density function is given by

$$f(z;(\pi,\rho)) = \binom{m}{z}\pi^z(1-\pi)^{m-z}\left(1+\rho\left\{\binom{m-z}{2}\frac{\pi}{1-\pi} - z(m-z) + \binom{z}{2}\frac{1-\pi}{\pi}\right\}\right). \tag{5}$$

The marginal parameters π and ρ play exactly the same role as in the BB model and can again be modeled as in (4). This Bahadur model will mainly be used to generate data in the simulation study described in Section 6.

1.2.3 The GEE2 method

Another interesting marginal approach is the use of generalized estimating equations (GEE). As a multivariate analog of quasi-likelihood it does not require a full likelihood specification (Liang and Zeger, 1986; Zeger and Liang, 1986). Here we consider a so-called GEE2 (Zhao and Prentice, 1990), using the first four moments of the BB model

$$\mu_i^{(1)} = \pi_i, \quad \mu_i^{(\ell)} = \frac{\pi_i(1-\rho_i)+(\ell-1)\rho_i}{1-\rho_i+(\ell-1)\rho_i}\mu_i^{(\ell-1)}, \quad \ell = 2,3,4,$$

and defined by the following equations

$$\sum_{i=1}^{N} \boldsymbol{X}_i^T (\boldsymbol{T}_i^T)^{-1} \boldsymbol{V}_i^{-1} (\boldsymbol{Z}_i - \boldsymbol{\mu}_i) = 0, \tag{6}$$

where z_i is the total number of malformed fetuses in the ith litter, N the total number of litters, $\boldsymbol{Z}_i = (z_i, \binom{z_i}{2})^T$, $\mu_i = (m_i\mu_i^{(1)}, \binom{m_i}{2}\mu_i^{(2)})^T$ (m_i the size of the ith litter), $\boldsymbol{V}_i = \mathrm{Cov}(\boldsymbol{Z}_i)$ determined by $\mu_i^{(3)}$ and $\mu_i^{(4)}$, $\boldsymbol{T}_i = \partial\boldsymbol{\eta}_i/\partial\boldsymbol{\mu}_i$ with $\boldsymbol{\eta}_i$ the left-hand side of (4) and $\boldsymbol{X}_i$ defined as in (4).

In the sequel, the resulting method is referred to as ''GEE2''.

2. BD determination via full likelihood or score

The idea is to reconsider existing likelihood approaches for developmental toxicity risk assessment from a non likelihood point of view; an important class of examples are the generalized estimating equations (Liang and Zeger, 1986). According to Williams and Ryan (1996) it is preferable to define the BD using the likelihood ratio statistic. This method is explained in Section 2.1. Of course, a full likelihood technique will perform best when the likelihood is correctly specified, but one might expect problems in case of misspecification; see Aerts *et al*. (1997). Therefore it is important to look at robust estimation methods like quasi-likelihood, GEE and pseudolikelihood.

Williams and Ryan (1996) also explain that the likelihood method is unavailable in quasi-likelihood settings, and hence also in GEE, since there is no analog to the likelihood ratio statistic. The approach we have in mind starts from the (robust) score statistic of a

GEE approach and defines a lower confidence limit by ''inverting'' this score test. This is the (robust) score approach, see Section 2.2, described by Carroll *et al*. (1998).

2.1 *A full likelihood-based BD*

Consider a dose-response model with parameter $\boldsymbol{\beta}$ and let ED be defined as the solution to

$$r(d;\boldsymbol{\beta}) = q, \tag{7}$$

for some small $q > 0$. For the simulations in Sections 4 and 6 and the data analysis in Section 7 we chose $q = 10^{-4}$.

When assuming a full likelihood model, the $100(1-\alpha)\%$ lower limit for the ED is calculated as (see Aerts *et al*., 1997)

$$\text{BD} = \min\Big\{d(\boldsymbol{\beta}) : r(d;\boldsymbol{\beta}) = q \text{ over all } \boldsymbol{\beta} \text{ such that } 2(\ell(\hat{\boldsymbol{\beta}})) - \ell(\boldsymbol{\beta})) \leq \chi^2_p(1-2\alpha)\Big\}, \tag{8}$$

with $\hat{\boldsymbol{\beta}}$ the ML estimator for $\boldsymbol{\beta}$ and $\ell(\tilde{\boldsymbol{\beta}})$ the value of the log likelihood when $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$, p the total number of parameters in the model and $\chi^2_p(1-2\alpha)$ the $(1-2\alpha)$-quantile of a χ^2_p distributed random variable. The dependence of the log likelihood function on the data is, for simplicity, omitted in the notation.

For the linear-constant BB model (4), there are three model parameters ($p = 3$). With a linear logit model (1), the excess risk function (2) is given by

$$r(d;\boldsymbol{\beta}) = \frac{1 - \exp\{-\beta_d d\}}{1 + \exp\{-(\beta_0 + \beta_d d)\}}.$$

It follows that the dose d such that $r(d;\boldsymbol{\beta}) = q$ is equal to

$$\text{ED} = \frac{\ln\left(\frac{1+qe^{-\beta_0}}{1-q}\right)}{\beta_d}. \tag{9}$$

Hence, Equation (8) can be rewritten as

$$\text{BD} = \min\left\{\frac{\ln\left(\frac{1+qe^{-\beta_0}}{1-q}\right)}{\beta_d} \text{ over all } (\beta_0, \beta_d) \text{ such that } 2(\ell(\hat{\boldsymbol{\beta}}) - \ell(\boldsymbol{\beta})) \leq \chi^2_p(1-2\alpha)\right\}. \tag{10}$$

Since the expression (9) only contains the two π-coefficients β_0 and β_d but the critical point is calculated from a chi-squared distribution with three degrees of freedom, this full likelihood approach is expected to be too conservative. Indeed, the parameter β_a can be considered as a nuisance parameter. A technique which treats (nuisance) parameters in a more parsimonious way is the profile likelihood method, which will be discussed in Section 3.1.

In the next section, we first modify definition (8) in a way allowing GEE based estimation. This approach is expected to be more robust against misspecification of the

probability model for clustered binary data. In Section 6 the effect of misspecification and robustification on the BD is examined by simulation.

