Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/140422

Model Averaged Benchmark Dose Analysis for Multiple Outcomes in Developmental Toxicity Experiments

Eman Khorsheed[1,*] *and Mehdi Razzaghi*[2]

[1]Department of Mathematics, College of Science, University of Bahrain, P.O. Box 32038, Sakhir, Kingdom of Bahrain
[2]Department of Mathematical and Digital Sciences, Bloomsburg University, Bloomsburg, PA, USA

Received: 7 Mar. 2020, Revised: 15 May. 2020, Accepted: 20 May. 2012
Published online: 1 Jul. 2020

Abstract: To assess the adverse effect of a toxin, animal bioassay experiments are frequently designed and performed on animals. In a typical experiment, pregnant mice are exposed to a dose of a chemical during a critical period of gestation. Animals are sacrificed before term and uterine content is examined for the number of fetuses that are dead/resorbed, malformed or normal. The outcomes are used to fit a dose-response relationship for risk assessment and determination of safe dosage levels. The choice of the dose-response model can severely affect the results. Although several models can fit the data well in the experimental dose ranges, when extrapolated to low human exposure levels, the estimate of safe exposure levels can vary substantially. To reduce uncertainty in the presence of multiple outcomes from developmental toxicity experiments, we propose the application of the Bayesian Model Averaging (BMA) technique whereby a series of candidate dose-response models are used, the safe exposure level is determined based on each model, and a weighted average is used for the final estimate. Simulation studies are presented to show that the methodology works well and can reliably be applied in practice. The methodology is further illustrated using a real experimental data set.

Keywords: Bayesian model averaging, MCMC, dose multinomial response data, risk assessment, Dirichlet-Trinomial model.

1 Introduction

Developmental toxicity studies are conducted on laboratory animals to assess the harmful effects and incidence of developmental disease as result of exposure to a toxic substance. Typically in such studies, pregnant female animals are exposed to a dosage of a chemical during a critical time of the gestation period. Animals are sacrificed just before term and the uterine contents are examined for a variety of developmental defects such as death, resorption, malformation and the fetal weight. Exposure generally occurs at dosage levels that are much higher than usual human to induce toxicity in a limited number of animals. A dose response model is then fitted to the proportion of affected fetuses. The model is used to estimate the risk at low exposure levels and the benchmark doses. Traditionally, a single outcome, such as the number of fetuses with malformation or the total number of affected offspring including dead/resorbed and malformed fetuses, were used for model fitting, see for example [1], [2] and [3]. This was the motivation behind [4] where the authors considered a model averaging approach to reduce uncertainty due to the choice of the dose-response model and demonstrate the application of Bayesian model averaging technique in developmental toxicity studies. However, it should be realized that consideration of a single outcome does not truly characterize developmental toxicity experiments. More realistically, a multiple combination of continuous and discrete outcomes, such as fetal weight, viability, and malformation status of the fetus, are simultaneously observed. Thus, concurrent outcomes should be considered simultaneously. Here, we consider the problem of risk assessment for multiple binary outcomes using the Bayesian Model Averaging (BMA) methodology.

A major difficulty in modeling outcomes from developmental toxicity experiments is the existence of the intra-litter correlation. Authors have studied modeling of multiple outcomes in developmental toxicity studies with varying approaches in incorporating the litter effects. For example, [5] considered a generalization of the beta-binomial model by applying the Dirichlet-trinomial distribution to multiple outcomes, [6] modelled the fetal

* Corresponding author e-mail: ekhorsheed@uob.edu.bh

death, fetal weight, and malformation, [7] used a multinomial dose-response model along with a modified Weibull distribution, and [8] proposed a double beta-binomial model using the quasi-likelihood method. The advantage of the multinomial approach is that it accounts for the hierarchical structure of the fetal development. More specifically, a fetus can become malformed only if resorption or death does not occur. [9] provide a useful discussion of various approaches. The authors further argue that their choice of the dose-response model might make more biological sense than the Weibull model of [2], since malformation and death can be associated with some underlying continuous phenomena.

Here, we consider the problem of modeling the trinomial responses from developmental toxicity studies. Our approach will be similar to that of [5] and [8]. Instead of using two single logistic distributions to express the dose-response relationships between the mean response probability of death/resorption and malformation with the dose, we propose a model averaging method to reduce uncertainty due to the choice of dose-response models. In the next section, we describe the modeling structure of the process.

