学　位　論　文

Simulation-based assessment of model selection criteria during the application of benchmark dose method to quantal response data
(ベンチマークドーズ法の用量反応関係への適用にかかるモデル選択基準に関するシミュレーション研究)

2022 年 12 月

北海道大学
吉　井　　啓　太
Keita Yoshii

学　位　論　文

Simulation-based assessment of model selection criteria during the application of benchmark dose method to quantal response data
(ベンチマークドーズ法の用量反応関係への適用にかかるモデル選択基準に関するシミュレーション研究)

2022 年 12 月

北海道大学
吉　井　　啓　太
Keita Yoshii

Contents

List of publications and presentations.....1
Abstract.....3
List of Abbreviations.....8
Introduction.....9
Materials and Methods.....18
Results.....31
Discussion.....41
Conclusion.....45
Acknowledgements.....46
Conflicts of interest.....47
References.....48

List of publications and presentations

Part of the research has been published as listed below:

1. Yoshii K, Nishiura H, Inoue K, Yamaguchi T, Hirose A. Simulation-based assessment of model selection criteria during the application of benchmark dose method to quantal response data. Theor. Biol. Med. Modell. 2020. 17:13

Part of the research has been presented in the conferences as listed below:

1. Yoshii K, Nishiura H, Inoue K, Hirose A. Simulation-based assessment of model selection criteria during application of benchmark dose method to quantal response data. 58th annual meeting and Toxexpo on Society of Toxicology (SOT), Baltimore Maryland USA, March 2019 (Poster)
2. Yoshii K, Nishiura H, Inoue K, Hirose A. Assessing model selection criteria during the application of benchmark dose method to quantal response data: Japanese perspectives. IUTOX 15th International Congress of Toxicology (ICTXV), Honolulu Hawaii USA, July 2019 (Poster)

Part of the research has been accepted to be presented but cancelled due to the pandemic of COVID-19:

1. Yoshii K, Nishiura H, Inoue K, Yamaguchi T, Hirose A. Evaluation of model selection criteria including model averaging during the application of benchmark dose method to quantal response data. 59th annual meeting and Toxexpo on Society of Toxicology (SOT), Anaheim California USA, March 2020 (Poster)
2. Yoshii K, Nishiura H, Inoue K, Yamaguchi T, Hirose A. 食品健康影響評価におけるベンチマークドーズ法適用上でのモデル選択基準や平均化の検討. The 90th Annual Meeting of the Japanese Society for Hygiene (JSH), Morioka Iwate Japan, March 2020 (Oral)

Abstract

【Introduction】

It has been popular to use NOAEL method for the determination of the reference dose of the chemical substances, such as the acceptable daily intake (ADI). However, we must deal with a serious statistical limitation that causes non-negligible sampling errors by determining NOAEL. The benchmark dose (BMD) method as an alternative can address the problems because all aspects of dose-response data which involves underlying biological uncertainties are taken account of using this method. Fitting various statistical models to the dose-response curve through this method, the benchmark dose (BMD) method determines the threshold dose (BMD) associated to the benchmark response (BMR), a specified level of response from the background response, by fitting various statistical models to the dose-response curve. A point of departure for the calculation of such reference dose using BMD method, which is empirically comparable to that based on NOAEL, can be the benchmark dose lower bound (BMDL), which is the lower (one-sided) limit of the 95% confidence interval (CI) of BMD. However, practical issues regarding to the application of BMD methods prevent from formulating an uniform guideline of technical use of the BMD method worldwide. It is critical to understand how BMDL for reference dose calculation is selected following statistical fitting procedures of multiple mathematical models to employ the BMD method in toxicological risk assessment. Although there are several technical problems, we believe that the uniform guidelines that specify the steps required to scrutinize fitting results and identify a single BMDL value for determining such reference dose is the key element for the wider application of the BMD method in various governmental settings. Among the remaining technical issues, limited number of model selection and exclusion criteria including model averaging have been verified and practiced on toxicological risk assessment.

While we cannot fully and immediately solve all issues surrounding the use of the BMD method for quantal response data, a simulation-based evaluation might help to identify a possible well-performing pathway of model exclusion and selection. A simulation study was conducted to compare the performance of each and various combinations of model exclusion and selection criteria as applied to three qualitatively different types of quantal response datasets, to support the formulation of technical guidelines for risk assessment practices for food safety in Japan.

【Materials and Methods】

We employed the BMD method using simulation-based evaluation of model exclusion and selection processes by comparing validity, reliability, and other model performance parameters. We analyzed three different empirical datasets for different chemical substances, each having qualitatively different characteristics of the dose-response pattern. Briefly, our analysis goes by: (a) identification of a “reference model” for each dataset by AIC (Akaike Information Criteria), (ii) generation of a total of 1,000 simulated datasets (each dataset includes fittings by 9 individual model) from the “reference model”, (iii) application of model exclusion criteria if available, (iv) application of one of the model selection criteria including methods using model averaging, and select or calculate one of the representative BMDL value from each dataset, and (v) BMDL values were evaluated in two aspects, the validity and the reliability. First, for each quantal dataset, the best-fit model was identified by selecting the model with the lowest AIC value out of the total of nine different distributions that consist of 2-4 unknown parameters. During the simulations, the identified best model for each chemical substance was regarded as the “reference model”. The known lower bound of the benchmark dose with response level at 10% (unbiased $BMDL_{10}$) was derived from the maximum likelihood estimates of the parameters. The likelihood function was defined

under the statistical assumption that the quantal response data at a given dose follows a binomial distribution. Through the simulation, we computed the 95% CI, including BMDL and BMD upper bound (BMDU) (i.e. one-sided upper 95% CI of BMD) using the Monte Carlo algorithm. Validity and reliability were used as evaluators of model selection criteria out of 1000 simulated values: Briefly, validity is the proportion that the simulated $BMDL_{10}$ value is the dose lower than the known benchmark dose with response level at 10% (unbiased BMD_{10}), and reliability is the relative distance between the simulated $BMDL_{10}$ value and the unbiased $BMDL_{10}$ value. We also assessed calculability of BMDL, the proportion of simulated datasets that yielded convergence, and "true dose-response", the proportion of the statistical model that was recovered by simulation to be identical to the reference model. A total of four possible model exclusion criteria and six possible model selection criteria were considered, including criteria using model averaging. Avoiding combinations of the excessive exclusion of the statistical model, a total of 18 possible combinations were tested and compared.

【Results】

It was turned out that the best performing criteria of model exclusion and selection were not the same across the different datasets. Generally, however, we did not observe the worst performance using one of the model selection criteria, model averaging over the three models with the lowest three AIC values (MA-3), in any of the three quantal datasets we used. Indeed, MA-3 without model exclusion produced the best results among the model averaging. The validity and reliability of the models were not improved by imposing model exclusion criteria including the use of the Kolmogorov-Smirnov test in advance of model selection.

【Discussion】

A simulation-based experiment was conducted to assess the model exclusion and selection process by comparing the validity, reliability, and other model performance indicators using all possible combinations of model exclusion and selection criteria, which is the part of the technical assessment for possible improvements in the guidelines. There are two take-home messages. First, although the best exclusion and selection criteria for the qualitatively differently distributed datasets were not uniformly identified, it was shown that model averaging over three models with the lowest three AIC values (MA-3) did not yield the worst result, and the best results was provided by MA-3 without prior model exclusion among all the model averaging results. Secondly, it was clearly observed that model exclusion using the KS test and the ratios of BMD or BMDU to BMDL did not necessarily yield better validity and reliability than non-exclusion. Particularly, all the model averaging options that we tested were found to perform well overall. Furthermore, we found that a better performance might possibly observed using averaging over some of the models than all converged models, considering that the uncertainties of well-fitted models might be far smaller than those of badly fitted models. While we still need to address numerous technical issues in applying the BMD methods to risk assessment, our conclusion was that the best guiding option is MA-3 to derive the reference dose when a single model exclusion and selection method were expected to be specified by the technical guidelines.

【Conclusion】

A simulation-based experiment was conducted to assess the model exclusion and selection process as part of the technical assessment for possible improvements in the guidelines. We compared the validity, reliability, and other model performance indicators using all possible combinations of model exclusion and selection criteria. If we need a uniform technical

suggestion for the guideline to choose the best performing model for exclusion and selection, using MA-3 is highly recommended whenever applicable in our study.

