Human and Experimental Toxicology
Volume 45: 1–26

DOI: 10.1177/09603271261464797
journals.sagepub.com/home/het

A new bayesian framework for dose-response modelling and benchmark dose determination

Marc Aerts[1], Cécile Kremer[1], Efisio Solazzo[2], and José Cortiñas Abrahantes[2]

Abstract

Introduction: A new Bayesian framework for dose-response modelling and benchmark dose determination is introduced, following guidelines from the World Health Organization and the European Food Safety Authority. It incorporates model averaging across a family of candidate models and defines prior distributions at the level of response distributions, dose-response median models, and model parameters. In addition to providing a fully probabilistic inferential framework, the Bayesian paradigm offers some methodological advantages over frequentist approaches. **Methods:** The proposed framework employs regularising default priors that are adaptive to study design and biologically plausible constraints. Informative priors can also be incorporated. The performance was evaluated against frequentist model averaging across four case studies using both real and simulated datasets. **Results:** In simulations based on the Bisphenol A study, only 3% of BMD estimates obtained using frequentist inference met the EFSA accuracy criterion (BMDU/BMDL < 50), compared with 83% obtained using Bayesian inference. Other analyses showed that design-adaptive and biology-respecting regularising priors improved estimation relative to frequentist hard constraints. In a small-sample case study with poorly informative experimental design, frequentist quantile bootstrap confidence intervals for the BMD were highly sensitive to the number of bootstrap replicates, whereas Bayesian MCMC-based inference remained stable and reliable. A final case study demonstrated that constructing an informative overarching BMD prior from multiple historical studies improved estimation performance. **Discussion:** The case studies illustrate improved estimation and the ability to incorporate historical information. These findings support the use of Bayesian methods as a powerful alternative for regulatory toxicology and risk assessment applications.

Keywords

Bayesian inference, benchmark dose, informative prior, model averaging, regularising prior, soft constraining

Received: 20 March 2026; Revised: 29 May 2026; Accepted: 11 June 2026

Introduction

Quantitative dose-response assessment forms the cornerstone of toxicological risk characterisation and regulatory decision-making. Historically, the No-Observed-Adverse-Effect-Level (NOAEL, the highest tested dose that does not differ statistically from controls with respect to an adverse effect) has served as the conventional point of departure (POD) for deriving health-based guidance values (HBGV) such as the Acceptable Daily Intake (ADI) or Tolerable Daily Intake

[1]Data Science Institute, I-BioStat, Hasselt University, Hasselt, Belgium
[2]Methodology and Scientific Support Unit, European Food Safety Authority, Parma, Italy

Corresponding author:
Marc Aerts, Data Science Institute, Hasselt University, Agoralaan Building D, 3590 Diepenbeek, Belgium.
Email: marc.aerts@uhasselt.be

(TDI).[1,2] Despite its intuitive simplicity, the NOAEL approach has long been criticised for its statistical and conceptual shortcomings, such as its dependence on arbitrary dose spacing, insensitivity to dose-response shape, and binary treatment of significance testing e.g.[3,4] Furthermore, it fails to provide a quantitative estimate of risk, which limits its comparability and reproducibility across studies and chemicals.[1] To address these limitations, Crump[3] introduced the Benchmark Dose (BMD) concept, which estimates the exposure level associated with a specified change in response relative to the control group, i.e., the benchmark response (BMR), and its lower confidence limit (BMDL) as a statistically defined conservative POD. The BMD approach integrates all dose-response data, allows interpolation between observed doses, and quantifies uncertainty, thus overcoming key weaknesses of the NOAEL paradigm. During the past two decades, it has become the preferred method for regulatory risk assessment, endorsed by agencies such as the U.S. Environmental Protection Agency (EPA), the European Food Safety Authority (EFSA), and the World Health Organization (WHO).[5–7] Empirical applications demonstrated the broad feasibility and scientific value of the BMD approach across diverse chemical classes, including pesticides, mycotoxins, and natural toxins.[1,2] Comparative evaluations by EFSA and the WHO have reinforced its utility, particularly when supported by high-quality experimental data and robust model-fitting criteria.[1,7]

Initial applications of the BMD approach used frequentist methods based on maximum likelihood and profile likelihood confidence limits.[8] These frequentist BMD approaches represented a major improvement over the NOAEL paradigm by providing a continuous, model-based estimate that captures experimental variability and sampling uncertainty. However, inference was based on a final selected model without acknowledging the additional uncertainty introduced by data-driven model selection, potentially leading to an underestimation of variability and overly optimistic confidence intervals and BMDLs. An alternative to estimation-past-selection that accounts for model uncertainty is multi-model inference or model averaging (e.g. Burnham and Anderson[9]). This next methodological step forward in the determination of the BMD, in which several dose-response models are fit to the data and the resulting BMD estimate is a weighted average across models, and thus accounting for uncertainty about the "correct" risk model,[10] was the major objective of EFSA's 2017 guidance update.[11] Determination of the BMDL using frequentist model averaging is implemented in the R package PROAST, version 71.1 (latest version at the time of publication).

During the process of upgrading the methodology for estimating the BMD in light of recent trends in the general statistical literature, a parallel methodological pathway emerged, considering the Bayesian inferential framework. Whereas a probability is used for representing a long-run frequency and a parameter is considered as an unknown constant in frequentist inference, the notion of probability is extended to reflect uncertainty of knowledge rather than variability of outcome and thus probability distributions are also attached to unknown parameters (such as the BMD) in the Bayesian approach. Whether the Bayesian or frequentist framework is superior in general is a never-ending debate. However, there is general recognition that each approach can contribute to statistical practice (e.g. Bayarri and Berger [12]). The Bayesian approach has the appeal of offering a very natural framework for determining the BMD. First of all, it is natural to offer fully probabilistic interpretations of quantities of interest, such as the probabilistic interpretation of a credible interval for the BMD (typically wrongly assigned to a frequentist confidence interval by toxicologists). Bayesian methods align with the broader evolution toward probabilistic risk assessment advocated by the WHO/IPCS,[7] enabling transparent, quantitative integration of prior toxicological evidence. Indeed, the Bayesian paradigm, by its core nature (Bayes' rule), "invites" the toxicological user to combine and accumulate historical data and knowledge with new data, using a prior distribution that is updated to a posterior distribution. This can be particularly relevant in the field of BMD estimation, as data can be scarce, the experimental design may not be optimal, and certain parameters may be only slightly or not at all informed by the data, among other aspects. Prior distributions play a central role, and even if the objective is not to apply informative priors, using "uninformative" priors can be challenging. The "obligation" to choose prior distributions can be perceived as a disadvantage, particularly for non-linear models, but on the other hand it offers an opportunity to regularise the estimation of poorly informed parameters. Last but not least, Bayesian inference does not rely on asymptotic approximations of sampling distributions. Whereas the construction of frequentist confidence intervals for a model-averaged BMD estimate relies on approximate bootstrap simulations, Bayesian credible intervals are not prone to asymptotic conditions. Bayesian inference treats model parameters, and even model structure, as random variables with prior distributions, thereby integrating prior knowledge and propagating all sources of uncertainty coherently.[13] Reparameterised dose-response models that explicitly include the BMD as a target parameter allow for direct estimation of its posterior distribution and credible intervals, while Bayesian model averaging (BMA) naturally incorporates model uncertainty by weighting multiple candidate models according to their posterior probabilities. Comparative analyses have shown that Bayesian BMD analysis yields fewer failed estimates and improved characterisation of uncertainty relative to frequentist software.[14] Relevant literature on Bayesian model-averaged BMD estimation includes[15–17] for quantal responses,[18,19] for continuous data, and[20] for an application to ecological data. Software for Bayesian model-averaged BMD estimation includes ToxicR,[21] BMDS,[14,22] BBMD,[14] and BMABMDR (https://github.com/cecilekremer/BMABMDR). A brief comparison of these implementations is provided in Supplementary Table S5.

As of 2022, the EFSA guidance for BMD estimation recommends the use of Bayesian model averaging to derive a reference point (RP) from the critical dose-response data to establish HBGVs and margins of exposure,[23] in alignment with the update of Chapter 5 on dose-response assessment from the WHO/IPCS Environmental Health Criteria 240.[7] Additionally, the set of default candidate models to include in model averaging was reviewed to obtain a common set of models for quantal and continuous data. This paper builds upon these methodological advances to explore the application and advantages of Bayesian BMD modelling in toxicological risk assessment, within the background of methodological and practical updates of the EFSA guidelines.[5,11,23] The objectives of this paper are multifold. The first objective is to present the recently developed Bayesian BMD framework as described in the EFSA guidelines,[23] the model formulation,[24] and implemented in the R-package BMABMDR and its web application.[25] The second objective is to illustrate, in various ways, the advancements enabled by the Bayesian approach compared to the frequentist approach. Real-life datasets formed the inspiration for four case studies, in which issues relevant to toxicological practice are presented, explored, and discussed from a methodological perspective. To demonstrate the Bayesian advancements, real datasets and simulated datasets are analysed within these four case studies using both frequentist and Bayesian model-averaged BMD analyses, and results are discussed and compared. Within the context of successive updates to the EFSA guidelines, the focus is on two R packages developed for this purpose: Proast 71.1 represents the frequentist framework, referred to as F-BMD, whereas BMABMDR 0.1.18 represents the Bayesian framework, referred to as B-BMD. The two approaches are among the most commonly used in Europe; moreover, the EFSA Bayesian app has replaced the Proast tool in the 2022 update of the EFSA guidance on BMD modelling.

Methods

As in statistical modelling in general, several model components have to be defined for BMD modelling, regardless of which framework is used. First, assumptions need to be made about the distribution $f(y|x)$ of the endpoint y of interest at a fixed dose x, which reflects the variation in the observed values of the endpoint for a given dose x. Next, one needs to specify how this distribution $f(y|x)$ is affected by the dose level x, i.e., specifying the dose response model DRM(x). These components define the suite of candidate models for model averaging. Then estimation and computational choices must be made, and these differ for F-BMD and B-BMD as the DRMs are embedded in different multi-model inferential paradigms. And finally, for B-BMD, choices must be made regarding prior distributions. These different ingredients of the Bayesian framework as implemented in the BMABMDR package are summarised in this Methods section, and, for adequate understanding of the results in the next section, essential differences with its frequentist alternative as implemented in Proast are briefly mentioned. For more details on the BMABMDR implementation, we refer to I-BioStat[24].

Distribution of the endpoint at fixed dose

For a continuous endpoint y, assumed to be strictly positive, BMABMDR includes the normal and lognormal distributions. The lognormal distribution is the principal candidate, but as the true distribution is unknown and all mathematical distributions are approximations, and as it is not certain that the right-skewed lognormal is the best approximating distribution, the inclusion of at least one other distribution is self-evident within the setting of multi-model inference. In BMABMDR, it was decided to include the normal distribution as well, which is symmetric and exhibits convenient mathematical and computational properties. Therefore, in BMABMDR, the candidate distributions $f(y|x)$ are the normal distribution $N(\mu(x), \sigma^2)$ with the mean and variance parameter on the original scale, and the lognormal distribution log $N(\mu(x), \sigma^2)$ where the mean and variance parameter refer to the endpoint on the log-scale. Model averaging, including averaging across continuous distributions, within the setting of BMD estimation, was studied in Wheeler et al.[19], demonstrating its superiority over single-distribution model averaging. In contrast, Proast only includes the lognormal distribution for a continuous endpoint Slob et al.[26]. As reported in Abrahantes et al.[27], an EFSA analysis of 3755 datasets from the National Toxicology Program (NTP) organ weight studies shows empirical evidence that the lognormal distribution should not be the sole exclusive choice.

For a quantal (i.e., binary) endpoint, both frameworks use the binomial distribution with the probability $\pi(x)$ on the adverse event as the only parameter. Unlike for continuous endpoints, this represents the only possible distribution and averaging across distributions is not applicable in this case.

Dose response models

As the notation $N(\mu(x), \sigma^2)$ for the normal and log $N(\mu(x), \sigma^2)$ for the lognormal distribution already suggests, it is assumed for a continuous endpoint that i) the variance is constant across dose levels in the normal case, ii) the coefficient of variation is constant across dose levels in the lognormal case, and iii) dose affects the median response (i.e.,the dose response model):

$$\mathrm{DRM}(x) = \mathrm{Median}(y \mid x) = \begin{cases} \mu(x) & \text{for the normal distribution,} \\ e^{\mu(x)} & \text{for the lognormal distribution.} \end{cases}$$

The set of candidate models for $\mu(x)$ was largely extended in BMABMDR based on the unifying model structure proposed in Aerts et al.[28]. Table S1 in the Appendix shows the eight (strictly) monotone dose-response models for continuous endpoints within the normally distributed family of models. We refer to I-BioStat[24] and More et al.[23] for more details and for the formulation for the lognormal distribution. The so-called 'm5' models in Proast are essentially a subset of the first family of models in BMABMDR. However, they also include versions of some models in this subset with the fold-change parameter fixed to specific values (the 'm3' models; see case study 1 for more details). We could not retrieve any published or publicly available technical report that clearly defines all models (with details on their implementation and parameter constraints) underlying the Proast software to date. The general guideline for model averaging is that the family of candidate models is "sufficiently rich" to include a member that approximates the true model sufficiently well, with "rich" referring not to similar models but rather to diverse ones. Challenging the current set of candidate models should remain a point of attention in BMD estimation, and revising it based on experiences in daily toxicological practice should remain an objective for the future. In conclusion, the set of models in BMABMDR is more inclusive, whereas that in Proast is rather restrictive.

The models in Table S1 are formulated in terms of parameters a, b, c, and d. Parameters a and c are related to two "natural" parameters: background response DRM(0) and the asymptotic response for an ever-increasing dose $\mathrm{DRM}(\infty) = \lim_{x\to\infty}\mathrm{DRM}(x)$. A third natural parameter is the potency parameter, which is the BMD parameter associated with a particular BMR, and is defined by the equation:

$$\frac{\mathrm{DRM(BMD)} - \mathrm{DRM}(0)}{\mathrm{DRM}(0)} = \pm\mathrm{BMR},$$

such that

$$\mathrm{BMD} = \mathrm{DRM}^{-1}(\mathrm{DRM}(0)(1 \pm \mathrm{BMR})), \tag{1}$$

with $\pm$BMR referring to an increasing (+) or a decreasing (-) dose-response curve and DRM^{-1} denoting the inverse dose-response model (well defined because of strict monotonicity).