2.2 *A score or robust score-based BD*

Because of the first order equivalence between likelihood ratio, Wald and score statistics, we can replace the likelihood ratio statistic in (8) by either the Wald or the score statistic; without affecting the first order asymptotic properties. The non-invariance of Wald tests to parameter transformations (see e.g., Phillips and Park, 1988) makes a replacement by the score statistic more appealing. Score inverted confidence intervals are described in, for example, Hinkley *et al.* (1991, p. 278).

Assume the parameter estimator $\hat{\boldsymbol{\beta}}$ is the solution of estimating equations

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{\psi}(d_i, z_i, m_i; \boldsymbol{\beta}) = 0, \tag{11}$$

where, as before, d_i is the dose, m_i the size and z_i the number of malformations of litter i and N is the total number of litters. For ML estimators $\boldsymbol{\psi}(d_i, z_i, m_i; \boldsymbol{\beta})$ is the derivative of the log likelihood with respect to $\boldsymbol{\beta}$ and for the GEE2 methods (11) corresponds to (6). The score statistic is defined as

$$\mathscr{S}(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{\beta})^T \boldsymbol{A}(\boldsymbol{\beta})^{-1} \boldsymbol{U}(\boldsymbol{\beta}),$$

where

$$\boldsymbol{A}(\boldsymbol{\beta}) = -\sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\psi}(d_i, z_i, m_i; \boldsymbol{\beta}).$$

A $100(1-\alpha)\%$ score based lower limit for the ED can be defined as

$$\min\{d(\boldsymbol{\beta}) : r(d; \boldsymbol{\beta}) = q \text{ over all } \boldsymbol{\beta} \text{ such that } \mathscr{S}(\boldsymbol{\beta}) \leq \chi_p^2(1-2\alpha)\}, \tag{12}$$

with $\mathscr{S}(\boldsymbol{\beta})$ the score statistic at parameter value $\boldsymbol{\beta}$.

Within a likelihood framework the asymptotic χ_p^2 distribution is not valid anymore in case the probability density (likelihood) function is misspecified; e.g., when there is some overdispersion not correctly accounted for. This holds for the score statistic as well as for the log likelihood ratio and Wald statistics. There exists an extensive literature on likelihood misspecification. For an overview and many related references, see White (1982, 1994).

When there is uncertainty about the correctness of the likelihood specification, it is better to use the so-called robust statistics, since these modified test statistics have an asymptotic chi-squared distribution, even when the assumed probability model is not correct. For full likelihood models, robust Wald and score tests can easily be modified by using the so-called sandwich variance estimator (Kent, 1982; Viraswami and Reid, 1996). We are not aware of such a simple modification of the likelihood ratio test. An alternative is the use of bootstrap methods, see Aerts and Claeskens (1999, 2001). As indicated before, we focus on score tests. Another advantage of the robust score tests is that they are

also defined in quasi-likelihood, GEE and pseudolikelihood models (Liang and Zeger, 1986; Rotnitzky and Jewell, 1990; Geys *et al*., 1999).

The robustified BD is defined similar as in (12) but with $\mathscr{S}(\boldsymbol{\beta})$ replaced by the robust score statistic

$$\mathscr{R}(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{\beta})^T \boldsymbol{A}(\boldsymbol{\beta})^{-1} \Big(\boldsymbol{A}(\boldsymbol{\beta})^{-1} \boldsymbol{B}(\boldsymbol{\beta}) \boldsymbol{A}(\boldsymbol{\beta})^{-1}\Big)^{-1} \boldsymbol{A}(\boldsymbol{\beta})^{-1} \boldsymbol{U}(\boldsymbol{\beta}),$$

with

$$\boldsymbol{B}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{\psi}(d_i, z_i, m_i; \boldsymbol{\beta}) \boldsymbol{\psi}(d_i, z_i, m_i; \boldsymbol{\beta})^T.$$

For details on the definition of $\mathscr{R}(\boldsymbol{\beta})$, we refer to the above mentioned papers. As explained before, all these methods (likelihood and score based) are expected to be too conservative. In the next section we turn to profile likelihood and profile score approaches.

3. BD determination via profile likelihood or score

In general, a profile log likelihood for a parameter vector is defined as follows (see, for example, Morgan, 1992, Section 2.7.3 or Barndorff-Nielsen and Cox, 1994, Section 3.4). Assume that $\boldsymbol{\lambda}$ is the parameter vector of interest and that $\boldsymbol{\eta}$ is a vector of nuisance parameters in the model. The partially maximized log likelihood function

$$\ell(\boldsymbol{\lambda}, \hat{\boldsymbol{\eta}}_{\lambda}) = \sup_{\eta} \ell(\boldsymbol{\lambda}, \boldsymbol{\eta}),$$

is the definition of the profile log likelihood for $\boldsymbol{\lambda}$.

3.1 *A profile likelihood-based BD*

The following profile likelihood based BD is defined in Aerts *et al*. (1997). Their starting point is that, for a specified excess risk q, three parameters fully specify the set $(\beta_0, \beta_d, \beta_a, d)$. Or, if β_0 and β_a are given, and if a monotonic relationship exists between β_d and the dose d, either member of the pair (β_d, d) contains the same amount of information. One such example is given by Equation (9). This was their motivation to construct a profile likelihood based confidence interval for β_d and to transform this interval into an interval for d via transformation (9).

There are two ''nuisance'' parameters in this approach: in order to obtain the log likelihood values, also β_0 and β_a need to be estimated. Let $\boldsymbol{\beta}_n(\beta_d)$ contain all components of $\boldsymbol{\beta}$ except for β_d. Here, $\boldsymbol{\beta}_n(\beta_d) = (\beta_0, \beta_a)$. This parameter vector is estimated by assuming a certain fixed value of β_d, which is exactly the meaning of a profile approach.

