Taylor & Francis
Taylor & Francis Group

Critical Reviews in Toxicology

ISSN: 1040-8444 (Print) 1547-6898 (Online) Journal homepage: www.tandfonline.com/journals/itxc20

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)

José Cortiñas Abrahantes, Efisio Solazzo, Marc Aerts, Peter Craig, Ursula Gundert-Remy, Thor Halldorsson, Laurentius Hoogenboom, Salomon Sand & Thomas Tietz

To cite this article: José Cortiñas Abrahantes, Efisio Solazzo, Marc Aerts, Peter Craig, Ursula Gundert-Remy, Thor Halldorsson, Laurentius Hoogenboom, Salomon Sand & Thomas Tietz (15 Dec 2025): 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, DOI: 10.1080/10408444.2025.2581776

To link to this article: https://doi.org/10.1080/10408444.2025.2581776

CRITICAL REVIEWS IN TOXICOLOGY
https://doi.org/10.1080/10408444.2025.2581776

Taylor & Francis
Taylor & Francis Group

LETTER TO THE EDITOR

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)

Dear Editor,

We commend *Critical Reviews in Toxicology* for supporting open and critical discussions in the field of benchmark dose (BMD) modeling and the recent publication by Slob et al. (2025) contributes to this dialogue. After thorough examination, we find that the article presents views and assumptions that are often insufficiently supported. It largely defends modeling approaches rooted in the early development of the BMD methodology, while overlooking well-known methodological limitations that have been highlighted by later applications. Moreover, it appears to dismiss important methodological developments that we consider essential for the appropriate and robust application of BMD modeling in regulatory toxicology, most importantly in relation to the Bayesian paradigm and model averaging. These latter two properties represent the state-of-the-science for BMD analysis, which international bodies agreed upon following the update of Chapter 5 on dose–response assessment of WHO/IPCS Environmental Health Criteria 240 (WHO 2020). The revised EFSA guidance on BMD (EFSA 2022) and the Bayesian EFSA app (https://r4eu.efsa.europa.eu/app/bmdbayesian) have been developed in alignment with these principles.

Below, we present a compilation of the issues we encountered in Slob et al. (2025).

On the interpretation of model structure and parameters

Understanding of "canonical models"

The authors assert that dose–response models must be "canonical" with parameters having identical interpretations across model structures. This premise exhibits several critical shortcomings:

All models in the Bayesian EFSA app are formulated in terms of the three so-called natural parameters (median background response at dose 0, median "maximum" response at very high dose, potency defined by the BMD being the dose corresponding to a certain BMR) and two technical parameters: *d* and the variance parameter. The natural parameters are well defined and have a clear biological interpretation for a particular endpoint. For monotonic dose–response curves, their values are unique but unknown and unrelated to any model for a specific dataset. The models are used to estimate these natural parameters, and model averaging provides unique estimates accounting for sampling variation in the data and model uncertainty. The EFSA (2022) guidance mentions this but, in place of background response, "maximum" response at very high dose, and potency, formulates the models using parameters *a*, *b*, *c*, and *d*. The main reason for this choice was not to confuse the reader or the user but to improve the readability and comparability of the extended family of models with the minimal family of models in the earlier 2017 guidance.

The fourth parameter *d*, however, is biologically related to steepness, but is not (mathematically) defined, and has no unique value or meaning across the different models. Different models with identical background, maximum and BMD and identical *d* parameter all have the same background, maximum and BMD but clearly show different "steepness", as illustrated in Figure 1. In other words, lacking a definition, *d* is not a biological but a model-specific parameter. Each model has its own *d* parameter. Hence, given that there is no unique *d* value across models, determining the model-averaged estimate for *d* has no meaning.

The authors' family of canonical models (exponential, Hill, inverse exponential, lognormal) was first included in the EFSA (2022) guidance as the "Family 1 of lognormal models", based on the work of Aerts et al. (2020). The Bayesian EFSA app does include all "canonical" models mentioned in their paper, but EFSA applies an inclusive rather than exclusive principle by extending the family with other plausible candidate models (other "canonical" models as well as others based on fewer prior assumptions). EFSA's approach is explicitly inclusive and data-driven: while including the canonical models of Slob et al. (2025), it acknowledges multiple valid mathematical representations of biological processes and emphasizes empirical validation. Furthermore, the Bayesian EFSA app allows the use of prior information at multiple levels, distributions, models, and parameters within each model and proposes the inclusion of sensitivity analyses to assess the robustness of results (e.g. by two optional default priors on the technical parameter *d* in the Bayesian EFSA app). The Bayesian EFSA app also allows the suite of candidate models to be restricted to any subfamily by using appropriate informative priors (at the level of the two distributions and at the level of the different median models, and at the level of the parameters). It is only required to motivate such restriction, as any informative prior requires justification.