2 Model Description

Suppose that a developmental toxicity experiment consists of a control and g nonzero dose levels with $0 = d_0 < d_1 < \ldots < d_g$. Assume that m_i, $i = 0,1,\ldots,g$ pregnant female animals are exposed to a dose d_i of a known teratogenic substance and let n_{ij}; $j = 1,\ldots,m_i$, $i = 1,\ldots,g$ be the total number of fetuses from jth dam in the ith dose group. Note that n_{ij} includes all normal fetuses as well as the dead/resorbed, and malformed ones. Generally, exposure occurs during a critical time when skeletal structure of the fetus is being formed and can be through gavage, diet, drink, inhalation, dermal, or any other route. For example, with murine, exposure usually is during gestation days of 5 through 12. Animals are sacrificed just before term and affected fetuses are counted upon examination of the uterus content for each animal. Now, fetuses can be affected in a variety of ways. One possible outcome is fetal death or resorption. Generally, in developmental toxicity experiments, no distinction is made between fetal death and resorption. Embryos can be born with skeletal defect or other malformations. Finally, another possible outcome is the normal healthy fetus. Many constructed models of developmental toxicity experiments considered only a dichotomy of responses as to whether or not a fetus was normal or abnormal with abnormality defined as including both malformation and death/resorption. Several models were used for this binary response approach. In fact, Khorsheed & Razzaghi [4] applied the model averaging technique to reduce the uncertainty due to the choice of models for dichotomous outcomes in developmental toxicity experiments. However, the realization of the multinomial nature of the outcomes is crucial. Following Chen et al. [5], several researchers considered the multinomial nature of the outcomes. Perhaps Chen et al [5] are the first to consider the responses as a trichotomy. It is noted that fetal weight is also a crucial indicator of toxicity. Other authors considered joint modeling of discrete and continuous outcomes. Specifically, Catalono and Ryan [10] defined bivariate latent variables to jointly model discrete and continuous outcomes. Regan and Catalano [11] proposed a likelihood-based model that is an extension of a correlated probit model to incorporate continuous outcomes . Najita and Catalano [12] studied the BMD determination for multiple outcomes from developmental toxicity experiments. In this paper, however, we only consider the advantages of the model averaging approach in the hierarchical structure described for discrete trinomial responses. Thus, one consideration may be the incidence of death/resorption, malformation, and normal outcomes as used in [5]. Another consideration, as applied in [13], is to dichotomize the fetal weight and consider the malformation, fetal weight status, and normal as the trinomial response. For simplicity, in the development of the model, we use the first consideration although the second consideration is identical. In fact, our simulation study will be based on the second consideration. Accordingly, let $Y_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})$ be the observed trinomial vector of the jth dam in the ith dose group, where the components of the vector Y_{ij} represent the number of dead/resorbed, malformed, and normal fetuses respectively. Clearly,

$$Y_{ij1} + Y_{ij2} + Y_{ij3} = n_{ij} \qquad j = 1,\ldots,m_i, i = 1,\ldots,g$$

Let $\theta_1(d_i)$ be the probability that at exposure level of $d_i; i = 0,1,\ldots,g$ an implanted embryo dies or is resorbed during the gestation. Thus, the probability $P_1(d_i)$ associated with Y_{ij1} is given by

$$P_1(d_i) = \mu_1(d_i)$$

Given that a fetus survives death/resorption, let the conditional probability that it becomes malformed be $\theta_2(d_i)$. Then, the probability associated with Y_{ij2} is given by

$$P_2(d_i) = [1 - \mu_1(d_i)]\mu_2(d_i)$$

Finally, the probability that a fetus is born normal, associated with Y_{ij3}, can be expressed as

$$P_3(d_i) = [1 - \mu_1(d_i)][1 - \mu_2(d_i)]$$

Clearly, the choice of the risk functions $\mu_1(d_i)$ and $\mu_2(d_i)$ can severely affect the risk estimates. [5] used the Weibull model

$$\mu_i(d) = 1 - exp(-exp(\alpha_i + \beta_i log(d)) \qquad i = 1,2$$

and [7] generalized the model by adding the number of implants of the litter as a covariate, while [9] applied the modified probit model. The fact is that there is not a single dose-response relationship that is universally accepted as producing the best results. We therefore believe that a model averaging approach can be promising in this regard.

3 Incorporating the Litter Effect

One of the main characteristics of the developmental toxicity experiments is the fact that fetuses from the same litter show more similarity in response than from different litters. Existence of this extra binomial variation results in overdispersed outcome data and ignoring this litter effect can result in severely erroneous conclusion.

To incorporate this litter effect in the case of binary responses, several approaches are possible. The most common is to assume that p_{ij}, the response probability of an adverse effect including death/resorption and malformation in the jth litter of the ith dose group follows a beta distribution. In that case, the unconditional distribution of the number of responses in the jth litter of the ith dose group becomes the familiar beta binomial distribution. This approach was first proposed by [14]. In fact, this was the approach adopted by [4] to illustrate the application of Bayesian model averaging in benchmark dose analysis for developmental toxicity experiments. As described in that paper, other methods have been proposed for modeling dichotomous responses in dispersed developmental toxicity outcomes. For example, Ryan [3] applied the Generalized Estimating Equation (GEE) approach, while [15] discussed the application of the quasi-likelihood method.

Now, for the multinomial response consideration, the natural extension of the beta-binomial distribution is the Dirichlet-Trinomial model, which was discussed and adopted by [5]. Accordingly, we assume a trinomial distribution for the response vector Y_{ij} that is

$$P(y_{ij1},y_{ij2},y_{ij3}|p_{ij1},p_{ij2},p_{ij3}) = \binom{n_{ij}}{y_{ij1},y_{ij2}} p_{ij1}^{y_{ij1}} p_{ij2}^{y_{ij2}} p_{ij3}^{y_{ij3}}$$

where p_{ij1},p_{ij2},p_{ij3} are respectively the probabilities that the jth litter of the ith dose group is dead/resorbed, malformed or normal with $p_{ij1}+p_{ij2}+p_{ij3}=1$ and

$$\binom{n_{ij}}{y_{ij1},y_{ij2}} = \frac{n_{ij}!}{y_{ij1}!y_{ij2}!(n_{ij}-y_{ij1}-y_{ij2})!}$$

