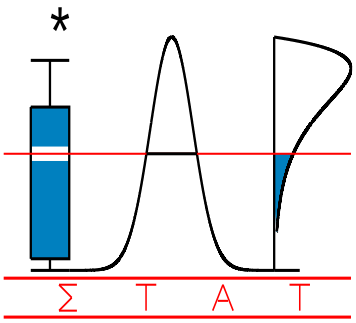
T E C H N I C A L

R E P O R T

0464

A HIERARCHICAL MODELING APPROACH FOR RISK ASSESSMENT IN DEVELOPMENTAL TOXICITY STUDIES

FAES, C., GEYS, H., AERTS, M. 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

A Hierarchical Modeling Approach for Risk Assessment in Developmental Toxicity Studies.

Christel Faes, Helena Geys, Marc Aerts and Geert Molenberghs

Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus,

B-3590 Diepenbeek, Belgium.

e-mail: christel.faes@luc.ac.be

Tel.: +32-11-26 82 85

Fax.: +32-11-26 82 99

Summary

Within the past decade, there has been an increasing interest in the problem of joint analysis of clustered multiple outcome data, motivated by developmental toxicity applications (Fitzmaurice and Laird 1995, Gueorguieva and Agresti 2001, Molenberghs and Ryan 1999, Regan and Catalano 1999, Aerts et al. 2002). So far, however, one has tackled the challenges in this setting only partly each time making different restricting assumptions (e.g., restriction to viable fetuses only). Ideally, a model should take the complete correlated hierarchical structure of the data into account. A hierarchical Bayesian method is proposed to this effect. Such a model can serve as a basis for quantitative risk assessment.

Keywords: Bayesian methods, Benchmark Dose, Hierarchical Model, Toxicology

1 Introduction

Developmental toxicity studies in laboratory animals are designed to assess potential hazardous effects of chemicals, drugs and other exposures on developing fetuses from pregnant dams. Such laboratory experiments play an important role in the regulation of adverse exposures for human health. A typical developmental toxicity study with a Segment II design includes a control group and several exposed groups, each involving 20 to 30 pregnant dams. Usually, exposure occurs early in gestation, during the period of major organogenesis and structural development of the fetuses. Just prior to normal delivery, the dams are sacrificed and the uterine contents are thoroughly examined for the occurrence of defects. The number of dead and resorbed fetuses is recorded. Viable offspring are examined carefully for the presence of malformations and also the fetal birth weights are measured.

The analysis of developmental toxicity data raises a number of challenges (Molenberghs et al. 1998). Since deleterious events can occur at several points in development, an interesting aspect lies in the staging or hierarchy of possible adverse fetal outcomes (Williams and Ryan 1996). Figure 1 illustrates the data structure. Because of the toxic insult, the developing fetus is at risk of fetal death. If the fetus survives the entire gestation period, growth reduction such as low birth weight may occur. The fetus may also exhibit one or more types of malformation.

In addition, because of genetic similarity and the same treatment conditions, offspring of the same mother behave more alike than those of another mother, i.e., the litter or cluster effect. Thus, responses on different fetuses within a cluster are likely to be correlated. There are several ways to handle the clustering of fetuses within litters. Several likelihood models for clustered binary data can be formulated, e.g., the beta-binomial model (Skellam 1948, Kleinman 1973), the Bahadur model (Williams

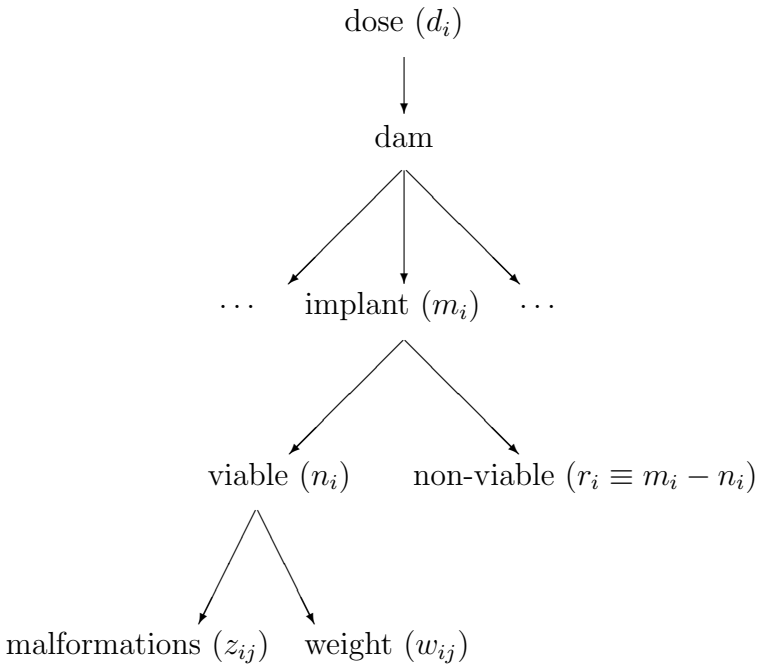


Figure 1: *Data Structure of Developmental Toxicity Studies.*

1975), the multivariate Dale model (Molenberghs and Lesaffre 1994) and the conditional exponential family model of Molenberghs and Ryan (1999). A thorough review is given in Aerts et al. (2002).

