

# Bootstrap Estimation of Benchmark Doses and Confidence Limits with Clustered Quantal Data

Yiliang Zhu,<sup>1</sup> Tao Wang, and Jenny Z.H. Jelsovsky

Department of Epidemiology and Biostatistics

University of South Florida

Last revised: November 1, 2006

<sup>1</sup>Address correspondence to Yiliang Zhu, Department of Epidemiology and Biostatistics, University of South Florida, 13201 Bruce B Downs Blvd, Tampa, FL 33612, email: yzhu@hsc.usf.edu. This material is based upon work supported in part by the National Science Foundation under Grant No.DMS9978370 to Zhu. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Science Foundation. The authors thank two anonymous referees for their thoughtful comments and suggestions on this article.

## Abstract

The benchmark dose (BMD) is an exposure level which would induce a small risk increase (BMR level) above the background. The BMD approach to deriving a reference dose for risk assessment of non-cancer effects is advantageous in that the estimate of BMD is not restricted to experimental doses and utilizes most available dose-response information. To quantify statistical uncertainty of a BMD estimate, we often calculate and report its lower confidence limit (i.e. BMDL), and may even consider it as a more conservative alternative to BMD itself. Computation of BMDL may involve normal confidence limits to BMD in conjunction with the delta-method. Therefore factors such as small sample size and nonlinearity of in model parameters can affect the performance of the delta-method BMDL, and alternative methods are useful. In this paper, we propose a bootstrap method to estimate BMDL utilizing a scheme that consists of a re-sampling of residuals after model fitting and a one-step formula for parameter estimation. We illustrate the method with clustered binary data from developmental toxicity experiments. Our analysis shows that with moderately elevated dose-response data, the distribution of BMD estimator tends to be left-skewed and bootstrap BMDLs are smaller than the delta-method BMDLs on average, hence quantifying risk more conservatively. Statistically the bootstrap BMDL quantifies the uncertainty of the true BMD more honestly than the delta-method BMDL as its coverage probability is closer to the nominal level than that of delta-method BMDL. We find that BMD and BMDL estimates are generally insensitive to model choices provided that the models fit the data comparably well near the region of BMD. Our analysis also suggests that, in the presence of a significant and moderately strong dose-response relationship, the developmental toxicity experiments under the standard protocol support dose-response assessment at 5% BMR for BMD and 95% confidence level for BMDL.

KEY WORDS: Benchmark dose; coverage probability; dose-response; generalized estimating equations; over-dispersed binary data; risk assessment.

# 1 Introduction

Dose-response assessment of non-cancer effects has largely relied on the paradigm of a reference dose (RfD) derived from a no-observed-adverse-effect-level (NOAEL). The NOAEL is the highest level of experimental exposure at which there is no statistically significant increase in risk compared with the spontaneous risk of a control group. Dividing safety (uncertainty) factors into a NOAEL yields an RfD. RfD is an estimate, with uncertainty of an order of magnitude, of a daily exposure to the human population that is likely to be without appreciable risk of adverse health effects during a lifetime (Barnes and Dourson, 1988). When a NOAEL cannot be identified, a lowest-observed-adverse-effect-level (LOAEL) is used instead, at which there is a significant increase in risk, and an additional order of 10 is included in the safety factor.

Restricted to be one of the experimental doses, a NOAEL does neither identify a zero-risk dose nor fully account for the underlying dose-response relationship. The associated risk is often non-negligible, varies widely with the experiments, and is sensitive to experiment (e.g. sample size) (Gaylor, 1992; Leisenring and Ryan, 1992). As a result, unacceptably high risks could occur at the RfD level even after applying the safety factors to the NOAEL. In view of these limitations, the use of benchmark doses (BMD) as an alternative to NOAEL has been recommended for non-cancer endpoints (Crump, 1984; Barnes *et al.*, 1995; EPA, 1991; 1998; 2000; 2003).

A BMD is the effective dose (ED) corresponding to a small increase of risk (i.e. BMR level), typically between 1% and 10% in excess of the background level. The BMD is estimated from the dose-response model fitted to data through appropriate statistical procedures. The estimator,  $\widehat{\text{BMD}}$ , is subject to data variation as well as model uncertainty. Thus a lower confidence limit for the BMD, i.e., BMDL, is computed to quantify the statistical variation towards the conservative end of health protection.

Estimation of BMD requires judgements on the adequacy of data and model, the choice of

statistical procedures, and the choice appropriate BMR level. The computation of BMDL typically requires further statistical approximation because exact solution to confidence intervals is usually unavailable.

EPA (2000) describes two common approaches to BMDL. One is likelihood-based and the other utilizes normal approximation to the distribution of  $\widehat{\text{BMD}}$  or a quantity that is a function of the dose-response model. The likelihood-based approach is restricted to where the likelihood function is available and the dose-response model can be parameterized by the BMD. Otherwise it is common to use normal-approximation confidence intervals, which require the estimation of the standard error of the statistical estimators under consideration. The delta-method is often used to estimate the standard error (USEPA, 2000). As a function of the model parameters, the estimator  $\widehat{\text{BMD}}$  is asymptotically normal when the estimators of the model parameters are asymptotically normal. This is the case for MLE or GEE estimators under certain regularity conditions (Liang and Zeger, 1986; Zhu et al, 1994; Xie and Yang, 2003). The performance of the normal approximation and delta method depends on several factors, particularly sample size and the degree of nonlinearity of the estimators in the model parameters. Biased or unprecise results may arise when the requirements are unmet. In view of the potential limitations of these common approaches to BMDL, alternative methods such as bootstrap are useful. Among many authors, for example, Budtz-Jørgensen *et al.* (2001) considered a parametric bootstrap of K-power model for a continuous response variable, in which the bootstrap residuals are simulated from a normal distribution. Bailer and Smith (1994) considered bootstrap upper confidence limit on risk of multi-stage tumor models. Moerbeek *et al.* (2004) illustrate likelihood-based, delta-method, and bootstrap confidence intervals for BMD with both continuous and quantal data; but they give examples and do not compare the performance of these methods.

In this paper, we propose a bootstrap method for computing BMDL utilizing the scheme of residuals sampling and a one-step estimation of the model parameters. We describe the methods in section 2, including modeling, estimation of BMD the delta-method, and the

bootstrap procedure. In section 3, we illustrate the methods through clustered quantal data arising from four developmental toxicity experiments. We compare in section 4 the coverage probabilities of the bootstrap and delta-method BMDLs through a simulation study. We present our conclusions as well as recommendations in section 5. The examples suggest that the bootstrap BMD is more accurate than to the delta-method BMD in terms of coverage probability, and is less sensitive to non-symmetric distribution of  $\widehat{\text{BMD}}$ .

## 2 Methods

### 2.1 Dose-Response Models

The first step in computing a BMD is to fit a dose-response model to data. Suppose  $Y^*$  is an outcome measure and  $\mathcal{A}$  is a collection of  $Y^*$  that represents undesirable or adverse outcomes. With a continuous outcome such as birth weight,  $\mathcal{A} = \{Y^* < y_\gamma^*\}$  consists of all values below  $\gamma$ -percentile of a reference population. This allows for dichotomization of the outcome  $Y$  into quantal  $Y = 1$  for  $Y^* \in \mathcal{A}$  and  $Y = 0$  for  $Y^* \notin \mathcal{A}$ . The risk of adverse outcome at exposure level  $d$  is the probability

$$\pi(d) = \text{Prob}(Y^* \in \mathcal{A}|d),$$

which also defines the dose-response relationship. In the absence of a mechanistic model, mathematical functions can be used to describe  $\pi(d)$ . We consider the following three popular dose-response models, the Weibull model,

$$\pi(d) = 1 - \exp(-\theta_1 - \theta_2 d^{\theta_3}),$$

and an extension of probit model,

$$\pi(d) = \Phi(\theta_1 + \theta_2 d^{\theta_3}),$$

and logit model,

$$\pi(d) = \frac{1}{1 + \exp(-\theta_1 - \theta_2 d^{\theta_3})}.$$

In these models, the parameters are constrained to ensure that  $0 < \pi(d) < 1$ . In particular,  $\theta_1$  characterizes the background risk, and  $\theta_2$  is the dose-slope, the power parameter  $\theta_3 > 0$ . The introduction of  $\theta_3$  to the probit and logit models affords an enhanced sigmoidal shape of the dose-response. The Weibull model is well known for its flexibility in approximating a possible threshold phenomenon. Ryan (1992) and Krewski and Zhu (1994) applied these models to binary developmental toxicity data. These three models affords an opportunity to investigate sensitivity of BMD estimate to model choice.

The illustration in this paper will focus on quantal outcomes of fetal death or any malformation in a live birth arising from developmental toxicity experiments. We use  $Y_{ijk}$  ( $i = 1, \dots, D; j = 1, \dots, N_i; k = 1, \dots, m_{ij}$ ) to denote the response of the  $k^{th}$  implant (fetus) from the  $j^{th}$  dam (litter) in the  $i^{th}$  dose group:  $Y_{ijk} = 1$  if the adverse event is present, and  $Y_{ijk} = 0$  otherwise. The total number of adverse events in a litter,

$$Y_{ij} = \sum_{k=1}^{m_{ij}} Y_{ijk}$$

has its mean

$$\mu_{ij} = E(Y_{ij}|m_{ij}) = m_{ij}\pi(d_i)$$

and variance

$$V_{ij} = \text{var}(Y_{ij}|m_{ij}) = m_{ij}\{1 + (m_{ij} - 1)\rho_i\}\pi(d_i)(1 - \pi(d_i)).$$

The intra-litter correlation  $\rho_i = \text{corr}(Y_{ijk}, Y_{ijl})$  ( $k \neq l$ ) characterizes the litter effects that littermates tend to respond more similarly because of biological similarity and the same growth environment. This correlation can be made dose-dependent to capture potential variation due to exposure (Krewski and Zhu, 1994).

## 2.2 Model Fitting and Generalized Estimating Equations

When it is difficult or uncertain to specify a full distribution for complex data such as the clustered quantal data, generalized estimating equations (GEEs)(Liang and Zeger, 1986) are attractive because they require specification of only the first few moments and the results are less sensitive to departure from distribution assumptions otherwise imposed. Many authors have adopted GEEs (Ryan, 1992; Zhu *et al.*, 1994, Krewski and Zhu, 1994) within the context of developmental toxicity data.