Also, if we express the joint distribution of p_{ij1},p_{ij2},p_{ij3} as the Dirichlet distribution given by

$$P(p_{ij1},p_{ij2},p_{ij3}) = \frac{\Gamma(\alpha_i+\beta_i+\gamma_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)\Gamma(\gamma_i)} p_{ij1}^{\alpha_i-1} p_{ij2}^{\beta_i-1} p_{ij3}^{\gamma_i-1}$$

where $\alpha_i,\beta_i,\gamma_i>0$, the unconditional distribution of Y_{ij} is given by

$$P(y_{ij1},y_{ij2},y_{ij3}) = \frac{n_{ij}!}{y_{ij1}!y_{ij2}!y_{ij3}!}\cdot \frac{\Gamma(\alpha_i+\beta_i+\gamma_i)\Gamma(y_{ij1}+\alpha_i)\Gamma(y_{ij2}+\beta_i)\Gamma(y_{ij3}+\gamma_i)}{\Gamma(n_{ij}+\alpha_i+\beta_i+\gamma_i)\Gamma(\alpha_i)\Gamma(\beta_i)\Gamma(\gamma_i)} \quad (1)$$

which is a generalization of the beta-binomial distribution, called the Dirichlet-Trinomial distribution, see [16]. Under the Dirichlet-Trinomial model, we have $E(Y_{ij1}) = n_{ij}\mu_{i1}$ and $E(Y_{ij2}|y_{ij1}) = (n_{ij}-y_{ij1})\mu_{i2}$ where $\mu_{i1} = \alpha_i\theta_i$ and $\mu_{i2} = \beta_i\theta_i(1-\alpha_i\theta_i)$ with $\theta_i = (\alpha_i+\beta_i+\gamma_i)^{-1}$. The intralitter correlation for the Dirichlet-Trinomial model is $\rho_i - \frac{\theta_i}{1+\theta_i}$. The parameters μ_{i1} and μ_{i2} can be interpreted as the means of p_{ij1} and p_{ij2}, respectively. As noted in [5], since in the model y_{ij1},y_{ij2}, and y_{ij3} are mutually negatively correlated, the Dirichlet-Trinomial model assumes that the correlations between number of deaths/resorptions, malformations, and normal fetuses are negative. Several other properties of the Dirichlet-Trinomial distribution in modeling the multiple outcomes from developmental toxicity experiments are discussed in [5]. Specifically, the authors show that the Dirichlet-Trinomial distribution can be expressed as the product of two beta-binomial models and that it is a special case of the more general class of distributions called the double beta-normal

$$P(y_{ij1},y_{ij2},y_{ij3}) = \frac{n_{ij}!}{y_{ij1}!(n_{ij}-y_{ij1})}\cdot \frac{\Gamma(\alpha_i+\delta_i)\Gamma(y_{ij1}+\alpha_i)\Gamma(n_{ij}-y_{ij1}+\delta_i)}{\Gamma(n_{ij}+\alpha_i+\gamma_i)\Gamma(\alpha_i)\Gamma(\delta_i)}\cdot \frac{(n_{ij}-y_{ij1})!}{y_{ij2}!(n_{ij}-y_{ij1}-y_{ij2})!}\cdot \frac{\Gamma(\beta_i+\gamma_i)\Gamma(y_{ij2}+\beta_i)\Gamma(n_{ij}-y_{ij1}-y_{ij2}+\gamma_i)}{\Gamma(n_{ij}-y_{ij1}+\beta_i+\gamma_i)\Gamma(\beta_i)\Gamma(\gamma_i)} \quad (2)$$

with constraint $\delta_i = \beta_i+\gamma_i$.

The maximum likelihood estimates of the parameters were derived and the model was used in risk assessment for an experiment on developmental effects resulting from exposure to hydroxyurea. No specific dose response model was used and μ_{i1} and μ_{i2} were treated as parameters rather than functions of dose. The double beta-binomial model was used by [8] for simultaneous modeling of multinomial responses in developmental toxicity experiments. The authors use the linear-logistic dose-response models

$$logit(\mu ir) = log(\frac{\mu_{ir}}{1-\mu_{ir}}) = a_r + b_r d_i \qquad r=1,2. \quad (3)$$

to express the mean functions. Rather than the maximum likelihood method, [7] applied the generalized

estimating equation (GEE) approach to estimate the parameters. An extended Dirichlet-Trinomial covariance function was used to express overdispersion. They showed that their method has some statistical and computational advantages over separate analysis of the end points. However, Gaylor [17] argues that because of the differences in the sensitivities of dams at a given dose within a group, more variability in the data is expected than could be explained by the trinomial distribution. The author suggests using the quasi-likelihood approach described in [18] to account for the overdispersion induced by the litter effect. The quasi-likelihood model was used by [8].

Since we aim to demonstrate the application of the Bayesian model averaging, we simply apply the Dirichlet-Trinomial model (1) similar to [5], but instead of treating the means μ_{i1} and μ_{i2} as parameters, we express them as functions of dose. Also, rather than using a single dose-response model to express μ_{i1} and μ_{i2}, as in [19] and [8], we use a weighted average of several models and determine the weights using a Bayesian model averaging approach.

4 Risk Assessment Using Bayesian Model Averaging

For risk assessment, we use a procedure described in [9] whereby the overall risk is measured by the probability of being affected (either dead/resorbed or malformed) i.e.

$$P(d) = 1 - P_3(d) = 1 - [1 - \mu_1(d)][1 - \mu_2(d)] \tag{4}$$

Now, as explained by Khorsheed and Razzaghi [4], if we define the measure of risk as the excess risk above the background,

$$\pi(d) = P(d) - P(0) \tag{5}$$

then the benchmark dose (BMD) is the dosage level that sets the above risk equal to a small negligible risk, such as 5% to 10%. The benchmark dose lower bound (BMDL) is the lower 95% confidence limit of BMD. An estimate of the variance of BMD may be obtained using approximate normality of BMD as suggested by [9] or by bootstrapping. Thus, in the Dirichlet-Trinomial model given in (1), we first derive a reparameterization by substituting

$$\alpha_i = \frac{\mu_{i1}}{\theta_i} \tag{6}$$

$$\beta_i = \frac{\mu_{i2}}{\theta_i(1-\mu_{i1})} \tag{7}$$

$$\gamma_i = \frac{\{(1-\mu_{i1})\}^2 - \mu_{i2}}{\theta_i(1-\mu_{i1})} \tag{8}$$

The mean functions μ_{i1} and μ_{i2} are then replaced by a dose-response function selected from a set of k candidate models $\mu_{rk}(d) = F_k(\omega_{1rk} + \omega_{2rk}d^h)$ for $r = 1,2$ and $k = 1,\ldots,K$. After estimating all model parameters, we can determine BMD and BMDL for each candidate model. We then use the Bayesian model averaging technique described below to derive the overall weighted BMD and BMDL.