Ultimately, analysis of the developmental toxicity data must account for the entire hierarchical, multivariate and clustered nature of the data. So far, one has tackled the challenges in this setting only partly each time making different restrictive assumptions, e.g., restricting to viable fetuses only. However, litters with a lot of malformed fetuses are likely to have more death fetuses than litters with good fetal health. As a result, litter size (number of viable fetuses) may be informative. A classical way to account for the litter size is to include it as a covariate in modeling the response rates (Williams 1987, Rai and Van Ryzin 1985, Catalano et al. 1993) and then calculating a safe dose at an "average" litter size, thereby avoiding the need for direct adjustment. However, several perspectives for modeling this data in a direct way can be considered. First, one may look at the hierarchical structure, and consider cluster size as

a random variable. Xu and Prorok (2003) developed a non-parametric procedure for the analysis of exchangeable clustered binary data when the cluster size is a random variable. As such, one acknowledges the stochastic nature of the litter size. Indeed, variation in the litter size is an extra source of variability in the data that must be accounted for. Dunson et al. (2003) proposed a general bayesian approach for joint modeling of cluster size and subunit-specific outcomes, although there method was not designed for quantitative risk assessment. Secondly, we may also consider a missing data model, because the unformed fetuses are not observable. In toxicity studies with pre-implantation exposure, the number of implants reduces with dose. Dunson (1998) proposed a multiple imputation scheme to estimate the number of missing fetuses. Also joint models for the number of implantations and fetus-specific outcomes in pre-implantation toxicity studies have been studies by several authors (Kuk et al. 2003, Allen et al. 2002). However, in Segment II studies, the random cluster size perspective seems more natural than does the missing data perspective. In this context, the dose effect is reflected in a reduction of the proportion of viable fetuses (or litter size) among the implanted fetuses. Alternatively, Williamson et al. (2003) proposed a weighted generalized estimating equation approach for fitting marginal models to clustered data when litter size is informative. Although this method accounts for the cluster size, it does not allow for modeling the cluster size as a function of covariates of interest.

We propose a Bayesian model dealing with the hierarchical structure in two stages. At the first stage, we express the probability that a fetus is non-viable. At the second stage, we model the probability that a viable fetus has a malformation and/or suffers from low birth weight as function of the litter size. At each stage we account for the intralitter correlation. The intractability of the likelihood function has led vari-

ous authors to propose a host of alternative estimation methods rather than carrying out maximum likelihood estimation. A full likelihood procedure can be replaced by quasi-likelihood methods (McCullagh and Nelder 1989), pseudo-likelihood (Arnold and Strauss 1991) or generalized estimating equations (Liang and Zeger 1986). Generalized linear random-effects models or hierarchical Bayesian models (McCulloch and Searle 2001) are attractive alternative modeling approaches. We opted for the latter approach and used Gibbs sampling (Zeger and Karim 1991) to deal with complex integrations.

2 Example

This article is motivated by the analysis of developmental toxicity of Ethylene Glycol (EG) in mice. EG is a high-volume industrial chemical with diverse applications. For instance, it can be used as an antifreeze, as a solvent in the paint and plastics industries, as a softener in cellophane, etc. The potential reproductive toxicity of EG has been evaluated in several laboratories. Price et al. (1985) for example, describe a study in which timed-pregnant CD-1 mice were dosed by gavage with EG in distilled water. Dosing occurred during the period of organogenesis and structural development of the fetuses (gestational days 8 through 15). Data are pictured in Figures 2. Table 1 summarizes the malformation rate and fetal birth weight per dose group. The data show clear dose-related trends for both malformation and fetal weight. The rate of malformation increases with dose, and the average fetal weight decreases monotonically with dose. The mean litter size is also tabulated, and shows a decrease with dose.

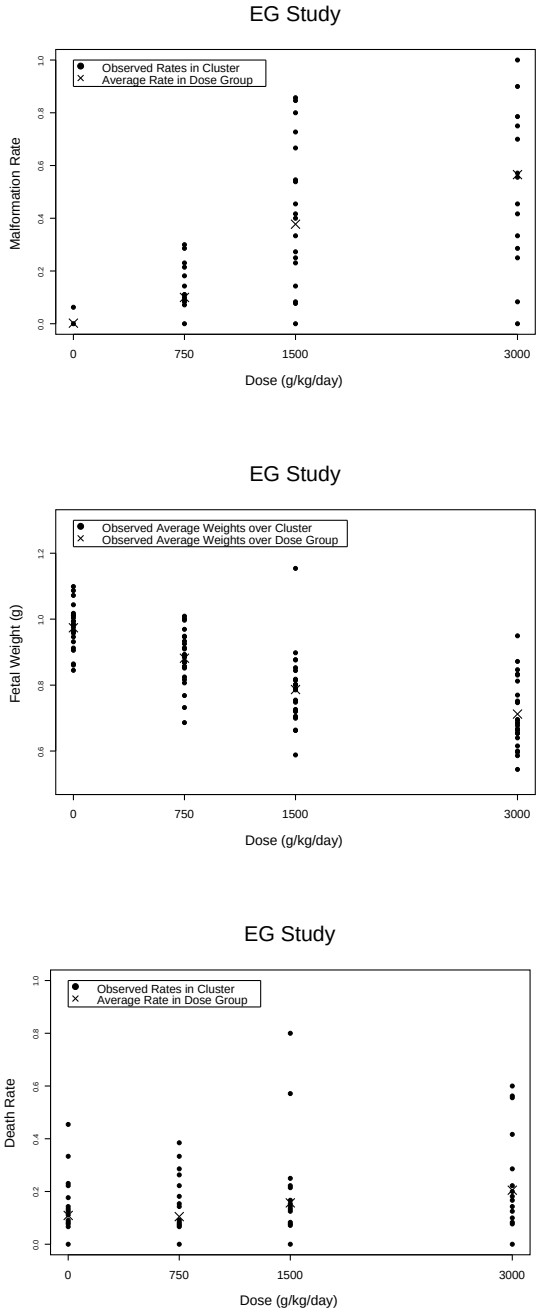


Figure 2: *EG data: From Top to Bottom: Observed malformation rates; Observed fetal weight; Observed death rate*

Table 1: Summary Data from an EG Experiment in Mice

Dose	Dams	Live	Litter Size		Malf.		Weight	
(g/kg/day)			Mean	(SD)	Nr.	(%)	Mean	(SD)
0	25	297	11.9	(2.45)	1	(0.25)	0.974	(0.065)
750	24	276	11.5	(2.38)	26	(10.0)	0.882	(0.082)
1500	22	229	10.4	(3.46)	89	(37.8)	0.787	(0.114)
3000	23	226	9.8	(2.69)	129	(56.5)	0.712	(0.105)

3 Modeling Approach

We propose a Bayesian hierarchical modeling framework for the joint analysis of fetal death and malformation/weight among the viable fetuses. Let L denote the total number of dams, and hence litters, in the study. For the ith litter ($i = 1, \ldots, L$), let m_i be the number of implants. Let r_i indicate the number of fetal deaths in cluster i. The number of viable fetuses, i.e., the litter size, is $n_i \equiv m_i - r_i$. The outcome measured on the viable fetuses is denoted $\boldsymbol{y}_{ij} = (w_{ij}, z_{ij}), j = 1, \ldots, n_i$, with w_{ij} the fetal birth weight and $z_{ij} = 1$ when fetus j in cluster i has a malformation, 0 otherwise.