List of Abbreviations

AIC - Akaike Information Criteria

BMD - Benchmark dose

BMDL - Benchmark dose lower bound

BMDU - Benchmark dose upper bound

CI - Confidence interval

MA - Model averaging

MA-3 - Model averaging over three models with the lowest three AIC

MA AIC<3 - Model averaging over models with the difference of AIC less than 3 from the best-fit model with the lowest AIC value

MA-all - Model averaging over all converged models

MLE - Maximum Likelihood Estimation

NOAEL - No observable adverse effect level

Introduction

A number of scientific approaches using dose-response experimental data have been used in practice to determine the reference dose of chemical substances, including food additives and agricultural chemicals, which cause the presence or absence of a harmful event (i.e. dichotomous outcome) so that the acceptable daily intake (ADI) can be specified. A popular toxicological method, NOAEL method, uses the responses to low dose exposures to confirm the absence of an outcome event. NOAEL is the term of the abbreviation, No Observable Adverse Effect Level, that the highest dose that does not cause the outcome event, below which no outcome event is expected. A point of departure, a reference value in toxicological context (e.g. ADI), has been determined in practice by multiplying this NOAEL with a specified uncertainty factor that addresses biological uncertainties, including the biological species barrier between experimental animals and humans (Barnes et al., 1988). However, NOAEL depends on low dose data alone, ignoring high dose data and the shape of overall dose-response curve; so, if the number of experimental animals per dose is limited, a serious statistical limitation that involves non-negligible sampling errors would follow the results we obtain (Brown et al., 1989; Leisenring et al., 1992). Crump had primarily implemented and formalized an alternative method to calculate the point of departure accounting for such errors (Crump, 1984). The benchmark dose method, abbreviated as BMD method, determines the threshold dose as called benchmark dose (BMD) associated to the benchmark response (BMR), a specified level of response from the background response, by fitting various statistical models to the dose-response curve. A point of departure that is empirically comparable to that based on NOAEL can be yielded by the benchmark dose lower bound with 10% extra risk (i.e. setting BMR to 10%) to the basement response ($BMDL_{10}$), which is the lower (one-sided) limit of the 95% confidence interval (CI) of BMD with 10% extra

risk to the basement response (BMD_{10}) (**Figure 1**) (Allen et al., 1994a; Wignall et al., 2014). **Table 1** summarizes the practical and technical points using BMD method to determine reference dose compared with NOAEL-based approach. Briefly, the BMD method can address the problems surrounding the use of NOAEL because it can account for the response data across all aspects of data including all points of doses and overall shape, unlike the NOAEL approach accounting only for the point of dose with the highest non-observable adverse effect, therefore it can quantify the dose-response data itself using confidence interval of BMD (European Food Safety Authority (EFSA) Scientific Committee et al., 2017). Some type of experimental dataset cannot yield NOAEL when they have some restriction, such as no response observed at all, but BMDL is often calculable under such "limited" dataset (Sand et al., 2008). Moreover, it is easy and obvious to assess the uncertainty of the true value (i.e. BMD) by calculating BMD confidence interval, whereas it is exceptionally difficult to address the uncertainty of the reference dose determined by NOAEL-based approach using a dataset produced by one single animal experiment (European Food Safety Authority (EFSA) Scientific Committee et al., 2017). BMDL is still retained its definition having a wider BMD confidence interval even when using the poor dataset (e.g. less sample size) (European Food Safety Authority (EFSA) Scientific Committee et al., 2017). Furthermore, the BMD method accounts for sample size appropriately to account for uncertainty; The higher reference dose is more likely to be produced by larger sample size in the experimental dataset due to less random variation, which is consistent with the principle of results provided by BMD method because it produces smaller BMD confidence interval when using larger sample size, whilst the results by NOAEL method moves the opposite direction because the significant response level will be shown by smaller doses when using larger sample size (Crump, 1984). Reference dose by NOAEL method is subject to sample size of a single animal experiment, in which the significant effect level may be differed from another

(Leisenring et al., 1992). Considering the limitations of NOAEL method, the BMD method can potentially be extremely useful in many scientific disciplines despite the fact that the BMD method has its two main limitations since it is an emerging technique, which are the lack of the established protocol how to apply the BMD method in practice and the inadequate amount of data and articles previously studied before (Bi, 2010; Kimmel et al., 1988). The US and the EU authorities have already used the BMD approach to determine the reference dose in a toxicological risk assessment. In the US, BMD approach had been acquired as a significant opportunity to improve scientific bases of noncancer risk assessment by The United States Environment Protection Agency (USEPA), founded in 1970 for the response of environmental problems through federal research, monitoring and enforcement activities (United States Environment Protection Agency (USEPA), 2012). It is also incorporated into the risk assessment guidance issued by European Food Safety Authority (EFSA), the organization founded in 2002 in response to food crisis in 1990s to be a source of scientific advice and communication on risks associated with the food chain (European Food Safety Authority (EFSA), 2009). Using BMD approach was recommended instead of NOAEL, as it could make a wider use of dose-response data (European Food Safety Authority (EFSA), 2009). In Japan, the BMD method for risk assessment has been used in the risk assessment of some chemical compounds (e.g. methylmercury (2004), inorganic arsenic (2013)) (食品安全委員会評価技術企画ワーキンググループ, 2018.). Nevertheless, due to the remaining technical issues applying BMD method in practice, it is problematic that no uniform guideline of practical use of the BMD method has been established worldwide neither in the US and EU, as seen in the difference between USEPA's and EFSA's criteria in model selection and exclusion from multiple mathematical models fitted to dose-response data. (食品安全委員会評価技術企画ワーキンググループ, 2018.)

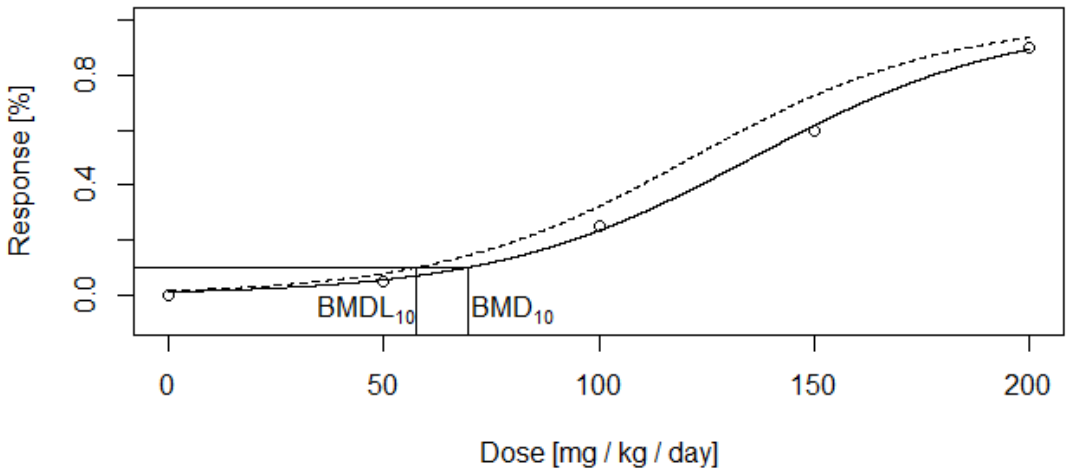

Figure 1. Example of benchmark dose (BMD), BMDL in Dose-response curve.

A dose-response curve illustrating relationship between BMD_{10} and $BMDL_{10.}$ Dots: fraction of animals affected in each dose group; Solid curve: Fitted model; BMD_{10}: BMD corresponding to 10% extra risk on this curve based on fitted model; Dashed line: the estimated lower bound on doses for a range of BMRs; $BMDL_{10}$: The lower bound on BMD_{10} based on the dashed curve.

Table 1. Characteristics of the BMD method on determination of reference dose compared to NOAEL-based approach.

Summarization of the points of the BMD method compared to NOAEL approach. 1. All aspects of dose-response curve, and 2. Uncertainty calculation.

Characteristics of BMD method compared to NOAEL-based
1. All aspects of dose-response curve
• The BMD method can include all aspect of dose-response curve including points of data in higher doses and overall shape, where NOAEL value can be only reflected to by the highest point of dose without extra risk (EFSA, 2017).
• BMDL can often be calculated using a dataset even when a NOAEL is not present in the dataset (Sand et al., 2008).
2. Uncertainty calculation
• The BMD method take the uncertainty of the true value into account by the calculation of the BMD confidence interval, where uncertainty associated with NOAEL cannot be evaluated by a single dataset (EFSA, 2017).
• When using poor or limited data, NOAEL value have larger uncertainty therefore it may not be objectively acceptable, but BMDL remains its definition although their value would be much lower than the true BMD value with a wider BMD confidence interval (EFSA, 2017).
• In larger study, BMDL would increase as the BMD confidence interval becomes smaller, which can reflect the less random variation, but NOAEL would decrease as such study has a more significant results in a set of smaller doses (Crump, 1984).
• NOAEL is sensitive to sample size: one experiment has a stronger power to detect the adverse effects, therefore significant effect level may vary between experiments (Leisenring et al, 1992).

It is critical to exclude the bad performing BMDL and select the best one (i.e. theoretically valid and reliable BMDL that is aligned with the reality) in order to employ the BMD method by following the statistical fitting procedures of multiple mathematical models. Multiple models (usually nine or more) are commonly fitted to the same experimental dataset as such parameterized models characterize reality alone. As a result, the reference dose is not usually determined at one point since many BMDL values can act as the candidate of preferred reference dose. However, the best performing BMDL should become the reference dose and it must be selected among the candidates after excluding some of the bad-fitting BMDL candidates properly; for example, the one that gives the best fitting results (United States Environment Protection Agency (USEPA), 2012). There are two additional issues in selecting or determining the BMDL described in **Table 2** in addition to the lack of the established criteria on model selection and exclusion as described above. First, a specified percentile point (e.g. 10% of the benchmark response, abbreviated as BMD_{10}) as the threshold for the reference value is used arbitrarily in the BMD method, but the "10% percentile point" is never strictly objective (Sand et al., 2002; Weterings et al., 2016; Allen et al., 1994b). This is quite similar to the rules of thumb that we use a p-value of 5% in many hypothesis tests or other arbitrarily chosen threshold values. Second, different parameter estimates were yielded when restrictions to the range of parameters were imposed in advance of the inference procedure using some fitted models (e.g. the Weibull model) (United States Environment Protection Agency (USEPA), 2012). It might be complicated and difficult for non-experts to understand quantitative guides for such restrictions.

Table 2. Technical issues surrounding application of the BMD method

The 4 main technical issues in application of BMD method to the risk assessment.