Canonical properties and distribution assumptions

The models should predict positive values only: This property refers to the distributional component of the model, reflecting the variation of the response values at a given dose. The

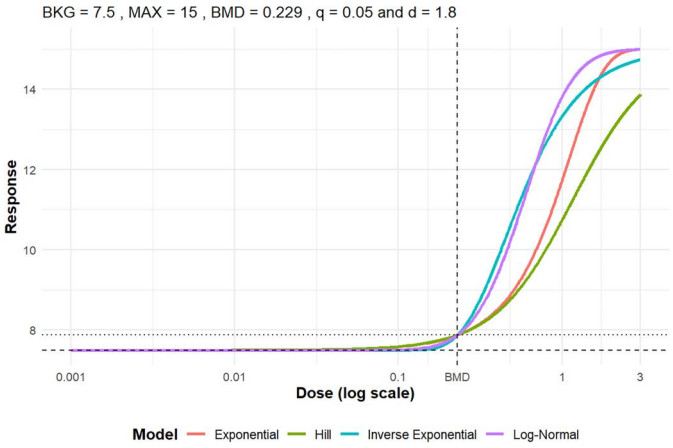

Figure 1. Graphical representation of ALL models considering the same background (BKG), fold change (MAX), BMD and *d* values, presented using log(Dose) scale.

authors state that this first property is self-evident. First, the models are not developed or used as predictive models, so there is no particular interest in predicting individual response values. Further, following the same rationale, this property should state that *the models should only predict values that can be observed in reality for the endpoint at hand*. Such a property would exclude any general distribution, including the lognormal distribution, as any distribution on the full positive range has a positive probability of predicting impossible values. The property would require a truncated distribution, which is truncated at unknown values (endpoint dependent). So, for instance, one could truncate the lognormal distribution at the right and truncate the normal distribution on both sides and use those truncated distributions. However, this would unnecessarily complicate the model, as such truncation is endpoint-specific and would imply computational costs.

In conclusion, one should consider plausible distributions for the within-dose-group variation, and the lognormal is a principal candidate. However, it is not certain that the lognormal is the best approximating distribution of the actual distribution, and model averaging has been developed to account for such uncertainty by including other plausible choices. EFSA has opted to include the normal and the lognormal distributions, representing the family of symmetric and right-skewed distributions. The Bayesian EFSA app first examines whether the probability of negative values is sufficiently low when applying the normal distribution. It does not determine whether the probability of unrealistically high values is sufficiently low for either the normal or lognormal. Again, EFSA prefers an inclusive approach, and the normal and lognormal distributions are computationally attractive.

An EFSA analysis of 3755 datasets from the National Toxicology Program (NTP) organ weight studies showed that 22.6% of the datasets were rejecting the log-normality assumption (Shapiro–Wilk test at level 5%), while 27.5% did not conform to the normality assumption, with 17.2% rejecting both distributions when tested using all responses from the different doses after subtracting the mean response for each dose and standardizing the residuals by the respective standard deviations. This empirical evidence directly contradicts the authors' assertion that any other distribution should be excluded, as the percentage of rejection for the lognormal (22.6%) is much larger than the expected nominal 5%.