A profile likelihood based BD is defined as follows:

$$BD_{p\ell 1} = \min\Big\{ d(\beta_d, \hat{\boldsymbol{\beta}}_n(\beta_d)) : r(d; \beta_d, \hat{\boldsymbol{\beta}}_n(\beta_d)) = q \text{ over all } \beta_d \text{ such that } 2\{\ell(\hat{\beta}_d, \hat{\boldsymbol{\beta}}_n(\hat{\beta}_d)) - \ell(\beta_d, \hat{\boldsymbol{\beta}}_n(\beta_d))\} \leq \chi_1^2(1-2\alpha) \Big\}. \quad (13)$$

This time a critical point from a chi-squared distribution with only one degree of freedom is used. Aerts *et al*. (1997) restricted attention to the profile likelihood. In the next section we extend this definition by use of (robust) score statistics.

3.2 *A profile score or profile robust score-based BD*

Similar to definition (12), we now define a profile score based BD as

$$BD_{ps1} = \min\Big\{ d(\beta_d, \hat{\boldsymbol{\beta}}_n(\beta_d)) : r(d; \beta_d, \hat{\boldsymbol{\beta}}_n(\beta_d)) = q \text{ over all } \beta_d \text{ such that } \mathscr{S}(\beta_d, \hat{\boldsymbol{\beta}}_n(\beta_d)) \leq \chi^2_1(1-2\alpha) \Big\}, \tag{14}$$

and, for the robust score approach

$$BD_{pr1} = \min\Big\{ d(\beta_d, \hat{\boldsymbol{\beta}}_n(\beta_d)) : r(d; \beta_d, \hat{\boldsymbol{\beta}}_n(\beta_d)) = q \text{ over all } \beta_d \text{ such that } \mathscr{R}(\beta_d, \hat{\boldsymbol{\beta}}_n(\beta_d)) \leq \chi^2_1(1-2\alpha) \Big\}. \tag{15}$$

All profile BD definitions (13)–(15) use only one degree of freedom for the chi-squared distribution, compared to three in (10). Therefore, the profile approach is expected to be more efficient. Only the robustified profile score BD definition (15) is theoretically valid in case of misspecification.

4. A simulation study comparing profile based BDs in correctly specified models

In this section we apply the profile methods to some simulated data sets in a typical setting of an animal experiment in toxicology. Data are generated and fitted with the same model. A detailed description of the settings is given in Section 4.1 and the results are in Section 4.2.

The purpose of this simulation study is two-fold:

1. We make a comparison of the likelihood, score and robust score approaches. It will turn out that all three yield very similar results.
2. Properties of the profile method itself are investigated. Some structural problems with the profile approach are discussed and an explanation is given for the undercoverage observed in the simulation study.

Another new aspect of these simulation results is that they are used to qualify the performance of methods to determine a BD. To our knowledge and according to an extensive literature search, this has not been done before.

4.1 *Details on the simulation settings*

A typical toxicological experiment includes one control group and two or three dosed groups. For the simulations we selected dose levels 0, 0.25, 0.5 and 1. Several parameter settings were investigated, all of these might occur in practical experiments. An equal number of $n_c = 20$ clusters was assigned to each dose group (unless mentioned otherwise). The number of fetuses m per cluster, i.e., the cluster size, is random. This cluster size m is assumed to follow a local linear smoothed version of the relative frequency distribution given in Kupper *et al*. (1986) (see Table 1 in Molenberghs *et al*., 1998).

For each setting 1000 datasets were generated using the GAUSS routine RNDU. For each dataset, the ED and BD were computed, the latter according to definitions (13), (14) and (15). Profile BD's were obtained by first calculating a 90% (two-sided) confidence interval for β_d. Next, each value of β_d in this interval is transformed to a value d using transformation (9). The smallest of these values is taken to be the BD $(\alpha = 0.05)$. Programming code was developed in GAUSS for Windows NT/95 (Version 3.2.32, © Aptech Systems, Inc. Maple Valley, WA) and is available from the authors on request.

4.2 *Simulation results and discussion*

Data were generated from a BB model with linear effect on logit scale and constant correlation. Realistic parameter values were chosen: the intercept term $\beta_0 = -2.5$, the dose effect $\beta_d = 1, 3$ or 5 and the association parameter $\beta_a = 0.1, 0.5$ or 0.9. The BB model was used to fit the data and to obtain the BD $(\alpha = 0.05)$. Note that this is an ideal situation in which the correct probability model is used for fitting.

Since this is a simulation study, we know the ED exactly; values are given in Table 1.

Note that the true values for ED are identical for settings (2), (4)–(7) which only differ by the amount of within cluster association or sample size.

Since the BD is by definition a lower confidence limit, we can look at coverage probabilities, i.e., how many of the simulated BDs are smaller than the true ED.

Simulation coverage probabilities are shown in Table 2. Clearly, they are much too small. Since we are dealing with one-sided intervals, we expect to get coverage percentages above 90%. On the other hand, the coverage probabilities for a 90% (2-sided) confidence interval for β_d, in that same simulation study, which were used to obtain the BD

Table 1. Parameter settings in the simulation study.

Parameter setting	*True ED*	$(\beta_0, \beta_d, \beta_a)$
(1)	0.001318	$(-2.5, 1, 0.5)$
(2)	0.000439	$(-2.5, 3, 0.5)$
(3)	0.000265	$(-2.5, 5, 0.5)$
(4)	0.000439	$(-2.5, 3, 0.1)$
(5)	0.000439	$(-2.5, 3, 0.9)$
(6)	0.000439	same as (2), $n_c = 10$
(7)	0.000439	same as (2), $n_c = 15$

Table 2. Simulated coverage probabilities for ED.

Parameter setting	*Pr. score*	*Pr. robust*	*Pr. likelihood*
(1)	0.722	0.729	0.722
(2)	0.708	0.698	0.708
(3)	0.848	0.823	0.841
(4)	0.760	0.752	0.760
(5)	0.670	0.659	0.670
(6)	0.652	0.636	0.652
(7)	0.690	0.676	0.693

values above, are (up to rounding) almost exactly equal to the nominal value of 90% (see Table 3).

A further discussion on these seemingly contradictory results is provided in the next sections.