The model parameter  $\theta$  in  $\pi(d)$  are estimated from solving the following set of equations

$$\sum_{i=1}^D \sum_{j=1}^{N_i} m_{ij} \frac{\partial \pi(d_i)}{\partial \theta} V_{ij}^{-1} (y_{ij} - m_{ij} \pi(d_i)) = 0. \quad (1)$$

An additional set of equations are used to estimate the intra-litter correlations,  $\rho_i$ ,

$$\sum_{j=1}^{N_i} \frac{\partial \nu_{ij}}{\partial \rho_i} W_{ij}^{-1} (y_{ij}^2 - \nu_{ij}) = 0, (i = 1, \dots, D) \quad (2)$$

where

$$\nu_{ij} = E(Y_{ij}^2 | m_{ij}) = (m_{ij} \pi(d_i))^2 + V_{ij},$$

and

$$W_{ij} = \text{var}(Y_{ij}^2 | m_{ij}) = E(Y_{ij}^4 | m_{ij}) - \nu_{ij}.$$

The use of  $y_{ij}^2$  as outcome in equations (??) requires the 3rd and 4th moments of  $Y_{ijk}$ , i.e. the joint probability of adverse outcomes of 3 or 4 fetuses in the same cluster. Because this joint probability is complex, approximations seem practical and necessary, and good approximations can improve statistical efficiency in estimating  $\rho_i$ . The approximation below relies on marginal pairwise correlation  $\rho = \rho^{(0)} = \text{corr}(Y_{ij1}, Y_{ij2})$  in replace of the conditional pairwise correlation

$$\rho^{(l)} = \text{corr}(Y_{ijl+1}, Y_{ijl+2} | Y_{ij1} = \dots = Y_{ijl} = 1)$$

between the  $(l + 1)$ th and  $(l + 2)$ th fetuses, conditioning on the first  $l$  fetuses being adverse in the same cluster. Using the notation

$$\pi^{(0)} = \pi$$

and

$$\pi^{(l)} = Pr(Y_{ijl+1} = 1 | Y_{ij1} = \dots = Y_{ijl} = 1),$$

and assuming an exchangeable joint probability, we can show that

$$Pr(Y_1 = \dots = Y_l = Y_{l+1} = 1) = \pi \pi^{(1)} \dots \pi^{(l)},$$

in which

$$\begin{aligned} \pi^{(l)} &= \pi^{(l-1)} + \rho^{(l-1)}(1 - \pi^{(l-1)}) \\ &= \pi + \{\rho + \rho^{(1)}(1 - \rho) + \rho^{(2)}(1 - \rho^{(1)})(1 - \rho) + \dots \\ &\quad + \rho^{(l-1)}(1 - \rho^{(l-2)}) \dots (1 - \rho)\}(1 - \pi). \end{aligned}$$

Replacing  $\rho^{(l)}$  ( $l = 1, 2$ ) with  $\rho$  results in

$$Pr(Y_{ijl+1} = 1 | Y_{ij1} = \dots = Y_{ijl} = 1) = \pi + (1 - (1 - \rho)^l)(1 - \pi)$$

This last equation affords a simplification of  $W_{ij}$  in the GEEs given in equation (??) which is originally used in Zhu et al. (1994).

Iteratively solving the two sets of equations until convergence yields the estimates for  $(\theta, \rho_i)$ . In particular, we can update the estimate of  $\theta$  by applying the Newton-Raphson algorithm to equations (??) while fixing  $\rho_i$ :

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + (G^T G)^{-1} G^T e, \tag{3}$$

where

$$G = V^{-1/2} \frac{\partial \mu}{\partial \theta^T},$$

$\mu = (\mu_{11}, \dots, \mu_{Dn_D})^T$ ,  $V = \text{diag}(v_{ij})$ , and  $e = (e_{11}, \dots, e_{Dn_D})^T$  with elements being a standardized Pearson's residual  $e_{ij} = (y_{ij} - \mu_{ij})/\sqrt{v_{ij}}$  ( $i = 1, \dots, D; j = 1, \dots, n_i$ ) or a deviance residual. The term  $G$  and  $e$  are both evaluated at the previous stage estimate  $\hat{\theta}^{(n)}$ . The final estimate  $\hat{\theta}$  is obtained when the algorithm converges, i.e., when  $\hat{\theta}^{(n+1)}$  and  $\hat{\theta}^{(n)}$  are sufficiently close. This iterative approximation formula is the foundation for a bootstrap sampling of residuals in section 2.4. The standard error of the estimate can be obtained using the sandwich estimates (Liang and Zeger, 1986) or the theoretical asymptotic variance (Zhu *et al.*, 1994).

## 2.3 Benchmark Doses and Lower Confidence Limits

A BMD is the dose at which exposure would induce a small increase  $\delta$  (benchmark response or BMR level) in risk from the background risk  $\pi(0)$  (USEPA, 2000). The extra risk

$$\frac{\pi(\text{BMD}_\delta) - \pi(0)}{1 - \pi(0)} = \delta \tag{4}$$

corresponds to an increase of risk  $\delta 100\%$  among those who would be otherwise free of the adverse effects in the absence of exposure. Upon fitting the dose-response model  $\pi(d)$ , we solve equation (??) for  $\text{BMD}_\delta$ . It is generally required to have a reasonable degree of goodness-of-fit of the model especially in the neighborhood of the BMR. Because  $\widehat{\text{BMD}}$  is a statistical estimator and is subject to statistical variation, a lower confidence limit  $\text{BMDL}_{1-\alpha}$  at  $(1 - \alpha)100\%$  level is calculated to quantify the variation and give a lower bound estimate.

Likelihood-based confidence limits (see Morgan, 1992, page 65; Crump and Howe, 1985) can be utilized to compute  $\text{BMDL}_{1-\alpha}$  when a likelihood function is available and can be expressed using BMD as a parameter. An alternative is to calculate confidence intervals of

a relevant quantity (i.e. BMD or a function of the dose-response model) using normal approximation. While both approaches require large sample to make the normal approximation satisfactory, the second approach typically uses the delta-method to estimate the standard error of the estimator (see Gart *et al.*, 1986). For example, BMDL derived directly from the normal confidence limit for BMD is given by

$$\text{BMDL}_{1-\alpha} = \widehat{\text{BMD}}_{\delta} - z_{1-\alpha}\sigma(\widehat{\text{BMD}}_{\delta}).$$

Here  $z_{1-\alpha}$  is the  $(1 - \alpha)100$  percentile of the standard normal distribution. However, delta-method may produce inaccurate results when the sampling distribution for  $\widehat{\text{BMD}}$  is asymmetric, or even lead to a negative BMDL when the distribution is heavily right-skewed or the estimate of  $\sigma(\widehat{\text{BMD}}_{\delta})$  is large. To circumvent this limitation, one can compute the lower confidence limit for  $\log(\widehat{\text{BMD}}_{\delta})$  first and then convert the confidence limit back to its original scale:

$$\text{BMDL}_{1-\alpha} = \exp[\log(\widehat{\text{BMD}}_{\delta}) - z_{1-\alpha}\sigma(\log(\widehat{\text{BMD}}_{\delta}))].$$

This gives rise to the idea of liberalization through transformation in order to obtain more accurate estimate of the standard error for the estimator under consideration. Gaylor *et al.*(1998) suggest first apply the delta-method to obtain an upper confidence limit  $U$  for a monotone (e.g. increasing) function  $H$  of the fitted model  $\pi(d)$

$$U(d) = H(\pi(\hat{d})) + z_{1-\alpha}\sigma(H(\hat{\pi})), \tag{5}$$

then equating this upper confidence limit at  $\text{BMDL}_{1-\alpha}$  with  $H(\text{BMD}_{\delta})$

$$U(\text{BMDL}_{1-\alpha}) = H(\pi(\widehat{\text{BMD}}_{\delta})),$$

the resulting  $\text{BMDL}_{1-\alpha}$  is an approximate  $(1 - \alpha)100\%$  confidence limit for  $\text{BMD}_{\delta}$ . It is preferable for  $H$  to be nearly linear in the model parameters  $\theta$ . If the upper confidence limit

in (??) is poor, however,  $\text{BMDL}_{1-\alpha}$  as a solution to the preceding equation may not exist.

## 2.4 The Bootstrap Method

The bootstrap method (Efron and Tibshirani, 1993, Chapter 13) is an attractive alternative because it neither imposes a normal distribution to, nor requires an explicit expression of the standard error for the BMD estimator. It relies on re-sampling from the observed data to simulate the variation of BMD estimator within the observed data range. By re-sampling the original data  $B$  times, we obtain a sample of bootstrap BMD estimates,  $(\text{BMD}^1, \dots, \text{BMD}^B)$ , which depicts the sampling variation of the BMD estimator. The  $\alpha$ -percentile of this bootstrap sample of BMDs serves as an estimate of  $\text{BMDL}_{1-\alpha}$ .

To implement a bootstrap scheme, one can re-sample either the original data or the residuals after fitting the model to the original data. Re-sampling the original data dictates that the model be fit to each bootstrap sample. For nonlinear models the chance of failure in model fitting (e.g. non-convergence) is considerable (Moulton and Zeger, 1991) as we also have experienced. For this reason, we follow the approach of Moulton and Zeger (1991) to re-sample the residuals and develop a one-step procedure to produce bootstrap estimates  $\{\hat{\theta}^b, b = 1, \dots, B\}$ . Within the context of clustered data, clustering is in fact mostly contained in the residuals. Thus sampling of residuals retains the clustering nature of the data. The one-step procedure yields consistent estimator of the model parameters and subsequently consistent estimator of the bootstrap variance. Extreme bootstrap samples, however, may influence the variance estimator (Moulton and Zeger, 1991). The scheme consists of the following steps:

**Step 1.** Upon fitting the model to data, compute the standardized residuals

$$e_{ij} = (y_{ij} - m_{ij}\hat{\mu}_{ij})/\sqrt{v_{ij}} \quad (j = 1, \dots, n_i; i = 1, \dots, D).$$

The adjusted residuals  $e_{ij}^* = e_{ij}/(1 - h_{ij,ij})^{1/2}$  can be used as an alternative to account for

possible correlation between the residuals  $e_{ij}$ . Here,  $h_{ij,ij}$  is the diagonal element of the matrix

$$H = G(G^T G)^{-1} G^T.$$

**Step 2.** Randomly draw, with replacement, a sample of residuals:

$$e^b = (e_{11}^b, \dots, e_{ij}^b, \dots, e_{Dn_D}^b)^T.$$

When the standardized residuals are non-homogeneous (e.g. non-constant variance) across dose groups, bootstrap sampling stratified within each dose group is recommended.