Claims that outcomes should not depend on the measurement unit: The authors discuss scale-invariance issues with the EFSA model family 2 in Section 8.2.2, with an illustration in Appendix 3. We agree with that claim. In our view, there is no issue with the family 2 probit model as formulated in formula (10), as a change in unit is fully absorbed by the *a* parameter only. The authors mention that the Bayesian EFSA app results in similar BMD's for differently scaled responses, and that they can only explain this by an internal normalization in the tool, and then mention that this dependency is hidden. And based on these speculations, it is concluded that family 2 is inappropriate. In the Bayesian EFSA app, no normalization is applied and therefore the claim that "by normalizing the response this dependency is just hidden" does not hold. All the models in the EFSA Bayesian app are scale-invariant, except for the two models of family 2 with the lognormal distribution. These models for the median response are of the structure $exp(aF(c + bx^d))$, with *F* the logistic or the normal cumulative distribution function. This absence of exact invariance was investigated earlier, and the findings led to the decision to keep the models included. Although not perfectly scale-invariant, it appeared that a scale modified model (multiplied with a scale factor) can be mathematically approximated very well by another model of the same family and the approximation did not affect the BMDL more than the MCMC variation as observed in any other model (including the lognormal family 1 model, being the "canonical" models). Referring to table A2 of Slob et al. (2025), it is not clear to which BMR the corresponding BMD estimates are shown, neither on which scale. The BMD20 row suggests a BMR = 20%, but it is more likely 5%. The authors claim that the BMD estimates were obtained by PROAST (although that model is not included in the regular PROAST package for continuous data). Table 1 shows the BMD estimates (point and lower limit) for a scale grid 0.001, 0.01, 0.1, 1, 10, 100, and 1000 for the probit model (member of family 2, with normal and lognormal distribution) and for the inverse exponential model (member of family 1, with normal and lognormal distribution, with the lognormal choice corresponding to the so-called canonical model). The results were obtained with the R package BMABMDR, underlying the EFSA Bayesian BMD app. Five rows with *scale = 1* were included to show the MCMC variation. Such variation is also present for the other scale values. The BMDL results with PROAST will also depend on the bootstrap variation. Both EFSA apps (both accessible at: https://r4eu.efsa.europa.eu/), based on the R-packages PROAST and BMABMDR (using Bridge sampling and default priors), do not show such variation when repeating the same analysis on the same data, by fixing the so-called generation seed. This is done to guarantee exact reproducibility. But in this exercise, it is of value to compare the variation between the estimates, not only for different scales but also for different runs with the same scale. Next to the results shown for the probit model (shown and discussed by Slob et al. (2025) in section 9.1 and Appendix 3), we also include the results of a "canonical model", the IE4LN model, being the lognormal inverse exponential model. This allows us to compare results across canonical and non-canonical models (being probit models, with

Table 1. BMD and BMDL estimates after scale modifications for the probit model (P4) and the inverse exponential model (IE4), with normal distribution (N) and lognormal distribution (LN) (in bold the BMDL estimates and in italic the BMD estimates analogues of the values in the last row of table A2 of Slob et al. 2025).

Scale	BMDL.P4N	BMD.P4N	BMDL.P4LN	BMD.P4LN	BMDL.IE4N	BMD.IE4N	BMDL.IE4LN	BMD.IE4LN
0.001	**0.021**	0.240	**0.018**	*0.158*	**0.048**	0.298	**0.040**	0.182
0.01	**0.021**	0.235	**0.018**	*0.156*	**0.043**	0.268	**0.037**	0.188
0.1	**0.024**	0.258	**0.017**	*0.166*	**0.044**	0.260	**0.040**	0.190
1	**0.025**	0.244	**0.020**	*0.175*	**0.044**	0.259	**0.036**	0.191
1	**0.022**	0.244	**0.017**	*0.179*	**0.040**	0.254	**0.041**	0.190
1	**0.023**	0.243	**0.018**	*0.167*	**0.041**	0.261	**0.038**	0.195
1	**0.020**	0.226	**0.017**	*0.175*	**0.041**	0.280	**0.039**	0.187
1	**0.024**	0.251	**0.019**	*0.180*	**0.043**	0.275	**0.036**	0.191
10	**0.025**	0.257	**0.017**	0.159	**0.043**	0.267	**0.040**	0.194
100	**0.021**	0.254	**0.017**	0.165	**0.043**	0.259	**0.039**	0.197
1000	**0.026**	0.255	**0.019**	0.177	**0.048**	0.295	**0.036**	0.186

normal and lognormal distribution, and inverse exponential model, with the normal distribution). The italic figures are the analogues of the BMDs in the last row of table A2 of Slob et al. (2025). But the results for the BMDLs (in bold) are more relevant. This simulation experiment is, of course, limited in many ways, and results should not be overinterpreted, but it does give some rough insights into the effect of changes in the scale of the endpoint for this dataset. The BMD estimates are somewhat lower for the probit than for the inverse exponential model, and somewhat higher for the normal distribution. The variation across scales is comparable to the variation within *scale = 1*, especially for the BMDL. The variation in the BMDL estimates for the non-canonical models is of the same magnitude as that of the canonical model. In conclusion, this exercise does not show convincing evidence that the non-canonical models suffer severely from varying scales, as alleged by Slob et al. (2025).