To define a model for the developmental toxicity data, the underlying hierarchy of the data is used. At the bottom level, the fetuses surviving the entire gestation period, are at risk for low birth weight and/or malformation. Assume that $\boldsymbol{y}_{ij}$ satisfies

$$\boldsymbol{y}_{ij}|n_i \sim F_i(\boldsymbol{y}_{ij}|\zeta, n_i), \tag{3.1}$$

i.e., conditional on the litter size, $\boldsymbol{y}_{ij}$ follows a pre-specified distribution F_i, possibly depending on covariates, such as the dose level, and parameterized through a vector ζ of unknown parameters. Further, the litter size n_i is a random variable, possibly depending on the dose level and other covariates of interest. Indeed, a toxic insult may result in a fetal death. The litter size n_i is modeled through the number of

non-viable fetuses $r_i \equiv m_i - n_i$. Assume the number of non-viable fetuses r_i to follow a distribution G depending on a vector ψ of unknown parameters, i.e.,

$$r_i \sim G(r_i|\psi, m_i). \tag{3.2}$$

Let $f_i(\boldsymbol{y}_{ij}|\zeta, n_i)$ and $g(r_i|\psi, m_i)$ denote the density functions corresponding to the distributions F_i and G, respectively.

Because of the hierarchy in the model, it lends itself naturally to estimate the parameters using Bayesian techniques (Box and Tiao 1992, Gelman 1995). In the Bayesian framework, unknown parameters are also considered as random, and all inference is based on their distribution conditional on the observed data, i.e., the posterior distribution.

It is obvious that different choices for F_i and G will lead to different models. The distribution G is crucial in the calculation of the marginal model for $\boldsymbol{y}_{ij}$. Next, a possible choice for the distributions F_i and G_i in the developmental toxicity setting is given.

A model for the death outcome

In the first step, a toxic insult early in gestation may result in a fetal death. This effect of dose d_i on cluster i with m_i implants can be described using the density $g(r_i|\psi, m_i)$. Considering the fetuses within a litter as independent, one could assume that r_i satisfies a binomial density

$$\binom{m_i}{r_i} \pi_{R_i}^{r_i} (1 - \pi_{R_i})^{m_i - r_i}, \tag{3.3}$$

with π_{R_i} the probability of a dead fetus in litter i, depending on the dose. To account for clustering, a random effects model in which each litter has a random parameter is considered. Skellam (1948), Kleinman (1973) and Williams (1975) assume the probability of death π_{R_i} of any fetus in litter i to come from a beta distribution with parameters a_i and b_i:

$$\frac{\pi_{R_i}^{a_i-1}(1-\pi_{R_i})^{b_i-1}}{B(a_i,b_i)}, \quad (0 \leq \pi_{R_i} \leq 1), \tag{3.4}$$

where $B(.,.)$ denotes the beta function. This leads to the well-known beta-binomial distribution.

The probability mass function $g(r_i|\psi, m_i)$ can be expressed directly in terms of the mean and correlation parameters, i.e., $g(r_i|\pi_{R_i}, \rho_{R_i}, m_i)$. The mean of this distribution is

$$\mu_{R_i} = m_i\pi_{R_i} = m_i\frac{a_i}{a_i+b_i}, \tag{3.5}$$

and the variance is

$$\sigma^2_{R_i} = m_i\pi_{R_i}(1-\pi_{R_i})[1+\rho_{R_i}(m_i-1)], \tag{3.6}$$

with ρ_{R_i} the intra-litter correlation, which is the correlation between two binary responses of litter i.

A model for malformation and weight

When a fetus survives the entire gestation period, it is still at risk for low fetal weight and malformation. A distribution for the combined continuous and binary outcomes, i.e., $f(w_{ij}, z_{ij}|\zeta, n_i)$ must be specified. Based on the mixed outcome probit model of

Regan and Catalano (1999), we propose the following model.

First, assume that littermates are independent. Under a probit model for the binary response Z_{ij}, the latent variable Z_{ij}^* is assumed to be normally distributed with mean $\gamma_{z_{ij}}$ and unit variance, so that

$$\pi_{z_{ij}} = P(Z_{ij} = 1) = P(Z_{ij}^* > 0) = \Phi(\gamma_{z_{ij}}), \tag{3.7}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function. The probability of malformation is related to covariates by expressing $\gamma_{z_{ij}}$ as some parameterized function of the predictors, e.g., the dose level, and the litter size.

For the bivariate response (W_{ij}, Z_{ij}), a bivariate normal distribution is assumed for the observed weight and the latent malformation variable for fetus j in litter i:

$$f(w_{ij}, z_{ij}^*) = \phi_2(w_{ij}, z_{ij}^* | \mu_{w_{ij}}, \sigma_{w_{ij}}^2, \gamma_{z_{ij}}, 1, \rho_{zw_{ij}}), \tag{3.8}$$

where $\rho_{zw_{ij}}$ is the intra-fetus correlation between the malformation and weight outcome. As a result, the joint distribution of the bivariate fetal weight and binary malformation outcome can be written as

$$\begin{aligned} f(w_{ij}, z_{ij}) &= f_w(w_{ij}) \times f_{z|w}(z_{ij}|w_{ij}) & (3.9) \\ &= \phi(w_{ij}|\mu_{w_{ij}}, \sigma_{w_{ij}}^2) \times \pi_{z|w_{ij}}^{Z_{ij}} (1 - \pi_{z|w_{ij}})^{1-Z_{ij}}, & (3.10) \end{aligned}$$

where $\pi_{z|w_{ij}} = \Phi(\gamma_{z|w_{ij}})$ is the conditional expectation of the binary malformation outcome $E(Z_{ij}|W_{ij})$. From bivariate normal theory,

$$\gamma_{z|w_{ij}} = \frac{\gamma_{z_{ij}} + \rho_{zw_{ij}} \frac{w_{ij} - \mu_{w_{ij}}}{\sigma_{w_{ij}}}}{(1 - \rho_{zw_{ij}}^2)^{1/2}}, \tag{3.11}$$

with $\pi_{z_{ij}} = \Phi(\gamma_{z_{ij}})$ the marginal expectation $E(Z_{ij})$.