**Step 3.** Compute the one-step bootstrap estimate of the model parameter using the equation

$$\hat{\theta}^b = \hat{\theta} + (G^T G)^{-1} G^T e^b, \tag{6}$$

where  $G$  and  $e^b$  are evaluated at original estimate  $\hat{\theta}$ . This equation is based on the iterative equation (??)

**Step 4.** Compute bootstrap estimate  $\widehat{\text{BMD}}_{\delta}^b$  using the model  $\pi(d)$  evaluated at  $\theta^b$ ;

**Step 5.** Repeat Steps 2-4  $B$  times to obtain a bootstrap sample of BMD estimates:  $(\text{BMD}_{\delta}^1, \dots, \text{BMD}_{\delta}^B)$ . Efron and Tibshirani (1993, pp. 275) recommend between 500 and 1000 replications. We choose  $B = 2000$  to account for the clustering effects seen in the developmental toxicity data. See our further discussion in section 3.4.

**Step 6.** Calculate the  $\alpha$ 100-percentile from the sample of bootstrap BMD estimates as the estimate of  $\text{BMDL}_{1-\alpha}$ .

## 3 Illustrations

### 3.1 Developmental Toxicity Experiments

To illustrate the bootstrap estimation of BMD and BMDL, we selected four data sets from a series of developmental toxicity experiments conducted under the National Toxicology Program. These data sets were analyzed previously by Krewski and Zhu (1994). The experiments included 3-4 dose groups in addition to a vehicle control, with between 20 to 30 pregnant dams in each group. The timed-pregnant animals were administered to the test agent during the period of organogenesis, beginning just after implantation. The effect on in-utero development was evaluated before parturition would normally occur. Measured outcomes include prenatal death among implants (data sets RBSM and MIEG), and any malformations among surviving fetus (data sets RTEG and RTSM). The summary statistics in Table 1 show that the incidence of the adverse effects in all four experiments was elevated by exposure. For example, the incidence of malformations in RTEG increased from a background rate of 4% to 68.6% at the highest dose of 5000 mg/kg.

(Table 1 about here)

### 3.2 Model Fitting

The logit, probit, and Weibull models were fit to each data set. The estimates of the model parameters are reported in Table 2. In all cases the fitted models demonstrate a positive dose-response with a statistically significant, positive slope  $\theta_2$ . For RBSM and RTSM, the value of  $\theta_3$  is greater than unity, but coupled by a larger standard error, indicating limited statistical power to ascertain the overall sigmoidal shape. For RTEG  $\theta_3$  is significantly greater than unity only under the Weibull model. Figure ?? displays the Weibull model fitted to all four data sets. Each circle in these plots represents observed incidences with its radius proportional to the multiplicity of clusters having identical fraction of affected fetuses. The plots show a rea-

sonable degree of goodness-of-fit of the Weibull model, and reveal dose-response relationships of varying shape and magnitude. RBSM and RTSM show a dose-response resembling the shape of a hockey-stick; RTEG illustrates a steady increase in response, reaching nearly 70% response rate at the highest dose level; MIEG depicts a shallow, nearly linear dose-response. The intra-litter correlation  $\rho$  (Table 2) is moderate in size, somewhat increases with dose (RTEG and RTSM), but lacks a consistent pattern (RBSM and MIEG). These examples are chosen purposely to illustrate potential issues associated of different shapes of dose-response. In particular, the weak dose-response of MIEG resulted in problematic behavior of BMDL estimates which we will discuss later.

(Table 2 and Figure ?? about here)

### 3.3 Benchmark Doses and Lower Confidence Limits

BMDs were computed for each data set under extra risk at BMR=0.01 and 0.05 for all three models and are reported in Table 3. Except for RTEG at BMR=0.01, the estimated BMDs are generally in close agreement among all three models. For RTEG the BMD estimate at BMR=0.01 under Weibull is 1.7-2 times as large as under the logit or probit model. This difference is clearly the result of a more pronounced nonlinear dose-response depicted by the Weibull model at the low dose range.

The median of  $B$  replications of bootstrap BMD is reported as "bootstrap BMD" in Table 3 for comparison purposes. Except for MIEG, the BMD estimates are in close agreement with the bootstrap BMD for all three models at both 1% and 5% BMR levels. For MIEG, the bootstrap BMDs are generally greater than the BMDs (about 20%-23% greater at BMR=5% and 58%-83% at BMR=1%). There were cases of bootstrap replications of MIEG which resulted in extreme estimates of the model parameters so that BMD<sup>b</sup> estimate did not exist (e.g. below zero). These cases were discarded from analysis and the actual number of replications  $B$  in Table 3 can be less than the targeted value of 2000. Bootstrap variation also

might result in attenuated dose-response in some cases, leading to estimates  $BMD^b$  beyond the maximum experimental dose, for instance. With MIEG especially the shallow dose-response (Figure ??) was sensitive to bootstrap variation and resulted in 493, 486, and 414 cases of excessive  $BMD^b$  estimates under the logit, probit, and Weibull models, respectively (Table 3, BMR=0.05, column B1). In contrast, such cases were less frequent under more elevated dose-responses: there was zero such case with RTEG, for instance. When increased response is seen only at the highest dose level, there can be considerable uncertainty about the overall shape of dose-response. This is the case of RBSM and RTSM where we observed a number of excessive  $BMD^b$  estimates.

The skewness measure reported in Table 3 suggests that the sampling distribution of the BMD estimator is somewhat left-skewed, and more so at 5% than at 1% BMR level. The degree of skewness is consistent across models. MIEG at BMR=1% is an only exception of which there is a slight right-skewness. The increased left-skewness at BMR = 0.05 is due in part to the right-shift of the  $BMD^b$  compared with at BMR=1%. The histograms in Figures 2a-d show the lack of symmetry of the distribution especially for the RBSM (Fig. 2a) and RTSM (Fig. 2c) at BMR=0.05. The second mode to the right of MIEG (Fig. 2d) is the result of setting the out-of-bound  $BMD^b$  to the highest experimental dose. These fixed values somewhat attenuated the skewness.

The 1- and 5-percentiles of the bootstrap sample  $BMD^b$  are given in Table 4 as the bootstrap  $BMDL_{99}$  and  $BMDL_{95}$ , respectively. The delta-method  $BMDL_{99}$  and  $BMDL_{95}$  are the lower confidence limits based on normal confidence interval to  $\log(BMD)$ . Several observations are noteworthy. First bootstrap BMDL is consistently smaller than that of the delta-method across data sets, models, BMR levels, and confidence levels. Except for Weibull model for RTEG where skewness is negligible, the bootstrap BMDL is multiple-time to orders of magnitude smaller than delta-method BMDL increasingly so as the confidence level increases and BMR level decreases. This is largely the consequence of left-skewed sampling distribution of BMD estimator.

Secondly, at 99%-confidence level and 1%-BMR level, both bootstrap and delta-method BMDL approach to 0, suggesting wider variation or greater uncertainty in estimating BMD and its associated risk at the required BMR level. To this end, MIEG illustrates a lack of adequate statistical power in the data to reliably establish the dose-response relationship or to reliably estimate BMD. Overall these observations also suggest that the typical developmental toxicity experiments may support dose-response assessment at or above  $\text{BMR} = 0.05$  for estimating BMD and confidence level at or below 95% for BMDL.

Third, while the BMDLs and their difference between the bootstrap and delta-method are generally insensitive to model choice, RTEG shows a sensitive case. There, the BMDL, is 581, 709, and 886 for the logit, probit, and Weibull model respectively at  $\text{BMR}=0.05$  and 95% confidence level, but is 28, 75, and 261 at  $\text{BMR}=0.01$  and 99% confidence level, considerably different among the three models. This sensitivity is the result of the more pronounced nonlinearity of the Weibull model.

Forth, the estimates of BMDL under additional risk essentially follow the same pattern of the BMDLs under the extra risk. The BMD and BMDL under additional risk are always somewhat greater than those under extra risk. These results are not reported here.

That the bootstrap BMDLs are consistently smaller than the delta-method BMDLs in the four data sets gives rise to the need to statistically assess the performance of the two methods. While both methods will perform well if sample size is sufficiently large, the delta-method depends heavily on the assumption of a symmetric distribution of the BMD estimator. But this assumption may not hold at BMR levels of interest. We conducted a simulation study to investigate coverage probability of the two BMDLs, which is reported in section 4.

### 3.4 Number of Bootstrap Replications

The bootstrap method usually requires a large number of replications ( $B$ ) in order to minimize variation introduced by bootstrap re-sampling. While a BMDL quantifies the variation of

the BMD estimator, a sufficient number of bootstrap replications is necessary in order to minimize the variation of BMDL itself. Is 1000 sufficient? To address this question we sampled with replacement  $B^*$  times ( $B^* = 250, 500, 750, 1000, 1250, 1500, 1750,$  and  $2000$ ) from the original pool of bootstrap replications to simulate an experiment with only  $B^*$  replications, and calculated the estimate of  $\text{BMDL}_{95}$  at  $\text{BMR} = 0.05$  of extra risk. For each  $B^*$  value we repeated this experiment 25 times. The variation among these 25 bootstrap BMDLs (y-axis) is summarized in a sequence of boxplots (Figure 3) against  $B^*$  (x-axis). The range of variation in BMDL narrows as  $B^*$  increases and stabilizes at and beyond  $B^* = 1250$ . The plots also show that the median of BMDLs stabilizes at  $B^* = 750$  for RBSM, RTEG, and RTSM, and at about  $B^* = 1000$  for MIEG. This suggests that a larger number of replications (e.g.  $B=1500$ ) is needed to achieve a stable estimate of BMDL with clustered quantal data.

## 4 Simulation Study

Our simulation study focused on the Weibull model, and used the Weibull model fitted to the original data set as truth for simulation purpose. Thus the estimated BMD under 5% BMR extra risk (Table 3: 1427.80, 1220.37, 715.54 and 1668.24 for RBSM, RTEG, RTSM, and MIEG, respectively) were also treated as the true BMD. For each data set, 2000 replications of dose-response data were simulated according to a correlated binomial distribution (Leisch *et al.* 1998), of which the mean response probability of adverse effects was the fitted Weibull model  $\pi(d)$  and correlation was given by dose-specific intra-cluster correlation  $\rho_i$  (Table 2). The size of dam or litter were identical to original data set. Data between dams/litters are treated as independent in the simulation.