All models in Table S1 are reformulated in terms of three natural parameters {DRM(0), DRM(∞), BMD} and two technical parameters: the parameter d which acts as a power of dose in all models, barring the quadratic exponential model, and the variance parameter σ^2. The issue is that the d parameter lacks a natural biological interpretation and serves as a power (a measure of steepness), but a specific value does not imply the same steepness across different models. Moreover, power parameters in nonlinear models require special attention during estimation, which is the topic of the second case study.

For quantal (i.e.,binary) endpoints in BMABMDR, the (increasing) dose response model is defined as the probability of an adverse event

$$\mathrm{DRM}(x) = \pi(x),$$

and the same dose-response models for $\pi(x)$ are considered as for $\mu(x)$ in Table S1, but with a and c specified such that the probability of the adverse event tends to 1 for an ever-increasing dose, i.e., $\mathrm{DRM}(\infty) = 1$, eliminating the parameter c from the models. All models are again reformulated in terms of two natural parameters {DRM(0), BMD} and one technical parameter d. In this case, the BMD is defined as the solution of:

$$\frac{\pi(\mathrm{BMD}) - \pi(0)}{1 - \pi(0)} = \mathrm{BMR},$$

or

$$\mathrm{BMD} = \pi^{-1}(\pi(0)(1 + \Psi(0)\mathrm{BMR}),$$

which is identical to (1), but with the BMR adjusted with the odds $\Psi(0) = (1 - \pi(0))/\pi(0)$ of being still at risk for the adverse event. Proast largely includes the same models, but also includes some models with a parameter representing $\pi(\infty)$. These models allow that, even for extremely high doses (mathematically infinite), there is a positive probability that the adverse event will not happen. In BMABMDR, these models were considered biologically implausible.

Extensions

All models in BMABMDR have been extended to handle hierarchically structured or clustered data (e.g., endpoints for fetuses within the same litter) and to incorporate categorical covariates. These extensions are not considered in this paper; for more details, we refer to I-BioStat.[24]

Inferential & computational aspects

The central idea of the Bayesian approach is to combine the data, through the likelihood expressing the plausibility of the observed data as a function of the parameters of a stochastic model, with prior knowledge in order to obtain the posterior probability as a revised, updated probability. The posterior distributions of the model-specific parameters yield model-specific BMD estimates. Using the posterior probabilities of the respective candidate models, these model-specific posterior distributions are mixed into a single averaged posterior BMD distribution. The 5% (BMDL) and 95% (BMDU) quantiles of this averaged posterior distribution provide the 90% posterior credible interval. A challenge arises in obtaining the posterior probability for each candidate model, as their marginal likelihoods are not analytically tractable and must be approximated numerically. An accurate but computationally intensive approach is Markov Chain Monte Carlo (MCMC) sampling. Bridge sampling is recognised as a reliable and relatively straightforward sampling method that can be used to obtain the marginal likelihoods of the candidate models. Bridge sampling was originally developed to directly estimate the Bayes factor, which is the ratio of the marginal likelihoods of two models.[29–31] However, a version of the method can also be used to approximate the marginal likelihood of a single model. The B-BMD framework used here has been implemented in RStan, the R interface to Stan. Stan is a probabilistic programming language for statistical inference that primarily uses No-U-Turn sampling (NUTS) to obtain posterior distributions given a user-specified model and data.[32]

In F-BMD modelling, a parametric bootstrap is often used to obtain a model-averaged confidence interval for the BMD. This approach explicitly reflects the repeated sampling philosophy of frequentist inference. The idea is to generate many new bootstrap datasets from a good approximation of the true but unknown underlying data-generating process. A BMD estimate is then determined from the model-averaged dose-response curve for each bootstrap sample, and these bootstrap replicates of the BMD estimate reflect the sampling distribution of the BMD. The left and right percentiles of this simulated bootstrap distribution are then taken as the BMDL and BMDU, respectively. A challenge is to define a good approximation of the true, unknown underlying data-generating process, and repeatedly re-analysing bootstrap samples can be very time-consuming. A useful percentile confidence interval may be based on 250 bootstrap samples, but it is generally recommended to generate at least 1000 bootstrap samples.[33] The default in Proast (version 71.1) is 200 bootstrap samples (see https://www.rivm.nl/documenten/proast-manual-webapp). We could not retrieve the motivation for this choice in any public reports, and EFSA has recommended using at least 1000 bootstrap samples. Additionally, it is essential to note that the bootstrap is an asymptotic method, meaning that the bootstrap distribution is only a valid approximation for a sufficiently large sample size (of the dataset at hand). Finally, some concerns have been expressed about the use of bootstrapping for model averaging in the statistical literature. For example, Hjort and Claeskens[34] showed that bootstrapping cannot be relied upon to consistently estimate the asymptotic distribution of averaging estimators. In the same vein, Liu[35] showed that the asymptotic distribution of the averaging estimator with data-dependent weights cannot be approximated by simulation. The above points of attention about bootstrap inference in F-BMD are the focus of the third case study.

Convergence criteria

In the B-BMD framework, there is a need to examine the convergence of the MCMC chains used to obtain samples from the posterior distribution of the parameters in each model to be used in the model averaging. To examine this, the potential scale reduction factor $\widehat{R}$, bulk and tail efficiency sampling for all model parameters, and the log density were computed.[36] $\widehat{R}$ greater than 1 is indicative of non-mixing MCMC chains, while bulk and tail efficiency samples less than $100 \times n_{chains}$, where n_{chains} is the number of chains used, is indicative of problems at the center (bulk) or tail of the posterior distribution. In that case, one cannot trust estimates of the center (means and medians) or tail (quantiles) from the posterior distribution. As an overall measure of convergence for each model, the criterion of $\widehat{R} < 1.01$ for the log density is used. By default, all models are

included in the model-averaging exercise, but there is an option to exclude the models that did not converge (recommended). In addition, warning messages are generated for model parameters that do not satisfy any of the three convergence criteria previously mentioned.

Prior distributions

Prior distributions are defined on three levels (Figure 1): i) on the level of the distribution of the endpoint at a given dose, ii) on the level of the set of candidate models, and iii) on the level of the parameters of a given dose-response model.

The default prior distribution for the set of 16 candidate models for a continuous endpoint or for the set of 8 candidate models for a quantal endpoint is the uniform discrete distribution, i.e., each model receives equal prior probability. Consequently, each distribution (normal and lognormal for continuous endpoints) also receives equal weight when using this default prior. Prior knowledge, such as past experience or a systematic reevaluation of existing data, may justify a different choice, leading to distinct prior probabilities for the candidate models and possibly unequal prior weights for the endpoint distributions. For instance, if there is scientific evidence against the normal distribution for a particular endpoint, the prior probabilities for the 8 models assuming normality can all be set to 0.

All dose-response models DRM(x) are formulated in terms of the 'natural' parameters, i.e., the background DRM(0), the asymptote DRM(∞) (only for continuous endpoints), and the BMD, and the technical parameters d and σ^2 (only for continuous endpoints). All these parameters need default prior distributions, preferably uninformative distributions for which BMD estimates are expected to be close to the F-BMD estimates. Defining and implementing such uninformative priors for the parameters appears to be a big challenge in nonlinear models. Uniform prior distributions or very flat uninformative prior distributions on the (transformed) parameter space can induce an informative distribution in the space of functional shapes[37,38]. Wheeler[39] states that flat uniform priors on the model parameters may add information and may introduce unexpected biases into the analysis, demonstrating through numerous empirical examples why prior distributions that are non-informative across all endpoints of interest do not exist for dose-response models; that is, other quantities of interest will be informed by choosing one inferential quantity not informed. Bornkamp therefore proposed the functional uniform priors for nonlinear modelling, a prior distribution that is uniform in the space of functional shapes of the underlying nonlinear function and then back-transforming to obtain a prior distribution for the original model parameters.[37,38] However, calculating the functional uniform densities explicitly is not straightforward in many cases.

In BMABMDR, the natural parameters are given a modified PERT prior distribution, which is a beta-distribution extended with minimum and maximum values (see Appendix). The other two parameters are the mode and a positive shape parameter, the latter controlling the distribution's flatness (i.e.,the smaller the shape parameter, the flatter the distribution). Both the default and informative priors for the natural parameters are implemented using modified PERT priors, since this distribution is easy to use with directly interpretable parameters. For the technical parameters, only a default lognormal prior is implemented, and no informative prior can be specified. For details about the default prior distributions on all parameters, we refer to More et al.[23] and I-BioStat[24], and results from an extensive simulation study can be found in Aerts et al.[40]. It should be noted that the default priors are not fully uninformative and are fine-tuned to regularise the estimation of some of the model parameters, as will be illustrated and discussed in the first two case studies.

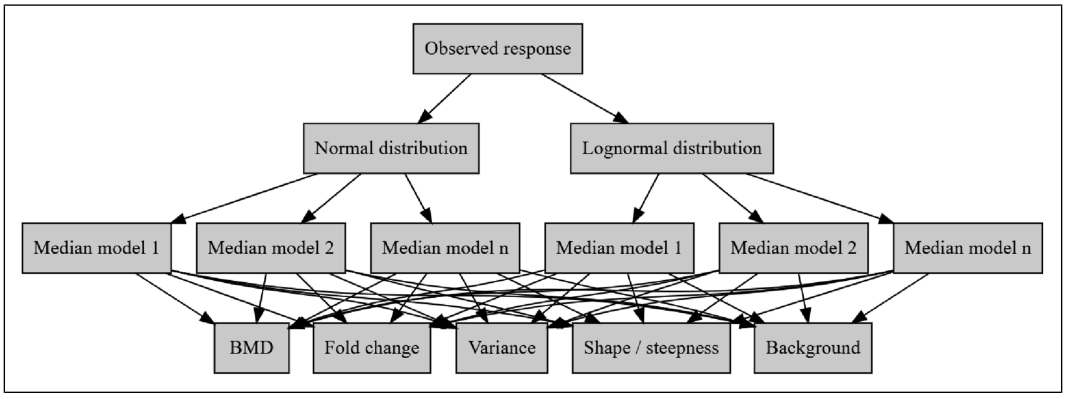

Figure 1. Three levels on which to define prior distributions, given an observed response.

Informative priors

The B-BMD framework provides the opportunity to reduce uncertainty and thereby increase the reliability of BMD estimation, by integrating relevant additional information (e.g., historical data, expert knowledge) into the dose-response model fitting process through informative PERT prior distributions for the natural parameters. If the construction of informative priors can be based on expert knowledge elicitation, implementing them is straightforward, given the PERT distribution's straightforward and easy interpretation of its parameters. On the other hand, the construction of informative PERT priors based on historical data requires analysing these historical data first using the default weakly informative priors, and then converting the obtained posteriors of the natural parameters into informative PERT priors. Given a posterior distribution for each of the natural model parameters (DRM(0), DRM(∞), BMD) obtained from historical data, the parameters of the corresponding informative PERT distribution for each of these model parameters can be estimated using maximum likelihood estimation (MLE).

This is a relatively new area in BMD estimation, and the literature on this topic is quite scarce. Shao[41] compared three methods for integrating historical information for Bayesian model-averaged BMD estimation. Fang et al.[42] illustrates the quantification of prior beliefs on an example from environmental carcinogenicity testing. The discussion about informative priors and related references in Jensen et al.[43] is also limited. The primary objective in Kremer et al.[44] was therefore to develop and compare strategies for constructing informative PERT priors from multiple historical studies, including different strategies for combining these studies and including the development of a methodology to create an overarching prior distribution for groups of chemicals. This will be illustrated in the last case study.

Preliminary tests and checks

In this section, we briefly introduce preliminary tests and checks as implemented in the BMABMDR package and its corresponding EFSA web application. It is essential to provide the user with additional statistical information: alerts about the absence of any dose effect, the suitability of the design, the appropriateness of the suite of candidate models, the assumption of normal or lognormal distribution, and the assumption of constant variance or constant coefficient of variation. We mention these tests and checks briefly below, for details we refer to More et al.[23]and I-BioStat[24]. Proast only includes alternative alerts based on AIC for no-dose effect and for appropriateness of the suite of candidate models.

Absence of any dose effect

In case the data at hand show no or a negligible dose effect, it will not be possible to reliably fit the dose-response models and accurately determine the BMD(L/U). Therefore, a Bayesian ANOVA test for no dose effect is performed beforehand. With x_i denoting the experimental dose levels and N the number of levels, the null model $\mathrm{DRM}(x_1) = \ldots = \mathrm{DRM}(x_N)$ is contrasted with a saturated ANOVA model. Bridge sampling with uninformative priors is used to estimate the marginal likelihoods and to approximate the Bayes factor. A simplified decision rule based on Jeffreys' classification is used, with a Bayes factor in favour of the saturated model greater than 10 indicating sufficient evidence of a "substantial" dose effect.[45] The user can continue in any case, but in case of insufficient evidence they are alerted that modelling might not result in a BMDL that is useful as an RP.

Suitability of design

The starting point for this check is the use of a one-sided hypothesis testing procedure for each pairwise comparison among dose levels. When at least three groups of responses are found to be significantly different from each other, it is expected that the study data contain enough information to characterise the dose-response relationship and that it therefore might contain enough information about the parameter of interest, i.e., the BMD. When fewer than three groups are detected, the user is alerted. This suitability check is only available for continuous endpoints in the EFSA web application. For more details and illustrations, see More et al.[23] (Section 2.6.3).

Appropriateness of the suite of candidate models

Since the performance of model averaging depends largely on the suite of candidate models and on the presence of a sufficiently good approximation model in that suite, a goodness-of-fit test is performed as part of the analysis. The model with the highest posterior probability is contrasted to the unconstrained, saturated model (not necessarily monotone) that fits

perfectly to the summary data. As a measure of evidence, the Bayes factor is used, which is computed approximately in this case based on the BIC values. The user is alerted if the best-fitting model does not fit sufficiently well.

Normal and lognormal assumption and homogeneity of variance

For continuous endpoints, several additional tests are performed to verify distributional assumptions. If individual data are available, a Shapiro-Wilk test for normality and log-normality is performed at each dose level and collapsing all data (after centering). Levene's test is used to test constant variance (for normal distribution) and constant coefficient of variation (for lognormal distribution). In the event of violations, the user is notified.

Experimental data, however, are mostly reported as summary statistics. More precisely, the dose-specific arithmetic sample means and standard deviations are usually available. Assuming a normal distribution, these summary statistics are jointly sufficient, implying that they contain all the information the individual data provide about the model parameters. Assuming a lognormal distribution, one needs the geometric sample means and standard deviations, which can only be approximated from the arithmetic statistics. The results in Shao et al.[46] suggest that this has minimal impact on the BMD estimates, in agreement with our own experiences. However, if only summary statistics are available, it is not possible to test for (log)normality and moreover, Levene's test cannot be applied. However, Bartlett's test can be used to test the homogeneity-of-variance assumption for summary data. In the event of a violation, the user is notified again.