A comparison of score, robust score and likelihood values shows that they yield very similar results. This confirms the applicability and usefulness of score and robust score definitions. Moreover, these latter two are computationally much easier to obtain since only estimates of the nuisance parameters need to be computed and not of the β_d parameter itself, which is necessary in the likelihood criterion (13). Note also that the robustification is not really needed here since the correct model has been fit to the data. Apparently, at least in these settings, there is not much of a price to pay for unnecessary robustification.

As the sample size increases, going from (6) to (7), to (2), the simulated coverage probabilities for β_d change relatively little. For ED, there seems to be a slight increase in coverage probability for case (2), 20 clusters per dose group.

For increasing within litter association, settings (4), (2) and (5), results are worst for the highest association, which is the case for the lowest amount of total information.

A possible explanation for the low coverage probabilities for the BD is the extra randomness due to estimation of β_0. Transformation (9) expresses the dosage d as a function of both β_0 and β_d. Since β_0 is unknown, in practice we have to substitute an estimator for the unknown true parameter. This additional source of variability might largely explain why a 90% confidence interval for β_d does not transform into an ''at least 90%'' one-sided confidence interval for the ED.

In the simulation study it frequently occurred that the confidence interval for β_d contains

Table 3. Simulated coverage probabilities for β_d.

Parameter setting	*Pr. score*	*Pr. robust*	*Pr. likelihood*
(1)	0.897	0.894	0.899
(2)	0.895	0.889	0.900
(3)	0.897	0.899	0.897
(4)	0.901	0.902	0.908
(5)	0.895	0.888	0.895
(6)	0.897	0.885	0.901
(7)	0.891	0.887	0.892

negative values. This results in two disjoined intervals after transformation to d, one part contains only negative values, the other only positive values. Therefore we restrict the search for the minimal d which satisfies Equation (7) to those d coming from positive β_ds in the confidence interval. This implies truncating the confidence interval for β_d. This approach is used throughout. Note that this does not change the coverage probabilities for β_d since the true values used in this simulation study were all positive. An important question which arises is whether it still makes sense to define a BD when the estimated dose effect parameter β_d is not significantly different from zero, i.e., when the confidence interval for β_d contains zero. One argument against defining a BD could be that when there is no statistically significant effect of the dose on the outcome, there is no interest in the value of a safe dose and one could just as well define the BD to be zero. On the other hand, it can still be of interest to have an idea on the value of the BD even if the dose effect happens to be statistically not significant.

Another case is when the estimated dose effect parameter $\hat{\beta}_d$ is negative. This implies a decreasing dose-response curve on logit scale. A problem with this situation is that possibly $\pi(d;\boldsymbol{\beta}) \leq \pi(0;\boldsymbol{\beta})$ for $d \geq 0$, which implies that $r(d;\boldsymbol{\beta}) \leq 0$. In other words, it could be ''healthy'' to be exposed to a certain level of the particular exposure. The above methods do not apply to this case.

5. Some alternative profile approaches

Here, we give alternative profile based definitions solving the problems mentioned in the previous section. The first one is explicitly based on the transformation (9), the second alternative is a two-dimensional profile method.

5.1 *Reparametrization in terms of d*

A first approach is to reparametrize the likelihood in terms of $d = \gamma$, i.e., replace β_d by the right-hand side of

$$\beta_d = \frac{\ln\left(\frac{1+q\mathrm{e}^{-\beta_0}}{1-q}\right)}{\gamma}.$$

This equation is obtained by solving $r(\gamma;\boldsymbol{\beta}) = q$ for β_d in the linear-logit model (4). The likelihood function is now maximized directly with respect to $(\beta_0, \gamma, \beta_a)$. This leads to definition (16), which is, however, not equivalent to the previous profile method (13)

$$\widetilde{BD}_{pl1} = \min\Big\{\gamma : 2[\ell\{\hat{\gamma}, \hat{\beta}_0(\hat{\gamma}), \hat{\beta}_a(\hat{\gamma})\} - \ell\{\gamma, \hat{\beta}_0(\gamma), \hat{\beta}_a(\gamma)\}] \leq \chi^2_1(1-2\alpha)\Big\}. \tag{16}$$

An advantage of this reparametrization is that the profile function is constructed directly in terms of dosage γ. The presence of β_0 is automatically taken care of by the chain rule while taking derivatives to obtain the solution to the likelihood equations. Also here, there is no reason why the likelihood equations automatically would yield a positive value for γ, this needs to be forced explicitly by the computational method. Similarly, by replacing the

Table 4. Simulated coverage probabilities for ED. Reparametrization in terms of d.

Parameter setting	*Pr. score*	*Pr. robust*	*Pr. likelihood*
(1)	0.984	0.934	0.953
(2)	0.975	0.926	0.945
(3)	0.978	0.930	0.946
(4)	0.965	0.927	0.946
(5)	0.980	0.921	0.945
(6)	0.979	0.895	0.935
(7)	0.981	0.926	0.946

profile likelihood by the profile (robust) score expression in (16), the quantities $\widetilde{BD}_{ps1}$ and $\widetilde{BD}_{pr1}$ can be defined and can be implemented for full likelihood as well as GEE response models.

Simulated coverage probabilities for the same parameter settings as given earlier are presented in Table 4.

Table 4 clearly shows that the reparametrization increases the coverage percentages significantly. All simulated coverage probabilities are now larger than 90% with the following ordering: score > likelihood > robust score. Clearly, this method is preferable to the simpler one-dimensional profile approach of Section 3.

The simulated coverage probabilities of the robust score approach are smaller than those of the other two approaches; this same phenomenon is not only observed for the reparametrization method, but also in the tables to follow (Tables 6 and 8). This might be related to the results of Carroll *et al*. (1998); they found the ''sandwich'' estimator, i.e., the robust variance estimator, to be inefficient. Further investigation is needed to fully understand how the sandwich construction affects the profile score function.

Another aspect of BD determination is the accuracy. Classically, one would look at the width of a two-sided confidence interval for ED. In this context, this is not of direct interest. Of importance here is the distance between the BD and the true ED. We need to consider two possible situations: either the BD is smaller than or equal to the true ED (that is what one aims at) or the BD is larger than the true ED. If this latter case occurs, a good method for BD determination would have a small mean/median distance between the BD and the true ED.