In fitting the Weibull model to the simulated data, non-convergence often occurred. To circumvent this problem, the parameter ( $\theta$  or  $\rho$ ) that was an identified source of non-convergence was set to the true value; this process continued and additional parameter also was fixed one at a time in sequence until convergence was achieved. There were 609, 46, 519, and 681

such cases for RBSM, RTEG, RTSM, and MIEG, respectively, in which one of  $\theta$  was fixed. Among them, 416, 38, 280, and 316 cases also involved setting one or more  $\rho$  parameter to its true value. There were two simulated RTSM replications in which two  $\theta$  parameters were fixed. We observed that the strong underlying dose-response of RTEG helped a higher rate of convergence.

For each simulation replication, the delta-method BMDL and bootstrap BMDL were calculated under 5% BMR extra risk and at 90%, 95%, and 99% confidence level as described in section 2. Coverage probability was estimated by the proportions of the BMDLs that were less than the corresponding true BMD out of 2000 simulation replications (Table ??). The boxplots in Figures ??-?? show the variation of estimated BMDLs for the four experiments.

(Table ?? and Figures ?? - ?? about here.)

The simulations revealed several interesting aspects about the performance of the bootstrap and delta-method BMDLs. With a moderate to strong dose-response relationship associated with RBSM, RTEG, and RTSM, the coverage probability of both the delta-method and bootstrap BMDLs are greater than the nominal levels, and furthermore the delta-method BMDL is nearly uniformly more conservative than the bootstrap BMDL. At 90% nominal level, for example, the bootstrap BMDL has coverage probability 0.91 for RTEG compared 0.95 for the delta-method (Table 4). For MIEG, on the other hand, the coverage probability is below the nominal level for bootstrap BMDL (70-76%), but is greater than 99% for delta-method BMDL. At 90% nominal level, for instance, the bootstrap BMDL has only .70 coverage probability compared with .99 for delta BMDL. The poor performance of both bootstrap and delta-method BMDL in this case is attributable to the shallow and weak dose-response of MIEG. The shallow dose-response, coupled with limited data, was responsible for between 15%-25% of the BMD estimates that exceeded the maximum dose. These out-of-bound BMD estimates led to an increasingly right-skewed sampling distribution of the BMD estimator, hence a larger standard error and conservative delta-method BMDL. At the same time, they resulted in extremely large bootstrap BMD, consequently extremely larger BMDL estimates,

hence the under-coverage.

It should be noted that we purposely chose data sets with different dose-response shapes to examine the performance of the bootstrap and delta-method under a variety of conditions. For data sets RBSM, RTEG and RTSM, the bootstrap BMDLs are more accurate than the delta-method BMDLs. In contrast, MIEG illustrates that data with a weak and shallow, moderately significant dose-response may not be appropriate for benchmark dose computation (EPA, 2000). When there is a high degree of uncertainty or variation about the underlying dose-response, the confidence limit reflects much of this uncertainty in addition to data variation. Note that the nearly 100% coverage probability of the delta-method BMDLs regardless of the nominal level indicates extremely conservative confidence limits which provide little useful information about the true BMD. In the context of BMD, for example, choosing BMDL=0 always achieves 100% conservative coverage. One can thus argue that conservative confidence intervals should not be accepted blindly as preferable to anti-conservative ones. Therefore we conclude that both the delta-method and bootstrap BMDLs are dissatisfactory, and are not appropriate for comparison under MIEG.

Skewness of the distribution of BMD estimator plays an important role in determining the performance of BMDLs. Figures 2a-2d display left-skewness of the distributions for BMD estimators at 5% BMR level of extra risk. The boxplots in Figures 4a-4d show the variation of 2000 simulated replications of BMDL to the BMD. For RBSM (Fig 4a) and RTSM (Fig 4c), the medians of the bootstrap BMDLs are generally smaller than the delta-method BMDLs, particularly at 99% level. This is consistent with the observation in Table 4 that the delta-method BMDLs are universally greater than the bootstrap BMDLs. However, the bootstrap BMDLs are also more variable than the delta BMDLs. Consequently slightly more of the bootstrap BMDLs are greater the "true" BMD, compared with the delta-method BMDLs, resulting in coverage probability of bootstrap BMDL closer to the nominal levels than delta-method BMDL. Where the distribution of BMD is somewhat symmetric for RTEG, bootstrap BMDLs and delta BMDLs are not only comparable but also close to the nominal levels in

their coverage probabilities (Figure ??).

The relatively smaller variation of the delta BMDL is due perhaps to two factors: setting parameters to their true value to gain convergence may have artificially reduced the variation of the BMD estimator; and the standard error of  $\log(\hat{\text{BMD}})$  estimator was evaluated at the true parameter value and may not fully reflect the simulation variation. The coverage probability of the delta-method BMDLs generally approaches to 100% as the nominal level increases to above 95%. Overall the bootstrap method appears to reflect more honestly the simulation variation.

We also calculated delta-method BMD and BMDL based on normal confidence limits directly on BMD(results not shown). The resulting medians of 2000 simulated BMDs, and BMDLs are universally smaller than those derived from  $\log(\text{BMD})$ ; the resulting coverage probabilities are slightly, but universally, greater than those of  $\log(\text{BMD})$ . Furthermore, negative BMDLs were likely based on direct confidence limits on BMD particularly in the case of MIEG.

## 5 Conclusion and Discussion

In this paper we have proposed a bootstrap procedure to estimate BMDL using a lower percentile of the bootstrap sample of BMD estimates. Our procedure utilizes re-sampling of residuals after first fitting model to original data, and then employs a one-step formula to obtain bootstrap estimate of model parameters. The bootstrap is a practical approach to BMDL compared with more conventional methods such as those based on normal confidence limit in conjunction with the delta-method. Performance of the delta-method depends heavily on the symmetry assumption of the distribution for the BMD estimator. In contrast, the bootstrap BMDL is not affected by this assumption.

We have illustrated the bootstrap method through four data sets. The fourth example, MIEG, is a case where dose-response assessment via BMD may not be suitable. When con-

ducting benchmark dose analysis caution should be exercised to examine whether or not data is adequate to support such analysis. The remaining three data sets show that the sampling distribution of BMD estimator is somewhat left-skewed, and more so at BMR=5% level. As a result, bootstrap BMDLs are generally smaller, but are more variable than those of the delta-method. Our simulations with RBSM, RTEG, and RTSM show that delta-method BMDLs are highly conservative with respect to coverage probability, and bootstrap BMDLs also are conservative but closer to the nominal levels than the delta-method BMDLs. The simulations also show the median and mean of the bootstrap BMDLs are generally smaller than their delta-method counterparts. Thus, the bootstrap BMDLs are more conservative in median and mean value from a risk assessment viewpoint. When data do not ensure approximately symmetric distribution for BMD estimator, the bootstrap BMDL is an attractive and advantageous alternative to delta-method BMDL.

Bootstrap BMDL can be implemented based on re-sampling of original data and then fitting the model for each bootstrap sample. This approach is often hindered by non-convergence in fitting a nonlinear model. For example, among 1000 bootstrap replications of the original data of these four data sets convergence rate in model fitting ranged between only 20% (MIEG) to 87% (RTEG) even when using the "true" parameter values as starting values. Our simulation study also encountered this difficulty. A fix-up strategy was adopted to fix some of the parameters to their true value and to achieve convergence.

BMDL estimates are generally insensitive to the model choice under consideration as long as the models fit data reasonably well near the BMR level. For RTEG, the Weibull model established a more nonlinear (concave) curve near the BMR level than the other two models. As a result, it produced considerably larger estimates of bootstrap BMD and BMDL. This observation underlines the importance of adequate model fitting at the region near the chosen BMR. In other cases, the three models used in illustration generally give similar BMD and BMDL estimates.

We also conducted analysis of BMD and BMDL based on additional risk

$$\pi(\text{BMD}_\delta) - \pi(0) = \delta. \quad (7)$$

The results are rather similar to those reported for extra risk, hence are not reported here.

Although risk assessment of non-cancer effects typically considers BMR level between 1% to 10%, amount and quality of data required for reliable estimation of BMD and BMDL vary considerably within this range. Our analysis and simulation showed that the bootstrap BMDLs are reasonably reliable at BMR=5% and 95% confidence level. At BMR=1% or 99% confidence level, the BMDLs are too variable and are suspiciously close to zero, rendering BMD estimates unreliable. These illustrations suggest that the existing protocol of developmental toxicity experiment generally supports dose-response assessment at BMR=5% for BMD and 95% confidence for BMDL. Only in exceptionally cases such as RTEG, is it feasible to use BMR=1% or 99% confidence level.

When toxic effects are only seen at the highest dose, reliable estimation of the dose-response curve would be difficult. EPA (2000) recommends the use of experiments that demonstrate dose effects at multiple levels ( $\geq 2$ ). As risk assessment stresses increasingly on quantitative dose-response assessment, effective dose-selection as well as other effective design strategies are called upon. Although the selection of a geometrically decreasing dose sequence from a top dose is common in toxicological experiments, such a default choice can be potentially deficient in generating adequate data for a formal dose-response modeling.

## 6 References

- Bailer, A.J. and Smith, R.J. (1994). Estimating upper confidence limit for extra risk in quantal multistage models. *Risk Analysis* **14**, 1001-1010.
- Barnes, D.G. and Daston, G.P., Evans, J.S., Jarabek, A.M., Kavlock, R.J., Kimmel, C.A., Park, C. and Spitzer, W.L. (1995). Benchmark dose workshop: Criteria for use of a benchmark dose to estimate a reference dose. *Regulatory Toxicol. Pharmacol.* **21**, 296-306.
- Budtz-Jørgensen, E., Keiding, N. and Grandjean, P. (2001) Benchmark Dose Calculation from Epidemiological Data. *Biometrics* **57**, 698-706.
- Barnes, D.G. and Dourson, M. (1988). Reference dose (RfD): Description and use in health risk assessments. *Regulatory Toxicol. Pharmacol.* **8**, 471-486.
- Crump, K.S. (1984). A new method for determining allowable daily intakes. *Fund. Appl. Toxicol.* **4**, 854-871.
- Crump, K.S., Howe, R. (1985). A Review of methods for calculating confidence limits in low dose extrapolation. Chapter 9. In *Toxicological Risk Assessment*. D.B. Clayson, D. Krewski, I. Munro, eds. Boca Raton: CRC Press, Inc.
- Efron, B. and Tibshirani, R.J. (1993). *An introduction to the Bootstrap*. Chapman & Hall: New York.
- Gaylor, D.W. (1992). Incidence of developmental defects at the no observed adverse effect level (NOAEL). *Regulatory Toxicol. Pharmacol.* **15**, 151-160.
- Gaylor, D., Ryan, L.M., Krewski, D. and Zhu, Y. (1998). Procedures for calculating benchmark doses for health risk assessment. *Regulatory Toxicol. Pharmacol.* **28**, 150-164.
- Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. and Wahrendorf, J. (1986). Statistical