In the BMA methodology, we begin by assuming that all the K models have a priori equal weights, that is

$$P(W_k) = \frac{1}{K} \qquad k = 1,2,\ldots,K$$

Then, the final weights are estimated by the posterior model probabilities, which by Bayes' theorem are given by

$$P(W_k|L) \propto K^{-1}P(L|W_k) \qquad k = 1,2,\ldots,K$$

where $P(L|W_k)$ represents the marginal distribution of the likelihood given each model. Now, as explained in [20], the computation of the marginal distributions for calculation of the posterior model probabilities in the implementation of BMA requires solving an integral that is difficult to calculate except for very simple cases. Indeed, in most cases, especially data related to environmental and epidemiological studies derivation of closed form solutions is infeasible and the use of the Bayesian Information Criteria (BIC) to approximate the marginal distributions has successfully been adopted. Specifically, [21] suggests the following approximation,

$$P(W_k|L) \propto exp(-\frac{1}{2}BIC(W_k)) \qquad k = 1,\ldots,K$$

with

$$BIC(W_k) = -2log(maxL|W_k) + a_k log(n)$$

where a_k is is the number of parameters for W_k, n is the sample size and $maxL$ is the maximum of the likelihood function. In developmental toxicity experiments, the sample size is the litter size. According to [22], this approximation works well in moderate sample sizes when the covariates are independent. Therefore, the weights may be obtained by

$$P(W_k|L) = \frac{exp(-\frac{1}{2}BIC(W_k))}{\sum_{r=1}^{K} exp(-\frac{1}{2}BIC(W_r))} \qquad k = 1,\ldots,K$$

This approach has been successfully implemented in several applications with dichotomous responses, see for example [23] and [24]. However, for calculating the weights, [25] suggests replacing $BIC(W_k)$ by

$$\Delta(k) = BIC(W_k) - min_{1\le r\le K} BIC(W_r) \tag{9}$$

The advantage of this technique is that the "Δ values are on a continuous scale of information and are interpretable regardless of the measurement scale and whether the data are continuous, discrete or categorical." Here, the technique is found to be computationally more stable especially when $BIC(W_k)$ is relatively large. Implementing this approach on the set of data analyzed by [26], the same weights are obtained. For further information, the reader is referred to [27] and [25].

5 Simulation

For a simulation of the developed methodology, we mimic an experiment that was conducted at the Research Triangle Park Institute under a contract to the National Toxicology Program. The experiment was designed to assess the development toxicity of the chemical ethylene glycol [28] and consisted of a control and three non-zero dose levels. Molenbergh and Ryan [13] used the same data to illustrate their model for multiple binary data. They considered the incidence of the following trinomial responses:

1.Malformation
2.Low fetal weight
3.Normal

for each fetus. The fetal weight was dichotomized and classified as low if it was below or equal to 0.75g, otherwise normal. The study conducted by [11] gave the means and standard deviations of the litter sizes along with the proportions of malformation.

For each dose group 0, 0.75, 1.5 & 3.0 g/kg, we simulated litter sizes using the empirical distribution associated with the means and standard deviations estimated by [11]. To generate the probabilities of response within each litter, we used the values displayed in Table 1 of [13] and the Dirichlet distribution. The individual pup-specific data are then determined via generating n_{ij} Trinomial random variables using the simulated probabilities of response.

Table 1 presents the numbers of (a) malformed, (b) low weight, and (c) normal outcomes generated for each dose group. To implement the adopted Bayesian Model Averaging technique, pairs of five dose-response models have been used to estimate the mean functions $\mu_{ir}(d)$ for $i = 1,\ldots,4$ and $r = 1,2$. These models are:

1.Logistic

$$P_1(d) = \{1 + exp[-(a_{1r} + b_{1r}d)]\}^{-1} \tag{10}$$

2.Probit

$$P_2(d) = \phi(a_{2r} + b_{2r}d) \tag{11}$$

where ϕ represents the commulative normal distribution.

3.Gamma

$$P_3(d) = c_{3r} + (1 - c_{3r})\Gamma(b_{3r}d, a_{3r}) \tag{12}$$

where Γ is the commulative gamma distribution, $a_{3r}, b_{3r} > 0$ and $0 < c_{3r} < 1$.

4.Quantal Quadratic:

$$P_4(d) = 1 - exp(-(a_{4r} + b_{4r}d^2)) \tag{13}$$

5.Weibull

$$P_5(d) = 1 - exp(-(a_{5r} + b_{5r}d_{5r}^h)) \tag{14}$$

To fit the response models, equation (1) and the reparameterization equations (6)-(8) are used during the application of MCMC Metropolis-Hasting sampling within a Bayesian framework.

Fig. 1 displays the negative log likelihood values of the fitted pairs of dose-response models for each simulated data set. As demonstrated in [5], each model minimum value of the log likelihood is calculated by summing up the individual log likelihood estimates obtained at each dose.