Main technical issues in application of BMD method to quantal response data
1. Model selection criteria
The criteria to determine the final model in the simulation, such as the model with lowest BMDL. Different criteria have been adapted between authorities, including USEPA and EFSA (食品安全委員会評価技術企画ワーキンググループ, 2018.).
2. Model exclusion criteria
The criteria to specify models to be excluded, such as goodness-of-fit test and BMD/BMDL ratio. Different criteria have been adapted between authorities, including USEPA and EFSA (食品安全委員会評価技術企画ワーキンググループ, 2018.).
3. Threshold (BMR)
The extra risk to the basement response as the threshold of the reference value, such as 10%. 10% BMR were empirically used, but this threshold is never objective (Sand et al., 2002; Weterings et al., 2016; Allen et al., 1994b).
4. Parameter restriction
The constraint on the range of the parameters in some of the mathematical models. Using constrains broadly considered to be consistent with biological processes (USEPA, 2012).

Although several technical problems exist, we believe that the lack of uniform guidelines that specify the steps required to scrutinize fitting results and identify a single BMDL value for determining ADI must be the biggest obstacle to the wider application of the BMD method in various governmental settings. To determine which candidate models should be included or excluded in the final evaluation report, an objective guidelines are required to be established. Goodness-of-fit testing and measuring arbitrarily defined marker of fit, e.g. the ratio of BMD to BMDL had been attempted to be established as part of model exclusion criteria (Hoeting et al., 1999; Shao et al., 2014; World Health Organization (WHO) & Inter-Organization Programme for the Sound Management of Chemicals, 2009). Nevertheless, we have not consistently practiced model exclusion criteria (i.e. sometimes it is not conducted) and the exclusion criteria have not been verified and/or harmonized across different studies. Further, more model selection methods have been discussed and developed by elsewhere (Burnham et al., 2004; Hjort et al., 2003) than exclusion methods. One conservative example of such methods is to use the modeling result that yields the lowest BMDL among all the fitted models (Sand et al., 2002; Sand et al., 2008), which was recommended by EFSA document issued in 2009 (the European Food Safety Authority (EFSA), 2009). However, the broadest uncertainty might be accompanied using the model with the lowest BMDL; i.e. such method is likely to produce a relatively wider confidence interval caused by a bad fit (e.g. even a fitted model can be rejected by Pearson's chi-squared test) (Sand et al., 2002). As alternative ways of measuring goodness-of-fit and selecting the model, the AIC (Akaike information criteria) (Burnham et al., 2004; Akaike, 1978) or BIC (Bayesian information criteria) (Burnham et al., 2004) could be used, both of which gives the lowest value when the model fitted the best. However, it is not theoretically supported by having the lowest AIC that an appropriate BMDL would not be yielded using the goodness-of-fit of the model around the low dose

response (Sand et al., 2002; Burnham et al., 2004). As a possible solution to address the uncertainties associated with the use of mathematical models, model averaging method has been proposed by previous studies to explain the dose-response data (Kang et al., 2000; Wheeler, 2008; Fletcher et al., 2012). A document issued by EFSA recommends the selection of multiple models with a certain value that indicates closer AIC values (within ± 2) and averaging the results from all the selected models (the European Food Safety Authority (EFSA) Scientific Committee et al., 2017). Various type of dose-response data under risk assessment settings has employed model averaging method (Wheeler et al., 2007; Wheeler et al., 2013; Wheeler et al., 2009), but a standard application method of model averaging has yet to be established, including the use of badly fitted model for averaging (e.g. model averaging over all converged models or averaging over well-fitted models only).

Although all issues described above might be a big obstacle to use BMD method for quantal response data properly, some of which cannot be fully and immediately resolved, a simulation-based evaluation might be a useful way to identify a possible well-performing pathway of model exclusion and selection. We have conducted a simulation study to compare the performance of each and various combinations of model exclusion and selection criteria to help the formulation of a unified technical guidelines for risk assessment practices for food safety in Japan, by applying to three qualitatively different types of quantal response datasets.

Materials and Methods

Quantal response data

Three datasets were selected that are qualitatively different in this simulation-based assessment; i.e., (i) a dataset with frequent testing at doses with high response rates, (ii) a dataset with frequent testing at doses with low response rates, and (iii) a dataset with doses involving both high and low response rates. Specifically, we retrieved the data from animal experiments with (i) 1-aminoanthraquinone with an outcome of eosinophilic droplet in proximal tubular epithelium in kidney in male rats (National Institute of Health Sciences (NIHS)), (ii) 2-ethylhexyl vinyl ether with an outcome of centrilobular hypertrophy in liver stem cells in male rats (National Institute of Health Sciences (NIHS)), and (iii) acrylamide with an outcome of axon degeneration in peripheral nerve in male rats (National Toxicological Program, 2012) as dataset (i), (ii), and (iii), respectively. In this study, the biological properties of the experimental results or interpretations for toxicological assessment were not of our interest, rather the qualitative patterns of the observed dose-response curves are of our focus when selecting dataset manually. The sample size for each determined dose was n=13, 6, and 48 each dataset, and the original study examined responses at 4, 4, and 5 different doses in order (thus involving a total of 52, 24 and 240 exposed animals in datasets (i), (ii), and (iii), respectively).

The BMD method was employed to analyze the quantal datasets using mathematical models with the total of nine different distributions that consist of 2-4 unknown parameters (EFSA, 2017). Here are some of the properties they have: always returning positive value, monotonic (always increase/decrease), suitable

for dose-response data limit to the maximum response, and established their position as they able to describe dose-response datasets by historical review (Slob and Setzer, 2014). Scale and shape parameters describing dose-response curve are generally included in all models.

The best-fit model was first identified per each dataset by selecting the model with the lowest AIC value, without imposing any parameter restrictions and without excluding any models in advance as of model selection. We used nine statistical models (shown below) in this study:

Logistic model: $\frac{1}{1+\exp(-a-bx)}$,

Probit model: $\Phi(a+bx)$,

Log-logistic model: $g+\frac{1-g}{1+\exp(-b-c\log(x))}$,

Log-probit model: $g+(1-g)\Phi\big(b+c\log(x)\big)$,

Gamma model: $g+(1-g)\frac{1}{\Gamma(a)}\int_0^{bx}(t^{a-1}\exp(-t))dt$,

Weibull model: $g+(1-g)\Big(1-\exp(-ax^b)\Big)$,

Multistage (quadratic) model: $g+(1-g)\exp(-ax-bx^2)$,

Multistage (cubic) model: $g+(1-g)\exp(-ax-bx^2-cx^3)$,

Quantal-linear model: $g+(1-g)\exp(-ax)$,

where a, b, and c represent unknown parameters, g is also an unknown parameter but it is used to represent the baseline response value (i.e. the response when dose 0) for $0 \leq g < 1$, x is the dose, $\Phi(x)$ is the cumulative distribution function of the standard normal distribution at dose x, and $\Gamma(x)$ is the gamma function at dose x.

During the simulations, the identified best-fit model for each chemical substance was regarded as the "reference model". The known lower bound of the benchmark dose with response level at 10% (i.e. unbiased $BMDL_{10}$) was calculated following to such a true

model, which derived from the maximum likelihood estimates of the parameters.

Statistical estimation

The statistical estimation was conducted using the maximum likelihood method, and the likelihood function was defined under the assumption that the quantal response data at a given dose follows a binomial distribution. Bootstrapping method (mentioned following section) was employed for computation of the 95% confidence interval (CI), including BMDL and BMD upper bound (BMDU) (i.e. one-sided upper 95% CI of BMD).

Specifically, the models were fit to the observational data in the following way. First, we identify initial values of parameters in each model which fit to the observational dose-response data either by hand calculation or BMDS version 2.7, an open-source software developed by USEPA providing us with easy access to numerous mathematical models that help risk assessors estimate the quantitative relationship between a chemical dose and the test subject's response. Due to our assumption that the quantal response data at a given dose follows a binomial distribution, the likelihood function for model i are formulated as follows:

$$L_i \coloneqq \prod_d {\pi_d}^{r_d}(1-\pi_d)^{n_d-r_d}$$

where π_d is the response rate calculated by the fitted dose-response curve of model i, r_d is the number of animals that responded positive among n_d, the number of animals used in the animal experiment at a dose d. negative log-likelihood function was used to simplify the calculation in the following way:

$$L_i \coloneqq -\sum_d (r_d\, log(\pi_d) + (n_d - r_d)\log(1-\pi_d)$$

Then, 2-steps optimizations were employed to get more concise values of the model parameters that we used in this study. Generally, optimization in a mathematical context is a mathematical technique to select the best element (Anonymous, 2014). In this analysis, a set of initial values was required, and a more appropriate set of values was expected to be produced. First optimization was conducted using Nelder-Mead method. Nelder-Mead method is an algorithm solving unrestricted parameter optimization issues that is widely employed in historical studies (Gao and Han, 2010; Nelder and Mead, 1965). We used the outputs identified by either BMDS or hand calculation as the initial values for the first optimization using Nelder-Mead method. Next, the second optimization was conducted using the limited memory BFGS (L-BFGS-B) method (Liu and Nocedal, 1989). It is a quasi-Newton method to optimize the large-scale optimization (Liu and Nocedal, 1989). Similarly, we used the outcomes of the first optimization (i.e. the results of the values of the parameters using Nelder-Mead method) as the initial values for the second optimization using L-BFGS-B method.

To calculate faster with less burden, we tried to compute gradients per model using wxMaxima version 17.10.1, a GUI of Maxima, free software computer algebra system. We calculated gradient to support the calculation by applying the gradient value to the optimization parameter since all models can be differentiable by any of parameters. However, we could not define gradient in Probit and Log-probit model due to the limitation on R to express the differential. Therefore, we decided not to compute gradient calculation to unify the application criteria.