In case of clustering, litter-specific parameters are considered to account for the correlation among the outcomes. Random effects on the mean fetal birth weight $\mu_{w_{ij}}$ and on the malformation parameter $\gamma_{z_{ij}}$ are introduced

$$\begin{aligned} \mu_{w_{ij}} &\sim N(\mu_{W_i}, \sigma^2_{\mu_i}) && (3.12) \\ \gamma_{z_{ij}} &\sim N(\gamma_{Z_i}, \sigma^2_{\gamma_i}), && (3.13) \end{aligned}$$

such that the bivariate distribution for fetal weight and binary malformation equals

$$\begin{aligned} f(w_{ij}, z_{ij}) &= f_w(w_{ij}) \times f_{z|w}(z_{ij}|w_{ij}) && (3.14) \\ &= \phi(w_{ij}|\mu_{W_i}, \sigma^2_{w_{ij}} + \sigma^2_{\mu_i}) \times \pi^{Z_{ij}}_{Z|W_{ij}}(1 - \pi_{Z|W_{ij}})^{1-Z_{ij}}, && (3.15) \end{aligned}$$

with $\pi_{Z|W_{ij}} = \Phi(\gamma_{Z|W_{ij}})$ the conditional expectation for the binary malformation outcome $E(Z_{ij}|W_{ij})$. We can derive that

$$\gamma_{Z|W_{ij}} = \frac{\frac{\gamma_{Z_i}}{\sqrt{1+\sigma^2_{\gamma_i}}} + \rho_{ZW_{ij}} \frac{w_{ij} - \mu_{W_i}}{\sigma^2_{w_{ij}} + \sigma^2_{\mu_i}}}{(1 - \rho^2_{ZW_{ij}})^{1/2}}, \qquad (3.16)$$

with $\rho_{ZW_{ij}}$ the intra-fetus correlation between the malformation and weight outcome. The marginal expectation for the binary malformation outcome $E(Z_{ij})$ equals $\pi_{Z_i} = \Phi(\gamma_{Z_i}/(1+\sigma^2_{\gamma_i}))$. The intra-litter correlation among the weight outcomes equals $\rho_{W_i} = \sigma^2_{\mu_i}/(\sigma^2_{\mu_i} + \sigma^2_{w_{ij}})$. The intra-litter correlation among the latent malformation outcomes equals $\rho_{Z_i} = \sigma^2_{\gamma_i}/(1 + \sigma^2_{\gamma_i})$. Further, the fetus-specific outcomes (malformation and weight) are modeled as function of dose d_i and litter size n_i, where $n_i \equiv m_i - r_i$ is a random variable with density $f(n_i|m_i) = g(r_i|\pi_{R_i}, \rho_{R_i}, m_i)$ as specified in previous section.

A dose-response model

Dose-response models are specified for the marginal outcomes of interest, i.e., the fetal weight, the probability of malformation, and the probability of death. Each of the univariate outcomes are allowed to vary as functions of dose and other covariates. The dose-response models can generally be written as

$$\mu_{W_i} = \boldsymbol{X}'_{a_{ij}}\boldsymbol{\alpha} + (n_i - \bar{n})\gamma_a, \tag{3.17}$$

$$\gamma_{Z_i} = \boldsymbol{X}'_{b_{ij}}\boldsymbol{\beta} + (n_i - \bar{n})\gamma_b, \tag{3.18}$$

$$\pi_{R_i} = \exp(\boldsymbol{X}'_{c_{ij}}\boldsymbol{\delta})/(1 + \exp(\boldsymbol{X}'_{c_{ij}}\boldsymbol{\delta})), \tag{3.19}$$

where $\{X_{a_{ij}}, X_{b_{ij}}, X_{c_{ij}}\}$ are the fetus- and/or litter-specific covariates with regression parameters $\theta = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}\}$. Often with developmental toxicity data, the assumption that variances and correlations are constant across dose groups is not appropriate. Therefore we allow the variances and correlations to vary with dose and possible other covariates as well. Thus, dose-response models for the parameters $\{\rho_D, \rho_{ZW}, \sigma_w, \sigma_\mu, \sigma_\gamma\}$ can be written as well, using appropriate transformations:

$$\rho_{ij} = (\exp(\boldsymbol{X}'_{t_{ij}}\boldsymbol{\tau}) - 1)/(\exp(\boldsymbol{X}'_{t_{ij}}\boldsymbol{\tau}) + 1), \tag{3.20}$$

$$\sigma_{ij} = \exp(\boldsymbol{X}'_{s_{ij}}\boldsymbol{\zeta}), \tag{3.21}$$

with $\rho_{ij} = \{\rho_D, \rho_{ZW}\}$ and $\sigma_{ij} = \{\sigma_w, \sigma_\mu, \sigma_\gamma\}$.

The need for numerical integration can be avoided by casting the model into a Bayesian framework and by resorting to the Gibbs sampler (Zeger and Karim 1991). In addition to the specified model, hyperprior distributions for the regression parameters need to be selected. We follow the recommendations of Besag, Green, Higdon and

Mengersen (1995) in using proper but highly dispersed hyperprior distributions. The hyperpriors chosen on the regression parameters for this analysis were $N(0, 10^6)$. We expect these priors to have minimal influence on the final conclusions of our analysis.