In case Levene's or Bartlett's test rejects constant variance or constant coefficient of variation, the EFSA web application automatically performs sensitivity analyses using scenarios with as summary data the observed dose-specific sample means and with constant variance or coefficient of variation according to the minimal and maximal observed sample variance or (approximate) coefficient of variation.

Case studies

Four case studies, including data analyses and simulations, are used to illustrate the main differences between the F-BMD (Proast in particular) and B-BMD (BMABMDR in particular) inferential frameworks for dose-response modelling and BMD determination. When the difference between both approaches is related to the specific implementation (Proast vs. BMABMDR), we explicitly mention this throughout the following sections. The first two case studies focus on estimating particular model parameters and the effect of study design. The third case study addresses the construction of confidence/credible intervals, and the fourth case study illustrates the use of informative priors. Throughout these case studies, the default sampling specifications in the BMABMDR package are used. More specifically, we use 3 MCMC chains, each with 3000 iterations of which the first 1000 are discarded as burn-in.

Case study 1

The median response at very high dose, i.e., the parameter DRM(∞), is one of the three natural parameters for a continuous endpoint. This first case study examines the estimation of DRM(∞) and its impact on BMD estimation, particularly when the design does not encompass sufficiently high dose levels. In such cases, the estimation of DRM(∞) needs some regularization. In F-BMD, this can be achieved by constraining parameters. In Proast, the so-called 'm3' models equal corresponding 'm5' models (e.g., exponential model) by putting the parameter c, representing the fold-change FCH = DRM(∞)/DRM(0), to 10^{18} for increasing and 10^{-18} for decreasing functions. In practice, at high dose levels, these 'm3' models tend to 0 and ∞ for decreasing and increasing functions, respectively. These 'm3' models compete with the 'm5' models in the set of candidate models, and the AIC-based weights determine how well each of them fit to the data. Next to the inclusion of these 'm3' models, Proast constrains the 'm5' models by putting FCH $\leq$ 1/1.3 for decreasing and FCH $\geq$ 1.3 for increasing functions as a way to manage the estimation of DRM(∞). We could not retrieve the biological or empirical basis for this specific threshold in any publicly available technical reports. Such hard constraints, however, may not adapt very well to a variety of datasets.

In B-BMD, the prior can be used to support the data in estimating DRM(∞). If prior information is available, there is the evident option to use an informative prior. If not, the challenge is to construct a default prior such that it is as weakly informative as possible, but informative enough to guide the estimation of DRM(∞) in the right direction. Table 1 defines how such a default prior is implemented in the BMABMDR package based on the observed change in response and depending on the 'flatness' of the dose-response curve by assessing whether the observed dose-response curve sufficiently flattens at the highest dose levels. A default fractional polynomial $P_F(x)$ is fit to the summary data and the relative differences (approximate 'slopes')

$$\Delta_i = \frac{|P_F(x_i) - P_F(x_{i-1})|}{\log(x_i + 0.0001) - \log(x_{i-1} + 0.0001)}, \quad i = 2, \ldots, N,$$

are computed. If the last slope is sufficiently small, relative to the maximal slope, i.e.,

$$\frac{\Delta_N}{\max_i \Delta_i} < 0.5,$$

the dose-response curve is considered sufficiently flattened to allow reliable estimation of DRM(∞). If this is not the case, the prior related to DRM(∞) is centred at a multiple of the observed responses at the highest dose level, ensuring that the fitted curve does not flatten at the observed highest dose. The choice 0.5 is based on experiences of the authors from simulations and analyses of regulatory datasets.

Case study 1 contrasts 'hard' constraining in Proast with 'soft' constraining in BMABMDR. It starts with the analysis of the Bisphenol A dataset from Hu et al.[47] and continues with a simulation study based on these data. These decreasing dose-response data do provide some information about DRM(∞), albeit some additional guidance is required. A second simulation setting presented in the Appendix provides further insights into the case of an increasing dose-response curve, with a design that lacks sufficiently high doses, and consequently, no data are available to estimate DRM(∞). The estimation then relies on extrapolating the candidate dose-response curves beyond the experimental dose range, along with additional guidance during the estimation process.

Case study 2

The second case study focuses on the d parameter in combination with properties of the experimental design. The parameter d is, besides the variance parameter for continuous endpoints, the only 'technical' parameter of the dose-response models. Although it is related to the 'steepness' of the dose-response curve, the parameter d lacks an unambiguous biological definition and a unique value across all models. In all models except the quadratic exponential model, the parameter d acts as a power on the dose. Together with the BMD parameter, it determines the pattern of dose-response from dose 0 to a very large dose. As for the estimation of the fold change, the study design may also interfere with the estimation of d. The estimation of d can become unstable, especially for large values of d. Moreover, values of $d < 1$ do not reflect observed biological behavior. Indeed, as, $\lim_{x\to 0+} d/dx(x^d) = \lim_{x\to 0+} dx^{d-1} = +\infty$ for $0 < d < 1$, the tangent line at 0 is vertical, implying that the dose-response is infinitely steep at the onset of exposure, which is considered biologically not plausible. This raises the question of whether one should impose the hard constraint $d > 1$.

Although values of the power parameter $d < 1$ imply an infinite slope at dose zero, and are therefore often regarded as biologically implausible, different views exist regarding whether such constraints should be imposed in practice.[48] For positive doses, the slope is finite, and models with $0 < d < 1$ may better capture strong low-dose curvature in the data. Consequently, imposing the restriction $d \geq 1$ may reduce model flexibility in some cases.

F-BMD relies on hard constraining to govern the estimation of d, and Proast seems to apply the hard constraint $0.2 < d < 5$. Such constraints exclude very low and very high values of d but do allow estimates less than 1. B-BMD provides again the option to regularise the estimation of the d parameter by applying soft constraining in terms of a prior on d, giving low probabilities to small and high values. Except for the quadratic exponential model, the default prior for d in BMABMDR is a lognormal(1,1) distribution truncated at 5, implying a probability of 0.218 for d to be smaller than 1. It follows the hard constraint at 5, but allows the median or even the 97.5%th quantile of the posterior to be below 1 if the data contain sufficient evidence to overrule this prior.

Table 1. Default PERT prior for DRM(∞) in the BMABMDR package; $\overline{y}_0$ is the observed sample mean at dose 0, $\overline{y}_N$ the observed sample mean at highest experimental.

parameter	min	Mode	Max	Shape
MAX flat	$\overline{y}_0(1.01 + \text{BMR})$	$\overline{y}_N$	$2\overline{y}_N$	4
MAX not flat	$\overline{y}_0(1.01 + \text{BMR})$	$3\overline{y}_N$	$6\overline{y}_N$	4
MIN flat	$0.5\overline{y}_N$	$\overline{y}_N$	$\overline{y}_0(1 - \text{BMR} - 0.01)$	4
MIN not flat	$0.1\overline{y}_N$	$0.5\overline{y}_N$	$\overline{y}_0(1 - \text{BMR} - 0.01)$	4

A simulation study was conducted in order to evaluate how the parameter d influences the estimation of the BMD and the corresponding interval produced by both frameworks. We considered two different scenarios, with d either larger or smaller than 1. In the first scenario, data were generated using a lognormal distribution with the DRM defined as

$$\mathrm{DRM}(x) = a(1 + (c-1)(1 - e^{-bx^d})),$$

with a= 7.5, b= 6,c= 2, and d= 1.2 and with geometric standard deviation $e^{0.05}$. Note that this model is not included in any suite of candidate models, not in BMABMDR nor Proast. The BMR was taken 5%, corresponding to a BMD of 0.0189. The simulated data follow a geometric dose spacing considering the following doses: control (0), 0.5, 5, and 50, with 20 observations at each dose level. This scenario has a less-than-optimal design, illustrating that estimating the d parameter can be cumbersome in such cases.

In the second scenario, data were generated using a lognormal distribution with the DRM defined as

$$\mathrm{DRM}(x) = e^{a(1+(c-1)(1-e^{-bx^d}))},$$

with a= log(7.5), b= 1.2, c= $log(2)/a$+ 1 = 1.344, (implying that DRM(0) = 7.5 and DRM(∞) = 2 × 7.5 = 15), and d= 0.4, and with geometric standard deviation $e^{0.05}$. Note that this model is now included in both suites of candidate models, for BMABMDR and Proast. The BMR was taken 5%, corresponding to a BMD of 0.0009. The simulated data follow a geometric dose space, considering the following doses: control (0), 0.01, 1, and 100, with again 20 observations at each dose level. The objective of this second scenario is to investigate how well a 'biologically implausible' value of d can be estimated. The two modelling paradigms were applied to 1000 simulated datasets, and the BMD confidence (based on 1000 bootstrap samples) and credible intervals were retrieved for each simulated dataset.

Case study 3

This case study gets to the heart of the difference between frequentist bootstrap and Bayesian inference. Several approaches for constructing confidence intervals for frequentist model-averaged parameters have been proposed in the literature (e.g. Kabaila et al.[49]). The approach of assuming a normal sampling distribution and constructing Wald intervals (i.e.,model-averaged estimate ± normal critical point × estimated standard error) is generally problematic, as the normality assumption is often incorrect. Indeed, it is well known that the distribution of frequentist model-averaged estimators is a mixture distribution (e.g. Claeskens and Hjort[50]) The bootstrap approach offers a general alternative inferential framework, and has been applied to the setting of model averaging as well (e.g. Garcia-Angulo and Claeskens[51]). In Proast, the basic percentile bootstrap interval is used to construct a confidence interval for the BMD parameter (utilising the quantiles of the bootstrap distribution) and to obtain the BMDL. However, the bootstrap is an asymptotic approach (similar to the normal approximation) and may perform poorly with small samples. Moreover, defining an appropriate mechanism for generating bootstrap data is challenging, and the number of bootstrap runs yielding bootstrap replicates must be sufficiently high in order to construct reliable confidence intervals. Finally, the percentile bootstrap interval is a very basic method, but it is known to be less optimal than other more complicated constructions (e.g. Garcia-Angulo and Claeskens).[51]

In contrast, B-BMD inference is fully embedded in the Bayesian inferential paradigm and does not rely on asymptotic approximations. The credible interval in BMABMDR is defined by the 5% and 95% quantiles of the BMD posterior distribution. Although there are alternative ways to obtain a 90% credible interval (e.g. Lesaffre and Lawson[52]), this simple choice is well-defined and guarantees a 95% probability that the unknown BMD is larger than the BMDL. This latter interpretation is also fully probabilistic, and toxicologists are often unfamiliar with the repeated-sampling interpretation of a frequentist confidence interval; they tend to interpret it in the above Bayesian way. The fully probabilistic nature of the Bayesian framework and ease of interpretation are certainly important assets in risk assessment.

In this case study, we compare both frameworks in the analysis of data on fetuses with skeletal and internal malformations and variations of SD rats exposed to BDE-99 *in utero* between GD6-19, testing on GD20.[53]

Case study 4

The possibility of incorporating prior information and knowledge alongside the data constitutes the primary difference between frequentist and Bayesian inference. Whereas Bayesian inference arises from probabilistic updating via Bayes' rule, a frequentist analysis can also be informed by other knowledge, albeit in a rather indirect, diffuse, and diverse manner. Historical data can be used to improve design, to determine sample size, to guide model choice, formulate hypotheses, and

suggest options in sensitivity analyses. However, many of these also apply to a Bayesian analysis, in addition to the direct use of historical data in constructing informative priors.

In this case study, we illustrate how multiple historical studies can be used to construct an informative prior for the BMD parameter. We consider three ways of combining the information from multiple studies to construct an informative prior for a particular (natural) parameter: (1) mixing posteriors, (2) pooling data, and (3) cumulative analysis. In a detailed previous simulation study,[44] mixing of posteriors was found to be the preferred approach, and we will therefore focus on this approach in this work. In the mixing posteriors approach, each historical dataset is analysed separately, resulting in a posterior distribution for each parameter of each model obtained from each historical dataset. These posterior distributions are then 'mixed' by sampling from each, which can be done using equal weights or weighted based on, for example, study importance or study quality. The informative prior distribution used to analyse a new study is then based on this mixed posterior distribution.

However, an issue when combining historical datasets is the use of different dose scales and dose units across different studies. Therefore, the following approach was used for standardising the dose scale of different datasets:

1. analyse each historical dataset using a BMR of $q = 0.05$ (i.e.,the most commonly applied BMR value for continuous endpoints),
2. obtain the model-averaged fold change estimate from each historical dataset,
3. set a new BMR as $q_f = f \times (\min(FCH) - 1)$, where min(FCH) is the lowest estimated fold change across the historical datasets,
4. analyse each historical dataset using the BMR q_f,
5. divide the original dose levels in each historical dataset by the BMD obtained from the analysis using BMR q_f.

Previous methodological work has found that results are, in general, not sensitive to the choice of f.[44] We therefore use a value of $f = 0.25$ in this illustration and the most recent dataset is used as the 'new' data to be analysed. The standardised datasets are used in the construction of an informative prior when mixing posteriors as described above, using the BMR of $q = 0.05$. In order to use the resulting overarching informative prior based on historical data in the analysis of a new dataset, the following approach was followed:

1. rescale the new dataset based on its BMD corresponding to q_f attached to the informative prior,
2. analyse the rescaled data with $q = 0.05$ and informative prior for the BMD,
3. transform the resulting BMD posterior back to the original scale of the data.

To illustrate the methodology for the construction of an informative prior, including how to combine historical studies as well as standardisation of the dose scale, we use a real-data example from the NTP organ database. Grouping of compounds based on mode of action (MoA) was performed in previous work.[44] For this example, we included five studies using compounds with a MoA of oxidative stress, administered via air inhalation in female B63CF1N mice. The endpoint of interest was relative lung weight.

EFSA guidance quality criteria and performance measures in data analyses and simulations

The 2022 EFSA guidance discusses the appropriateness of the BMD modelling outcome, and more precisely of the BMDL, for deriving an RP to establish HBGVs and margins of exposure (MOE). The uncertainty of the BMD estimation, as characterised by the BMDU/BMDL ratio or by the ratio between the estimated BMD (the median of the posterior BMD distribution as point estimate) and the BMDL, was suggested as a quality criterion by US EPA.[1,23] The ratio BMDU/BMDL can be considered as the geometric width of the credible interval [BMDL,BMDU], whereas the ratio BMD/BMDL is the geometric width of the interval at the left of the BMD point estimate.