Bootstrapping

We performed case resampling using the Monte Carlo algorithm generated either by data bootstrapping, parametric bootstrapping or profile likelihood. Briefly, data bootstrapping is the resampling method generating the 'artificial' experimental data to build confidence intervals (Joshi et al., 2006). Parametric bootstrapping is another resampling method on parameters of mathematical models by applying covariance matrix between parameters (Ripley, 1987). Finally, profile likelihood focuses on parameters of interest, which are the ones we want to estimate, by estimating the nuisance parameters, those other than parameters of interest, using MLE method given the parameters of interest fixed (Murphy and Van Der Vaart, 2000). Compared to the full likelihood function (examples described above section), this approach is beneficial in handling higher-dimensional parameters (Murphy and Van Der Vaart, 2000).

Profile likelihood was examined to decrease the volume of calculation. One of the parameters were replaced by using the definition formulation of BMD: $r = f(d)$, where d is the BMD, $f(d)$ is one in nine mathematical models shown above at dose d, and r is BMR. Here, calculation of d is of our interest, and hence the other parameters (i.e. a,b,c,g, etc.) are the parameters of nuisance. As a result of replacement, new formulations were generated describing the dose-response curve including dose d describing each model. However, some of the parameters in the simulation did not converged or generated unreasonable results due to the smaller sample sizes in the original datasets, therefore the use of multivariate normal distribution was not fully supported.

Parametric bootstrapping was also examined using multivariate normal distribution. Mean and sigma were identified by the values

of the parameters in the initial fitting and the covariance matrix. However, some simulation resulted in a too conservative confidence interval (CI) (Data not shown).

Therefore, data bootstrapping sounded to be the best approach to adapt any type of dose-response data in our main simulation study. Parametric bootstrapping and profile likelihood follow asymptotic normality, which is the theorem that it is more likely that the original distribution can be approximated to normal distribution when it has greater number of the sample size (Mammen, 1992; White, 1982). Therefore, the broader CI of the parameters in the models caused by the experimental dataset with a limited number of samples would induce bias against the calculation of the confidence interval (CI) of BMD, including BMDL and BMDU.

Simulation-based evaluation

A simulation-based assessment of model performance was conducted using the three "reference models" with three different dose response curves. Briefly, our analysis goes by: (i) identification of a reference model for each dataset by AIC (described previous section), (ii) generation of a total of 1,000 simulated datasets (each dataset includes fittings by 9 individual model) from the reference model, (iii) application of model exclusion criteria if available, (iv) application of one of the model selection criteria including methods using model averaging, and select or calculate one of the representative BMDL value from each dataset, and (v) the BMDL value was evaluated in two aspects, the validity and the reliability.

Because of the statistical estimation described above, we assumed we knew the unbiased BMD_{10} and unbiased $BMDL_{10}$ value specifically

to the experimental dataset respectively that should be recovered through the simulation (i.e. they have to be calculated using the simulated datasets). Specifically, a total of 1,000 simulated datasets from the reference model was randomly generated (**Figure 2**).

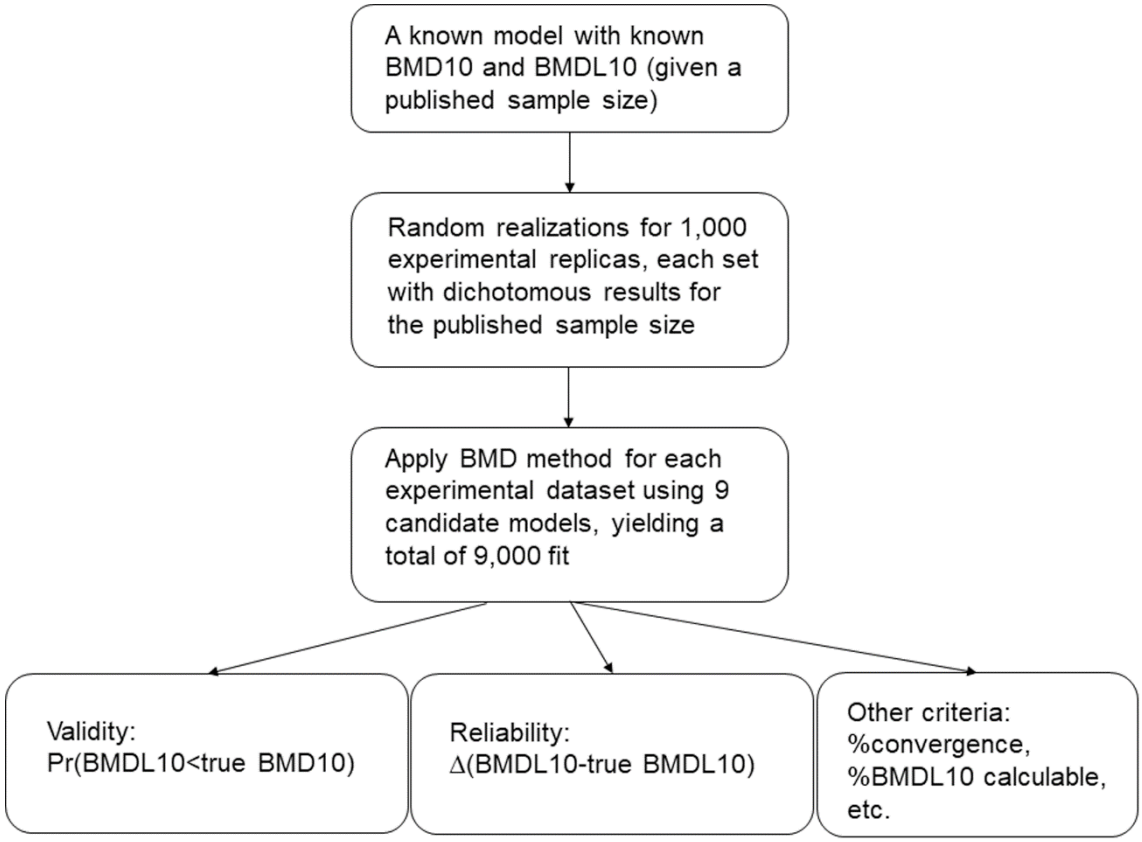

Figure 2. Simulation-based assessment of model selection criteria using the benchmark dose (BMD) method.

We used three different datasets, 1-aminoanthraquinone, 2-ethylhexyl vinyl ether, and acrylamide. For each dataset, estimates for unbiased BMD and BMDL values were obtained for the benchmark dose which were 10% of the benchmark response (BMD_{10}) and the lower bound of the benchmark dose with response level at 10% (unbiased $BMDL_{10}$). We performed random realizations for generating 1,000 replicas of the experimental data and applied the BMD method with nine models for each replica, which generated 9,000 different model fits in total that were to be evaluated.

We randomly generated the response outcome data in an assumption that the response outcome follows a binomial distribution for each examined exposure dose for the number of samples that were originally allocated for the given dose (i.e. n=13, 6, and 48 dichotomous responses in each observation dose for the substances in datasets (i), (ii), and (iii)). For each replicated dataset, a total of nine standard distributions of the BMD method were fitted and examined whether the appropriate $BMDL_{10}$ value (i.e. unbiased $BMDL_{10}$ that was identified in previous section) could be recovered. To recover $BMDL_{10}$ values, different combinations of model exclusion and selection criteria were imposed, which allowed us to assess which combination of criteria would likely produce a valid and reliable estimate of $BMDL_{10}$. Model averaging is also an option for model selection criteria. To quantify how much valid and reliable the simulated criteria were, we evaluated the performance as follows:

(i) Validity

Because the statistical role of BMDL is to act as the one-sided 95% lower bound of BMD, the simulated $BMDL_{10}$ value must be lower than the unbiased BMD_{10}., the validity was measured in a following way, as the number of simulations that the simulated $BMDL_{10}$ value

was lower than the unbiased BMD_{10} out of the total of 1,000 simulations for each chemical substance:

$$Validity := \frac{1}{1000}\sum_{i=1}^{1000} I_{ij} \times 100\ (\%),$$

$$I_{ij} = \begin{cases} 1 & \text{if } l_{ij} < B \\ 0 & \text{if } l_{ij} \geq B \end{cases},$$

where l_{ij} is the $BMDL_{10}$ value based on the i-th simulated data and determined model exclusion and selection criteria j used, and B is the unbiased BMD_{10} value.

(ii) Reliability

Similar results must be reproduced by repeating the same experiments for the criteria to be reliable. That is, the simulated $BMDL_{10}$ value must be close to the unbiased $BMDL_{10}$ value, and a bad combination (i.e. unreliable) criteria can be the one that yield a "distant" $BMDL_{10}$ value from the unbiased one. Reliability was measured quantitatively as the relative distance from the unbiased $BMDL_{10}$ as

$$Reliability := \frac{1}{1000}\sum_{i=1}^{1000} \frac{(l_{ij}-L)^2}{L},$$

where L is the unbiased $BMDL_{10}$ value.

The calculability of the BMDL value was also assessed in addition to validity and reliability. This is the proportion of simulated datasets that yielded convergence and thus the BMDL value out of the total simulated datasets (1,000). Moreover, the calculability that survived the model exclusion process was also generated to remove the impact of substantial exclusions before model selection on the calculability assessment. Lastly, as a potential pitfall of simulation-based studies, we have to keep in mind that we could be likely to recover the original true model more often than other models, especially when a higher number of samples is tested for each dose. To avoid overoptimistic interpretation of

the simulated results, the proportion of the same statistical model out of the nine candidate models that was recovered to be identical to the original (i.e. reference) model among 1,000 simulated datasets was tallied (i.e. proportion which the statistical model having selected $BMDL_{10}$ value by combination of criteria is identical to the reference model per each experimental dataset) under the name of “true dose-response”. If the selected statistical model was the same as the reference model, the corresponding result may be more likely to be recovered due to the computational nature of the simulation study (e.g. random simulations using the Weibull model may lead to the choice of the Weibull model in each simulation run).