- Methods in Cancer Research Volume III - The design and analysis of long-term animal experiments. International Agency for Research on Cancer, Lyon.
- Krewski, D., and Zhu, Y. (1994). Applications of multinomial dose-response models in developmental toxicity risk assessment. *Risk Analysis* **14**, 613-627.
- Leisch, F., Weingessel A. and Hornik K. (1998). On the generation of correlated artificial binary data. Working Paper 13, *SFB Adaptive Information Systems and Modeling in Economics and Management Science*.
- Leisenring, W. and Ryan, L. (1992). Statistical properties of the NOAEL. *Regulatory Toxicol. Pharmacol.* **15**, 161-171.
- Liang, K.Y., Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **73**, 13-22.
- Moerbeek, M., Piersma, A.H., and Slob W. (2004) A Comparison of three methods for calculate confidence intervals for the benchmark dose. *Risk Analysis* **24**, 31-40.
- Morgan, B.J.T. (1992). *Analysis of quantal response data*. Chapman and Hall: London.
- Moulton L.H., Zeger S.L. (1991). Bootstrapping generalized linear models. *Computational Statistics & Data Analysis* **11**, 53-63.
- Ryan, L.M. (1992). Quantitative Risk assessment for developmental toxicity. *Biometrics* **48**, 163-174.
- U.S. Environmental Protection Agency (1991). Guidelines for developmental toxicity risk assessment. *Federal Register* **56**, 63797-63826.
- U.S. Environmental Protection Agency (1998). Guidelines for neurotoxicity risk assessment. *Federal Register* **63**(93):26926-26954
- U.S. Environmental Protection Agency (2000). Benchmark Dose Technical Guidance Document (External Review Draft).
- U.S. Environmental Protection Agency (2003). Draft Final Guidelines for Carcinogen Risk

## Assessment

- Xie, M. and Yang Y. (2003) Asymptotics for generalized estimating equations with large cluster sizes. *Annals of Statistics* **31:1**, 310-347.
- Zhu, Y., Krewski, D. and Ross, W.H. (1994). Multinomial models for developmental toxicity experiments. *Appl. Statist.* **43**, 583-598.

Table 1: Summary Data of Four Developmental Toxicity Experiments

a. RBSM: Fetal deaths in rabbits exposed to sulfamethazine (SM)				
<i>Dose(mg/kg)</i>	<i>Dams</i>	<i>Implants</i>	<i>Deaths<sup>a</sup></i>	<i>Death Rate</i>
0	24	192	28	0.146
600	26	201	33	0.164
1200	25	205	23	0.112
1500	28	237	64	0.270
1800	23	197	63	0.320

b. RTEG: Malformations in rats exposed to ethylene glycol (EG)				
<i>Dose(mg/kg)</i>	<i>Litters</i>	<i>Fetuses</i>	<i>Malformation<sup>b</sup></i>	<i>Malformation Rate</i>
0	28	379	5	0.013
1250	28	357	21	0.059
2500	29	345	86	0.249
5000	26	287	197	0.686

c. RTSM: Malformations in rats exposed to sulfamethazine (SM)				
<i>Dose(mg/kg)</i>	<i>Litters</i>	<i>Fetuses</i>	<i>Malformation<sup>b</sup></i>	<i>Malformation Rate</i>
0	25	289	12	0.042
545	22	251	8	0.032
685	24	278	24	0.086
865	24	259	52	0.201

d. MIEG: Fetal deaths in mice exposed to ethylene glycol (EG)				
<i>Dose(mg/kg)</i>	<i>Dams</i>	<i>Implants</i>	<i>Deaths<sup>a</sup></i>	<i>Death Rate</i>
0	25	333	36	0.108
750	24	310	34	0.110
1500	23	266	37	0.139
3000	23	283	57	0.201

<sup>a</sup> Number of dead or resorbed implants.

<sup>b</sup> Number of fetuses with any types of malformation.

Table 2: Estimates (Standard Error) of Dose-Response Model Parameters

<i>Study</i>	<i>Model</i>	Mean Parameters			Correlation Parameters				
		$\theta_1$	$\theta_2$	$\theta_3$	$\rho_1$	$\rho_2$	$\rho_3$	$\rho_4$	$\rho_5$
a. RBSM	Weibull	0.144 (0.025)	0.299 (0.098)	7.603 (4.808)	0.032 (0.065)	0.480 (0.344)	-0.039 (0.082)	0.529 (0.252)	0.295 (0.245)
	Logit	-1.855 (0.189)	1.254 (0.329)	6.307 (4.832)	0.028 (0.063)	0.470 (0.341)	-0.040 (0.083)	0.574 (0.265)	0.318 (0.245)
	Probit	-1.103 (0.102)	0.729 (0.194)	6.479 (4.820)	0.029 (0.063)	0.472 (0.342)	-0.040 (0.083)	0.568 (0.263)	0.315 (0.245)
b. RTEG	Weibull	0.013 (0.010)	1.282 (0.208)	2.282 (0.278)	0.130 (0.042)	0.068 (0.067)	0.344 (0.196)	0.383 (0.431)	
	Logit	-4.489 (0.832)	5.483 (0.861)	0.770 (0.216)	0.173 (0.078)	0.046 (0.063)	0.446 (0.206)	0.352 (0.432)	
	Probit	-2.278 (0.307)	2.888 (0.342)	0.935 (0.218)	0.167 (0.073)	0.049 (0.063)	0.434 (0.205)	0.359 (0.432)	
c. RTSM	Weibull	0.034 (0.014)	0.190 (0.055)	6.889 (3.651)	0.111 (0.066)	0.007 (0.063)	0.224 (0.149)	0.197 (0.178)	
	Logit	-3.360 (0.445)	1.964 (0.515)	4.612 (3.556)	0.108 (0.060)	0.003 (0.062)	0.266 (0.172)	0.202 (0.177)	
	Probit	-1.833 (0.193)	0.987 (0.243)	4.894 (3.528)	0.109 (0.062)	0.003 (0.062)	0.258 (0.167)	0.201 (0.177)	
d. MIEG	Weibull	0.110 (0.024)	0.121 (0.048)	1.455 (1.309)	0.061 (0.068)	0.040 (0.072)	0.344 (0.241)	0.080 (0.116)	
	Logit	-2.149 (0.233)	0.797 (0.307)	1.219 (1.234)	0.061 (0.068)	0.038 (0.072)	0.365 (0.251)	0.082 (0.116)	
	Probit	-1.257 (0.120)	0.435 (0.166)	1.253 (1.244)	0.061 (0.068)	0.038 (0.072)	0.362 (0.249)	0.081 (0.116)	

Table 3: Benchmark Dose (mg/kg/day) under Extra Risk

<i>Study</i>	<i>Model</i>	BMR=0.01				BMR=0.05			
		Bootstrap				Bootstrap			
		BMD	BMD <sup>†</sup>	Skewness	B(B <sub>1</sub> ) <sup>‡</sup>	BMD	BMD <sup>†</sup>	Skewness	B(B <sub>1</sub> ) <sup>‡</sup>
a. RBSM	Weibull	1152.29	1219.87	-0.921	1999(153)	1427.80	1466.68	-1.968	1999(155)
	Logit	1144.37	1255.48	-0.820	1999(244)	1455.77	1512.97	-2.040	1999(244)
	Probit	1145.33	1247.99	-0.841	2000(226)	1451.55	1504.17	-2.039	2000(227)
b. RTEG	Weibull	597.51	596.18	-0.076	2000(0)	1220.37	1218.97	-0.305	2000(0)
	Logit	311.47	293.26	-0.154	2000(0)	1132.28	1067.55	-0.947	2000(0)
	Probit	364.46	354.11	-0.125	2000(0)	1148.99	1112.87	-0.768	2000(0)
c. RTSM	Weibull	564.78	578.03	-1.214	2000(59)	715.54	723.38	-2.596	2000(64)
	Logit	559.39	596.34	-0.881	2000(201)	737.79	756.44	-2.683	2000(205)
	Probit	559.30	591.23	-0.914	1999(174)	733.17	749.99	-2.605	1999(176)
d. MIEG	Weibull	544.35	857.16	0.888	1988(310)	1668.24	2000.13	-0.490	1988(414)
	Logit	511.82	935.72	0.739	1983(379)	1733.09	2136.55	-0.659	1983(493)
	Probit	516.92	928.65	0.754	1984(374)	1722.31	2122.32	-0.635	1985(486)

<sup>†</sup> Bootstrap BMD is the median of bootstrap replications

<sup>‡</sup> B<sub>1</sub> is the number of replications in which BMD exceeded and was subsequently set to the highest experimental dose