Based on those criteria, EFSA recommends that the risk assessor should consider alternatives to the BMDL as an RP when

1. none of the candidate models fit the data sufficiently well (More et al.[23] , Section 2.5.3), or
2. BMD/BMDL > 20,or
3. the BMD is 10 times lower than the lowest non-zero dose, or
4. BMDU/BMDL >50.

According to EFSA, these criteria should be used as 'indicators', and it is also mentioned that the cut-off values for these criteria may be reconsidered after further experience with their use has been accumulated. To measure and compare the

performance of Proast and BMABMDR in simulations, the EFSA criteria 2-4 are used because of their practical importance, together with the following more theoretical measures:

left coverage (% runs with BMDL < true BMD): most relevant due to the prominent role of the BMDL as candidate RP; the nominal level is 95%,
total coverage (% runs with BMDL < true BMD < BMDU): incorporating the performance of the BMDU; total coverage equals at most the left coverage; the nominal level is 90%,
median left width (median of the values BMD/BMDL): always larger than 1 and the closer to 1, the more the 45% probability in the left tail of the BMD-posterior (between BMDL and BMD) is concentrated in a smaller region on the dose scale,
median total width (median of the values BMDU/BMDL): always larger than 1 and larger than the median left width, and the closer to 1, the more the 90% probability in the center of the BMD-posterior (between BMDL and BMDU) is concentrated in a smaller region on the dose scale.

As the BMD point estimate plays no central role in the daily practice of benchmark dose analysis, we do not consider typical theoretical measures such as bias, variance, and mean squared error for the BMD point estimate.

Results

Case study 1

To illustrate the effect of hard and soft constraining of the fold change parameter on the resulting BMD estimation, we use the data from Hu et al.[47] shown in Table 2. The standard error of the mean (SEM) was converted to the standard deviation using $SD = SEM \times \sqrt{N}$. A plot with the dose-specific sample means (with vertical bars reflecting ± one standard deviation) is shown in Figure 2.

Figure 3 shows the fits of the selected candidate models when using Proast with a BMR of 5%. None of these models seem to fit the data sufficiently well. Note that the data in Figure 3 give a false impression of homoscedasticity (not reflecting the data in Table 2), as the vertical bars around the dose-specific sample means seem to suggest constant variance. Moreover, the assumed lognormal distribution implies a constant coefficient of variation (implying heteroskedasticity). Ignoring these violations of assumptions, the frequentist model-averaged BMD estimate is 15.92 with 90% confidence interval [0.917, 333], and the model weights are 0.044, 0.163, 0.505, and 0.288 for the exponential, Hill, inverse exponential, and lognormal models, respectively. When using BMABMDR, only the normal distribution is given model weights above 0, and the models seem to provide a much better fit to the data (left panel of Figure 2). The B-BMD tool embedded in the EFSA web application also warns the user that Bartlett's test rejects the assumptions of constant variance and constant coefficient of variation (p-value 0.0000 for both). In such a situation, the B-BMD app performs a sensitivity analysis assuming different variances. As only summary data are available, the app does not report violations of the normal or lognormal assumption. Using the data as they are, the Bayesian model-averaged BMD estimate is 16.41 with 90% credible interval [6.33, 147.88]. These results differ substantially from the F-BMD results. For all models, the estimated fold change parameter is approximately 0.91, whereas in the F-BMD analysis, the estimate hits the default constraint of 1/1.3 = 0.7692 for all models. The left panel of Supplementary Figure S1 shows the poor fits of all 1000 bootstrap samples used to construct the confidence interval in Proast, questioning the validity of its inference. When removing the default constraint of 1/1.3 on the fold-change parameter, the inverse exponential and lognormal models, and consequently the model-averaged curve, provide a much better fit to the data (Supplementary

Table 2. Case study 1. Summary data from the Hu et al.[47] study. SEM = standard error of the mean.

Dose	Mean	SEM	N
0	0.684	0.003	8
1	0.677	0.001	8
10	0.649	0.011	8
100	0.630	0.012	8
1000	0.625	0.008	8
10000	0.618	0.002	8

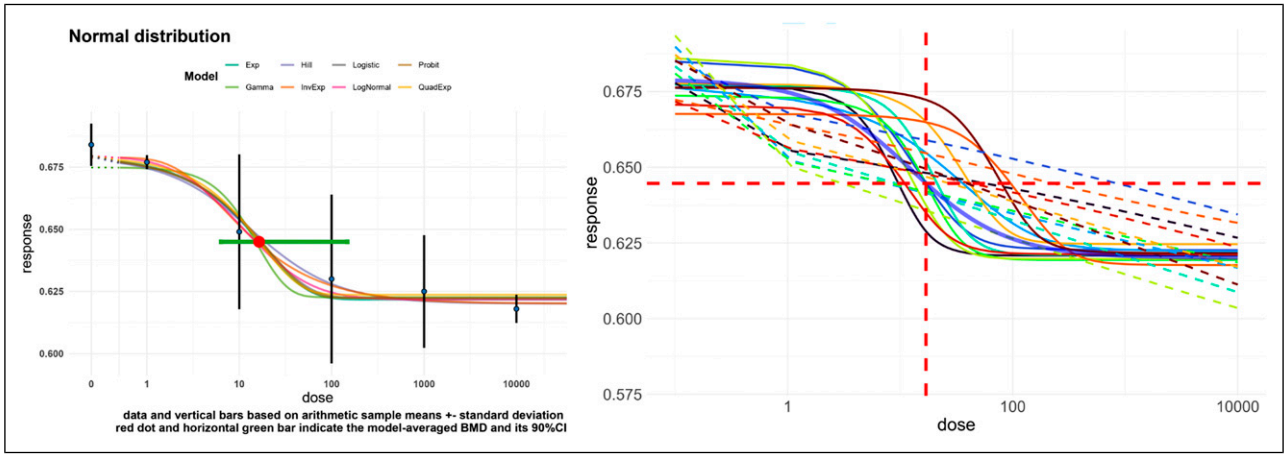

Figure 2. Case study 1, Bisphenol A data. The left panel shows the individual model fits from the BMABMDR analysis. The right panel shows the model-averaged median response curves for the first 10 simulated datasets (blue transparent solid curve shows the true underlying response curve derived using the Hill model with parameters estimated by BMABMDR, solid colored lines show the BMABMDR model averaged fits, and dashed colored lines show the Proast model averaged fits in corresponding colors). The horizontal and vertical red dashed lines in the right panel represent the true BMD at the level of BMR 5%.

Figure S2 and right panel of Supplementary Figure S1). The resulting model-averaged BMD estimate in this unconstrained analysis is 10.10 with 90% confidence interval [2.45, 37.3], and the model weights are 0.002, 0.002, 0.541, and 0.456 for the exponential, Hill, inverse exponential, and lognormal models, respectively. Note that the fold change in the exponential and Hill 'm3' model is fixed at 10^{-18}.

To gain further insights, 1000 datasets were generated according to the design of the Bisphenol A study, using the Hill model with parameters estimated by BMABMDR. To comply with the assumptions of both frameworks, the lognormal distribution was used to simulate the data, with the control-group variance set to the pooled estimate of the observed variances across all dose levels. As in the analysis of the original data, the BMR was set to 5%, corresponding to a true BMD value of 16.95. The right panel of Figure 2 shows the fitted model averaged median response curves for the first 10 simulated datasets for both approaches, on top of the true response curve. As observed in the original analysis, the B-BMD curves appear to fit the data much better. Approximately 35% of the datasets (353 runs) did not exhibit a dose-response relationship (i.e. absence of dose effect as defined in Methods) and were excluded from the results for both frameworks. Across the remaining 647 runs, the left coverage of the intervals was 90% for BMABMDR and 96% for Proast, and total coverage was 87% and 94%, respectively. Compared to the nominal levels of 95% and 90%, coverage was slightly too low for BMABMDR and somewhat too high for Proast.

Figures 4(a) and 4(b) show that the BMDLs for BMABMDR are much closer to the true BMD value than the Proast ones, and that those for Proast are often extremely small. Figures 4c-(f) confirm that the main reason for the surprisingly good performance of Proast in terms of coverage is the result of very low BMDL estimates. Consequently, the ratios for BMDU/BMDL and BMD/BMDL are extremely high for Proast. Applying the EFSA accuracy thresholds for applicability of the BMDL as RP,[23] 99%, 83%, and 84% of the BMABMDR estimates satisfy the thresholds BMD/BMDL < 20, BMDU/BMDL < 50, or both, respectively. On the other hand, only 67%, 3%, and 3% of the Proast estimates meet these thresholds. The median left width is 2.86 for BMABMDR and 14.61 for Proast (Figures 4e-(f)), and the median total width of the BMD intervals is 10.25 for BMABMDR and 267.34 for Proast (Figures 4c-(d)). These geometric width measures indicate a much more accurate estimation when using BMABMDR. For both methods, no BMD estimates are 10 times lower than the lowest non-zero dose. Applying these EFSA thresholds would thus imply that almost no BMDL estimates resulting from the Proast analysis would be considered useful as an RP for setting HBGVs.

Another simulation scenario with an increasing dose-response curve and a design range that does not include sufficiently high dose levels for estimating the fold change or DRM(∞) parameter is presented in the Appendix (left panel of Supplementary Figure S3). In contrast to the BPA example and simulations, where the observed decreasing data pattern nearly reached DRM(∞), this second illustration covers the case where the data clearly do not. The main conclusion from these illustrations is that there are instances in which the estimation of the fold change or DRM(∞) parameter needs some regularisation. The F-BMD implementation in Proast includes some hard thresholding, likely based on experiences with certain datasets, but this may be insufficiently adaptive to other datasets. Statistical checking of sufficient 'flatness' at the highest dose levels of the dataset at hand, combined with an adaptive prior for the DRM(∞) parameter, offers a more flexible

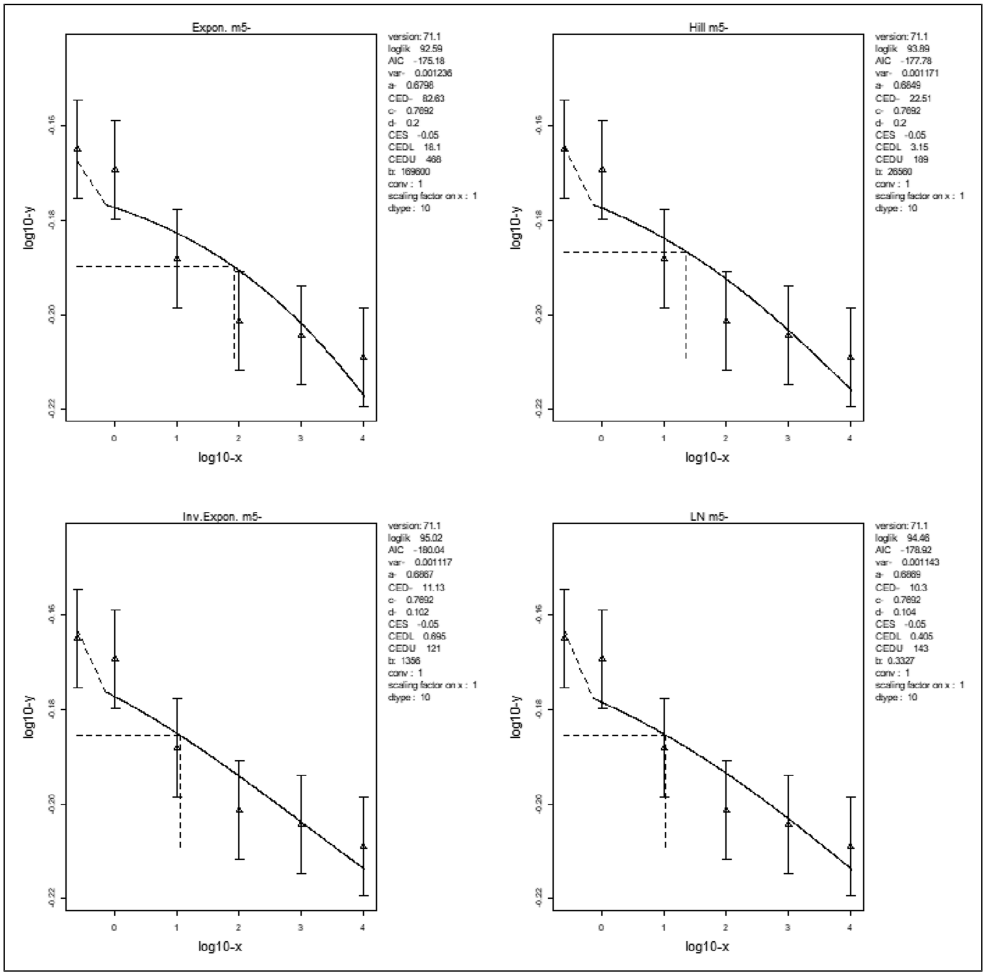

Figure 3. Case study 1, Bisphenol A: Individual model fits from the Proast analysis.

alternative. No approach will handle all possible data and design situations perfectly, but the adaptive B-BMD approach implemented in BMABMDR shows improved performance characteristics. Additionally (not illustrated here), the use of an informative prior for the DRM(∞) parameter, based on historical data or expert knowledge elicitation, provides an additional asset in B-BMD.

Case study 2

Simulation setting considering d *to be larger than 1 in combination with a poor design*

An example of a simulated dataset is shown in the left panel of Figure 5 and in Table 3. Supplementary Figure S7 shows the fits of the selected candidate models for this particular dataset when using Proast with a BMR of 5%. The model-averaged BMD estimate is 0.00001 with 90% confidence interval [1e-6, 0.115], and the model weights are 0.19, 0.27, 0.27, and 0.27 for the exponential, Hill, inverse exponential, and lognormal models, respectively. When using BMABMDR, the log-normal distribution is assigned 99.5% of the model weights. The model-averaged BMD estimate is 0.143 with 90% credible interval [0.014, 0.336]. The results from Proast differ substantially from those of BMABMDR, with the latter showing a more precise BMD interval, a higher BMD point estimate, and a higher BMDL. Using Proast, the BMDL hits the lower constraint. The left panel in Figure 6 displays the individual model fits from BMABMDR, indicating that most of them accurately fit the data. In the right panel, the fits of a subset of the bootstrap samples used to construct the Proast confidence interval are shown. Supplementary Figure S9 shows that most BMABMDR models correctly estimate the d parameter, whereas the Proast models estimate d to be 0.266, 0.355, 0.316, and 0.209 for the exponential, Hill, inverse exponential, and lognormal models, respectively (Supplementary Figure S7).