Model exclusion and selection criteria

A total of four possible model exclusion criteria and six possible model selection criteria were considered. Avoiding excessive combinations of the two (i.e. multiple exclusion criteria plus model averaging over preferred models only), a total of 18 possible combinations were tested and compared.

The four model exclusion criteria were (i) no exclusion, (ii) implementing goodness-of-fit testing using the Kolmogorov-Smirnov test (KS test) to avoid models with $p < 0.10$, (iii) KS test to exclude models with $p < 0.10$ and also exclusion of models with the BMD/BMDL ratios >10 (Massey, 1951), and (iv) KS test to exclude models with $p < 0.10$ and also exclusion of models with BMDU/BMDL ratios >10 (Massey, 1951). The KS test were employed rather than Pearson’s chi-squared or Fisher’s exact test because the experimental sample sizes were very small, which might cause the technical issues in calculating test statistics (Massey, 1951; Justel et al., 1997; David et al., 1948). We excluded ratios of BMD/BMDL and BMDU/BMDL >10 because such models

with either of ratios that exceed 10 have been regarded as imprecise models enough to yield a proper confidence limit (Muri et al., 2009; Moffat et al., 2015; Matsumoto et al., 2019). We used the models that survived these exclusion procedures in the model selection process.

Among the six model selection criteria, three were single selection criteria and three were model averaging methods. The single selection criteria include: (i) select the model with the lowest BMDL value to be conservative as part of risk assessment practice (Lowest BMDL) (Sand et al., 2002; Sand et al., 2008); (ii) select the model with the lowest BMD value, not necessarily relying on the lower uncertainty bound (Lowest BMD) (Shao et al., 2014); or (iii) select the model with the lowest AIC value as the best fit model (Lowest AIC) (Shao et al., 2014). Model averaging results were also computed by averaging all or part of the fitted models for each resampled data instead of taking the average BMDL value. The model uncertainty was considered in using model averaging by integrating results from all or part of selected models (Shao et al., 2011; Walter et al., 2013; Bailer et al., 2005; Morales et al., 2006). Three different patterns of model averaging were candidates: (i) model averaging over all nine models (MA-all) (Wheeler et al., 2007); (ii) model averaging over three models with the lowest three AIC values (MA-3) (Wheeler et al., 2007); and (iii) model averaging over all models that yielded AIC values within 3 of the lowest AIC value (MA-AIC). Let $\pi_i(d)$ the dose-response curve of i-th model and d the given dose, MA-all was calculated as

$$\pi_{\mathrm{MAall}}(d) = \sum_{i=1}^{9} w_i \pi_i(d) \text{ where } w_k = \frac{\exp(-I_k/2)}{\sum_{i=1}^{9} \exp(-I_i/2)}$$

where I_k is the AIC value of model k.

We calculated MA-3 using the same formula with normalization above, over the three best-fit models as judged by AIC (i.e. three models with lowest AIC were selected and put into the weight

function above). Similarly, we computed MA-AIC using the arithmetic average of models (i.e. averaging without weight function) and adhering to rules of thumb, averaging all models with AIC within +3 of the lowest AIC of the best-fit model (Hjort et al., 2003). We did not use the weight function above for MA-AIC, because we regarded models with similar AIC values as equally well-fitted models. MA-3 and MA-AIC were intended to conduct averaging over well-fitted models alone compared with MA-all, therefore combinations of model exclusion with MA-3 or MA-AIC were not examined to avoid similar removal of bad-fit models multiple times (i.e. MA-3 and MA-AIC has already exclude models that were bad-fit).

All calculations were conducted using R version 3.5.0 (CRAN) or versions above.

Results

Quantal response data

Table 3 listed the validity and reliability of the simulation results for the 1-aminoanthraquinone dataset, which contained frequent testing at doses with high response rates. $BMDL_{10}$ was 0.92 and BMD_{10} was 7.67 under the selection of Probit model as the reference model with the lowest AIC value (**Figure 3**), and we performed resampling-based simulations given the results. The best validity result was yielded using lowest BMDL or lowest BMD, except when applying the exclusion criteria of the KS test and BMDU/BMDL ratio >10 in advance of model selection. The best reliability result was yielded using lowest BMD following model exclusion criteria of both the KS test and BMD/BMDL ratio. Although the lowest AIC had a best "true dose-response" of approximately 1/3 of the simulation results were produced by selecting Probit model which was identical to the reference model, it is the worst selection criteria in terms of validity and reliability. Model averaging results yielded intermediate ranks among all model exclusion and selection criteria, and MA-3 produced the best reliability and validity results among the model averaging techniques.

Similarly, **Table 4** shows the simulation results for the 2-ethylhexyl vinyl ether dataset, which contained frequent testing at doses with low response rates. $BMDL_{10}$ was 24.69 and BMD_{10} was 28.65 under the selection of the Probit model as the unbiased model (**Figure 3**). Highest validity was produced under the selection criteria of the lowest BMDL or the lowest BMD whether or not any of model exclusion criteria was applied in advance. The best reliability performance was observed using model averaging, especially MA-3. Validity was not improved by applying any of model exclusion criteria, rather the reliability estimates of MA-all were decreased when applied. Calculability of BMDL was

not changed by any of model exclusion criteria applied, and they only improved a little in the reliability of MA-all.

Table 5 shows the simulation results for the acrylamide dataset, which contained doses involving both high and low response rates. $BMDL_{10}$ was 0.79 and BMD_{10} was 0.94 under the selection of the Logistic model as the unbiased model (**Figure 3**). The highest validity was observed using the lowest BMDL for model selection criteria, and MA-3 had a best reliability. The Logistic model, the unbiased dose-response curve for acrylamide, was selected for about every 1 in 3 selected models when selection by lowest AIC. Since this dataset had a larger sample size than the other two datasets, the models did successfully converge at a higher frequency than those in the other datasets and were rarely excluded by the exclusion criteria, especially when applying BMD/BMDL or BMDU/BMDL ratio.

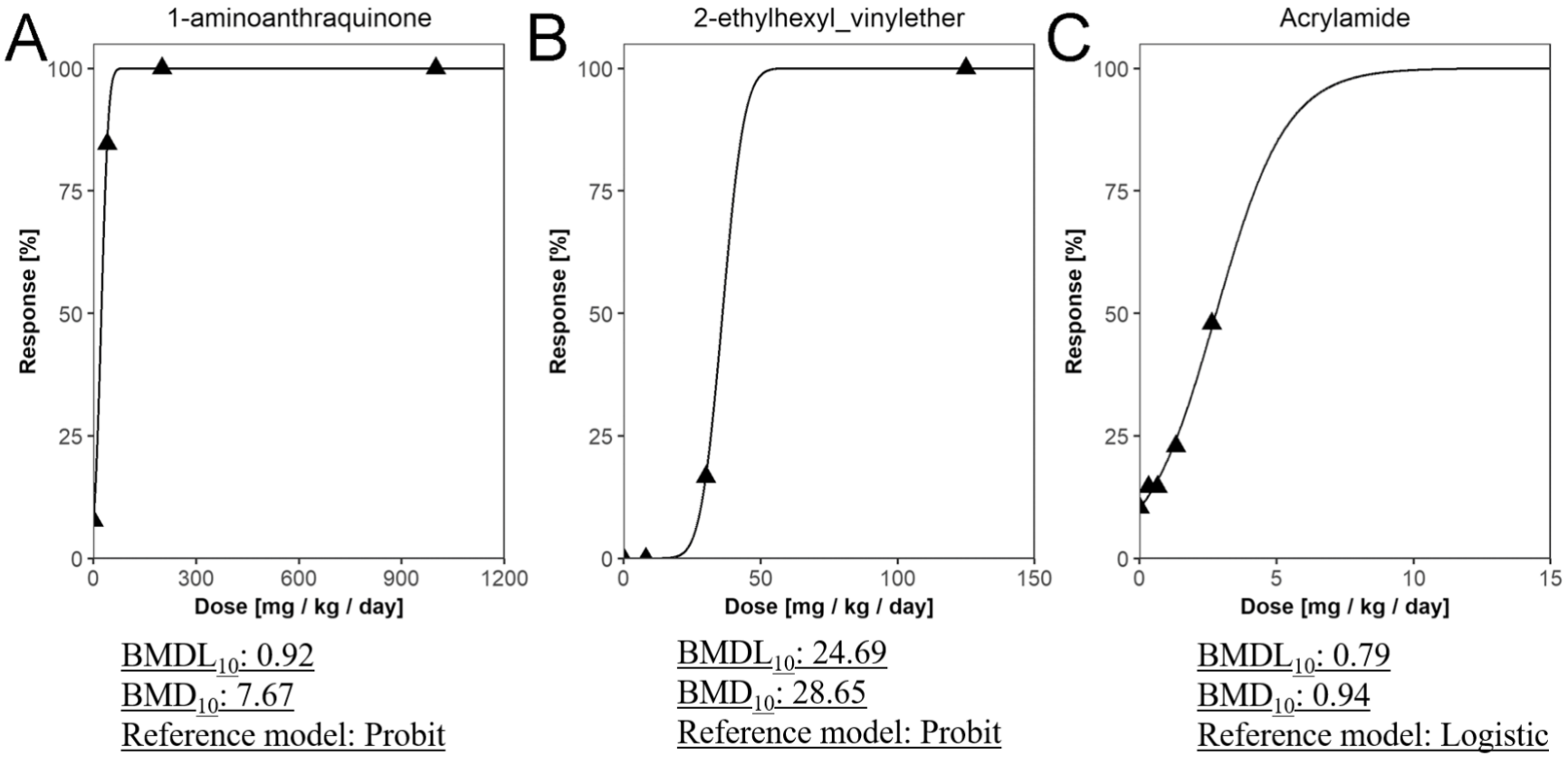

Figure 3. Observed and predicted dose-response relationships for the 1-aminoanthraquinone, 2-ethylhexyl vinyl ether, and acrylamide datasets.