4 Application to Quantitative Risk Assessment

The primary goal of these studies is to determine a safe level of exposure. Recent techniques for risk assessment in this area are based on fitting dose-response models and estimating the dose corresponding to a certain increase in risk of an adverse effect over background, i.e., benchmark dose (Crump 1984). In case of multiple outcomes, the outcomes are often examined individually, using appropriate methods to account for the correlation, and regulation of exposure is based on the most sensitive outcome. It has been found, however, that a clear pattern of correlation exists between all outcomes (Ryan et al. 1991), so that risk assessment based on a joint model is more appropriate. The model must both incorporate the correlation between the outcomes, as well as the correlation due to clustering.

We define the combined risk due to a toxic effect as the probability that a fetus is dead or a viable fetus is malformed and/or suffers from low birth weight. This risk can be expressed as

$$\begin{aligned} r(d) &= P(\text{death}|d) + P(\text{viable}|d) \times P(\text{malformed or low weight}|\text{viable}, d) \quad (4.1) \\ &= P(R = 1|d) + (1 - P(R = 1|d)) \times P(M = 1 \text{ or } W < W_c|N \geq 1, d) \\ &= \pi_{dth} + (1 - \pi_{dth}) \times P(M = 1 \text{ or } W < W_c|N \geq 1, d). \quad (4.2) \end{aligned}$$

The joint probability of a malformation or low birth weight is equal to:

$$P(M = 1 \text{ or } W < W_c | N \geq 1, d) \tag{4.3}$$

$$= \int_1^\infty P(M = 1 \text{ or } W < W_c | N = n, d) P(N = n | d) dn \tag{4.4}$$

and

$$P(M = 1 \text{ or } W < W_c | N = n, d) \tag{4.5}$$

$$= 1 - \int_{-\infty}^{-\tau} \int_{W_c}^{\infty} \phi_2(W_{ij}, M_{ij}^*; \mu_W(d), 0, \sigma_w(d)^2 + \sigma_\mu^2, 1 + \sigma_\gamma^2, \rho_{ZW}(d)) dW_{ij} dM_{ij}^*$$

$$= \Phi(\tau) + \Phi_2(-\tau, \omega; \rho(d)), \tag{4.6}$$

where $\tau = \gamma_z(d)/\sqrt{1 + \sigma_\gamma^2}$ and $\omega = (W_c - \mu_w(d))/\sqrt{\sigma_w(d)^2 + \sigma_\mu(d)^2}$ and Φ_2 is the standard bivariate normal distribution function.

The benchmark dose is defined as the level of exposure corresponding to an acceptably small excess risk over background, i.e., the dose satisfying

$$r^*(d) = \frac{r(d) - r(0)}{1 - r(0)} = q, \tag{4.7}$$

with q the prespecified level of increased risk over background, typically specified as 0.01, 1, 5, or 10% (Crump, 1984). In the frequentist framework, the benchmark dose calculation is based on the estimated dose-response curve. In the Bayesian approach one could choose to base the benchmark dose calculation on the mean posterior risk curve. This method is illustrated in Figure 3. The full line corresponds with the mean posterior risk. But, benchmark dose calculations are no more precise than the data on which they are based. Therefore, rather than calculating a point estimate of the safe dose, one might be interested in the entire posterior distribution of the safe dose. In

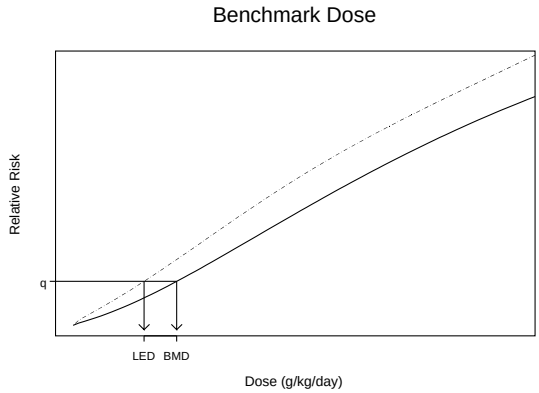


Figure 3: *Definition of benchmark dose (BMD) and lower effective dose (LED) at $q\%$ of increased risk over background.*

this way, the researcher could get an idea of the precision of the estimate. Often, one is interested in an upper bound of the benchmark dose to set a safe level of exposure. One can construct a 95% upper credibility limit of the risk function and base the safe dose calculation upon this upper limit. In analogy with the frequentist approach, the lower effective dose is defined as the dose such that the 95% upper credibility limit of the excess risk is equal or greater than the predefined level q. This is illustrated in Figure 3. The dashed line corresponds to the 95% upper credibility limit of the posterior risk.

5 Data Analysis

Dose-Response Modeling

For risk assessment to be reliable, the dose-response model should fit the data well in all respects. A frequently used predictor model in literature is the linear model. With the aim on low dose extrapolation, more flexible predictor models are investigated. High order polynomials offer a wide range of curves, but often fit badly at the extremes. Royston and Altman (1994) introduced fractional polynomials as a generalization of

Table 2: Deviance Information Criterion for the best first and second order fractional polynomials to model the malformation parameter γ_Z.

$m = 1$		$m = 2$	
tranformation	DIC	transformation	DIC
$1/d^2$	678.4	$(\sqrt{d}, d)$	677.8
$1/d$	679.9	(d, d^2)	678.4
$1/\sqrt{d}$	681.0	$(\ln(d), \ln^2(d))$	678.0
$\ln(d)$	682.2	$(1/\sqrt{d}, 1/d)$	677.9
$\sqrt{d}$	678.1	$(1/d, 1/d^2)$	678.1
d	685.3	$(d, d\ln(d))$	678.5
d^2	692.0	$(\sqrt{d}, \ln(d))$	677.2
d^3	693.6	$(d, \ln(d))$	678.5

the conventional polynomials. A fractional polynomial of degree m is defined as

$$\beta_0 + \sum_{j=1}^{m} \beta_j d^{p_j}, \tag{5.1}$$

where $d^0 \equiv \ln(d)$ and the powers $p_1 \leq \ldots \leq p_m$ are positive or negative integers or fractions. Royston and Altman (1994) argue that polynomials of degree higher than 2 are rarely required in practice and suggested to choose the value of the powers from the set $\{-2, -1, -0.5, 0, 0.5, 1, 2, \ldots, \max(3, m)\}$. Fractional polynomials are shown to be very useful both in the context of dose-response modeling and quantitative risk assessment (Geys et al. 1999, Faes et al. 2003).