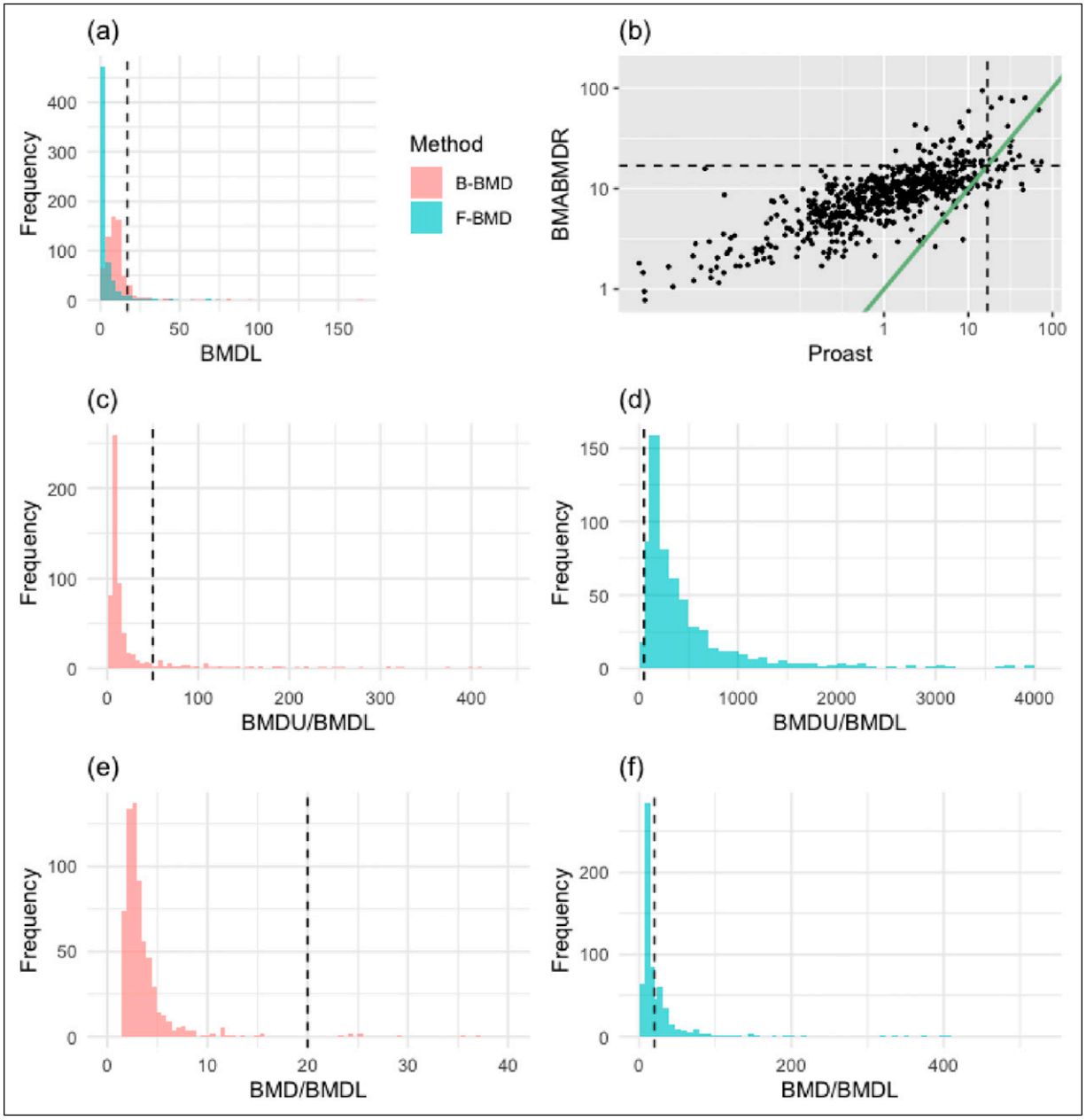

Figure 4. Case study 1, Bisphenol A simulations: (a) histogram of BMDL values, (b) scatterplot of BMDL values, (c,d) histogram of BMDU/BMDL ratios, (e,f) histograms of BMD/BMDL values; red histograms for B-BMD (BMABMDR), blue for F-BMD (Proast). Dashed lines show the (a-b) true BMD and (c-f) EFSA accuracy thresholds for applicability of the BMDL.

Figures 7a-(b) show that the BMDLs for BMABMDR are quite close to the true BMD value, and that those for Proast are extremely small, with 69% equal to 10^{-6}, which is the lower constraint in Proast. The left coverage of the intervals was 100% for both BMABMDR and Proast, and total coverage was 100% and 99.9%, respectively. Both approaches perform very well on these coverages, albeit too conservatively. Figures 7c-(f) reveal that the ratios for BMDU/BMDL and BMD/BMDL are extremely high for Proast, implying highly inaccurate estimation in the vast majority of runs. Applying the EFSA accuracy thresholds for applicability of the BMDL as RP, 100%, 99.2%, and 99.2% of the BMABMDR estimates satisfy the thresholds BMD/BMDL < 20, BMDU/BMDL < 50, or both, respectively. On the other hand, only 22.5%, 0%, and 0% of the Proast estimates meet these thresholds. The median left width is 10.9 for BMABMDR and 73918.4 for Proast (Figures 7e-(f)); the median total width is 27.9 for BMABMDR and 122386.1 for Proast Figures 7c-(d)). No BMABMDR estimates are 10 times lower than the lowest non-zero dose, whereas 45.1% of the Proast estimates are. Applying these EFSA thresholds would thus imply that no BMDL estimates resulting from the Proast analysis would be considered useful as an RP for setting HBGVs.

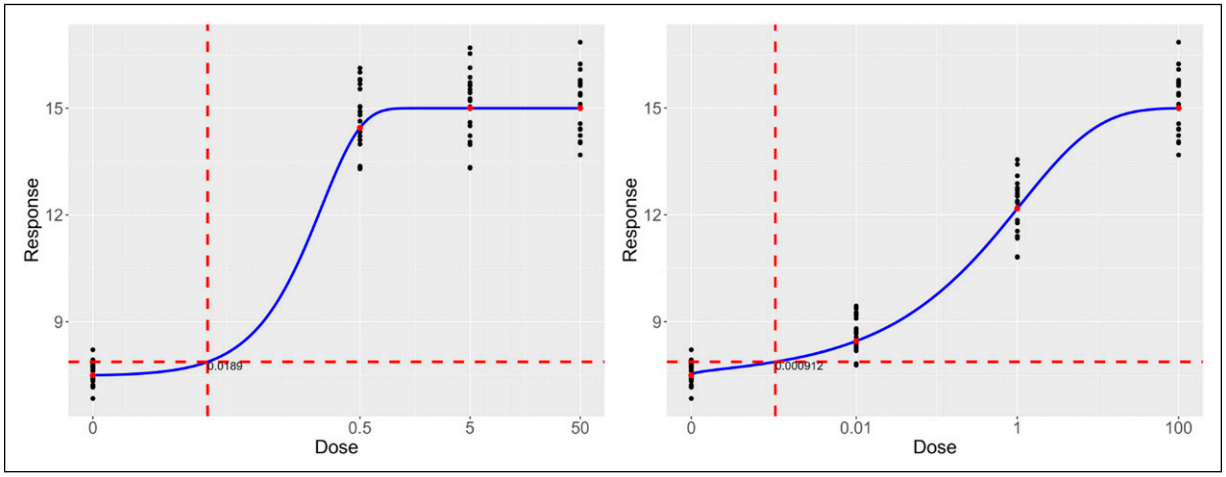

Figure 5. Case study 2: Example of a simulated dataset considering $d = 1.2$ in combination with a poor design (left panel) and $d = 0.4$ (right panel).

Simulation setting considering d *to be smaller than 1*

An example of a simulated dataset is shown in the right panel of Figure 5 and in Table 4. Supplementary Figure S8 shows the fits of the selected candidate models when using Proast with a BMR of 5%. The model-averaged BMD estimate is 0.002 with 90% confidence interval [0.001, 0.005], and the model weights are 0.25 for each of the four models. When using BMABMDR, the lognormal distribution is given all of the model weights. The model-averaged BMD estimate is 0.003 with 90% credible interval [0.001, 0.005]. These results are similar to those obtained using Proast. The left panel in Figure 8 displays the individual model fits from BMABMDR, showing that most of them accurately fit the data. In the right panel, the fits of a subset of the bootstrap samples used to construct the confidence interval in Proast are shown. Supplementary Figure S9 shows that most BMABMDR models correctly estimate the d parameter, as do the Proast models, which estimate d to be 0.45, 0.59, 0.38, and 0.36 for the exponential, Hill, inverse exponential, and lognormal models, respectively (Supplementary Figure S8).

For this simulation scenario, Figures 9(a) and 9(b) show that the BMDLs for BMABMDR and Proast are quite similar. In terms of coverage, neither framework reached nominal levels. Left coverage was 77% for BMABMDR and 72.6% Proast, and total coverage was 77% and 72.3%, respectively. Figures 9c-(f) reveal that the ratios for BMDU/BMDL and BMD/BMDL are also quite similar for both frameworks. Applying the EFSA accuracy thresholds for applicability of the BMDL as RP, 100%, 99.7%, and 99.7% of the BMABMDR estimates satisfy the thresholds BMD/BMDL < 20, BMDU/BMDL < 50, or both, respectively; and 99.7%, 99.7%, and 99.7% of the Proast estimates meet these thresholds. The median left width is a bit lower for Proast (2.30 as compared to 3.00, see Figures 9e-(f)), as is the case for the median total width (5.25 as compared to 6.70, see Figures 9c-(d)). For BMABMDR, 3.6% of the BMD point estimates are 10 times lower than the lowest non-zero dose, and for the Proast point estimates, this percentage is 15.4%. The BMABMDR estimates therefore comply somewhat better with the EFSA criteria, but overall the performance of both frameworks is quite similar. Supplementary Table S2 in the Appendix presents the results for all performance measures across the four scenarios of case studies 1 and 2, demonstrating the superior performance of BMABMDR.

Table 3. Case study 2. Summary data from the selected simulated dataset considering d larger than 1 and poor design.

Dose	Mean	SD	N
0	7.7	0.397	20
0.5	14.3	0.794	20
5	14.9	0.892	20
50	15.2	0.698	20

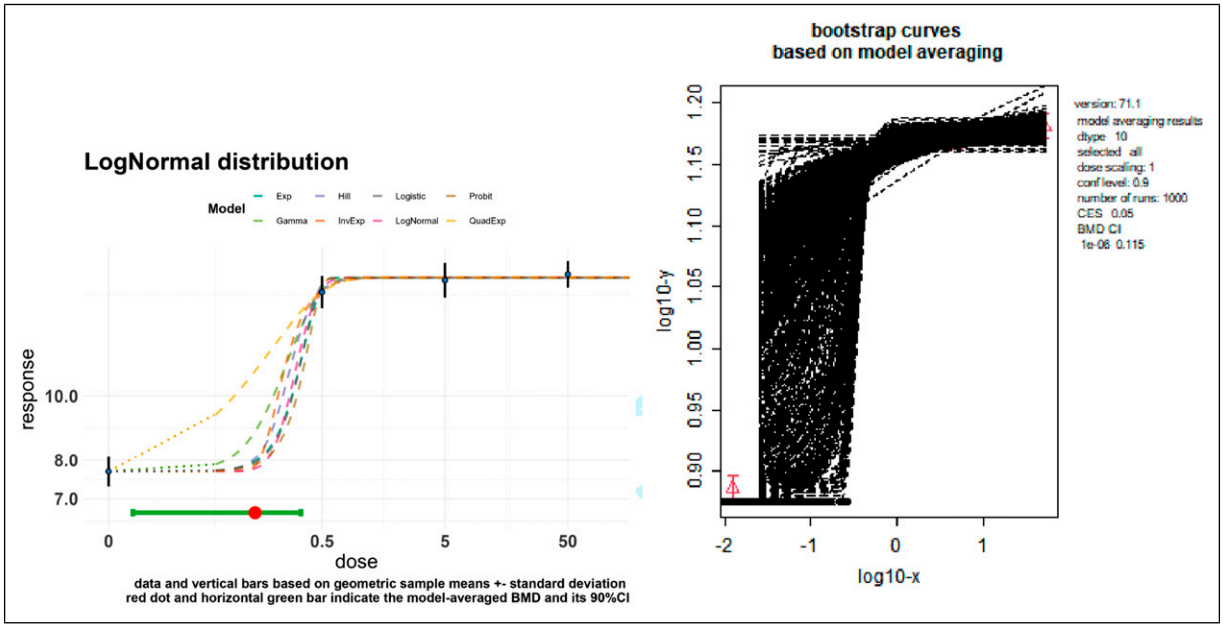

Figure 6. Case study 2, BMABMDR (left) and Proast (right) model averaging result of the simulated dataset with $d > 1$ and poor design.

Case study 3

In this case study we consider the quantal endpoint of the number of skeletal and internal malformations and variations from Blanco et al.,[53] analysed in Annex E of EFSA CONTAM Panel [54] and shown in Table 5. The BMR used is a 10% change in mean response compared to the controls. The total number of 8 observations per dose is small and it is very questionable that it is large enough for asymptotic approximations to be valid, especially for quantal data. Indeed, a major limitation of the application of the bootstrap as used by F-BMD is that for such small datasets, the number of distinct datasets that can be generated is very limited, resulting in a highly discrete bootstrap distribution. The standard choice for the number of bootstrap runs in Proast is 200, which is considered to be rather small for obtaining accurate estimates of the quantiles of the bootstrap distribution, and it is generally recommended to use 1000 or even 2000 or more bootstrap runs. However, if the number of possible bootstrap samples and resulting bootstrap replicates for the BMD estimate is limited, even increasing the number of bootstrap runs will not necessarily yield more accurate quantile estimates. In contrast, in the B-BMD framework, MCMC methods are used to approximate the posterior distribution of the BMD parameter and the credible interval is directly inferred from this posterior distribution Gilks et al.[55.]

The left panel of Figure 10 shows the model-averaged dose-response curve and BMD posterior obtained in the BMABMDR analysis, resulting in the 90% credible interval [0.176, 1.434] and BMD point estimate of 0.56. In the default setting of the BMABMDR package, the number of draws from the posterior distribution is 30000, the number of MCMC chains is 3, the number of MCMC iterations per chain is 3000, and the number of MCMC iterations discarded as burn-in per chain is 1000. Doubling these numbers will increase precision and result in very similar estimates: a BMD point estimate of 0.575 and 90% credible interval [0.205,1.481].