(A) 1-aminoanthraquinone with a substantial weight applied for doses with high response rates (n=13 per observed dose). (B) 2-ethylhexyl vinyl ether with a weight applied for doses with low response rates (n=6 per dose). (C)

Acrylamide with an approximately linear dose-response relationship (n=48 per dose). Original outcomes were eosinophilic droplet in renal proximal tubular epithelium in male rats for 1-aminoanthraquinone, centrilobular hypertrophy in liver stem cells in male rats for 2-ethylhexylvinyl ether, and axon degeneration in peripheral nerve in male rats for acrylamide. The best fit models selected using only the Akaike information criterion were the Probit model for 1-aminoanthraquinone and 2-ethylhexyl vinyl ether, and the Logistic model for acrylamide. Unbiased $BMDL_{10}$ and BMD_{10} were estimated as 0.9 and 7.7 respectively, for 1-aminoanthraquinone, 24.7 and 28.7, respectively, for 2-ethylhexylvinyl ether, and 0.8 and 0.9, respectively, for acrylamide.

Table 3. Simulation results for the 1-aminoanthraquinone dataset using the benchmark dose method (reference model: Probit)

[a]Exclusion criteria: KS, Kolmogorov-Smirnov test of goodness-of-fit; BMD/BMDL, ratio of benchmark dose (BMD_{10}) to benchmark dose lower bound ($BMDL_{10}$) with values >10 excluded; BMDU/BMDL, ratio of benchmark dose upper bound ($BMDU_{10}$) to $BMDL_{10}$ with values >10 excluded. [b]Model selection criteria: Lowest BMDL, model with the lowest value of $BMDL_{10}$; Lowest BMD, model with the lowest value of BMD_{10}; Lowest AIC, model with the lowest AIC value; MA-all, model averaging of all converged models; MA-3, model averaging of three models with the three lowest AIC values; MA-AIC, model averaging of all models with AIC values <3 compared with the best model that yielded the minimum AIC. [c]Reliability (Mean distance), measured as the mean distance between unbiased $BMDL_{10}$ and calculated $BMDL_{10}$ followed by rank. [d]Validity (%), measured as the iterations that satisfied calculated $BMDL_{10}$ lower than unbiased BMD_{10} followed by rank. [e]BMDL calculability (%), measured as the iterations that yielded BMDL in the model selection criterion. [f]Non-exclusion and BMDL calculation (%), measured as the iterations that yielded BMDL in the model selection criterion along with exclusion criteria. [g]True dose response (%), measured by the default model selected by the model selection criterion. Note: Validity (%), BMDL calculability (%), non-exclusion and BMDL calculation (%), and true dose response (%) were converted into rates of iterations divided by 9000, nine models in 1000 simulation data. NA, not applicable.

Exclusion criteria[a]	Selection[b]	Reliability (Mean distance)[c]	Rank	Validity[d] (%)	Rank	BMDL calculability[e] (%)	Non-exclusion and BMDL calculation[f] (%)	True dose-response[g] (%)
None	Lowest BMDL	0.4	5	100.0	1	95.6	95.6	0.1
	Lowest BMD	0.3	2	100.0	1	95.6	95.6	0.1
	Lowest AIC	120.9	15	88.4	15	95.6	95.6	34.1
	MA-all	6.2	9	99.6	8	95.6	95.6	NA
	MA-3	4.7	7	99.8	7	100.0	100.0	NA
	MA AIC <3	9.0	11	98.8	11	100.0	100.0	NA
KS	Lowest BMDL	0.4	5	100.0	1	95.6	95.6	0.1
	Lowest BMD	0.3	2	100.0	1	95.6	95.6	0.1
	Lowest AIC	120.9	15	88.4	15	95.6	95.6	34.1
	MA-all	6.1	8	99.6	8	95.6	95.6	NA
KS, BMD/BMDL	Lowest BMDL	0.3	4	100.0	1	95.6	79.1	0.5
	Lowest BMD	0.2	1	100.0	1	95.6	79.1	0.5
	Lowest AIC	121.0	17	88.4	15	95.6	79.1	38.4
	MA-all	6.3	10	99.3	10	95.6	79.1	NA
KS, BMDU/BMDL	Lowest BMDL	27.2	13	91.0	12	95.6	49.0	18.4
	Lowest BMD	27.1	12	91.0	12	95.6	49.0	18.4
	Lowest AIC	148.3	18	79.4	18	95.6	49.0	35.3
	MA-all	33.9	14	90.2	14	95.6	49.0	NA

Table 4. Simulation results for the 2-ethylhexyl vinyl ether dataset (reference model: Probit)

[a]Exclusion criteria: KS, Kolmogorov-Smirnov test of goodness-of-fit; BMD/BMDL, ratio of benchmark dose (BMD_{10}) to benchmark dose lower bound ($BMDL_{10}$) with values >10 excluded; BMDU/BMDL, ratio of benchmark dose upper bound ($BMDU_{10}$) to $BMDL_{10}$ with values >10 excluded. [b]Model selection criteria: Lowest BMDL, model with the lowest value of $BMDL_{10}$; Lowest BMD, model with the lowest value of BMD_{10}; Lowest AIC, e model with the lowest AIC value; MA-all, model averaging of all converged models; MA-3, model averaging of three models with the three lowest AIC values; MA-AIC, model averaging of all models with AIC values <3 compared with the best model that yielded the minimum AIC. [c]Reliability (Mean distance), measured by the mean distance between unbiased $BMDL_{10}$ and calculated $BMDL_{10}$ followed by rank. [d]Validity (%), measured as the iterations that satisfied calculated $BMDL_{10}$ lower than unbiased BMD_{10} followed by rank. [e]BMDL calculability (%), measured as the iterations that yielded BMDL in the model selection criterion. [f]Non-exclusion and BMDL calculation (%), measured as the iterations that yielded BMDL in the model selection criterion along with exclusion criteria. [g]True dose response (%), measured by the default model selected by the model selection criterion. Note: Validity (%), BMDL calculability (%), non-exclusion and BMDL calculation (%), and true dose response (%) were converted into rates of iterations divided by 9000, nine models in 1000 simulation data. NA, not applicable.

Exclusion[a]	Selection[b]	Reliability[c] (Mean distance)	Rank	Validity[d] (%)	Rank	BMDL calculability[e] (%)	Non-exclusion and BMDL calculation[f] (%)	True dose-response[g] (%)
None	Lowest BMDL	17.8	7	100.0	1	85.2	85.2	0
	Lowest BMD	17.8	7	100.0	1	85.2	85.2	0
	Lowest AIC	20.7	15	66.7	9	85.2	85.2	33.4
	MA-all	6.2	6	66.7	9	85.2	85.2	NA
	MA-3	3.6	1	66.7	9	100.0	100.0	NA
	MA AIC<3	5.1	2	66.7	9	100.0	100.0	NA
KS	Lowest BMDL	17.8	7	100.0	1	85.2	85.2	0
	Lowest BMD	17.8	7	100.0	1	85.2	85.2	0
	Lowest AIC	20.7	15	66.7	9	85.2	85.2	33.4
	MA-all	6.2	5	66.7	9	85.2	85.2	NA
KS, BMD/BMDL	Lowest BMDL	17.8	7	100.0	1	85.2	85.2	0
	Lowest BMD	17.8	7	100.0	1	85.2	85.2	0
	Lowest AIC	20.7	15	66.7	9	85.2	85.2	33.4
	MA-all	6.1	4	66.7	9	85.2	85.2	NA
KS, BMDU/BMDL	Lowest BMDL	17.8	7	100.0	1	85.2	85.2	0
	Lowest BMD	17.8	7	100.0	1	85.2	85.2	0
	Lowest AIC	20.7	15	66.7	9	85.2	85.2	0
	MA-all	6.1	3	66.7	9	85.2	85.2	33.4

Table 5. Simulation results for the acrylamide dataset (reference model: Logistic)

[a]Exclusion criteria: KS, Kolmogorov-Smirnov test of goodness-of-fit; BMD/BMDL, ratio of benchmark dose (BMD_{10}) to benchmark dose lower bound (BMDL10) with values >10 excluded; BMDU/BMDL, ratio of benchmark dose upper bound ($BMDU_{10}$) to $BMDL_{10}$ with values >10 excluded. [b]Model selection criteria: Lowest BMDL, model with the lowest value of $BMDL_{10}$; Lowest BMD, model with the lowest value of BMD_{10}; Lowest AIC, model with the lowest AIC value; MA-all, model averaging of all converged models. MA-3, model averaging of three models with three smallest AIC values. MA-AIC, model averaging of all models with AIC values <3 compared with the best model that yielded the minimum AIC. [c]Reliability (Mean distance), measured as the mean distance between unbiased $BMDL_{10}$ and calculated $BMDL_{10}$ followed by rank. [d]Validity (%), measured as the iterations that satisfied calculated $BMDL_{10}$ lower than unbiased BMD_{10} followed by rank. [e]BMDL calculability (%), measured as the iterations that yielded BMDL in the model selection criterion. [f]Non-exclusion and BMDL calculation (%), measured as the iterations that yielded BMDL in the model selection criterion along with exclusion criteria. [g]True dose response (%), measured as the default model selected by the model selection criterion. Note: Validity (%), BMDL calculability (%), non-exclusion and BMDL calculation (%), and true dose response (%) were converted into rates of iterations divided by 9000, nine models in 1000 simulation data. NA, not applicable.