In order to select a parsimonious model for the data we select a suitable set of dose transformations for each of the three outcomes separately. Model selection is performed using the deviance information criterion (DIC) as proposed by Spiegelhalter et al. (1998, 2002):

$$DIC = D + 2P_D, \tag{5.2}$$

with D a point estimate of the deviance and P_D the effective number of parameters. Smaller values of DIC indicate a better fitting model. Table 2 shows that a fractional polynomial of degree $m = 1$, whether represented by $1/d^2, 1/d, 1/\sqrt{d}, \ln(d), d, d^2$ or d^3, is unacceptable as opposed to a fractional polynomial of degree $m = 2$ to model the malformation parameter γ_{Z_i}. Table 2 tabulates only a selection of the considered two-degree fractional polynomials. None of the other combinations provided a substantial improvement. The fractional polynomial represented by $(\ln(d), \sqrt{d})$ yields the smallest DIC. A similar approach, applied to the death outcome and weight outcome, suggest a d^2 trend on π_{R_i} and a $\ln(d)$ trend on μ_{W_i}. The resulting set of transformations is then used to construct more elaborate models that can be scrutinized further by means of the DIC. The most complex model we considered (Model 1) allows the following trends on the malformation, weight and death outcomes:

$$\begin{aligned} \mu_{W_i} &= \beta_{0W} + \beta_{1W} \ln(d+1) + \beta_{2W}(n - \bar{n}), && (5.3) \\ \gamma_{Z_i} &= \beta_{0Z} + \beta_{1Z}\sqrt{d} + \beta_{2Z} \ln(d+1) + \beta_{3Z}(n - \bar{n}), && (5.4) \\ \text{logit}(\pi_{R_i}) &= \beta_{0R} + \beta_{1R} d^2. && (5.5) \end{aligned}$$

Further, linear d trends on the association parameters

$$\rho_{R_i}, \sigma^2_{\gamma_i}, \sigma^2_{\mu_i}, \rho_{ZW}, \qquad (5.6)$$

are considered. From Table 3 summarizes the model selection procedure on the association parameters. Based on the deviance information criterion, there is evidence for choosing a model with a constant association between weight and malformation and a constant malformation variance (Model 3). In contrast, there is evidence for choosing a model with a d trend on the weight variance (Model 4). Finally, there seems to be

Table 3: Model Selection on the Association Parameters. A '*' indicates a linear d trend on that parameter. All other effects are kept constant.

Model	ρ_{dth}	σ^2_μ	σ^2_γ	ρ_{ZW}	DIC
1	*	*	*	*	-1356.870
2	*	*	*	·	-1358.310
3	*	*	·	·	-1358.490
4	*	·	·	·	-1357.950
5	·	*	·	·	-1360.470
6	·	·	·	·	-1359.540

no evidence for the linear d trend on the correlation among death outcomes. As such, we choose Model 5.

Parameter estimates obtained from fitting the final model are displayed in Table 4. The dose coefficient is significantly negative for fetal birth weight, and the negative coefficient of litter size suggests that larger litters had a higher risk of low fetal birth weight which is not unexpected due to competition for food resources. The intralitter correlation for weight is substantial, and increases from 0.441 in the control group to 0.644 in the highest dose group. For malformation, there is an increasing dose effect, and there appears to be little effect of litter size on malformation. The intralitter correlation for malformation is also large. The correlation between malformation and birth weight appears to be negative, indicating that fetal malformations are associated with lower fetal weights. For fetal death, there is a significantly positive effect with dose. The intralitter correlation for fetal death is also significantly positive. Figure 4 shows the posterior mean curves together with the 95% credibility intervals of the univariate dose-response curves. All the univariate fits are acceptable.

Table 4: Posterior Mean and Standard Deviation of the Parameters in the Final Model.

	Effect	Mean	(StDev)
Fetal Weight:			
Mean:	intercept	0.980	(0.013)
	log(dose)	-0.433	(0.037)
	$n_i - \bar{n}$	-0.011	(0.003)
Correlation:	0.000	0.441	(0.063)
	0.250	0.492	(0.047)
	0.500	0.545	(0.043)
	1.000	0.644	(0.069)
Malformation:			
Mean:	intercept	-3.857	(0.631)
	$\sqrt{\text{dose}}$	6.062	(2.529)
	log(dose)	-2.664	(2.995)
	$n_i - \bar{n}$	-0.002	(0.049)
Correlation:		0.691	(0.048)
Fetal Weight / Malformation:			
Correlation:		-0.018	(0.005)
Fetal Death:			
Mean:	intercept	-2.099	(0.153)
	dose^2	0.730	(0.258)
Correlation:		0.069	(0.024)