The right panel of Figure 10 shows bootstrap model-averaged curves obtained when using Proast. The quantile bootstrap interval based on 1000 bootstrap runs equals [0.000021, 0.62], resulting in a BMDU/BMDL ratio of 29523 (compared to a ratio of 8.15 for BMABMDR). According to the EFSA accuracy thresholds, this implies that the BMD analysis is highly inaccurate and the BMDL cannot be used as an RP. The figure also shows the discrete nature of the generated dose-response curves. The left panel of Figure 11 depicts the highly discrete bootstrap distribution based on 3000 bootstrap samples, and the right panel shows the bootstrap quantile intervals [BMDL,BMDU] for an increasing number of bootstrap samples. Whereas the BMDU seems to be rather stable, the BMDL is highly sensitive to the number of bootstrap replicates and varies from about 0 to above 0.25.

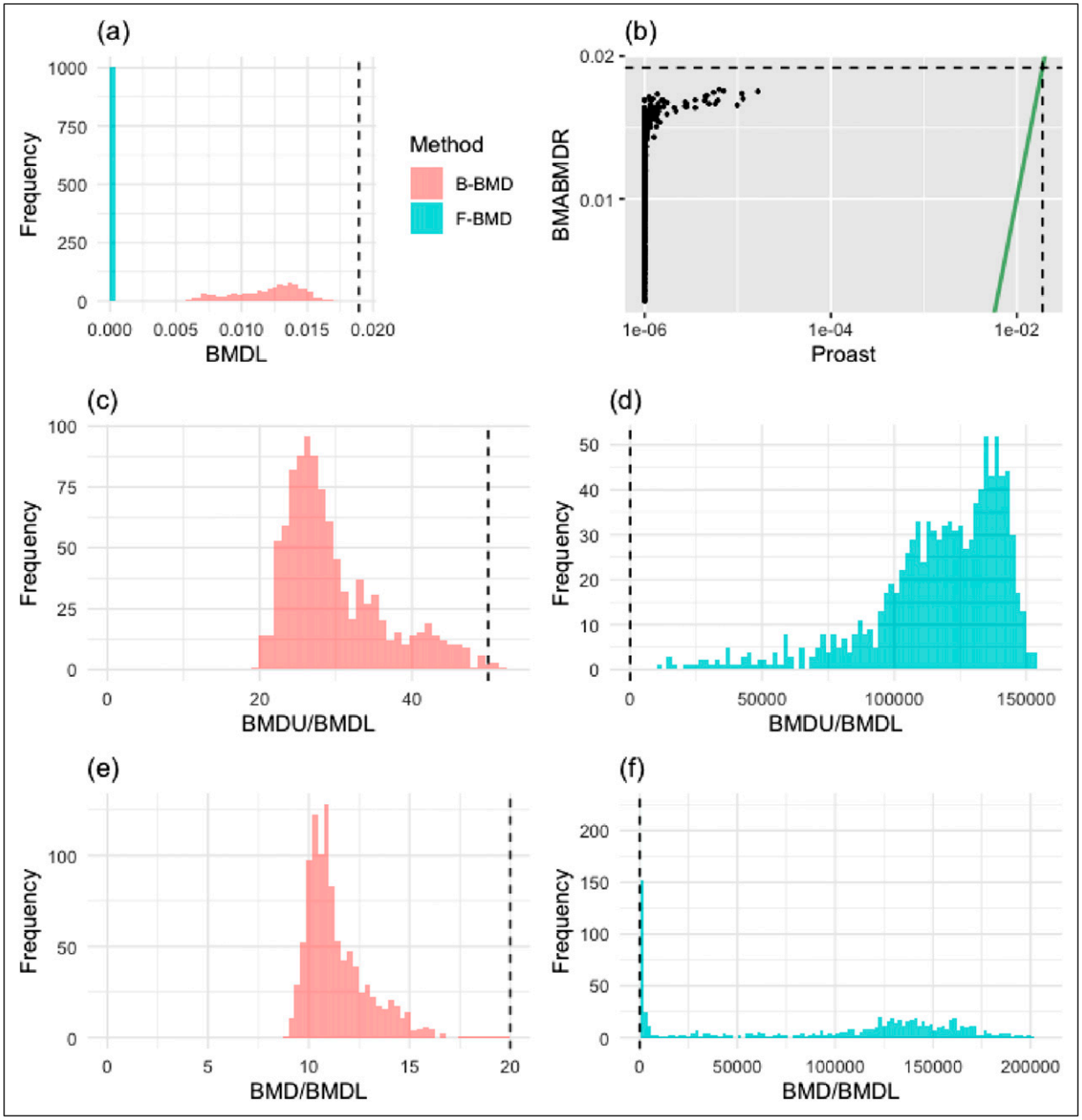

Figure 7. Case study 2, simulation setting considering $d = 1.2$ in combination with a poor design: (a) histogram of BMDL values, (b) scatterplot of BMDL values, (c,d) histogram of BMDU/BMDL ratios, (e,f) histograms of BMD/BMDL values; red histograms for B-BMD (BMABMDR), blue for F-BMD (Proast). Dashed lines show the (a-b) true BMD and (c-f) EFSA accuracy thresholds for applicability of the BMDL.

Case study 4

To illustrate the methodology for the construction of an overarching informative BMD prior for groups of chemicals, including how to combine historical studies as well as standardisation of the dose scale, we use a real-data example from the NTP organ database. For this example, we included 5 studies using compounds where the MoA was oxidative stress, and the compound was administered using air inhalation in female B63CF1N mice. The endpoint of interest was relative lung weight. The data of these 5 studies are shown in the left panel of Figure 12, and we consider the most recent study as the 'new' study (i.e.,Study 5) to which an informative prior based on Studies 1 to 4 will be applied for the BMD parameter. The summary data for each study are shown in Supplementary Table S4 in the Appendix. The estimated fold changes from the historical studies

Table 4. Case study 2, summary data from the selected simulated dataset considering d smaller than 1.

Dose	Mean	SD	N
0	7.5	0.29	20
0.01	8.4	0.366	20
1	12.6	0.695	20
100	14.8	0.79	20

were 1.78, 3.60, 2.19, and 4.01, and using $f = 0.25$ this resulted in a BMR $q_f = 0.1942$ used for standardization of the dose scale. The right panel of Figure 12 shows the standardised data from the historical studies. After analysing each standardised historical dataset with $q = 0.05$, the BMD posteriors were converted back to their original scale and it was confirmed that these were consistent with the posteriors resulting from analysis of the original data. The left panel of Figure 13 shows the individual BMD posteriors from the analysis of each historical study on the standardised dose scale, with $q = 0.05$. The right panel of Figure 13 shows the informative PERT prior for the BMD based on mixing the posteriors of the historical studies (using equal weights), on the standardised dose scale. The 'new' dataset was then analysed using this informative prior, and the results of this analysis, as well as the default analysis, are shown in Table 6. As can be seen, applying an informative prior based on historical studies resulted in a similar BMD point estimate but with improved precision as indicated by the smaller ratios BMDU/BMDL and BMD/BMDL. Figure 14 shows the individual model fits of the default (left) and informative (right) analysis.

Discussion

Based on guidelines from EFSA and the WHO, a new implementation of Bayesian model averaging for determining benchmark doses is presented. The tool is developed in R using rstan and is available as an R package and as an EFSA web application. The dose-response models are defined in terms of natural and technical parameters. Prior distributions act on three levels: the distribution for the endpoint, the models for the median response within a particular distribution, and the parameters of the median response model. At the first two levels, the priors are categorical; at the third level, the natural

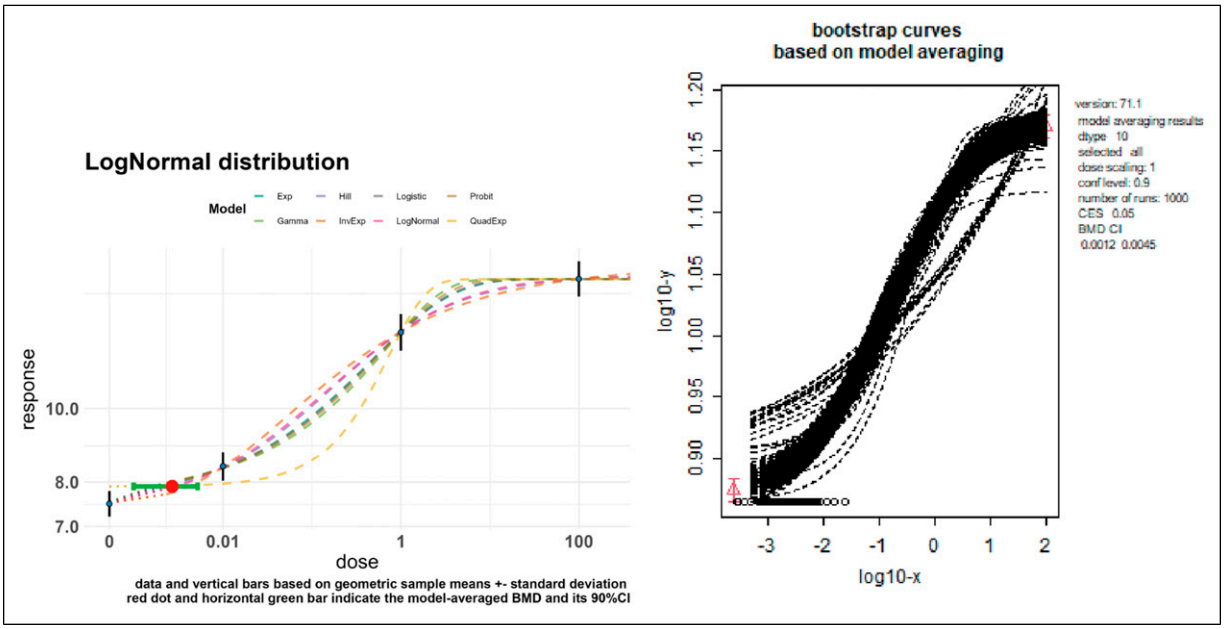

Figure 8. Case study 2, BMABMDR (left) and Proast (right) model averaging result of the simulated dataset with $d < 1$.

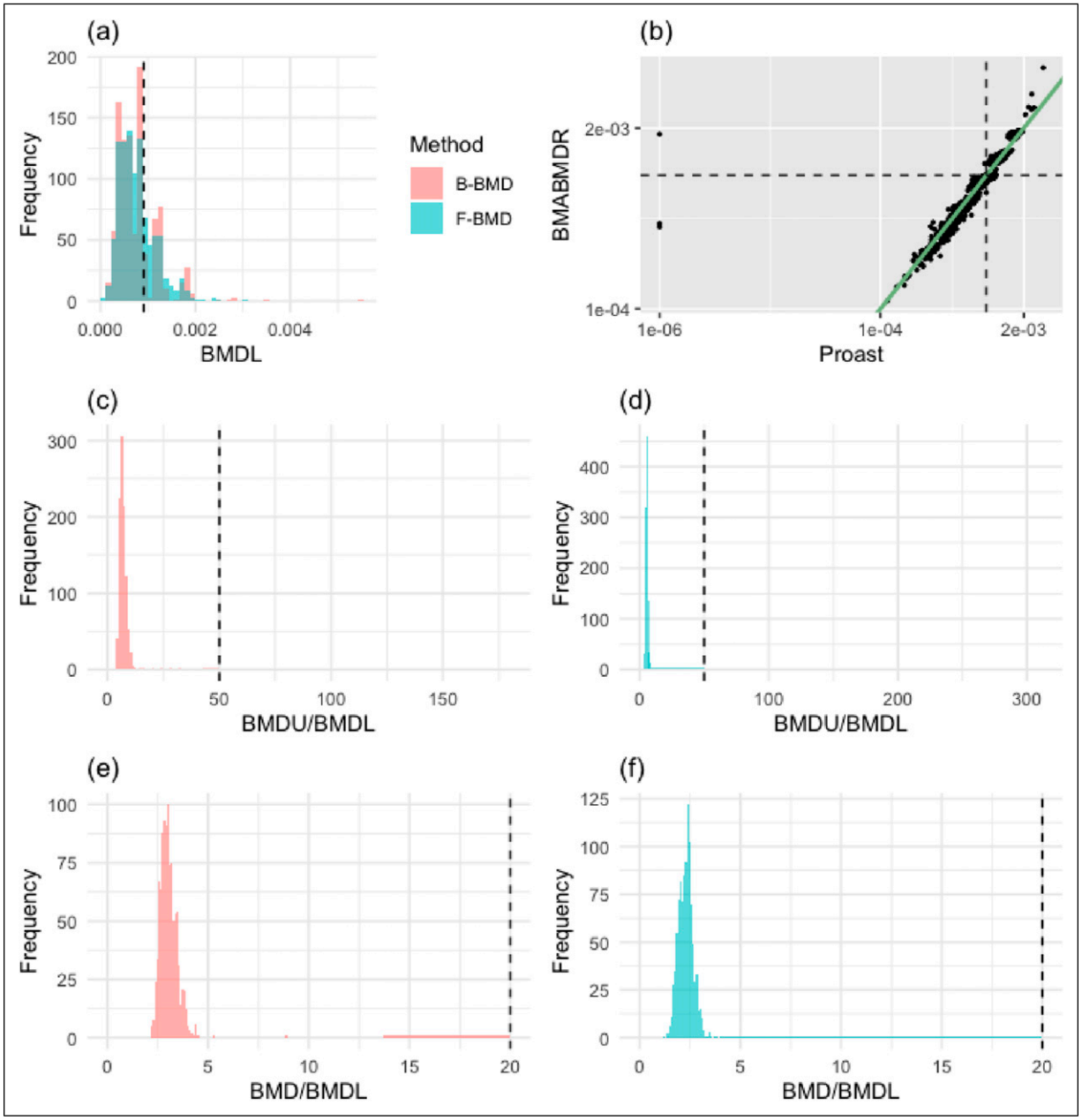

Figure 9. Case study 2, simulation setting considering $d = 0.4$: (a) histogram of BMDL values, (b) scatterplot of BMDL values, (c,d) histogram of BMDU/BMDL ratios, (e,f) histograms of BMD/BMDL values; red histograms for B-BMD (BMABMDR), blue for F-BMD (Proast). Dashed lines show the (a-b) true BMD and (c-f) EFSA accuracy thresholds for applicability of the BMDL.

Table 5. Case study 3. Data from Blanco et al.[53]: dose (col 1), number of adverse events (col 2), total number of experimental units (col 3).