Exclusion[a]	Selection[b]	Reliability[c] (Mean distance)	Rank	Validity[d] (%)	Rank	BMDL calculability[e] (%)	Non-exclusion and BMDL calculation[f] (%)	True dose-response[g] (%)
None	Lowest BMDL	0.4	17	99.9	1	99.8	99.8	0.0
	Lowest BMD	0.3	13	99.7	5	99.8	99.8	0.1
	Lowest AIC	0.1	2	89.0	16	99.8	99.8	38.3
	MA-all	0.2	10	98.2	11	99.8	99.8	NA
	MA-3	0.1	1	93.9	14	100.0	100.0	NA
	MA AIC<3	0.2	8	97.8	13	100.0	100.0	NA
KS	Lowest BMDL	0.4	17	99.9	1	99.8	99.8	0.0
	Lowest BMD	0.3	13	99.7	5	99.8	99.8	0.1
	Lowest AIC	0.1	2	89.0	15	99.8	99.8	38.3
	MA-all	0.2	9	98.4	10	99.8	99.8	NA
KS, BMD/BMDL	Lowest BMDL	0.3	16	99.9	1	99.8	98.7	0.0
	Lowest BMD	0.3	12	99.7	5	99.8	98.7	0.2
	Lowest AIC	0.1	5	88.7	17	99.8	98.7	38.7
	MA-all	0.2	7	98.7	9	99.8	98.7	NA
KS, BMDU/BMDL	Lowest BMDL	0.3	15	99.9	1	99.8	93.9	0.0
	Lowest BMD	0.2	11	99.6	8	99.8	93.9	0.2
	Lowest AIC	0.1	4	88.5	18	99.8	93.9	39.0
	MA-all	0.2	6	98.2	11	99.8	93.9	NA

Discussion

The BMD method is now widely used to determine the reference dose for toxicological risk assessment in food chemicals, agricultural chemicals, and environmental hazards. However, several ambiguous parts of model assessment are often concern for governmental, especially the model exclusion and selection processes. As part of the technical assessment for possible improvements in the guidelines, a simulation-based experiment was conducted to assess the model exclusion and selection process by comparing the validity, reliability, and other model performance indicators using all possible combinations of model exclusion and selection criteria. For the exposition, three different empirical datasets were examined, each of which had different characteristics of the dose-response pattern (i.e. the datasets had either rich information about high or low response rates, or approximately linear dose-response pattern). By replicating 1,000 sets of hypothetical experimental data computationally in a random manner, the best criteria of model exclusion and selection were identified to be different across the chemical substances in each dataset. Further, the best criteria achieving good validity were not necessarily the best for ensuring good reliability. For instance, to achieve higher validity, the lowest BMDL outperformed the other criteria, but did not always yield the best reliability. The lowest AIC yielded the best reliability result for the acrylamide dataset, but for the 1-aminoanthraquinone dataset it produced the worst reliability. Besides, the model averaging results always ranked at an intermediate level both in validity and reliability among all possible criteria, and did not yield the worst results at all.

There are two take-home messages we would like to highlight. First, although the best exclusion and selection criteria for

the qualitatively differently distributed datasets could not be identified, it is discovered that model averaging over three models with the lowest three AIC values (MA-3) did not yield the worst result, and the best results among all the model averaging results were produced by MA-3 without prior model exclusion. For instance, the best reliability result was yielded by MA-3 for the 2-ethylhexyl vinyl ether dataset. If we need a uniform guideline to implement model exclusion and selection, MA-3 would be suggested in our results as the recommended option whenever applicable. Second, it was turned out that model exclusion using the KS test and the ratios of BMD or BMDU to BMDL did not necessarily yield better validity and reliability than non-exclusion. By imposing any of exclusion criteria, both for validity and reliability results were even worse using 1-aminoanthraquinone dataset in particular. For example, reliability (mean distance) of Lowest BMDL has been increased from 0.4 to 27.2 as compared with non-exclusion by applying the exclusion criteria of KS test and the ratio of BMDU and BMDL (Table 3). Furthermore, validity (rate of "successful" calculation) of MA-all has also got worse: it decreased from 98.8 without exclusion to 90.2 as applied of KS test and the ratio of BMDU and BMDL (Table 3). Thus, visual assessment might be enough in model exclusion to proceed to model selection.

Model averaging has previously been demonstrated as a useful option to determine the point of departure (Wheeler et al., 2007), especially for datasets that do not necessarily exhibit a sigmoidal dose-response curve. It was observed that the performances of all the model averaging options that we tested were well overall. However, we still have to consider how to account for the distance metric (AIC) across different models and how to quantify model uncertainty of each parametric assumption in the process of averaging. Based on the fact that

at least nine models are fitted to the same dataset and some of them share similar properties while others do not, we have to consider which models should be averaged, e.g. averaging over all models or only some of them. It was discovered that averaging over all converged models did not necessarily yield a better performance than, considering that averaging over some of the models. The uncertainty of well-fitted models might be far smaller than those of badly fitted models. It is still needed to discuss if it is valid to use averaging over the three best models judged by AIC (e.g., averaging over two best models rather than three) and the number of models that we can use as a best model (we used 3) might change according to the total number of models to be tested (e.g. more than nine models could be tested) (Wheeler et al., 2007). However, it must be noted that reliance only on manual determination by AIC during the averaging might not be a good option given that averaging over the three best models (MA-3) outperformed all the models with close AIC values (MA-AIC). For now, MA-3 is the method that is recommended in our study, and the programing code and a package used in this procedure are to be open in the public in the future.

We have to note that the total number of converged models requires at least 3 for recommended option to work properly; indeed, we occasionally observed the convergence of one model alone. In such an instance, we have to consider other criteria including using lowest BMDL or lowest AIC, although both lowest BMDL and lowest AIC did not act as the unique best method for model selection, even if lowest BMDL can ensure good validity, which is understandable from the conservative nature of this method (i.e. this can be the best in terms of validity as it select the lowest BMDL). We have to note that the worst results for two of the datasets by (the exception was the acrylamide

dataset) were produced by the use of lowest AIC when it comes to validity and reliability.

Five technical limitations should be considered. First, only three different chemical substances were examined as source of information and qualitative differences among them were addressed. Additional insights into ranking the model selection criteria would be revealed by more number of datasets. Second, there should be a corresponding unique criterion that is best suited to its analysis if a specific dataset behaved uniquely. Our objective was to identify fairly acceptable model selection criteria across qualitatively different datasets (which found MA-3 was acceptable overall), but it was not achieved to classify dose-response curves into several different groups for better fitting. Third, computer simulation was only the method used and the relationship was not able to define and quantify between outcomes and reference model prior to simulations during the estimation process. Although we counted this bias by the index “true dose-response” in Tables 3-5, it is not known how this had impacted on out examined criteria. Fourth, parameter constraints were not explored in this study. Fifth, other percentile cutoff levels were not examined, i.e. the benchmark response was fixed at 10%.

While further research would be required to solve numerous technical issues in applying the BMD methods to risk assessment, it could be concluded that MA-3 is the best guiding option to derive the reference dose when uniform guidelines are expected to specify a single model exclusion and selection method.

Conclusion

A simulation-based experiment was conducted to assess the model exclusion and selection process as part of the technical assessment for possible improvements in the guidelines. We compared the validity, reliability, and other model performance indicators using all possible combinations of model exclusion and selection criteria. Our analysis suggested the following bullet points:

- The best model selection criteria were varied across the types of dose-response, but the model averaging over three-best AIC models (MA-3) produced better results in all datasets used.
- The model exclusion criteria did not necessarily yield better validity and reliability than non-exclusion ones.

These findings would be the first step to establish the BMD method to the toxicological risk assessment in Japan. While there are some limitations exist, if we need a uniform technical suggestion for the guideline to choose the best performing model for exclusion and selection, using MA-3 is highly recommended whenever applicable in our study. Further research would be expected on exploring different datasets to find more appropriate criteria, or technical issues such as parameter restrictions and threshold value.

Acknowledgements

First, I must express a great thank to my supervisor Hiroshi Nishiura for being a constant source of ideas, enthusiasm, and commitment. Hiroshi provided a number of insights and opportunities since I started working with him in 2018 at Hokkaido University to ensure my Ph.D. degree. He opened his laboratory for someone who tried to change approach from laboratory-based science to data-based, with basic ideas and knowledges on public health and epidemiology. His contagious energy motivated me to develop my career on the area of health using data analysis, and to keep working for 4 years on this dissertation at Hokkaido University. This dissertation would have never been finished without hours of discussions with him in the lab, and periodically while jogging.

I’ve been fortunate to publish scientific articles with numerous talented scientists: Akihiko Hirose, Andrei Akhmetzhanov, Hyojung Lee, Kaoru Inoue, Katsuma Hayashi, Kazuki Shimizu, Ryo Kinoshita, Sung-mok Jung, and Takayuki Yamaguchi. I thank all of them for so many inspirations and lessons to learn. I would also like to thank all the amazing lab mates (past and present), especially Ayako Suzuki and Shinya Tsuzuki, who were mentor to discuss and proceed this research. Moreover, I would appreciate the administrative staffs who made things easier, especially Hisae Hirama, Keiko Saito and Miwako Inagi. Finally, I’m so grateful to my family – Takashi, Naoko, Yuzuki, and Akiho – for all their love and support over the years that enable me to concentrate on Ph.D research.