Table 4: Lower Confidence Limit of Benchmark Dose under Extra Risk

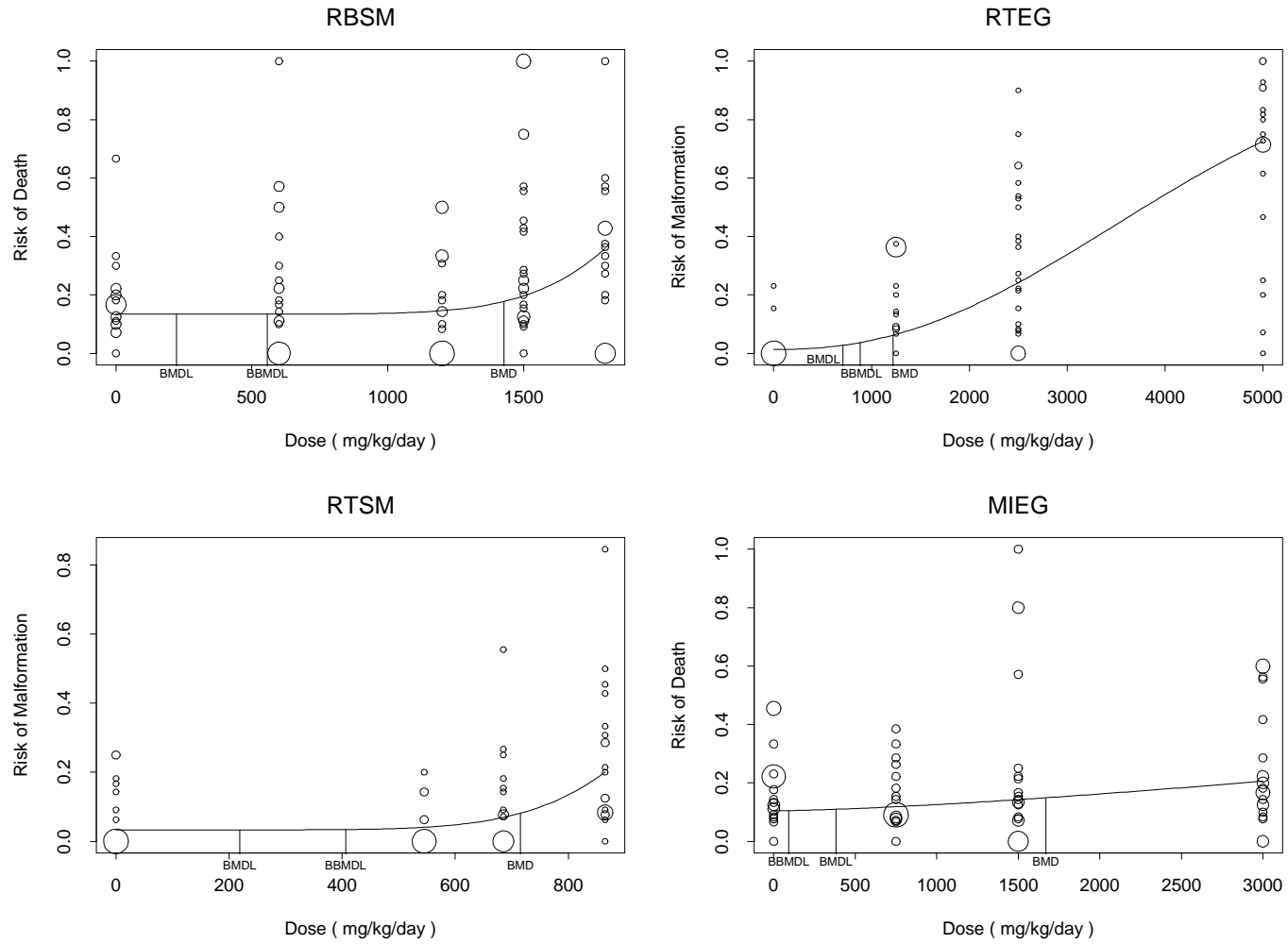
<i>Study</i>	<i>Model</i>	BMR=0.01				BMR=0.05			
		Bootstrap		Bootstrap		Bootstrap		Bootstrap	
		BMDL <sub>95</sub> <sup>†</sup>	BMDL <sub>95</sub>	BMDL <sub>99</sub> <sup>†</sup>	BMDL <sub>99</sub>	BMDL <sub>95</sub> <sup>†</sup>	BMDL <sub>95</sub>	BMDL <sub>99</sub> <sup>†</sup>	BMDL <sub>99</sub>
a. RBSM	Weibull	725.99	193.79	599.52	0.04	1118.72	557.50	1011.17	6.54
	Logit	652.84	259.96	517.39	0.13	1117.13	689.55	1001.06	28.40
	Probit	663.71	250.29	529.42	0.02	1117.02	676.56	1002.13	4.38
b. RTEG	Weibull	411.53	357.28	352.62	261.38	962.03	886.31	871.74	736.83
	Logit	180.55	99.97	144.04	28.38	853.81	581.23	759.60	299.48
	Probit	223.58	157.98	182.60	74.81	875.16	709.23	781.84	458.14
c. RTSM	Weibull	389.99	155.13	334.51	0.44	602.56	406.46	561.15	28.89
	Logit	337.30	128.16	273.52	0.02	612.36	433.93	566.87	19.21
	Probit	346.72	142.18	284.41	0.42	609.65	448.78	564.79	56.46
d. MIEG	Weibull	40.20	0.27	13.65	0.00	605.53	92.10	397.91	0.00
	Logit	25.96	0.35	7.55	0.00	621.56	149.60	406.42	0.15
	Probit	27.96	0.48	8.35	0.00	618.48	166.59	404.62	0.14

<sup>†</sup> BMDL based on a normal confidence limit for log(BMD)

Table 5: Coverage probability of BMDL based on simulation

<i>Study</i>	<i>Method</i>	Nominal Level		
		99%	95%	90%
RBSM	Delta	1.000	0.998	0.985
	Bootstrap	1.000	0.997	0.973
RTEG	Delta	0.994	0.978	0.949
	Bootstrap	0.997	0.969	0.913
RTSM	Delta	1.000	0.993	0.983
	Bootstrap	1.000	0.997	0.966
MIEG	Delta	1.000	0.999	0.993
	Bootstrap	0.762	0.744	0.697

Figure 1: Fitted dose-response curve under Weibull model and 5% BMR<sup>†</sup>



<sup>†</sup>RBSM: BMD=1427.80, BMDL<sub>95</sub>=223.44, Bootstrap BMDL<sub>95</sub>=557.50    RTEG: BMD=1220.37, BMDL<sub>95</sub>=708.97, Bootstrap BMDL<sub>95</sub>=886.31

RTSM: BMD=715.54, BMDL<sub>95</sub>=219.01, Bootstrap BMDL<sub>95</sub>=406.46    MIEG: BMD=1668.24, BMDL<sub>95</sub>=381.71, Bootstrap BMDL<sub>95</sub>=92.10

The radius of a circle is proportional to the square root of the number of identical observations observed at its location.