Table 5 gives the number of simulated datasets where the BD is smaller than or equal to, respectively greater than, the ED, and the corresponding median lengths, for each of the seven parameter settings (see Table 1).

Several observations can be made from this table. First, as the sample size increases, i.e., going from (6) to (7) to (2), the median distances decrease. This is intuitively expected, the larger the sample size, the more accurate the results will be. Also the number of times that the BD was larger than the true ED decreases.

If the within-cluster association increases, the accuracy decreases. We observe increasing median distances while going from parameter setting (4) to (2) to (5). Also the number of times that BD > ED increases. This is consistent with the previous observation, since an increasing within-cluster association is comparable to a decreasing sample size.

Table 5. Distances of BD to the true ED. Reparametrization in terms of d.

Setting	*Method*	$\# \leq ED$	*Median length*	$\# > ED$	*Median length*
(1)	Score	979	0.0004818	16	0.0000389
	Robust	929	0.0003923	66	0.0001000
	Likelihood	948	0.0004158	47	0.0000985
(2)	Score	975	0.0001073	25	0.0000192
	Robust	926	0.0000924	74	0.0000221
	Likelihood	945	0.0000962	55	0.0000240
(3)	Score	978	0.0000748	22	0.0000175
	Robust	930	0.0000627	70	0.0000196
	Likelihood	946	0.0000658	54	0.0000156
(4)	Score	949	0.0000759	34	0.0000173
	Robust	911	0.0000684	72	0.0000147
	Likelihood	930	0.0000701	53	0.0000134
(5)	Score	979	0.0001266	20	0.0000223
	Robust	920	0.0001076	79	0.0000271
	Likelihood	944	0.0001116	55	0.0000263
(6)	Score	971	0.0001428	21	0.0000230
	Robust	888	0.0001195	104	0.0000238
	Likelihood	928	0.0001243	64	0.0000268
(7)	Score	979	0.0001209	19	0.0000205
	Robust	924	0.0001014	74	0.0000274
	Likelihood	944	0.0001062	54	0.0000217

5.2 *A two-dimensional profile approach*

A second solution is a higher dimensional profile approach, where all parameters in the function $\pi(\cdot)$ are taken into account. For a model logit $= \beta_0 + \beta_d d$, this means both (β_0, β_d). This leads to the following definition

$$BD_{p\ell 2} = \min\Big\{ d(\beta_0, \beta_d, \hat{\beta}_a) : r(d; \beta_0, \beta_d, \hat{\beta}_a) = q \text{ over all } (\beta_0, \beta_d) \text{ such that } 2\{\ell(\hat{\beta}_0, \hat{\beta}_d, \hat{\beta}_a) - \ell(\beta_0, \beta_d, \hat{\beta}_a(\beta_0, \beta_d))\} \leq \chi^2_2(1-2\alpha) \Big\}. \quad (17)$$

We now use one more degree of freedom; the degrees of freedom are taken to correspond to the length of that part in the $\boldsymbol{\beta}$-vector which occurs in the dose-response model.

For the robust score approach, a profile version reads as follows:

$$BD_{pr2} = \min\Big\{ d(\beta_0, \beta_d, \hat{\beta}_a) : r(d; \beta_0, \beta_d, \hat{\beta}_a) = q \text{ over all } (\beta_0, \beta_d) \text{ such that } \mathscr{R}(\beta_0, \beta_d, \hat{\beta}_a) \leq \chi^2_2(1-2\alpha) \Big\}. \quad (18)$$

A similar definition can be given for the score approach by replacing $\mathscr{R}$ in (18) by $\mathscr{S}$.

This approach seems to be more natural because it uses two degrees of freedom and in this sense it tries to find a compromise between the full likelihood methods which use the full length of $\boldsymbol{\beta}$ and the one-dimensional profile approach which only takes β_d into account in its degrees of freedom determination.

Table 6. Simulated coverage probabilities for ED. Two-dimensional profile approach.

Parameter setting	*Pr(2). score*	*Pr(2). robust*	*Pr(2). likelihood*
(1)	0.933	0.934	0.931
(2)	0.891	0.881	0.925
(3)	0.880	0.878	0.935
(4)	0.882	0.881	0.922
(5)	0.878	0.864	0.921
(6)	0.861	0.841	0.905
(7)	0.892	0.881	0.927

The coverage probabilities in Table 6 are much higher than the corresponding ones in Table 3 where a χ^2_1 quantile has been used. Except for setting (1), coverage percentages are highest for the likelihood-based method. For the latter method the simulated coverage probabilities exceed 90% although most other percentages are not statistically significantly different from 90% at the 1% level of significance. For parameter setting (6), score and robust score values are significantly different from 90%.

We compare the results of the two-dimensional profile approach (Table 7) with those of the reparametrized one-dimensional approach. The median lengths of the two-dimensional profile approach are smaller than the corresponding ones of the reparametrized one-dimensional profile approach in case the BD is smaller than or equal to the true ED (see Table 5). The BD obtained via the two-dimensional method tends to be larger. For most

Table 7. Distances of BD to the true ED. Two-dimensional profile approach.

Setting	*Method*	$\# \leq ED$	*Median length*	$\# > ED$	*Median length*
(1)	Score	936	0.0003896	64	0.0001135
	Robust	936	0.0003919	64	0.0000781
	Likelihood	933	0.0003793	67	0.0001089
(2)	Score	896	0.0000898	104	0.0000294
	Robust	884	0.0000884	116	0.0000259
	Likelihood	936	0.0000914	64	0.0000310
(3)	Score	880	0.0000599	107	0.0000174
	Robust	873	0.0000582	114	0.0000199
	Likelihood	927	0.0000647	60	0.0000153
(4)	Score	899	0.0000612	101	0.0000238
	Robust	893	0.0000598	107	0.0000242
	Likelihood	931	0.0000644	69	0.0000177
(5)	Score	885	0.0001048	114	0.0000350
	Robust	872	0.0001014	127	0.0000367
	Likelihood	931	0.0001057	68	0.0000297
(6)	Score	821	0.0001174	127	0.0000519
	Robust	800	0.0001144	148	0.0000490
	Likelihood	861	0.0001195	87	0.0000445
(7)	Score	900	0.0000974	100	0.0000362
	Robust	884	0.0000971	116	0.0000362
	Likelihood	934	0.0000996	66	0.0000370

parameter settings, the median length of the distance between the BD and the true ED when BD > ED is larger for the two-dimensional profile method than for the reparametrized one-dimensional profile method. In other words, when the BD is at the correct, left side of ED, the distance is smaller, but when it is at the wrong side, the distance is larger compared to the reparametrization method.