Conflicts of interest

The author declares no conflict of interest.

References

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. Ann. Inst. Stat. Math. *30*(1), 9-14.

Allen, B.C., Kavlock, R.J., Kimmel, C.A., and Faustman, E.M. (1994a). Dose-response assessment for developmental toxicity II. Comparison of Generic benchmark dose estimates with no observed adverse effect levels. Fundament. Appl. Toxicol. *23*(4), 487-495.

Allen, B.C., Kavlock, R.J., Kimmel, C.A, and Faustman, E.M. (1994b). Dose-response assessment for developmental toxicity III. Statistical models. Fundament. Appl. Toxicol. *23*, 496-509.

Anonymous, 2014, "The Nature of Mathematical Programming Archived 2014-03-05 at the Wayback Machine," *Math. Program.* INFORMS Computing Society.

Bailer, A.J., Noble, R.B., and Wheeler, M.W. (2005). Model uncertainty and risk estimation for experimental studies of quantal responses. Risk Anal. *25*(2), 291-299.

Barnes, D.G., and Dourson, M. (1988). Reference dose (RfD): Description and use in health risk assessments. Regul. Toxicol. Pharmacol. *8*(4), 471-486.

Bi, J. (2010). Using the benchmark dose (BMD) methodology to determine an appropriate reduction of certain ingredients in food products. J. Food Sci. *75*(1), R9-R16.

Brown, K.G., and Erdreich, L.S. (1989). Statistical uncertainty in no-observed-adverse-effect level. Fundament Appl Toxicol. *13*, 235-244.

Burnham, K.P., and Anderson, D.R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. Sociol. Methods Res. *33*(2), 261-304.

Crump, K. (1984). A new method for determining allowable daily intakes. Fundament. Appl. Toxicol. *4*(5), 854-871.

David, F.N., and Johnson, N.L. (1948). The probability integral transformation when parameters are estimated from the sample. Biometrika. *35*(1/2), 182-190.

European Food Safety Authority (EFSA). (2009). Guidance of the scientific committee on a request from EFSA on the use of benchmark dose approach in risk assessment. EFSA J. *7*(6), 1-72.

European Food Safety Authority (EFSA) Scientific Committee, Anthony, H., Diane, B., Thorhallur, H., Michael, J.J., Katrine, H.K., Simon, M., Alicja, M., Hanspeter, N., Hubert, N., Colin, O., et al. (2017). Update: use of the benchmark dose approach in risk assessment. EFSA J. *15*(1):4658.

Fletcher, D., and Turek, D. (2012). Model-averaged profile likelihood intervals. J. Agr. Biol. Environment Stat. *17*(1), 38-51.

Gao, F., and Han, L. (2010). Implementing the Nelder-Mead simplex algorithm with adaptive parameters. Comput. Optim. Appl. *51*, 259-277.

Hjort, N.L., and Clarskens, G. (2003). Frequentist model average estimators. J. Am. Stat. Assoc. *98*(464), 879-899.

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial. Stat. Sci. *14*, 382-401.

Joshi, M., Seidel-Morgensterna, A., and Kremling, A. (2006). Exploiting the bootstrap method for quantifying parameterconfidence intervals in dynamical systems. Metab. *8*: 447-455.

Justel, A., Peña, D., and Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. Stat. Prob. Letters. *35*, 251-259.

Kang, S.H., Kodell, R.L., and Chen, J.J. (2000). Incorporating model uncertainties along with data uncertainties in microbial risk assessment. Regul. Toxicol. Pharmacol. *32*(1), 68-72.

Kimmel, C.A., and Gaylor, D.W. (1988). Issues in qualitative and quantitative risk analysis for developmental toxicology. Risk Anal. *8*(1), 15-20.

Leisenring, W., and Ryan, L. (1992). Statistical properties of the NOAEL. Regul. Toxicol. Pharmacol. *15*, 161-171.

Liu, D.C., and Nocedal, J. (1989). ON THE LIMITED MEMORY BFGS METHOD FOR LARGE SCALE OPTIMIZATION. Math. Program. *45*, 503-528.

Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. Probab. Theory Relat. Fields *93*, 439-455.

Massey, F. (1951). The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. *46*(253), 68-78.

Matsumoto, M., Hirata-Koizumi, M., Kawamura, T., Sakuratan, S., Ono, A., and Hirose, A. (2019). Validation of the statistical parameters and model selection criteria of the benchmark dose methods for the evaluation of various endpoints in repeated-dose toxicity studies. Fundament. Toxicol. Sci. *6*(4), 125-136.

Morales, K.H., Ibrahim, J.G., Chen, C.J., and Ryan, L.M. (2006). Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. J. Am. Stat. Assoc. *101*(473), 9-17.

Muri, S.D., Schlatter, J.R., and Brüschweiler, B.J. (2009). The benchmark dose approach in food risk assessment: Is it applicable and worthwhile? Food Chem. Toxicol. *47*, 2906-2925.

Murphy, S.A., and Van Der Vaart, A.W. (2000). On profile likelihood. J. Am. Stat. Assoc. *95*:450, 449-46.

Moffat, I., Chepelev, N.L., Labib, S., Bourdon-Lacombe, J., Kuo, B., Buick, J.K., Lemieux, F., Williams, A., Halappanavar, S., Malik, A., et al. (2015). Comparison of toxicogenomics and traditional approaches to inform mode of action and points of departure in human health risk assessment of benzo[a]pyrene in drinking water. Crit. Rev. Toxicol. *45*, 1-43.

National Institute of Health Sciences (NIHS). Dose-response data on 1-aminoanthraquinone. http://dra4.nihs.go.jp/BMD/RawData/BMD_82451_kidney_eosinophilic_droplet_in_proximal_tubular_epithelium.txt. (Accessed on 18 December 2019).

National Institute of Health Sciences (NIHS). Dose-response data on 2-ethylhexylvinylether. http://dra4.nihs.go.jp/BMD/RawData/BMD_103446_liver_centrilobular_hypertrophy_m.txt.com. (Accessed on 18 December 2019).

National Toxicological Program (NTP). (2012). NTP technical report on the toxicology and carcinogenesis studies of acrylamide (CAS No. 79-06-1) in F344/N rats and B6C3F1 mice (feed and drinking water studies). (Durham: National Toxicological Program).

Nelder, J.A., and Mead, R. (1965). A simplex method for function minimization. Comput. J. *7*, 308-313.

Ripley, B.D. (1987). Stochastic Simulation. (New York; Wiley).

Sand, S., Falk, F.A., and Victorin, K. (2002). Evaluation of the benchmark dose method for dichotomous data: model dependence and model selection. Regul. Toxicol. Pharmacol. *36*, 184-197

Sand, S., Victorin, K, and Filipsson, A.F. (2008). The current state of knowledge on the use of the benchmark dose concept in risk assessment. J. Appl. Toxicol. *28*, 405-421.

Shao, K., and Gift, J.S. (2014). Model uncertainty and Bayesian model averaged benchmark dose estimation for continuous data. Risk Anal. *34*(1), 101-120.

Shao, K., and Small, M.J. (2011). Potential uncertainty reduction in model-averaged benchmark dose estimates informed by an additional dose study. Risk Anal. *31*(10), 1561-75.

Slob, W., and Setzer, R.W. (2014). Shape and steepness of toxicological dose-response relationships of continuous endpoints. *Crit. Rev. Toxicol. 44*, 270-297.

United States Environment Protection Agency (USEPA). (2012). Benchmark Dose Technical Guidance. (Washington DC: United States Environment Protection Agency (USEPA)).

Walter, W.P., An, L., Wickens, A.A., West, R.W., Peña, E.A., and Wu, W. (2013). Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. Environmetrics *24*, 143-157.

Weterings, P.J.J.M., Loftus, C., and Lewandowski, T.A. (2016). Derivation of the critical effect size/benchmark response for the dose-response analysis of the uptake of radioactive iodine in the human thyroid. Toxicol. Letters. *22*, 38-43.

Wheeler, M.W. (2008). Model averaging software for dichotomous dose response risk estimation. J. Stat. Software. *26*(5), 1-15.

Wheeler, M.W., and Bailer, A.J. (2007). Properties of model-averaged BMDLs: A study of model averaging in dichotomous risk estimation. Risk Anal. *27*(3), 659-670.

Wheeler, M.W., and Bailer, A.J. (2009). Comparing model averaging with other model selection strategies for benchmark dose estimation. Environ. Ecol. Stat. *16*(1), 37-51.

Wheeler, M.W., and Bailer, A.J. (2013). An empirical comparison of low-dose extrapolation from points of departure (PoD) compared to extrapolations based upon methods that account for model uncertainty. Regul. Toxicol. Pharmacol. *67*(1), 75-82.

White, H. (1982). Maximum likelihood estimation of misspecified model. Econometrica *50*:1.

Wignall, J.A., Shapiro, A.J., Wright, F.A., Woodruff, T.J., Chiu, W.A., Guyton, K.Z., and Rusyn, I. (2014). Standardizing benchmark dose calculations to improve science-based decisions in human health assessments. Env. Health Perspect. *122*(5), 499-504.

World Health Organization (WHO) & Inter-Organization Programme for the Sound Management of Chemicals. (2009). Principles for modelling dose-response for the risk assessment of chemicals. (Geneva: WHO).

食品安全委員会評価技術企画ワーキンググループ. (2018). 新たな時代に対応した 評価技術の検討 ～BMD 法の更なる活用に向けて～. http://www.fsc.go.jp/senmon/sonota/index.data/wg_gijyutsukikaku_houkoku_2.pdf (Accessed on 30 July 2018).