Figure 2a: Sampling distribution of bootstrap BMD of extra risk with RBSM

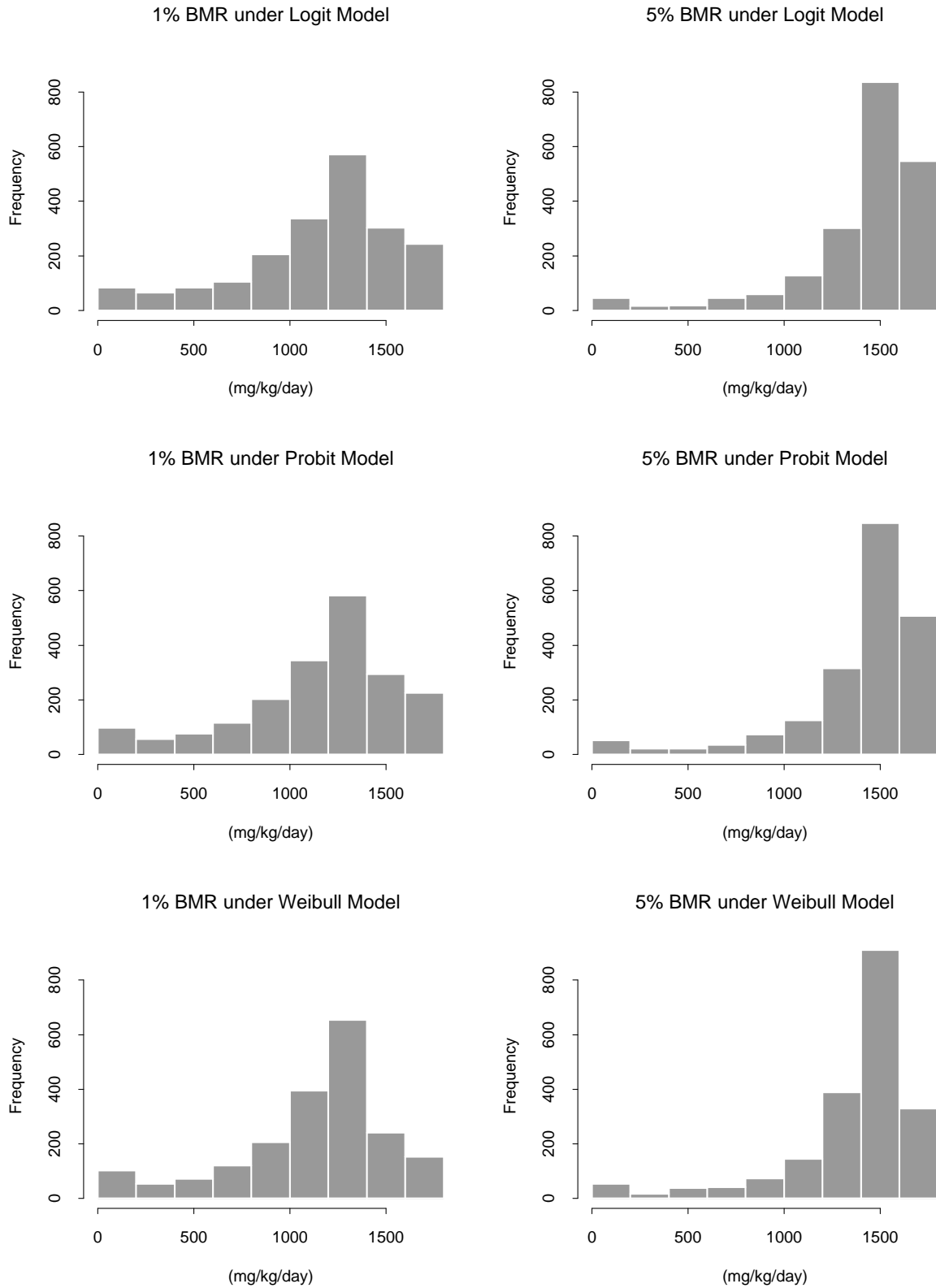


Figure 2b: Sampling distribution of bootstrap BMD of extra risk with RTEG

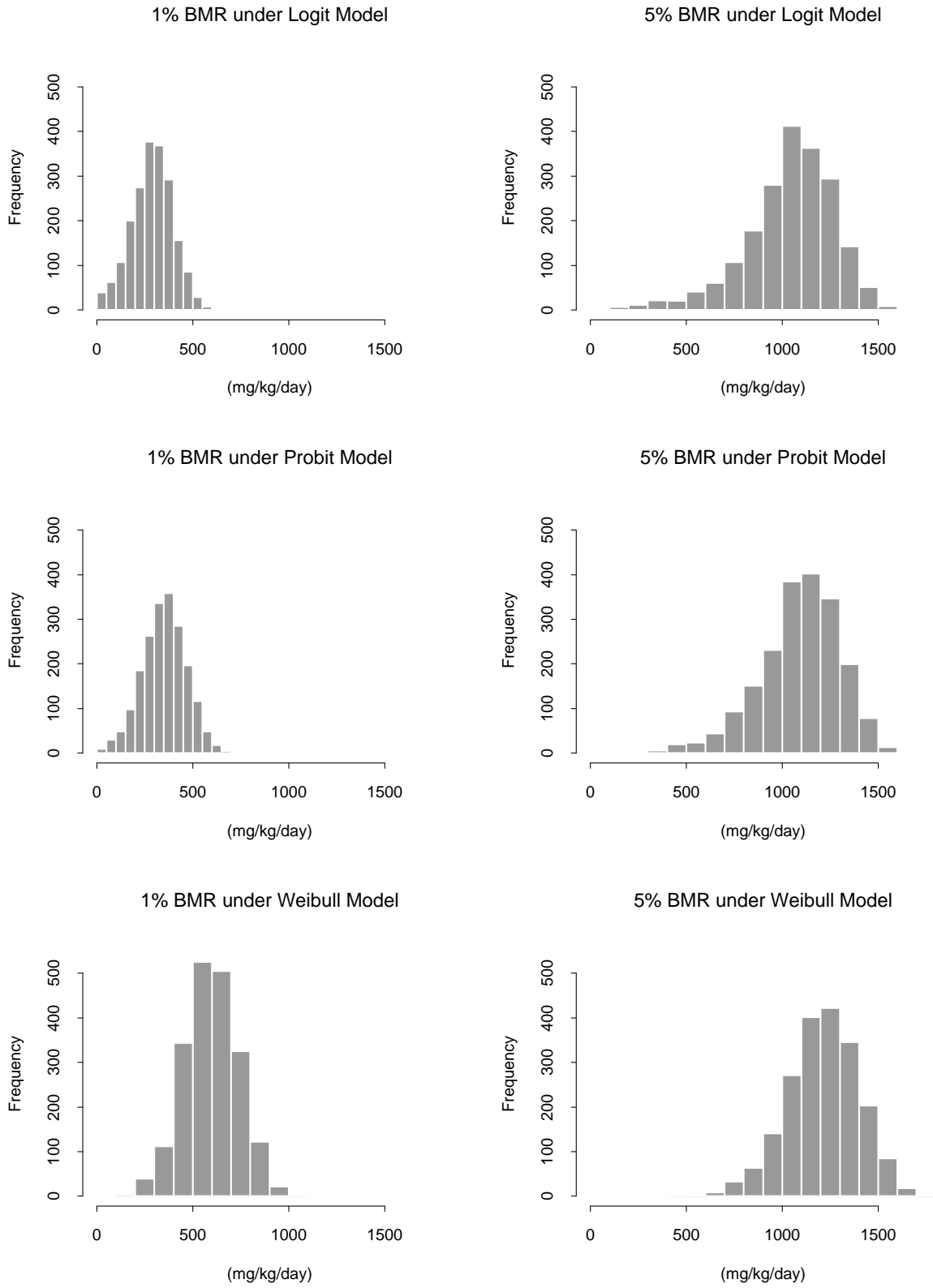


Figure 2c: Sampling distribution of bootstrap BMD of extra risk with RTSM

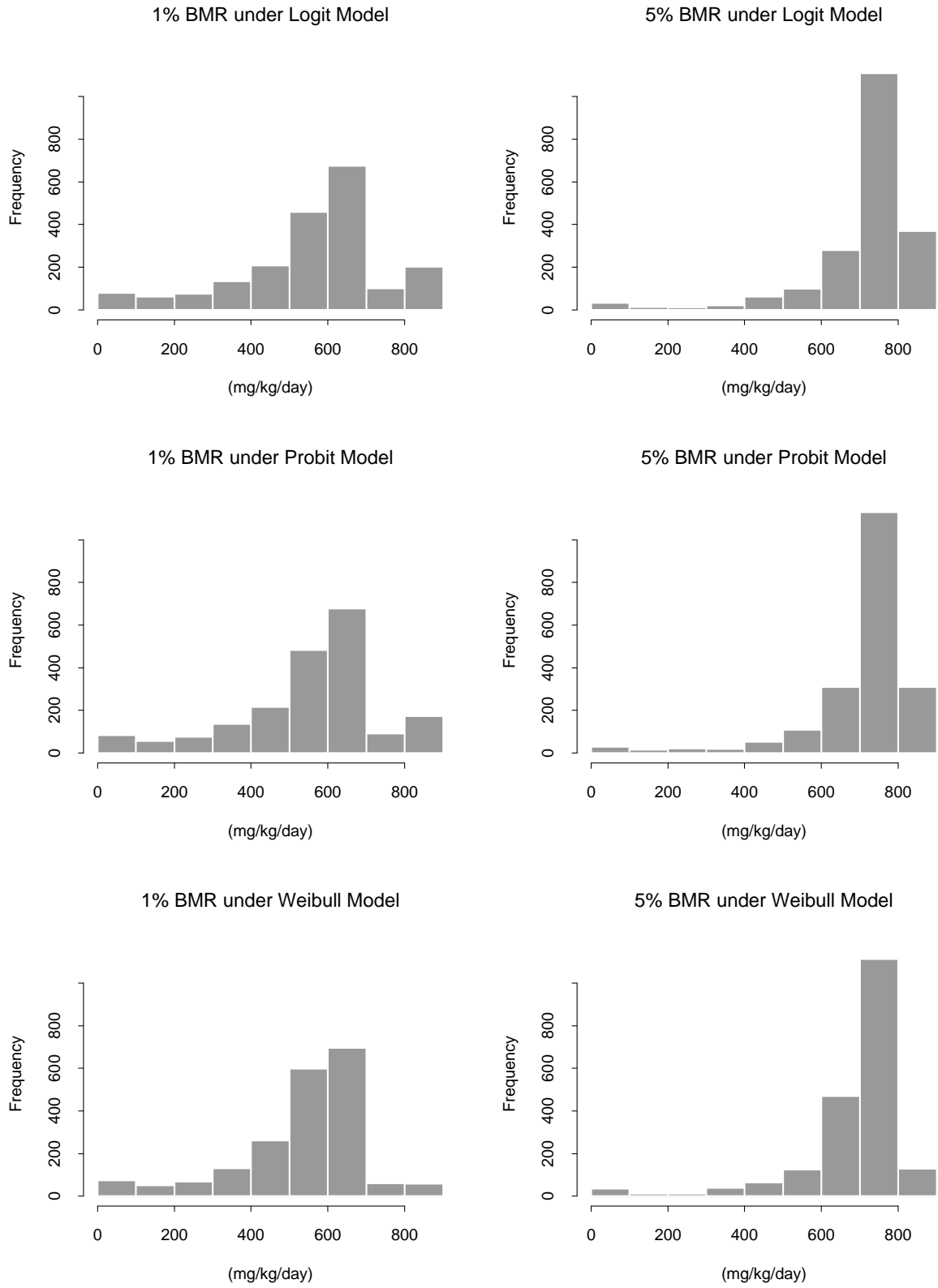


Figure 2d: Sampling distribution of bootstrap BMD of extra risk with MIEG

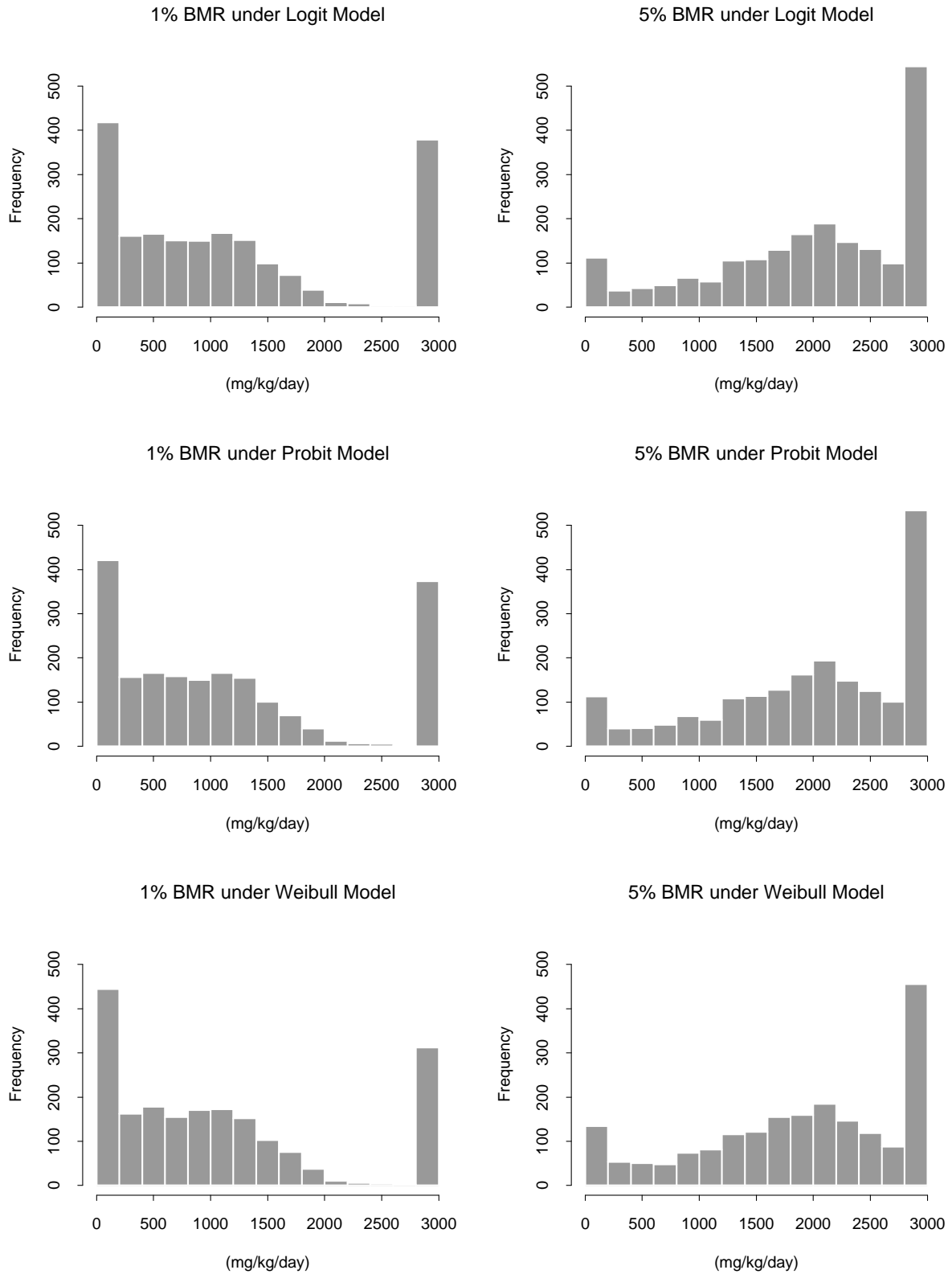


Figure 3: Variation of Bootstrap BMDL<sub>95</sub> of Extra Risk at BMR=0.05 under Weibull model