BDE-99 (mg/Kg bw/day)	Skeletal and internal malformations and variations	Litters
0	6	8
0.5	5	8
1	8	8
2	8	8

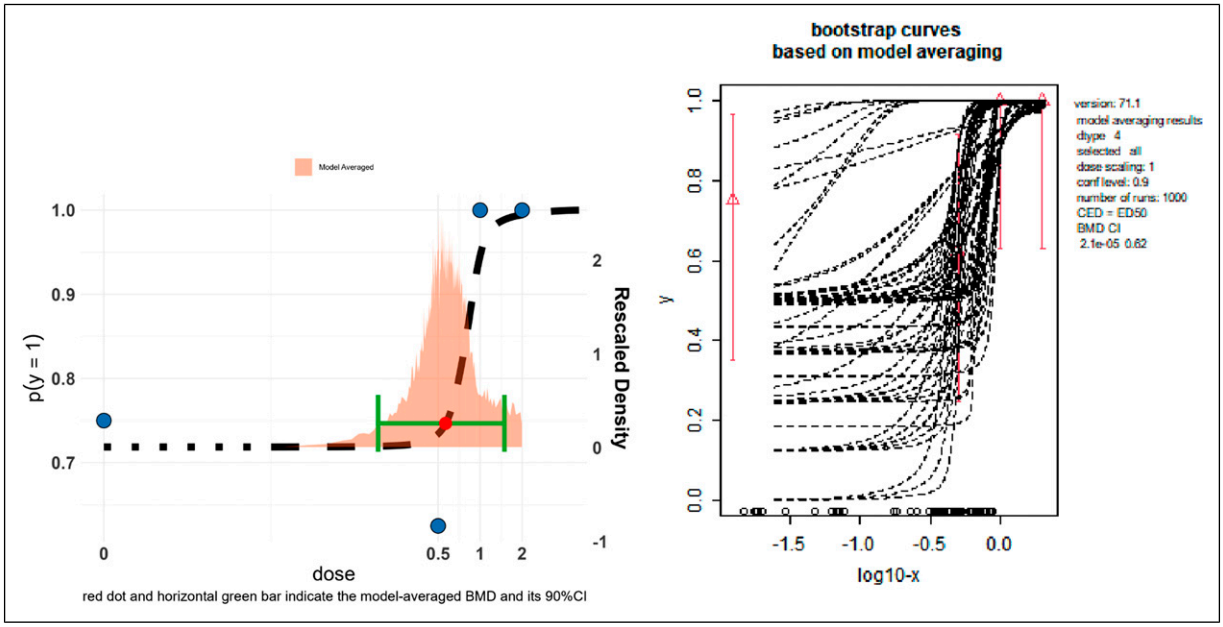

Figure 10. Case study 3. Data from Blanco et al..[53] The left panel shows the BMABMDR model-averaged dose-response curve, BMD posterior and BMD credible interval (green). The right panel shows a subset of repeated bootstrap model-averaged dose-response curves from Proast, generating the bootstrap distribution for the BMD (open dots on horizontal axis).

parameters receive modified PERT priors, and the technical parameters receive lognormal priors. Default uniform discrete priors are operating at the first two levels, and a continuous uniform PERT prior is used for the BMD parameter. For the asymptotic median response DRM(∞), a data-adaptive weakly informative prior is applied, and a weakly informative lognormal prior is defined for the technical power parameter d. The presented case studies illustrate the merits of regularising default priors and of using informative priors. In the first and second case studies, we showed that the maximum response or fold change parameter, as well as the d parameter, require regularisation when the experimental design is not optimal in order to obtain reliable BMD estimates that can be used as an RP. We show that hard constraining, as used in the frequentist framework, is not optimal for controlling the fold change parameter, as it may result in underestimation of the BMD. Checking for flatness and regularising with prior distributions for the model parameters performs better, enabling an adaptive, data-driven approach. Applying an informative prior based on historical data or expert knowledge elicitation is an additional asset in the Bayesian framework, as illustrated in the last case study. In the third case study, we showed that the discreteness of the bootstrap distribution often used to construct confidence intervals in the frequentist framework might be an issue for small samples of quantal data.

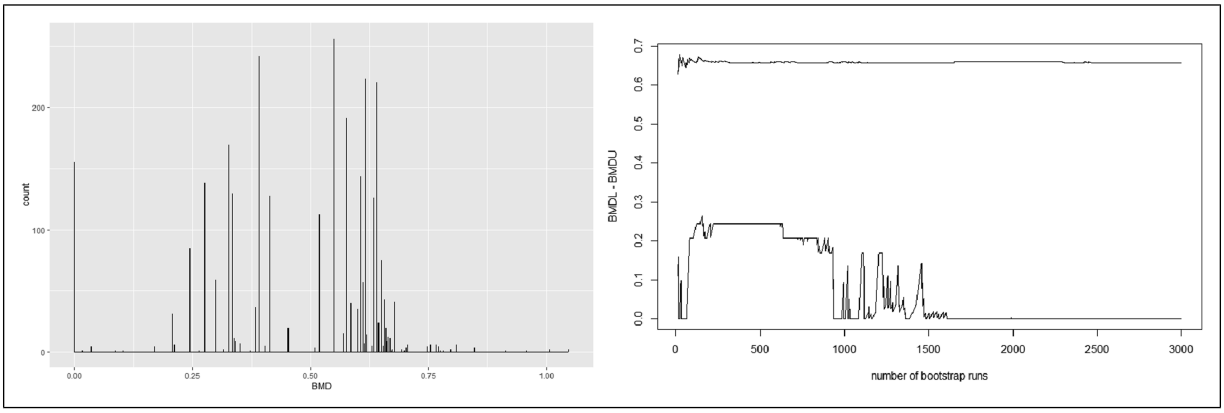

Figure 11. Case study 3. Proast analysis of the data from Blanco et al..[53] The left panel shows the bootstrap distribution of the BMD using 3000 replicates. The right panel shows the change in bootstrap confidence interval [BMDL,BMDU] by the number of bootstrap replicates.

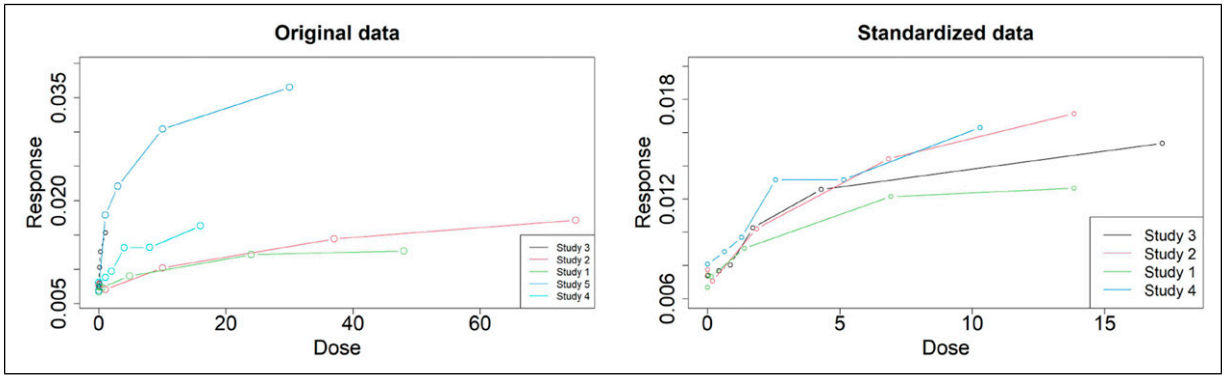

Figure 12. Case study 4. The left panel shows the summary data of the historical studies and the 'new' study (Study 5). The right panel shows the summary data of the historical studies on the standardised dose scale.

Relevance for regulatory decision-making

The case studies show that, in certain settings, the frequentist approach as implemented in Proast may yield BMDL values that are practically unsuitable for regulatory decision-making, as they may be unrealistically low and reflect a methodological artefact (a hard constraint or a bootstrap failure) rather than a biologically meaningful result. Such values may lead to overly conservative HBGVs, potentially resulting in disproportionate risk-management measures. In these settings, a fully Bayesian probabilistic approach with regularising and data-adaptive components seems to yield BMDL values that align with expectations derived from simulation settings and biological plausibility. Also, the potential to translate the accumulation of data evidence across multiple datasets into informative priors that increase the precision of the BMD estimate enhances confidence in the estimates and supports their use in regulatory contexts. Finally, both approaches, F-BMD and B-BMD, are computationally intensive. The computational burden of constructing frequentist bootstrap intervals in Proast can, however, be considerably higher than that of MCMC-based credible intervals in BMABMDR. Routine use across a large collection of datasets is therefore expected to be more feasible with B-BMD. In conclusion, the B-BMD approach, consistent with EFSA's 2022 guidance update, is a promising step forward in regulatory toxicology.

Limitations

The value of the default prior and the potential of informative priors have been highlighted in this paper. Although the current default priors, balancing uninformativeness with soft constraints, have been developed based on extensive simulation studies and have been shown to be appropriate for the analysis of many datasets, it is recommended to perform sensitivity analyses to assess the robustness of the results to alternative choices. The EFSA app and the R package BMABMDR allow the user to examine the effects of two predefined priors for the d parameter (assigning different probabilities to $d < 1$) and two priors for the BMD parameter (allowing the BMD to extend beyond the experimental range or not). They also allow the user to easily

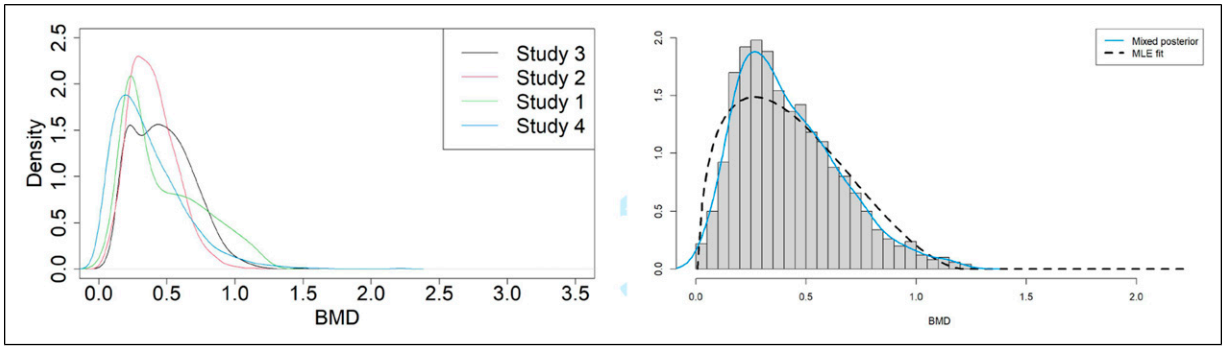

Figure 13. Case study 4. The left panel shows the model-averaged BMD posteriors from each historical study, on the standardised dose scale. The right panel shows the histogram of mixed posteriors (blue line: density) and the fitted informative PERT prior (black dashed line) based on historical studies, on the standardised dose scale.

Table 6. Case study 6. Results from the analysis of the 'new' dataset using the default uninformative BMD prior and the informative priors based on the mixing of posteriors from historical datasets.

Analysis	BMDL	BMD	BMDU	BMDU/BMDL	BMD/BMDL
Default	0.0006	0.0098	0.0437	69.434	15.551
Informative	0.0015	0.0074	0.0192	12.696	4.912

change the prior around its default values. If the experimental range is not sufficiently wide, a sensitivity analysis to confirm the parameter DRM(∞)'s robustness to its prior is a recommended exercise.

In case study 3, the comparison of Bayesian MCMC-generated credible intervals with frequentist bootstrap intervals is limited to the basic percentile bootstrap interval. Improved bootstrap methods have however been proposed for frequentist model averaging (e.g. Garcia-Angulo and Claeskens[51]). Also, other types of Bayesian credible intervals beyond equal-tailed intervals could be considered (e.g., highest posterior density intervals or shortest probability intervals). An implementation and comparative study of such improved intervals in the context of BMD estimation, however, is beyond the scope of this paper. Moreover, given the prominent role of the BMDL (defined as the 5% quantile), the percentile interval or equal-tailed credible interval seems the most natural choice.

In case study 4, we limited ourselves to an illustration of constructing an informative prior based on historical studies. However, many important questions need to be addressed in future research, such as criteria for sufficient exchangeability of contributing studies, management of the risk of prior-data conflict, and bias propagation. In addition, recommendations need to be formulated before the construction of informative priors becomes a daily practice.

Experiences from case studies and results from simulations provide important insights into the settings considered, but, of course, also have limitations. There are many variations in dataset characteristics and simulation scenarios (design, sample size, etc.), implying that results and conclusions primarily hold for the settings considered and generalisation beyond these needs to be made with caution. A comprehensive evaluation across a broader range of regulatory toxicology datasets, encompassing diverse endpoints and substances, remains an important area of future research.

Future research

Current ongoing research is pursuing the topics presented in case studies 2 and 4. The parameter d is the only parameter of the median dose-response models that lacks a unique biological interpretation and a corresponding mathematical definition. The use of informative priors would benefit from dose-response models that are fully described by common natural parameters with a clear biological interpretation. Criteria for selecting historical studies that are sufficiently similar (in terms of species, endpoint, mechanism of action, study protocol, etc.), how to combine the selected studies, dealing with differences in dose scaling, and ways to standardise dose scales were topics touched upon in Kremer et al.[44]. Further research on these important topics, including non-exchangeability and bias propagation, is part of a running follow-up project that also includes the

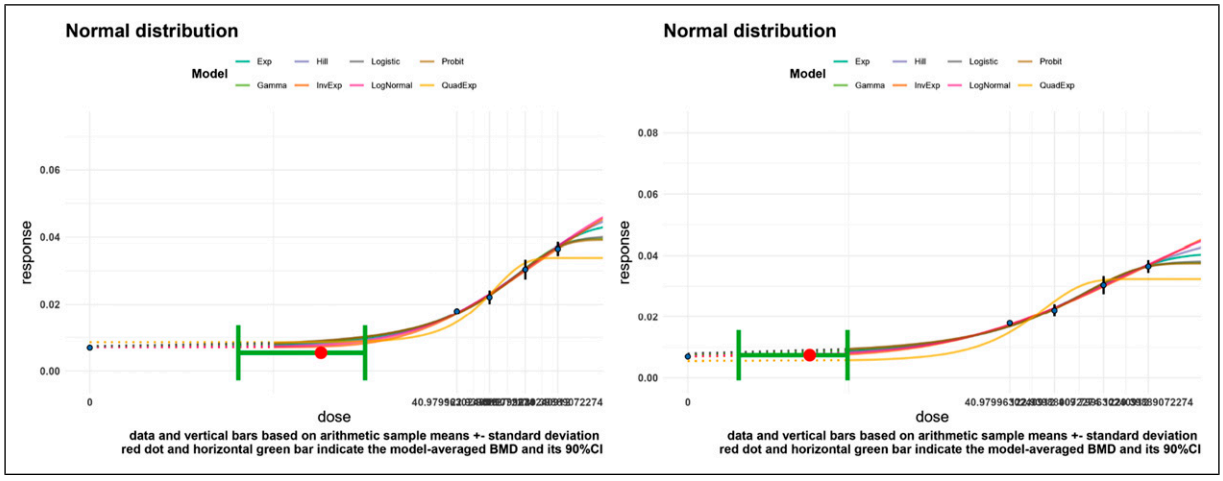

Figure 14. Case study 4. Individual models fits (normal distribution) for the default (left) and informative (right) analysis.

creation of an inventory of informative priors based on the repository of posteriors (available at Zenodo) and an implementation in EFSA's web application, ready for use by toxicologists.