Claims about BMD requirements

Independence from background response: The authors claim BMDs should not depend on the background response. But for none of EFSA's models, the BMDs do explicitly depend on the background. The BMDs only depend on the foldchange (ratio of median response at very high dose and the median response in the control group) for models in family 1. For two models (log-normal family 2), this is the ratio of the log median response at very high dose and the log median at zero dose. Dependency of the BMD on such a ratio is evident in our view. Even if the BMD is functionally independent from the background for each model in the mathematical sense, the estimates of all parameters, including the background and the BMD, are statistically correlated, and in nonlinear models in general, that correlation can depend on the true values of all parameters (Seber and Wild 1989).

Moreover, this requirement is not achievable in a model averaging estimation framework. Indeed, the final BMD estimates are the result of averaging across all models.

1. In the frequentist paradigm of PROAST, the BMD estimate is derived from applying the BMD definition with the particular BMR to the averaged model (a weighted mean of all models). This averaged model will inevitably depend on all estimated backgrounds from all constituent models, as the estimates for the background vary across models, to varying degrees, depending on the data at hand. So, the BMD derived from such an averaged model will also depend on all estimates of the background. The lower limit BMDL is obtained from bootstrap simulation. The bootstrap is actually an asymptotic method, implying it might have rather poor performance for small sample sizes (not infrequently occurring in practice).
2. In the Bayesian paradigm, even when restricting to the canonical models, the posterior of the BMD for a particular model is based on the joint Bayesian estimation of all parameters of that model. The final BMD point estimate and credible interval are derived from the weighted mixture of all model-specific BMD posteriors, and therefore, they may depend on the Bayesian estimates of all parameters of all models. Finally, the priors for all parameters also have an important impact on the model-specific BMD posteriors and the final model-averaged BMD posterior.

Methodological considerations and empirical evidence

While the authors state that there is "ample empirical evidence" supporting their assumptions of parallel dose–response on log-scale (at the end of section 2) and claim that dose–response relationships should be multiplicative rather than additive, the authors advocate for visual inspection of parallelism rather than statistical tests. They state statistical tests are "not reliable as [they assume] random scatter". Yet the same assumption of random scatter is implicitly accepted when they fit models and calculate BMD confidence intervals – which, in our view, undermines the coherence of the methodological framework.

Moreover, the authors argue for exclusive use of lognormal distributions without adequate justification. Above (end of section "Canonical properties and distribution assumptions – The models should predict positive values only") we provided quantitative empirical evidence that directly contradicts the authors' assertion that normal distributions should be excluded, as the percentage of rejection for the lognormal (22.6%) is much larger than the nominal 5% expected. This

raises questions about the selectivity of the empirical basis used to support their argument.

In addition, the manuscript does not include an evaluation of the robustness of the proposed modeling framework. Despite suggesting that a more restricted model set offers advantages, no comparative analysis is presented to demonstrate improved performance relative to broader, more inclusive model sets.

Regarding the discussion on model averaging, the recommendation to limit the model set appears to overlook widely accepted statistical principles. The strength of model averaging lies in its ability to incorporate a range of models that may reflect different aspects of the underlying process. By doing so, it helps mitigate the effects of model misspecification. The benefits of this ensemble approach – particularly its capacity to balance varying model strengths and weaknesses – do not seem to be fully acknowledged in the manuscript.

Implementation and transparency

We are also concerned about the endorsement of the PROAST tool in the reviewed paper. While PROAST has played a key role in BMD development over the years, performance issues with the tool have been observed in specific scenarios and communicated to the authors of the paper in writing as well as in several meetings. In particular, EFSA has observed situations where PROAST underperforms and introduces results that are difficult to interpret due to limited transparency, including: the use of bootstrap procedures whose limitations are not clearly stated, hard-coded constraints within the software that are not visible or accessible to users; the lack of sufficient documentation and openness regarding these limitations. More details are available upon request.

EFSA's Bayesian BMD app, is open-source, fully transparent with documentation (Hasselt University 2022), and publicly available for scrutiny and validation. It uses Bayesian model averaging to account for model uncertainty and incorporates sensitivity analyses to provide a nuanced view of data and assumptions. This transparency is a strength – it invites improvement and fosters confidence.