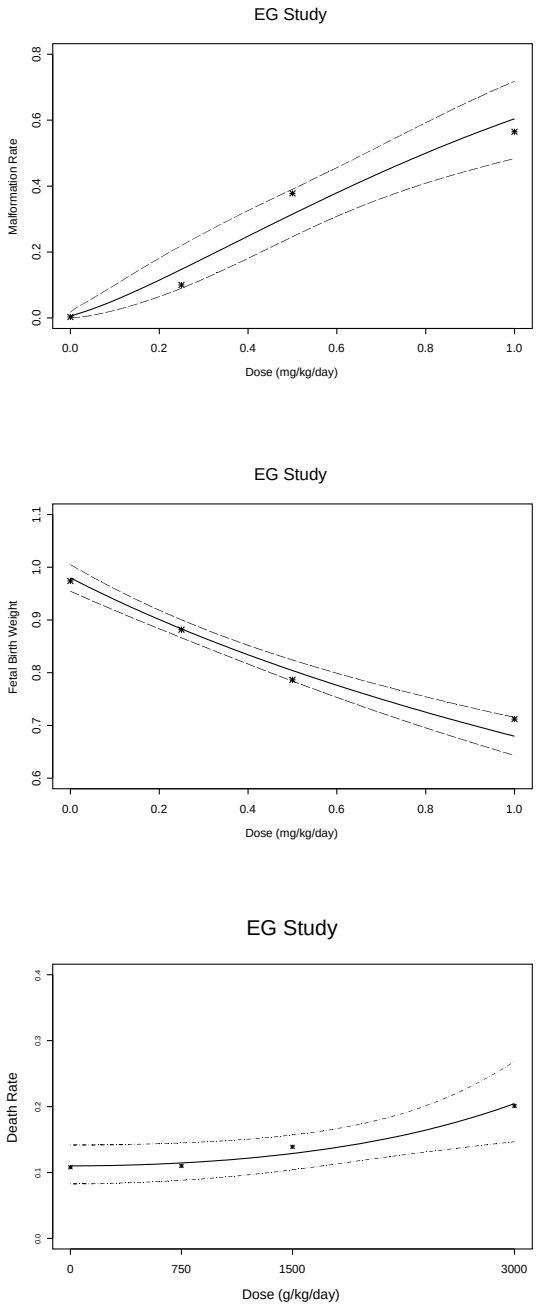


Figure 4: *EG data: From Top to Bottom: Estimated malformation rates; Estimated fetal weight; Estimated death rate*

Table 5: Risk Assessment for EG Study in Mice.

	BMD		LED	
Model	$q = 0.01$	$q = 0.05$	$q = 0.01$	$q = 0.05$
Joint	37	190	17	126
Malf	56	299	19	163
Weight	106	383	77	312
Death	1055	2209	878	1843

Quantitative Risk Assessment

To calculate the risk of low birth weight, we need to define a weight below which a fetus can be considered as being of "low fetal weight". Because of the arbitrariness of the cutpoint, estimating a benchmark dose from a continuous response has led to much discussion (Bosch et al. 1996, Crump 1984). We specify the cutoff point W_c as two standard errors below the control average fetal weight (Catalano and Ryan 1992). By means of this definition, fetuses that weighed less than 0.777g are considered to be of low fetal weight, which corresponds to a 3.4% rate in the control animals. The posterior density of the combined risk due to a fetal death, a malformation or low fetal weight is pictured in Figure 5. The risk gradually increases when dams are exposed to larger quantities of the toxic substance, before finally reaching an asymptote.

Table 5 shows the benchmark dose and lower effective dose corresponding to a 1% and 5% excess risk over background, respectively based upon the posterior mean and 95% upper credibility limit of the risk curve. We also added the corresponding quantities, calculated from univariate risks. The joint model yields more conservative doses. Often, a safe level of exposure is determined separately for each outcome and the lower of the individual outcomes is used as an overall benchmark dose. It is clear that this approach would yield too high estimated safe doses. Therefore, it is necessary to model the full hierarchical data structure when searching for a safe level

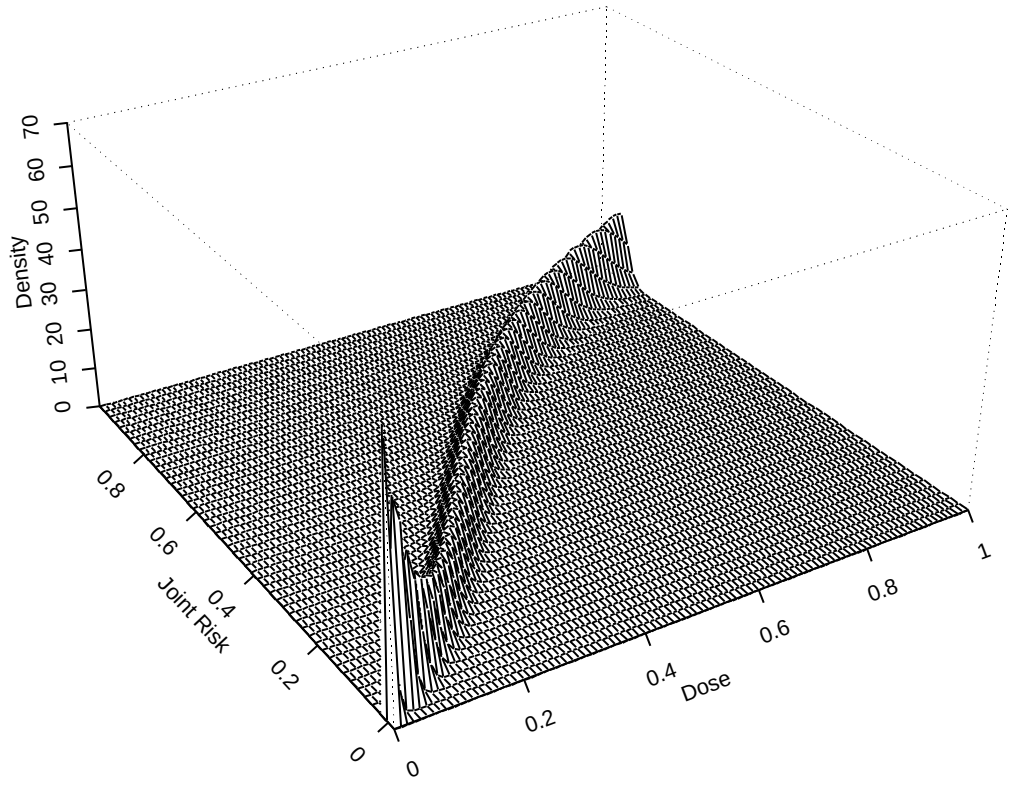


Figure 5: *EG data: Posterior Density of Combined Risk Due to a Fetal Death, a Malformation or Low Fetal Weight.*

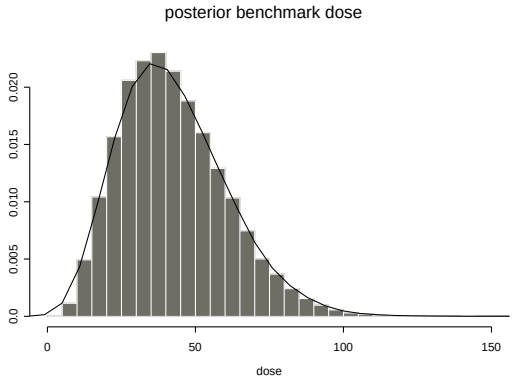


Figure 6: *EG data: Posterior distribution of benchmark dose, corresponding to a 5% increase of risk over background.*

of exposure.