The family of candidate models extends essentially all models used by EFSA and Proast in the past, and these have proven to cover all use cases in daily practice at EFSA over the last few decades. However, future datasets of new substances and/or new endpoints might show different or unexpected dose-response relationships. If BMABMDR's goodness-of-fit test indicates that none of the models fit sufficiently well, this might indicate that the current suite of models is not rich enough and some expansion is needed. We have not encountered this need to date, but it might surface in the future.

Further extensions of interest are BMD estimation for other types of data, including count data, zero-inflated data, ordinal or nominal multi-category responses, and longitudinal data. More methodological research is also needed to translate the BMD approach for determining an RP from experimental animal studies to other types of data, including in vitro cytotoxicity experiments, omics data, and observational epidemiological studies. Central issues are related to the lack of control data and the appropriate definition of the BMR. Haber et al.[1] suggested that additional research would be useful on methods for incorporating biological considerations into the weights used in the model averaging. Although converting biological insights into an informative prior for the model weights is not trivial, the B-BMD framework already allows the use of such priors.

Acknowledgements

The authors thank the reviewers and the editor for their constructive comments, which have improved the manuscript. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation Flanders (FWO) and the Flemish Government department EWI.

ORCID iD

Marc Aerts https://orcid.org/0000-0002-1803-9072

Ethical considerations

Ethical approval was not required for the present manuscript.

Author Contributions

All authors contributed equally to this work.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The BMABMDR R package is freely available at https://github.com/cecilekremer/BMABMDR. All data used for analyses are publicly available. R code for the simulation studies can be made available upon request. All analyses were performed using R version 4.4.3 and rstan version 2.32.7.

Disclaimer

The authors E. Solazzo and J. Cortiñas Abrahantes are currently employed by the European Food Safety Authority (EFSA) in the Methodology and Scientific Support Unit (MESE), which provides scientific and administrative support to the Scientific Committee, its working groups and EFSA's scientific networks for safety assessments and for the development of cross-cutting methodologies and guidance. This article reflects their opinions. While the opinions expressed in this article are consistent with previous work published by EFSA, this article cannot be considered as an EFSA scientific output.

Supplemental material

Supplemental material for this article is available online.

References

1. Haber L, Dourson M, Allen B, et al. Benchmark dose (BMD) modelling: current practice, issues, and challenges. *Critical Reviews in Toxicology* 2018; 48(5): 387–415. https://doi.org/10.1080/10408444.2018.1430121
2. Muri S, Schlatter J and Brüschweiler B. The benchmark dose approach in food risk assessment: is it applicable and worthwhile? *Food and Chemical Toxicology* 2009; 49(12): 2906–2925. https://doi.org/10.1016/j.fct.2009.08.002
3. Crump K. A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology* 1984; 4(5): 854–871. https://doi.org/10.1016/0272-0590(84)90107-6
4. Travis K, Pate I and Welsh Z. The role of benchmark dose in a regulatory context. *Regulatory Toxicology and Pharmacology* 2005; 43(4): 280–291. https://doi.org/10.1016/j.yrtph.2005.07.003
5. EFSA Scientific Committee. Guidance of the scientific committee on a request from EFSA on the use of the benchmark dose approach in risk assessment. *EFSA Journal* 2009; 7(6): 1–72. https://doi.org/10.2903/j.efsa.2009.1150
6. US EPA. Benchmark dose (BMDS) technical guidance document. *Technical report, U.S. Environmental Protection Agency*, 2012. EPA/100/R-12/001.
7. WHO Chapter 5: Dose-response assessment and derivation of health-based guidance values. *Principles and Methods for the Risk Assessment of Chemicals in Food. 2nd ed* 2020; Environmental Health Criteria 240. WHO/IPCS.
8. Crump KS. Calculation of benchmark doses from continuous data. *Risk Analysis* 1995; 15(1): 79–89. https://doi.org/10.1111/j.1539-6924.1995.tb00095.x
9. Burnham K and Anderson D. *Model selection and multimodel inference: a practical information-theoretic approach (2nd ed)*. Springer-Verlag, 2002.
10. West W, Piegorsch W, na EP, et al. The impact of model uncertainty on benchmark dose estimation. *Environmentrics* 2012; 23(8): 706–716.
11. EFSA Scientific Committee. Hardy A, Benford D, et al. Update: Guidance on the use of the benchmark dose approach in risk assessment. *EFSA Journal* 2017; 15(1): 1–41. https://doi.org/10.2903/j.efsa.2017.4658
12. Bayarri M and Berger J. The interplay of Bayesian and frequentist analysis. *Statistical Science* 2004; 19(1): 58–80. https://doi.org/10.1214/088342304000000116
13. Fang Q, Piegorsch WW and Barnes KY. Bayesian benchmark dose analysis. *Environmetrics* 2015; 26(5): 373–382. https://doi.org/10.1002/env.2339
14. Shao K and Shapiro A. A web-based system for Bayesian benchmark dose estimation. *Environmental Health Perspectives* 2018; 126(1): 017002. https://doi.org/10.1289/ehp1289
15. Simmons S, Chen C, Li X, et al. Bayesian model averaging for benchmark dose estimation. *Environmental and Ecological Statistics* 2015; 22: 5–16. https://doi.org/10.1007/s10651-014-0285-4
16. Wheeler M and Bailer A. Properties of model-averaged BMDLs: a study of model averaging in dichotomous response risk estimation. *Risk Analysis* 2007; 27(3): 659–670. https://doi.org/10.1111/j.1539-6924.2007.00920.x
17. Wheeler MW, Blessinger T, Shao K, et al. Quantitative risk assessment: Developing a Bayesian approach to dichotomous dose-response uncertainty. *Risk Analysis* 2020; 40(9): 1706–1722. https://doi.org/10.1111/risa.13537
18. Shao K and Gift J. Model uncertainty and Bayesian model averaged benchmark dose estimation for continuous data. *Risk Analysis* 2014; 34(1): 101–120. https://doi.org/10.1111/risa.12078
19. Wheeler MW, Cortiñas Abrahantes J, Aerts M, et al. Continuous model averaging for benchmark dose analysis: Averaging over distributional forms. *Environmetrics* 2022; 33(5): e2728. https://doi.org/10.1002/env.2728
20. Morales KH, Ibrahim JG, Chen CJ, et al. Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. *Journal of the American Statistical Association* 2006; 101(473): 9–17. https://doi.org/10.1198/016214505000000961
21. Wheeler MW, Lim S, House JS, et al. Toxic R: A computational platform in R for computational toxicology and dose-response analyses. *Computational Toxicology* 2023; 25: 100259. https://doi.org/10.1016/j.comtox.2022.100259
22. Gift J and Davis A (2022) *Benchmark dose software version 3.3 (BMDS3.3)*. u.s. Environmental Protection Agency, Washington, DC.
23. EFSA Scientific Committee. More SJ, Bampidis V, et al. Guidance on the use of the benchmark dose approach in risk assessment. *EFSA Journal* 2022; 20(10): 7584. https://doi.org/10.2903/j.efsa.2022.7584
24. Interuniversity Institute for Biostatistics and statistical Bioinformatics. EFSA platform for Bayesian benchmark dose analysis. *EFSA Supporting Publications* 2022; 19(12): 7740E. https://doi.org/10.2903/sp.efsa.2022.EN-7740
25. Kremer C, Oluwafemi O, Shkedy Z, et al. Bayesian benchmark dose modelling Web app. EFSA. *Technical report* 2024; https://doi.org/10.5281/zenodo.10523648
26. Slob W, Bakker MI, Bokkers BGH, et al. The use of canonical dose-response models for benchmark dose analysis of continuous toxicological data. *Critical Reviews in Toxicology* 2025; 55(4): 437–461. https://doi.org/10.1080/10408444.2025.2464067
27. Abrahantes J, Solazzo E, Aerts M, et al. Letter to the editor: Critical analysis of "The use of canonical dose-response models for benchmark dose analysis of continuous toxicological data" by Slob et al. (2025). *Critical Reviews in Toxicology* 2025; 1–5. https://doi.org/10.1080/10408444.2025.2581776

28. Aerts M, Wheeler M and Abrahantes J. An extended and unified modeling framework for benchmark dose estimation for both continuous and binary data. *Environmentrics* 2020; 31(7): 15. https://doi.org/10.1002/env.2630
29. Bennett CH. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* 1976; 22(2): 245–268. https://doi.org/10.1016/0021-9991(76)90078-4
30. Gronau QF, Sarafoglou A, Matzke D, et al. A tutorial on bridge sampling. *Journal of Mathematical Psychology* 2017; 81: 80–97. https://doi.org/10.1016/j.jmp.2017.09.005
31. Gronau QF, Singmann H and Wagenmakers EJ (2020) bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software* 92(10): 1–29. https://doi.org/10.18637/jss.v092.i10. URL. https://www.jstatsoft.org/index.php/jss/article/view/v092i10
32. Guo J, Gabry J, Goodrich B, et al. (2020) Package 'rstan'. https://cranr-projectorg/web/packages/rstan2020
33. Efron B and Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1986; 1(1): 54–75. https://doi.org/10.1214/ss/1177013815
34. Hjort N and Claeskens G. Frequentist model average estimators. *Journal of the American Statistical Association* 2003; 98(464): 879–899. https://doi.org/10.1198/016214503000000828
35. Liu CA. Distribution theory of the least squares averaging estimator. *Journal of Econometrics* 2015; 186(1): 142–159. https://doi.org/10.1016/j.jeconom.2014.07.002
36. Vehtari A, Gelman A, Simpson D, et al. Rank-Normalization, Folding, and Localization: An improved R for assessing convergence of MCMC (with Discussion). *Bayesian Analysis* 2021; 16(2): 667–718.
37. Bornkamp B. Functional uniform priors for nonlinear modeling. *Biometrics* 2012; 68(3): 893–901. https://doi.org/10.1111/j.1541-0420.2012.01747.x
38. Bornkamp B. Practical considerations for using functional uniform prior distributions for dose-response estimation in clinical trials. *Biometrical Journal* 2014; 56(6): 947–962. https://doi.org/10.1002/bimj.201300138
39. Wheeler MW (2023) An investigation of non-informative priors for bayesian dose-response modeling. *Regulatory Toxicology and Pharmacology* 141: 105389. https://doi.org/10.1016/j.yrtph.2023.105389. https://www.sciencedirect.com/science/article/pii/S0273230023000570
40. Aerts M, Kremer C, Oluwafemi O, et al. Simulation results for the EFSA platform for Bayesian Benchmark Dose Analysis. *Technical report, EFSA* 2022; https://doi.org/10.5281/zenodo.7118583
41. Shao K. A comparison of three methods for integrating historical information for Bayesian model averaged benchmark dose estimation. *Environmental Toxicology and Pharmacology* 2012; 34(2): 288–296. https://doi.org/10.1016/j.etap.2012.05.002
42. Fang Q, Piegorsch W, Simmons S, et al. Bayesian model-averaged benchmark dose analysis via reparameterized quantal-response models. *Biometrics* 2015; 71(4): 1168–1175. https://doi.org/10.1111/biom.12340
43. Jensen SM, Kluxen FM and Ritz C. A review of recent advances in benchmark dose methodology. *Risk Analysis* 2019; 39(10): 2295–2315. https://doi.org/10.1111/risa.13324
44. Kremer C, Saenen N, Bijnens K, et al. (2026) Compilation of NTP and published dose-response data: study of informative priors in the EFSA platform for Bayesian benchmark dose analysis. *EFSA Supporting Publications* 23(2):EN9902. 129. https://doi.org/10.2903/sp.efsa.2026.EN-9902
45. Jeffreys H. *The theory of probability*. OuP, 1998.
46. Shao K, Gift JS and Setzer RW. Is the assumption of normality or log-normality for continuous response data critical for benchmark dose estimation? *Toxicology and Applied Pharmacology* 2013; 272(3): 767–779. https://doi.org/10.1016/j.taap.2013.08.006
47. Hu Y, Yuan D, Wu Y, et al. Bisphenol A initiates excessive premature activation of primordial follicles in mouse ovaries via the PTEN signaling pathway. *Reproductive Sciences* 2018; 25(4): 609–620. https://doi.org/10.1177/1933719117734700
48. Slob W and Setzer RW. Shape and steepness of toxicological dose-response relationships of continuous endpoints. *Critical Reviews in Toxicology* 2014; 44(3): 270–297. https://doi.org/10.3109/10408444.2013.853726
49. Kabaila P, Welsh A and Abeysekera W. Model-averaged confidence intervals. *Scandinavian Journal of Statistics* 2016; 43(1): 35–48. https://doi.org/10.1111/sjos.12163
50. Claeskens G and Hjort N. The Focused Information Criterion. *Am Stat Assoc* 2003; 98: 900–916. https://doi.org/10.1198/016214503000000819
51. Garcia-Angulo A and Claeskens C. Bootstrap for inference after model selection and model averaging for likelihood models. *Metrika* 2025; 88: 311–340. https://doi.org/10.1007/s00184-024-00956-2
52. Lesaffre E and Lawson A. *Bayesian Biostatistics*. John Wiley & Sons, Ltd, 2012.
53. Blanco J, Mulero M, Domingo J, et al. Gestational exposure to BDE-99 produces toxicity through upregulation of CYP isoforms and ROS production in the fetal rat liver. *Toxicological Sciences* 2012; 127: 296–302. https://doi.org/10.1093/toxsci/kfs082
54. EFSA CONTAM Panel. Update of the risk assessment of polybrominated diphenyl ethers (PBDEs) in food. *EFSA Journal* 2024; 22(1): e8497.
55. Gilks WR, Richardson S and Spiegelhalter D. Markov chain Monte Carlo in practice. CRC Press.