Table 2 displays the corresponding BIC values and the associated model weights. The results in Table 2 reveal that the quantal quadratic model accounts for more than 58% of the weights followed by the probit (21.1%) and logistic (19.9%) models.

The BMA estimates of $P(0)$, overall risk $P(d)$, and excess risk $\pi(d)$ are obtained using the weighted averages of $\mu_{ir}(d)$ for $r = 1\&2$ for all fitted models through (4). Using parameter estimates for the control dose $(d = 0)$ models, the risk at some small doses such as $d = 0.01, 0.02, 0.05$ and 0.1 is derived. Furthermore, by considering 1000 MCMC replications, as demonstrated in Fig. 2, the associated variances and 95% upper confidence values of risk are determined. Table 3 summarizes the results. Expectedly, the later table confirms that the overall risk increases with dose level.

Moreover, we estimated the benchmark dose (BMD) for the following excess risk values: 0.01, 0.05, and 0.1 over the control using equations (4) & (5). The BMDL is derived via exploiting the asymptotic normality of the BMD estimates associated with the 1000 MCMC replications and constructing 95% lower confidence limit. Table 4 displays the approximated BMDs, their standard deviations, and the BMDLs at the different excess risk values. Obviously, the BMD and BMDL values increase as the added risk is enlarged.

Table 1: The generated data with dams = 25, 24, 22, and 23, respectively.

Dose = 0	Dose = 0.75	Dose = 1.5	Dose = 3.0
(a) (b) (c)	(a) (b) (c)	(a) (b) (c)	(a) (b) (c)
0 0 9	0 0 11	2 2 11	1 6 2
0 0 8	3 4 6	3 3 5	1 4 1
0 2 10	0 0 9	6 3 4	4 1 2
0 1 8	0 3 11	2 5 0	4 6 4
0 0 14	0 3 10	2 2 4	2 3 3
0 0 13	1 0 8	3 4 4	3 4 2
0 0 13	0 2 7	2 1 13	0 3 4
0 0 15	0 0 9	2 2 2	2 7 0
0 0 15	0 0 9	5 5 5	4 5 1
0 0 9	0 0 14	0 0 6	3 5 2
0 0 16	0 0 10	2 7 2	1 6 2
0 0 11	3 1 7	2 5 6	6 2 2
0 0 11	1 2 10	4 2 4	3 2 5
0 0 14	2 1 11	5 2 5	5 5 3
0 0 13	0 0 13	2 2 7	3 4 4
0 0 11	2 2 7	4 2 2	1 5 5
0 0 11	2 2 8	3 3 7	4 7 0
0 1 10	0 2 10	3 0 5	2 6 4
0 0 9	1 1 10	0 2 4	3 3 2
0 1 9	0 2 12	1 1 9	2 6 2
0 0 12	5 1 8	1 2 3	4 4 5
0 0 15	2 0 6	3 5 3	3 5 1
0 0 12	1 3 10		4 5 0
0 1 7	2 1 7		
0 0 15			

Table 2: BIC values and the corresponding weights of the fitted dose-response models for the simulated data set.

Model	BIC	Weight
Logistic	2084.149	0.1986526
Probit	2084.032	0.2106204
Gamma	2093.441	0.001906841
Quantal Quadratic	2081.992	0.5841498
Weibull	2091.649	0.004670414

Table 3: The estimated overall risk at some selected low dose levels and the corresponding 95% confidence values.

Dose(d)	P(d)	P(d) 95% upper confidence value
0.01	0.2495788	0.2506823
0.02	0.2518427	0.2529466
0.05	0.2600474	0.2611539
0.1	0.2783115	0.2794271

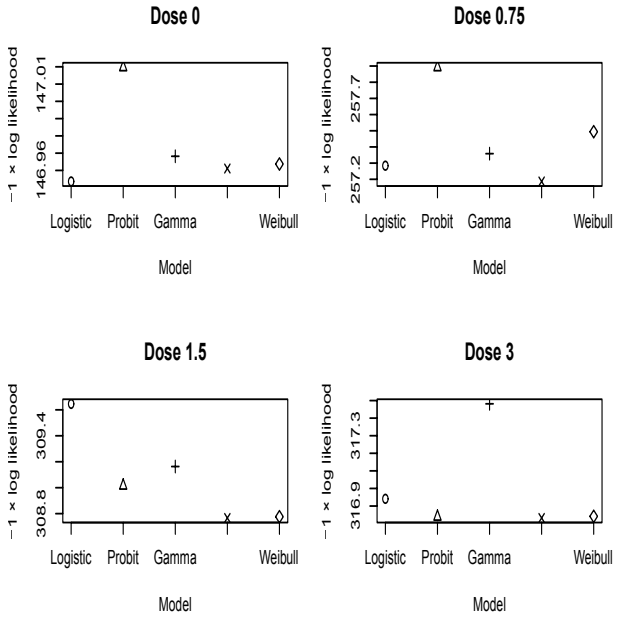

Fig. 1: The $-log$ likelihood values associated with the MCMC estimates for the fitted models of the simulated datasets at the different dose levels.

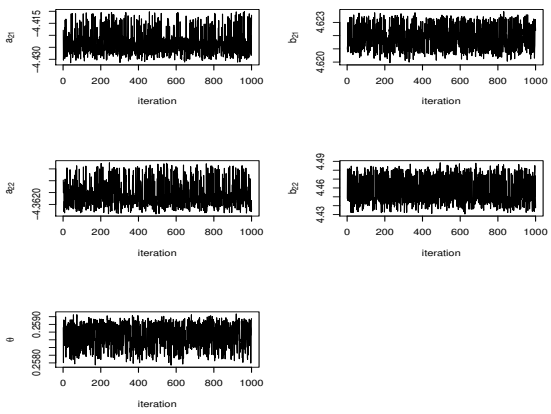

Fig. 2: Trace plots of pairs of the Probit model parameters with the associated θ estimated values derived from MCMC long run of 10000 iterations generated for the simulated dataset at dose level 0.75. The posterior mean estimates with the corresponding standard deviations in parentheses are: $a_{21} = -4.423(0.005), b_{21} = 4.622(0.0009), a_{22} = -4.362(0.0005), b_{22} = 4.458(0.016)$, and $\theta = 0.259(0.003)$.