6. Simulation results under misspecification

In this section focus is on parameter setting (6):

$$\begin{pmatrix} \ln(\frac{\pi}{1-\pi}) \\ \ln(\frac{1+\rho}{1-\rho}) \end{pmatrix} = \begin{pmatrix} -2.5+3d \\ 0.5 \end{pmatrix}, \tag{19}$$

and 10 clusters for each of the four dose values. A comparison is made between the three profile methods for BD determination in the following cases:

a. Data are generated from a BB distribution (3) according to model (19) and this BB model is used for BD determination.
b. Data are generated from a BB distribution according to model (19) and the GEE2 method, see Section 1.2, is used for BD determination.
c. Data are generated from a Bahadur distribution (5) according to model (19) and the BB model is used for BD determination.
d. Data are generated from a Bahadur distribution (5) according to model (19) and the GEE2 method, see Section 1.2, is used for BD determination.

In all cases a linear logit/constant association model is used for model fitting. Only case (a) uses the fully correct model specification. In case (c) the wrong likelihood function is used for BD determination and in cases (b) and (d) we use generalized estimating equations instead of a full likelihood approach. For each of the four situations we will apply the three profile methods:

Pr(1): One-dimensional profile method with χ^2_1 critical value, see Section 3.
Pr(1,d): One-dimensional profile method with χ^2_1 critical value, reparametrization in terms of d, see Section 5.1.
Pr(2): Two-dimensional profile method with χ^2_2 critical value, see Section 5.2.

Simulated coverage probabilities for ED based on, for each case, 1000 datasets are shown in Table 8.

For all four cases coverage probabilities for the simple one-dimensional profile method Pr(1) are much smaller than the nominal ''at least 90%''; the two-dimensional profile method already performs much better, although the 90% coverage is not attained in all cases. For case (c) where data are generated from a Bahadur model, but fitted using the BB likelihood, especially both score and robust score simulated coverage probabilities are still too small. This, however, is not true in case (d) where the GEE2 method is used for model fitting and BD determination. In this latter case, coverage probabilities for the one-dimensional reparametrization method are also very large.

Table 8. Simulated coverage probabilities for ED, parameter setting (6). Data are generated from a beta-binomial model in (a)–(b) and from a Bahadur model in (c)–(d). Beta-binomial fitting in (a), (c) and GEE2 fitting in (b), (d).

Case	*Method*	*Score*	*Robust*	*Likelihood*
(a)	Pr(1)	0.652	0.636	0.652
	Pr(1,d)	0.979	0.895	0.935
	Pr(2)	0.861	0.841	0.905
(b)	Pr(1)	0.684	0.669	
	Pr(1,d)	0.987	0.946	
	Pr(2)	0.919	0.895	
(c)	Pr(1)	0.633	0.604	0.632
	Pr(1,d)	0.981	0.899	0.992
	Pr(2)	0.850	0.828	0.896
(d)	Pr(1)	0.752	0.742	
	Pr(1,d)	0.996	0.974	
	Pr(2)	0.964	0.957	

In Table 9 we show the results on distances of BDs to the true ED-value, the latter value can be calculated exactly since the data generating mechanism is known. In all four cases, the median simulated distance from the BD to the ED when BD $\leq$ ED is slightly smaller for the two-dimensional profile method Pr(2), although comparable to the distances obtained by Pr(1,d). In the unfavorable case that BD $>$ ED, median distances for Pr(2) are larger than those for Pr(1,d). For case (d) where data are generated from a Bahadur model and fitted by GEE2, the median simulated distance of the robust score Pr(2) values is very comparable to the corresponding distance obtained by Pr(1,d).

Comparing likelihood and GEE2 methods, it is of interest to note that, by comparing (d) and (c), for the robust score approach, for all three profile methods, the distances obtained by GEE2 in the case that BD $>$ ED are smaller than those obtained by use of the BB likelihood method. This comparison does not hold for (b) versus (a), but recall that the model used for fitting in (a) is the perfect one. In practice, the correct likelihood is usually unknown.

Although the focus is on BD determination, it is also worthwhile to look at some aspects of the ED estimation. Note that the ED estimator is not directly used to obtain a BD, this in contrast to the Wald-based methods (delta method). Table 10 shows simulated bias, variance and mean squared error (MSE) of the ED estimators. The parameter estimates $\hat{\boldsymbol{\beta}}$ depend on the model which is used for fitting and, obviously, not on the way a confidence interval for ED is constructed. Hence, we now only distinguish the four cases in which data were generated.

At least in this setting, for cases (a)–(c) the ED estimator is biased upwards. The bias, in absolute value, for the GEE2 estimator is smaller than for the full likelihood model, under both data generation methods, Bahadur and BB. For case (d), data generation from a Bahadur model, and fitting with GEE2, the simulated bias of the GEE2 estimator is negative. Not only the bias of the GEE2 estimator is smaller, also its variance is smaller than the variance of the BB estimators. As a consequence, also the mean squared error of the GEE2 estimators is smaller.

Table 9. Distances of BD to the true ED, setting (6). Data generated from a beta-binomial model in (a)–(b) and from a Bahadur model in (c)–(d). Beta-binomial fitting in (a), (c); GEE2 fitting in (b), (d).