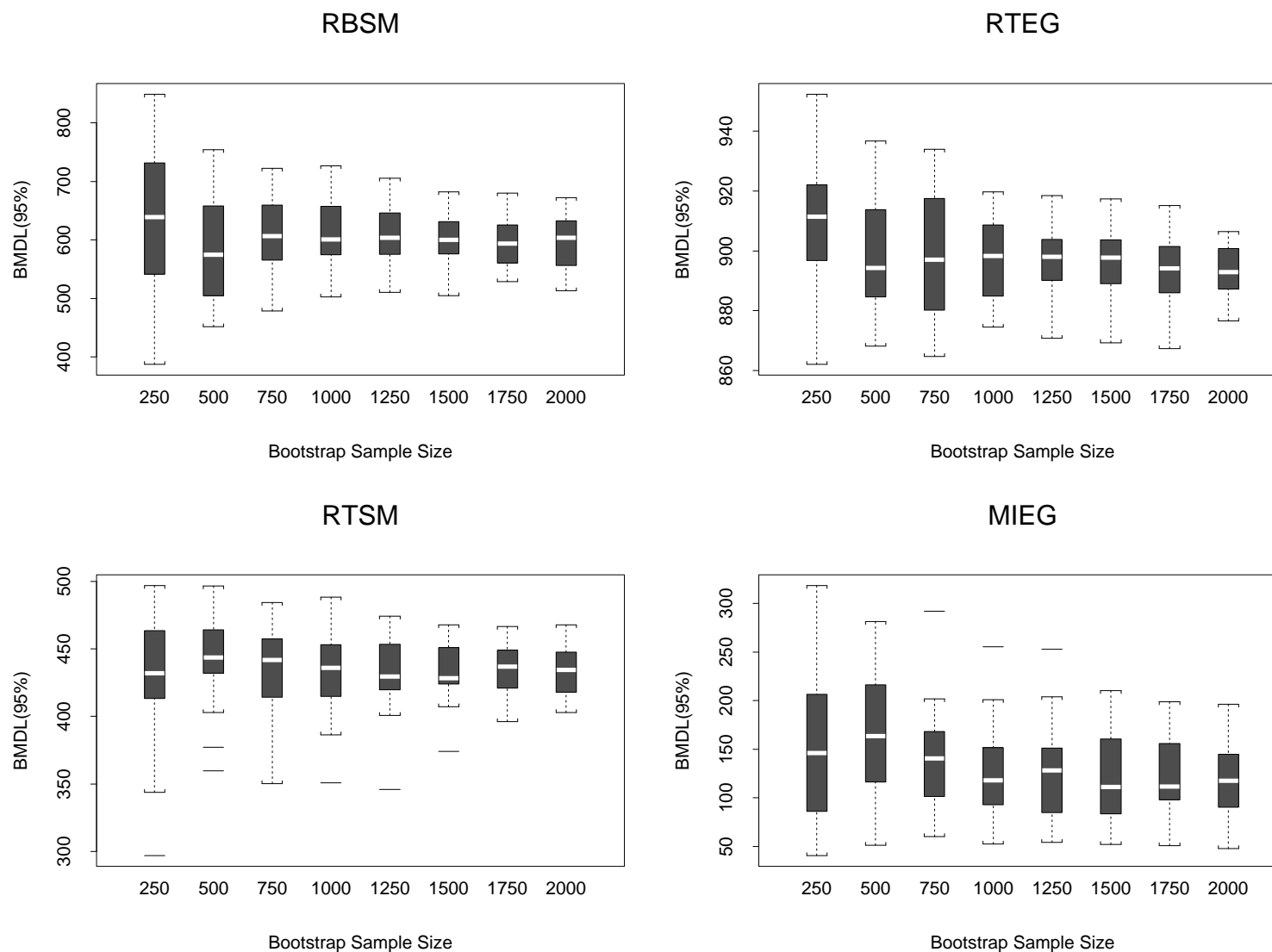


Figure 4a: Variation of BMDL based on 2000 simulations of RBSM: extra risk at BMR=0.05

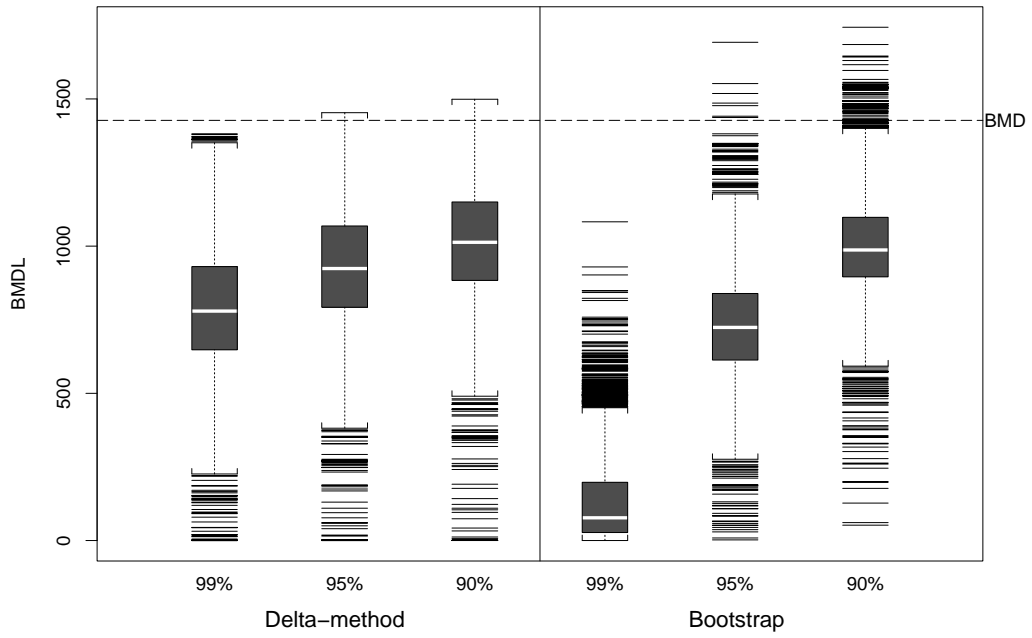


Figure 4b: Variation of BMDL based on 2000 simulations of RTEG: extra risk at BMR=0.05

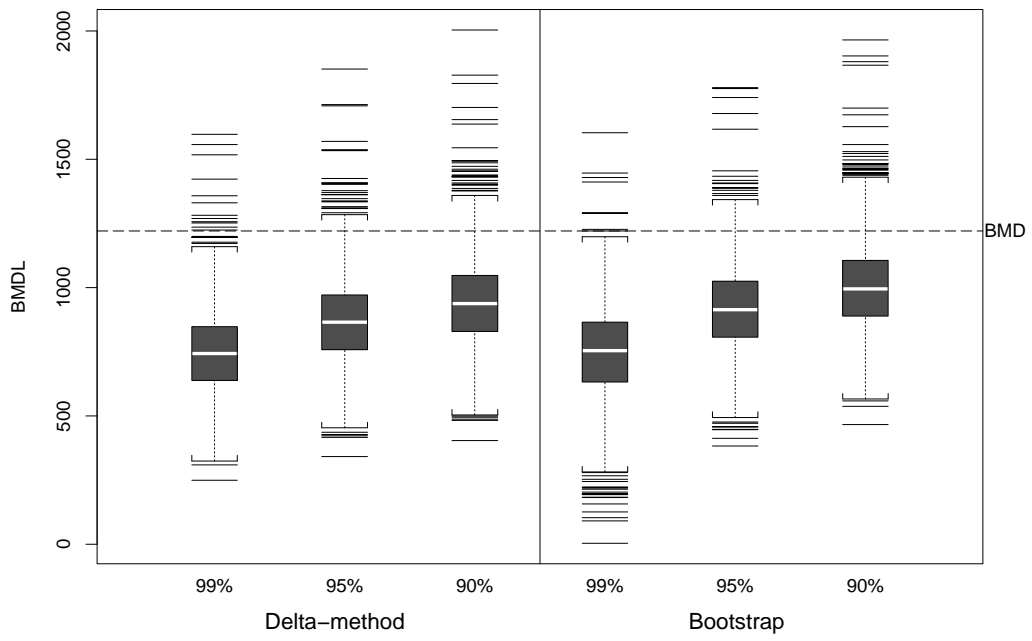


Figure 4c: Variation of BMDL based on 2000 simulations of RTSM: extra risk at BMR=0.05

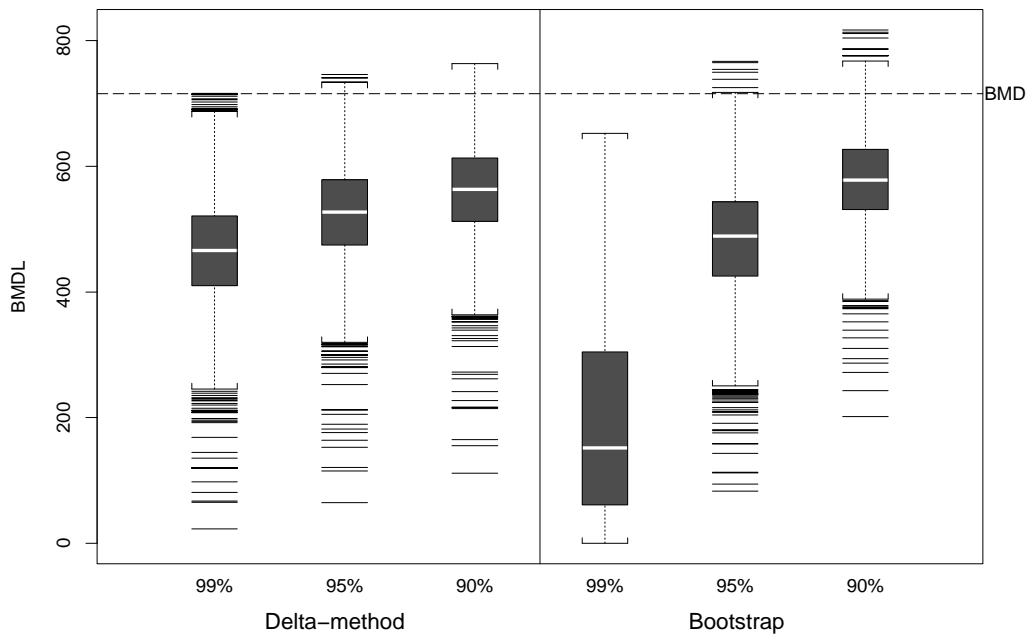


Figure 4d: Variation of BMDL based on 2000 simulations of MIEG: extra risk at BMR=0.05