Illustrative experimental application

In this section, we illustrate our BMA methodology for multiple outcomes using data obtained from [5]. The experiment was conducted to assess the development effects resulting from exposure to hydroxyurea. It consisted of a control group (0 dose) and three treatment groups classified as: low dose (150), medium dose (200), and high dose (250). The hydroxyurea data in each treatment group include the number of implementation litters, dead/resorbed, malformed, and total number of

Table 4: BMA estimates for BMD with their corresponding standard deviations in parentheses and BMDL at different excess risk values.

Excess risk	0.01	0.05	0.1
$\widehat{BMD}$	0.04155 (0.0001290994)	0.14135 (0.0030335410)	0.22675 (0.0063510702)
$\widehat{BMDL}$	0.04133763	0.1363598	0.2163025

fetuses. No data about the control group are made available. Therefore, only the non-zero dose data set will be used in this study for the purpose of demonstration.

[8] fitted a pair of logistic models for this set of dead/resorbed and malformation data. In this research, pairs of five response models given in equations (10-14) are fitted using MCMC techniques and a Bayesian approach. The resulting parameter estimates and the corresponding log-likelihood values are presented in Table 5. The weight of each model is determined according to the BIC criterion as described earlier. Table 6 displays the associated BIC values and model weights. Fig. 3 and Table 7 present the averaged dose-response models μ_{ir} and the associated risks with each dose level, respectively.

Because no data are available for the control dose ($d = 0$) and the associated excess risk with the low dose is very small (< 0.005), as revealed in Table 7, the estimated low dose model is used as a control model.

The approximated benchmark doses (BMDs) for excess risk values of 0.05 and 0.1 over the background are 191.8 and 228.8, respectively.

Using the response model parameter estimates of 1000 MCMC replications, estimates of BMD mean, standard deviation and BMDL at the selected excess risk values are obtained and displayed in Table 8.

The results shown in Table 5 reveal that estimates of the reparameterization parameter, θ_i, for all three treatment doses are between 0.2 and 0.27. This result matches the expectations of experts in fitting teratological data for example, see [5]. The model weights listed in Table 6 indicate that the Logistic model has the highest account of weights (44.2%), which is in line with [8] choice of dose-response model that best fits the hydroxyurea data. Fig.3 reveals that the averaged dead/resorbed response increases, while averaged malformation response decreases with higher dose levels because the exposure to elevated levels of hydroxyurea

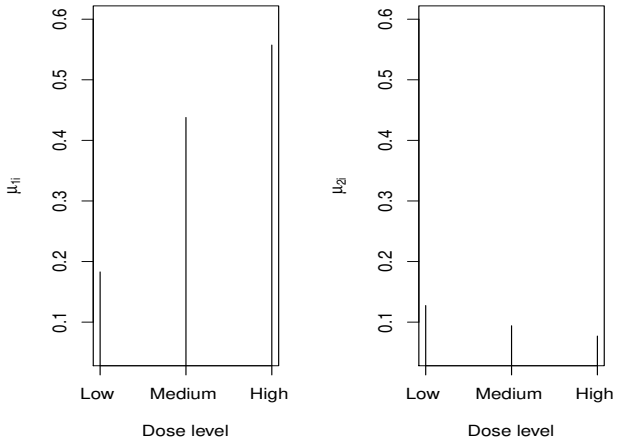

Fig. 3: The averaged dose-response models for hydroxyurea data.

minimizes the total number of living fetuses and consequently fewer malformed fetuses remain. Moreover, the $P(0)$ estimates in Table 7 give an evidence of similar initial overall risk of about 0.28 regardless of the associated dose level. As expected, the BMD and BMDL values increase with increments of added risk.

6 Conclusion

Concern over the protection and health of a developing embryo has been grown rapidly over the last several decades. Bioassay experiments on laboratory animals have been designed and conducted to assess the potential developmental risks in unborn children as result of maternal exposure to environmental and industrial toxins. Several statistical models have been developed to provide a mathematical framework for the estimation of risk when the mother is exposed during a critical time of pregnancy. These models are structured to facilitate extrapolation from high experimental doses to low human exposure levels. Many models assume a dichotomous response from developmental toxicity experiments, but some are more realistic and allow for multiple outcomes. One approach for estimating a threshold for safe exposure level that has found widespread popularity is the benchmark dose (BMD) methodology. The advantage of the benchmark dose method is that it can be applied to both discrete and continuous outcomes. However, the method is known to be model dependent and can vary substantially based on the assumed dose-response relationship. To reduce this model dependence, a methodology based on model averaging has been proposed and successfully applied. Several potential

Table 5: The log-likelihood values, L, parameter estimates for pairs of dose-response models and estimated θ_i values for $i = 1,2,3$.