Constructive alternative approach

We stress that inclusive modeling frameworks do not simply accept models uncritically but allow divergent perspectives and assumptions to coexist and be empirically tested. They offer opportunities to diagnose when assumptions fail, when data are insufficient, or when biological systems defy a-priori categorization. We advocate for a more comprehensive/inclusive approach that:

1. Incorporates a diverse set of models reflecting different possible dose–response relationships.
2. Uses model averaging techniques to account for model uncertainty.
3. Employs appropriate prior distributions based on biological knowledge and empirical evidence.
4. Conducts thorough sensitivity analyses to evaluate robustness.
5. Considers both normal and lognormal distributions driven by data characteristics.

This approach provides a more reliable process to estimate the relevant BMD while acknowledging inherent uncertainty in dose–response modeling. To these continuous developments, it also follows our commitment for the continuous improvements of our guidance and tool to support the users navigating the BMD framework with evolving scientific knowledge.

Declaration of interest

Authors J. Cortiñas Abrahantes and M. Aerts initiated the letter in response to Slob et al. (2025), whose original article contained claims and statements regarding the EFSA guidance and the EFSA Bayesian app that were inaccurate and unsubstantiated. E. Solazzo contributed to the editing and content of the letter and coordinated its circulation among the coauthors for their comments. All feedback received was reviewed and incorporated into the version of the letter initially submitted to this journal.

All authors are members of the EFSA Working Group on Benchmark Dose Modelling. Authors J. Cortiñas Abrahantes and E. Solazzo are employees of EFSA, while the other coauthors were selected based on their expertise. Authors J. Cortiñas Abrahantes and M. Aerts coauthored the BMD EFSA Guidance (2022). Author M. Aerts is affiliated with the University of Hasselt, which is a beneficiary of an EFSA grant for the development of the Bayesian app for Benchmark Dose Modelling.

None of the authors received any compensation or external assistance for the preparation of this letter. Furthermore, none of the authors is currently, or has recently been, involved as an expert witness in litigation or formal government rulemaking on the subject of this article.

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.

References

Aerts M, Wheeler MW, Abrahantes JC. 2020. An extended and unified modeling framework for benchmark dose estimation for both continuous and binary data. Environmetrics. 31(7):15. https://doi.org/10.1002/env.2630

[EFSA] EFSA Scientific Committee et al. 2022. Guidance on the use of the benchmark dose approach in risk assessment. EFSA J. 20(10):e07584, 67 pp. https://doi.org/10.2903/j.efsa.2022.7584

Hasselt University. 2022. EFSA platform for Bayesian benchmark dose analysis. EFSA Supporting Publication 2022: EN-7740. 91 pp.

Seber GAF, Wild CJ. 1989. Nonlinear regression. Wiley series in probability and statistics. Vol 1989. p 3. John Wiley & Sons Inc.

Slob W et al. 2025. The use of canonical dose–response models for benchmark dose analysis of continuous toxicological data. Crit Rev Toxicol. 55(4):437–461. https://doi.org/10.1080/10408444.2025.2464067

[WHO] World Health Organization. 2020. Chapter 5: dose–response assessment and derivation of health-based guidance values. In: Principles and methods for the risk assessment of chemicals in food. 2nd ed. Environmental Health Criteria 240. WHO/IPCS. https://cdn.who.int/media/docs/default-source/food-safety/publications/chapter5-dose-response.pdf?sfvrsn=32edc2c6_5

José Cortiñas Abrahantes, Efisio Solazzo
Methodology and Scientific Support Unit, European Food Safety Authority (EFSA), Parma, Italy
efisio.solazzo@efsa.europa.eu

Marc Aerts
I-BioStat, Hasselt University, Diepenbeek, Belgium

Peter Craig
Department of Mathematical Sciences, Durham University, Durham, UK

Ursula Gundert-Remy
Institute of Clinical Pharmacology and Toxicology, Charité-Universitätsmedizin Berlin, Berlin, Germany

Thor Halldorsson
Faculty of Food Science and Nutrition, School of Health Sciences, University of Iceland, Reykjavik, Iceland

Ron (Laurentius) Hoogenboom
Department of Toxicology, Wageningen Food Safety Research, Wageningen University & Research, Wageningen, Netherlands

Salomon Sand
Department of Risk and Benefit Assessment, Swedish Food Agency, Uppsala, Sweden

Thomas Tietz
Department Chemicals and Product Safety, German Federal Institute for Risk Assessment, Berlin, Germany

Received 23 June 2025; Accepted 20 October 2025