Case	*Method*		$\# \leq ED$	*Median length*	$\# > ED$	*Median length*
(a)	Pr(1)	Score	645	0.0000612	345	0.0000478
		Robust	630	0.0000602	360	0.0000508
		Likelihood	645	0.0000612	345	0.0000468
	Pr(1,d)	Score	971	0.0001428	21	0.0000230
		Robust	888	0.0001195	104	0.0000238
		Likelihood	928	0.0001243	64	0.0000268
	Pr(2)	Score	821	0.0001174	127	0.0000519
		Robust	800	0.0001144	148	0.0000490
		Likelihood	861	0.0001195	87	0.0000445
(b)	Pr(1)	Score	590	0.0000640	273	0.0000419
		Robust	577	0.0000639	286	0.0000455
	Pr(1,d)	Score	846	0.0001507	11	0.0000337
		Robust	811	0.0001264	46	0.0000322
	Pr(2)	Score	822	0.0001279	72	0.0000523
		Robust	800	0.0001240	94	0.0000459
(c)	Pr(1)	Score	629	0.0000483	364	0.0000405
		Robust	600	0.0000485	393	0.0000438
		Likelihood	628	0.0000483	365	0.0000401
	Pr(1,d)	Score	997	0.0001168	19	0.0000120
		Robust	895	0.0000921	101	0.0000304
		Likelihood	918	0.0001028	78	0.0000241
	Pr(2)	Score	847	0.0000928	150	0.0000491
		Robust	826	0.0000890	171	0.0000521
		Likelihood	893	0.0000970	104	0.0000308
(d)	Pr(1)	Score	734	0.0000569	242	0.0000318
		Robust	724	0.0000567	252	0.0000319
	Pr(1,d)	Score	839	0.0001382	3	0.0000156
		Robust	819	0.0001162	22	0.0000209
	Pr(2)	Score	952	0.0001147	36	0.0000272
		Robust	946	0.0001121	42	0.0000298

Table 10. Simulated bias, variance and mean squared error (MSE) of the ED estimator obtained from Equation (9). True ED equals 0.0004392.

Case	*Bias*	*Variance*	*MSE*
(a)	0.0000433	$1.7932 \cdot 10^{-8}$	$1.9807 \cdot 10^{-8}$
(b)	0.0000196	$1.2515 \cdot 10^{-8}$	$1.2900 \cdot 10^{-8}$
(c)	0.0000511	$1.2404 \cdot 10^{-8}$	$1.5016 \cdot 10^{-8}$
(d)	-0.0000350	$0.3370 \cdot 10^{-8}$	$0.4592 \cdot 10^{-8}$

7. Toxicity study on ethylene glycol

This data set (Price *et al*., 1985) resulted from a study conducted by the Research Triangle Institute under contract to the National Toxicology Program of the United States. The effects in mice of the chemical ethylene glycol was investigated. There is one control group (zero dose level) and three active dose levels, 750, 1500 and 3000 mg/kg/day, which were standardized in the analysis such that the highest dosage level is one. The number of (female) mice in each group was between 22 and 25. Each of these mice was impregnated and exposed to a certain amount of the chemicals. A few weeks later, the mice were sacrificed and for each mouse, the fetuses were examined. Litter sizes varied between 3 and 19. The outcomes of interest are indicators for external, skeletal and visceral malformation on each of the fetuses. Also a collapsed binary outcome is considered, indicating whether at least one of these malformation types occurs. We will take a univariate approach and estimate ED and determine BDs for each of these malformation types in turn using the two-dimensional profile approach and the one-dimensional reparametrization method.

For GEE2, only the robust profile BD's should be considered. Note that for GEE2, the last column is empty since the likelihood ratio based BD does not exist.

Table 11 shows ED and BD values, which, for visceral malformation are much higher than for the other types of malformations. For visceral and skeletal malformation, and for a collapsed outcome variable (any of those three malformation types), profile likelihood and score BD values nearly coincide for the BB model. For skeletal and collapsed outcomes, the results by GEE2 (both profile methods) do not differ much from the corresponding BB-values. For skeletal malformation and a collapsed outcome, the largest BD values are obtained by the GEE2 profile robust score method; for external malformation by the BB profile robust score method and for visceral by the BB profile likelihood approach. For skeletal malformation and the collapsed outcome, the BD values obtained by the robust score approach are, up to rounding, the same for both profile GEE2 methods. For visceral malformation, the one-dimensional profile method Pr(1,d) yields a somewhat smaller value, in comparison with the Pr(2) method. The reverse is true for external malformation, although both values are rather comparable.

Table 11. ED and BD determination for the dataset EG.

Model	*Outcome*	*Effective dose*	*Score*	*Robust*	*Likelihood*
BB	External	0.00734	0.00447	0.00504	0.00385
Pr(2)	Visceral	0.03674	0.02244	0.01730	0.02242
	Skeletal	0.00056	0.00043	0.00045	0.00043
	Collapsed	0.00051	0.00041	0.00041	0.00040
GEE2	External	0.00824	0.00450	0.00486	
Pr(2)	Visceral	0.03797	0.01581	0.01683	
	Skeletal	0.00058	0.00044	0.00046	
	Collapsed	0.00053	0.00040	0.00042	
GEE2	External	0.00824	0.00347	0.00495	
Pr(1,d)	Visceral	0.03797	—	0.01228	
	Skeletal	0.00058	0.00041	0.00046	
	Collapsed	0.00053	0.00038	0.00042	

8. Discussion and topics of further research

Throughout this report we considered a linear dose-response function on the logit scale, which for the marginal models under study, resulted in an explicit formula for the ED, see Equation (9). In general, one might have to deal with more flexible functional relationships $\pi(d;\boldsymbol{\beta})$, such as non-linear dose-response models, and/or with other link functions. The ED is defined in the same way as before, or equivalently, as the solution d to the equation

$$\pi(d;\boldsymbol{\beta}) = q + (1-q)\pi(0;\boldsymbol{\beta}), \tag{20}$$

which, in general, will require numerical methods. In case expression (20) has more than one solution, one could define the ED to be the smallest positive solution. BD determination in these models can proceed through a similar profile method as in Section 5.2, we only need to adapt the degrees of freedom of the chi-squared distributed random variable according to the number of components of $\boldsymbol{\beta}$ actually appearing in (20). A generalization of the profile method in Section 5.1 is not so obvious since the estimating equations need to be rewritten in terms of the ED, the latter which might only be defined implicitly.