Instead of calculating a point estimate of the benchmark dose, we can derive the full posterior distribution of the benchmark dose. The posterior distribution of the benchmark dose corresponding to a 1% increase in risk over background is pictured in Figure 6. The posterior mode equals 35. Previously obtained estimate of the benchmark dose based on the posterior mean of the risk curve (37) lies in between the mode (35) and mean (43) of the posterior distribution of the benchmark dose. The 95% lower credibility limit of the posterior benchmark dose is equal to 17. Calculation of the posterior distribution of the benchmark dose does not only give information about the estimated safe level of exposure, but also its uncertainty and shape of the distribution.

6 Discussion

Developmental toxicity studies are complicated by the hierarchical, clustered and multivariate nature of the data. As a consequence, a multitude of modeling strategies

have been proposed in literature. Often, focus is only on the outcomes measured on the viable fetuses. However, as observed from the data sets, the number of viable fetuses in a dam, i.e., the litter size, also decreases with increasing dose levels. Thus, a method that acknowledges the stochastic nature of the litter size is in demand. In this setting, a Bayesian random effects model was proposed. All outcomes of the developmental toxicity study were analyzed simultaneously. The main advantage of the proposed methodology is the flexibility in which all stages of the data can be modeled.

The model was applied to a developmental toxicity study (EG in mice), and used for quantitative risk assessment. When interested in a safe level of exposure, it is important to account for all possible adverse effects. Often however, focus is only on the outcome that is most sensitive to the exposure when performing quantitative risk assessment. But, use of univariate methods to determine a safe dose level can yield unreliable, and thus unsafe, dose levels. This acknowledges the importance of a model that accounts for the full data structure.

A Bayesian estimation of a safe level of exposure provides an attractive alternative to the commonly used frequentist approaches. The posterior distribution of the benchmark dose does not only give a point estimate, but reflects also the uncertainty associated with this estimate.

Although the method is presented in the specialized field of developmental toxicity, the methodology is applicable in a general clustered or even general correlated data setting with a continuous and binary outcome. Thus, use of the proposed modeling approach extends far beyond the developmental toxicity context.

Acknowledgment

We gratefully acknowledge support from the Institute for the Promotion of Innovation by Science and Technology (IWT) in Flanders, Belgium, from the Fund for Scientific Research Flanders and from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy).

References

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). *Topics in Modeling of Clustered Data*, Chapmann and Hall.

Allen, A.S., Barnhart, H.X. (2002). Joint Modles for Toxicity Studies with Dose-Dependent Number of Implantations. *Risk Analysis*, **22**, 1165–1173.

Arnold, B.C., and Strauss, D. (1991). Pseudolikelihood estimation: Some examples. *Sankhya B*, **53**, 233–243.

Besag, J., Green, P.J., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, **10**, 3–66.

Bosch, R.J., Wypij, D., and Ryan, L.M. (1996). A semiparametric approach to risk assessment for quantitative outcomes. *Risk Analysis*, **16**, 657–665.

Box, G.E.P., and Tiao, G.C. (1992). *Bayesian Inference in Statistical analysis*, John Wiley and Sons: New York.

Catalano, P., Ryan, L., and Scharfstein, D. (1994). Modeling fetal death and malformation in developmental toxicity. *Journal of the American Statistical Association*, **87**, 651–658.

Crump, K.S., and 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, 187–203.

Crump, K. (1984). A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology*, **4**, 854–871.

Dunson, D.B. (1998). Dose-dependent number of implants and implications in developmental toxicity. *Biometrics*, **54**, 558–569.

Dunson, D.B., Chen, Z. and Harry, J. (2003). A Bayesian approach for Joint Modeling of Cluster Size and Subunit-Specific Outcomes. *Biometrics*, **59**, 521–530.

Fitzmaurice, G.M., and Laird, N.M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, **90**, 845–852.

Gueorguieva, R.V., and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, **96**, 1102–1112.

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

Kuk, A.Y.C. (2003). A generalized estimating equation approach to modelling foetal responses in developmental toxicity studies when the number of implants is dose dependent. *Applied Statistics*, **52**, 51–61.

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

McCullagh, P. and Nelder, J.A. (1994). *Generalized Linear Models*. London: Chapman and Hall.

McCulloch, C.E. and Searle, S.R. (1989). *Generalized, linear and mixed models*. Willey: New York.

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

Molenberghs, G., and Lesaffre, E. (1999). Marginal Modeling of Multivariate Categorical Data. *Statistics in Medicine*, **18**, 2237–2255.

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

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.

Rai, K., and Van Ryzin, J. (1985). A dose-response model for teratological experiments involving quantal responses. *Biometrics*, **47**, 825–839.

Regan, M.M. and Catalano, P.J. (1999). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics*, **55**, 760–768.

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

Ryan, L.M., Catalano, P.J., Kimmel, C.A., and Kimmel, G.L. (1991). Relationship between Fetal Weight and Malformation in Developmental Toxicity Studies, *Teratology*, **44**, 215–223.

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

Spiegelhalter, D.J., Best, N.G., and Carlin, B.P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Research Report 98-009, Division of Biostatistics, University of Minisota*.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linder, A. (2002). Bayseian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 1–34.

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

Williams, P.L., and 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.

Williamson, J.M., Datta, S., and Satten, G.A. (2003). Marginal Analyses of Clustered Data When Cluster Size is Informative, *Biometrics*, **59**, 36–42.

Zeger, S.L., and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–102.

List of Figures

List of Tables