Model	Parameter	Low dose	Medium dose	High Dose
Logistic	a_{11}	0.02390101	-0.7904562	0.1364078
	b_{11}	-0.9703013	0.2639519	0.02887366
	a_{12}	-2.217992	-0.9504916	-2.474379
	b_{12}	0.2426339	-0.6432312	0.004821484
	θ_i	0.2613715	0.2181258	0.2438043
	L	-224.7381	-260.6348	-562.2774
Probit	a_{21}	-1.406183	-1.675734	-1.6712727
	b_{21}	0.1976464	0.750125	0.7320323
	a_{22}	-2.409988	-2.646601	-1.9864809
	b_{22}	0.7179207	0.6607857	0.2319488
	θ_i	0.2613632	0.2183735	0.2433926
	L	-226.1509	-260.6872	-562.0023
Gamma	a_{31}	1.786545	1.580096	1.353346
	b_{31}	0.2767687	0.3327715	0.2280071
	c_{31}	0.08916573	0.2373231	0.3686591
	a_{32}	1.735629	1.796754	3.177613
	b_{32}	0.2391621	0.08086992	0.2340125
	c_{32}	0.0444897	0.05983448	0.06160946
	θ_i	0.2611724	0.2182088	0.2433884
	L	-225.144	-261.5022	-562.7745
Quantal Quadratic	a_{41}	0.0001952265	0.02579366	0.003313833
	b_{41}	0.09258587	0.1400943	0.1312673
	a_{42}	0.04403493	0.04639637	0.01963903
	b_{42}	0.04262925	0.01255803	0.009270703
	θ_i	0.2613239	0.2186908	0.2439582
	L	-224.7639	-260.6143	-562.3277
Weibull	a_{51}	0.06147006	0.1478446	0.1600666
	b_{51}	0.09814872	0.2103591	0.2607989
	h_{51}	0.9883212	1.054518	0.0320832
	a_{52}	0.03247388	0.03412807	0.02300154
	b_{52}	0.07440968	0.03174825	1.017156
	h_{52}	0.9847963	0.9409696	0.78945
	θ_i	0.2614004	0.2180075	0.2439062
	L	-224.7299	-260.6272	-562.1716

Table 6: BIC values and the corresponding weights of the fitted dose-response models for hydroxyurea data.

Model	BIC	Weight
Logistic	2117.800	0.4415892
Probit	2120.180	0.1343274
Gamma	2130.340	0.0008354126
Quantal Quadratic	2117.911	0.417707
Weibull	2126.556	0.005540975

models are used and a weighted average of the BMDs is determined. We have demonstrated the application of the model averaging technique in developmental toxicity experiments with multiple outcomes. The Bayesian model averaging (BMA) procedure was used to estimate the weights. Both the simulation study and the illustrative example showed that the method works well and can successfully be applied. In fact, it is encouraging to note that the logistic and the quantal quadratic models appear to pick up most of the weights in the averaged model.

The traditional approach for animal models in testing of developmental effects have relied on experimenting with mammals especially mice and rats. However, the application of high throughput models using zebrafish and other creative systems such as limb bud culture, hydra assay and whole embryo culture were investigated as alternatives to the traditional methods for cost reduction and acceleration of the testing process. Although these methods are rapidly evolving, the validation and their application in teratogenic testing is still in an early stage. No statistical models have been developed for risk assessment. For a description of these new approaches, we refer to [29] and [30]. The aim of this work is to demonstrate that the powerful BMA technique can produce reliable results in developmental toxicity experiments.

Table 7: BMA estimates of $P(0)$, overall risk $P(d)$, and excess risk $\pi(d)$ obtained using MCMC parameter estimates of dose-response models associated with the three dose groups.

Dose Level	$P(d)$	$P(0)$	$\pi(d)$
Low	0.2870439	0.2825706	0.0044733
Medium	0.4907544	0.2762921	0.2144623
High	0.5914754	0.2792948	0.3121806

Table 8: BMA estimates for BMD with their corresponding standard deviations in parentheses and associated BMDL at different excess risk values.

Excess risk	0.05	0.1
$\widehat{BMD}$	180.3778 (0.3032146)	216.0361 (0.6183924)
$\widehat{BMDL}$	179.879	215.0188

References

[1] K. Rai, and J. Van Ryzin, A Generalized multi-hit dose-response model for low-dose extrapolation, Biometrics **39**, 341- 352, (1981).

[2] J. J. Chen, and R. L. Kodell, Quantitative risk assessment for terotological effects, Journal of the American statistical Association **84**, 966-971 (1989).

[3] L. M. Ryan, Quantitative risk assessment for developmental toxicity, Biometrics **48**, 163-174 (1992).

[4] E. Khorsheed, and M. Razzaghi, Bayesian Model Averaging for Risk Assessment in Developmental Toxicology, Applied Mathematics and Information Sciences **13(1)**, 1-10 (2019).

[5] J. J. Chen, R. L. Kodell, R. B. Howe, and D. W. Gaylor, Analysis of trinomial responses from reproductive and developmental toxicity experiment, Biometerics **47**, 1049-1058 (1991).

[6] P. J. Catalano, D. O. Scharfstein, L. M. Ryan, C. A. Kimmel, and G. I. Kimmel, Statistical model for fetal death, fetal