Another interesting topic would be a study on the impact of the design on the BD determination. Weller *et al*. (1995) use an additive excess risk to calculate the BD as $ED - 1.645\hat{\sigma}_{ED}$, where $\hat{\sigma}_{ED}$ is the estimated standard deviation of the ED, in their comparison of several dose allocation schemes.

To address the coverage percentage issue, one possibility to try to increase the coverage percentages, in particular for the two-dimensional profile method and for the robust score approach, is to use bootstrap critical points instead of the χ^2 critical values. This could be advantageous if the reason for small coverage is the use of the critical value from the asymptotic distribution. A naive application of bootstrap methods for the construction of confidence intervals would perform a bootstrap test at each value of the grid, which would make the method computationally very unattractive. Alternatively, other methods for obtaining confidence intervals could be considered, see Davison and Hinkley (1997) for an overview. None of those techniques have been studied in the context of quantitative risk assessment, and their theoretical and practical properties in this context are yet unknown.

For full likelihood-based methods, Declerck *et al*. (2000) defined litter-based risks and compared fetus- with litter-based EDs. It is interesting to investigate whether profile methods based on score statistics, as presented in Section 5, can be used in this context.

Other relevant questions are whether one should study malformation indices (as, for example, in the NTP data) jointly, separately or collapse them into a single indicator? Should one study death and malformation jointly? According to Ryan (1992), methods which are based on joint modeling of multivariate outcomes take more appropriately the risks from several sources into account, and provide more conservative estimates of risk than the univariate approach of regulating on the most sensitive endpoint. Multivariate outcomes require different models, it would be very interesting to study properties of BD determination methods for these models.

Further, as already noted in Section 4.2, the benchmark dose determination methods as described above, do not apply to decreasing dose response curves on logit scale, or any situation which results in a negative risk $r(d;\boldsymbol{\beta}) \leq 0$. Further investigation is required to deal with situations like this.

While this work is motivated by experiments where the proportion of malformed fetuses is the main quantity of interest, it could also be of interest to study the proposed profile score method in experiments requiring continuous dose-response modeling.

Acknowledgments

This project was supported by the NATO Collaborative Research Grant 950648.

References

Aerts, M. and Claeskens, G. (1999) Bootstrapping pseudolikelihood models for clustered binary data. *Annals of the Institute of Statistical Mathematics*, **51**, 515–30.

Aerts, M. and Claeskens, G. (2001) Bootstrap tests for misspecified models, with application to clustered binary data. *Journal of Computational Statistics and Data Analysis*, **36**, 383–401.

Aerts, M., Declerck, L., and Molenberghs, G. (1997) Likelihood misspecification and safe dose determination for clustered binary data. *Environmetrics*, **8**, 613–27.

Bahadur, R.R. (1961) A representation of the joint distribution of responses of *n* dichotomous items, in *Studies in item analysis and prediction*, H. Solomon (ed.), Stanford Mathematical Studies in the Social Sciences VI. Stanford, California, Stanford University Press.

Carroll, R.J., Wang, S., Simpson, D.G., Stromberg, A.J., and Ruppert, D. (1998) The sandwich (robust covariance matrix) estimator. Unpublished manuscript.

Crump, K. (1984) A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology*, **4**, 854–71.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge.

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

Gad, S.C. (1999) *Statistical and Experimental Design for Toxicologists*, Third Edition, CRC Press, Boca Raton.

Geys, H., Molenberghs, G., and Ryan, L. (1999) Pseudo-likelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94**, 734–45.

Hinkley, D.V., Reid, N., and Snell, E.J. (1991) (eds) *Statistical Theory and modeling*, in honour of Sir David Cox, FRS, Chapman and Hall, London.

Kent, J.T. (1982) Robust properties of likelihood ratio tests. *Biometrika*, **69**, 19–27.

Kimmel, C.A. and Gaylor, D.W. (1988) Developmental toxicity studies, in *Handbook of In Vivo Toxicity Testing*, pp. 271–300, Academic Press.

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

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

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

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

Molenberghs, G., Declerck, L., and Aerts, M. (1998) Misspecifying the likelihood for clustered binary data. *Journal of Computational Statistics and Data Analysis*, **26**, 327–49.

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

Phillips, P.C.B. and Park, J.Y. (1988) On the formulation of Wald tests of nonlinear restrictions. *Econometrica*, **56**, 1065–83.

Price, C.J., Kimmel, C.A., Tyl, R.W., and Marr, M.C. (1985) The developmental toxicity of ethylene glycol in mice. *Toxicology and Applied Pharmacology*, **81**, 113–27.

Rotnitzky, A. and Jewell, N.P. (1990) Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, **77**, 485–97.

Ryan, L.M. (1992) Quantitative risk assessment for developmental toxicity. *Biometrics*, **48**, 163–74.

Skellam, J.G. (1948) A probability distribution derived from the binomial distribution by the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, **10**, 257–61.

Viraswami, K. and Reid, N. (1996) Higher-order asymptotics under model misspecification. *The Canadian Journal of Statistics*, **24**, 263–78.

Weller, E.A., Catalano, P.J., and Williams, P.L. (1995) Implications of developmental toxicity study design for quantitative risk assessment. *Risk Analysis*, **15**, 567–74.

White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–26.

White, H. (1994) *Estimation, Inference and Specification Analysis*, Cambridge University Press, Cambridge.

Williams, P.L. and Ryan, L.M. (1996) Dose response models for developmental toxicity, in *Handbook of Developmental Toxicology*, R.D. Hood (ed.), CRC Press, Boca Raton, pp. 609–40.

Zeger, S. and Liang, K.-Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–30.

Zhao, L.P. and Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–8.

Biographical sketches

Gerda Claeskens is assistant professor at Texas A&M University. Marc Aerts and Geert Molenbergs are associate professors at Limburgs Universitair Centrum. Prof. Louise Ryan works on statistical methods related to environmental risk assessment for cancer, developmental and reproductive toxicity, and other non-cancer endpoints such as respiratory disease. She also works on cancer clinical trials and epidemiological methods for the study of environmental effects on cancer, reproduction and respiratory health.

The authors have collaborated in the framework of a NATO Collaborative Research Grant on ‘‘Statistical Research for Environmental Risk Assessment’’.