weight, and malformation in developmental toxicity studies, Teratology **47**, 281-290 (1993).
[7] D. Krewski, and Y. Zhu, Applications of multinomial dose-response models in developmental toxicity risk assessment, Risk Analysis **14**, 613-627 (1994).
[8] J. J. Chen, and Lung-An Li, Dose-response modeling of trinomial responses from developmental experiments, Statistica Sinica **4**, 265-274 (1994).
[9] P. J. Catalano, L. M. Ryan, and D. Scharfstein, Modeling fetal death and malformation in developmental toxicity studies, Risk Analysis **14**, 629-637 (1994).
[10] P. J. Catalano, and L. M. Ryan, Bivariate latent variable models for clustered discrete and continuous outcomes, Journal of the American Statistical association **87**, 651-658 (1992).
[11] M. Regan, and P. J. Catalano, Likelihood models for clustered binary and continuous outcome: application to developmental toxicology, Biometrics **55**, 760-768 (1999).
[12] J. S. Najita, and P. J. Catalano, On determining the BMD from multiple outcomes in developmental toxicity studies when one outcome is intentionally missing, Risk Analysis **33(8)**, 1500-1509 (2013).
[13] G. Molenbergs, and L. M. Ryan, An exponential family model for clustered multivariate binary data, Environmetrics **10**, 279-300 (1999).
[14] D. A. Williams, The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, Biometrics **31**, 949-952 (1975).
[15] K. -Y. Liang, and J. Hanfelt, On the use of quasi-likelihood method in tertological experiments, Biometrics **50**, 872-880 (1994).
[16] N. L. Johnson, and S. Kotz, Discrete Distributions, Wiley, New York, (1969).
[17] D. W. Gaylor, Dose-response modeling, In Developmental Toxicology, 2nd edition, C. A. Kimmel, and J. Buelke-Sam (Eds), Raven Press, New York, (1994).
[18] P. McCullagh, and J. A. Nelder, Generalized Linear Models, Chapman and Hall, CRC Press, London, (1989).
[19] H. Moon, K. Hyun-Joo, J. J. Chen, and R. L. Kodell, Model averaging using the Kullback information criterion in estimating effective doses for microbial infection and illness, Risk Analysis **25**, 1147-1159 (2005).
[20] M. Whitney, and L. M. Ryan, Quantifying dose-response uncertainty using Bayesian model averaging, In: Uncertainty Modeling in Dose-Response, R.M. Cooke (ed.). Wiley, N. J. Hoboken, 165-179 (2009).
[21] A. E. Raftery, Bayesian model selection in social research, Sociological Methodology **25**, 111-163 (1995).
[22] L. Wasserman, Bayesian model selection and model averaging, Journal of Mathematical Psychology **44**, 92-107 (2000).
[23] A. J. Bailer, R. B. Noble, and M. W. Wheeler, Model uncertainty and risk estimation for experimental studies of quantal responses. Risk Analysis **25**, 291-299 (2005).
[24] S. J. Simmons, C. Chen, X. Li, Y. Wang, W. W. Piegorsch, Q. Fang, B. Hu, and G. E. Dunn, Bayesian Model Averaging for benchmark dose estimation. Environmental and Ecological Statistics **22**, 5-16 (2015).
[25] K. P. Burnham, and D. R. Anderson, Model selection and multimodel inference: a practical information-theoretic approach, 2nd edition, Springer, New York, (2002).
[26] M. W. Wheeler, and M. J. Bailer, Properties of model-averaged BMDLs: A study of model averaging in dichotomous response risk estimation, Risk Analysis **27(3)**,659-670 (2007).
[27] D. R. Anderson, Model based inference in the life sciences: a practical information- theoretic approach, 2nd edition, Springer, New York, (2008).
[28] C. J. Price, C. A. Kimmel, R. W. Tyl, and M. C. Marr, The developmental toxicity of ethylene glycol in rats and mice, Toxicology and Applied Pharmacology **81**, 113-127 (1985).
[29] A. H. Piersma, Alternative methods for developmental toxicity testing, Basic and Clinical Pharmacology and Toxicology **98**, 427-431 (2006).
[30] J. M. DeSesso, Future of developmental toxicity testing, Current opinion in toxicology **3**, 1-5 (2017).

Eman Khorsheed received her PhD degree in Statistics from University of Bath, UK. She is an Assistant Professor at University of Bahrain- Department of Mathematics. In 2011, she became a Fellow of the Higher Education Academy (HEA), UK. Her research interests are in the areas of applied Bayesian Statistics including spatial, spatio-temporal and hierarchical models, Markov chain Monte Carlo (MCMC) techniques, image analysis, Tomography, Demography, Biostatistics, Data Analytics, and Operations Research. She has published research articles in reputed international journals and is referee and editor of several statistical journals. She is a member of the American Statistical Association and also the International Statistical Institute.

Mehdi Razzaghi is a Professor of Statistics at Bloomsburg University and has served in this capacity for 32 years. He holds a BS in mathematics from the University of Sussex in England, a BS in Computer Science from Bloomsburg University and a PhD in statistics from the University of London, England which he received in 1977. He held faculty positions at Marburg University in Germany, University of Kentucky, and California State University Chico, before joining Bloomsburg University faculty in 1987. Prof. Razzaghi is the recipient of the Faculty Recognition Award (2003) and the Outstanding Scholarship Award (2012) from Bloomsburg University. During the 2014-2015 academic year, he was a Fulbright fellow at the University Warsaw in Poland. In 2018, he served as a Fulbright Specialist at the Medical University of Silesia in Poland. Prof. Razzaghi has collaborated extensively with the Food and Drug Administration (FDA) as a Fellow of the National Center for Toxicological Research. In his role, he assisted in the mathematical development of dose-response models and statistical risk assessment procedures in animal bioassay experiments. He further studied methods for extrapolation of the procedures for human exposure to toxic chemicals in the environment. Prof. Razzaghi has also served as a consultant with the US Environmental Protection Agency (EPA) where he was a member of the peer review panel on the effects of perchlorate environmental contamination. Further, he has consulted with the National Institutes for Health (NIH) and Geisinger Medical Center. He has been the recipient of grants from the NIH and International Life Science Institute. Professional affiliations include membership with the American Statistical Association, Society for Risk Analysis, American Mathematical Society, and International Biometric Society, among others. He is also a Fellow of the Royal Statistical Society in England.