Fitting the model to the data

In this guide

In this guide

1. Benchmark dose modelling in a UK chemical risk assessment framework - cover
2. Benchmark Dose Modelling in a UK Chemical Risk Assessment Framework - Executive Summary
3. Benchmark dose modelling in a UK chemical risk assessment framework - Recommendations
4. Benchmark dose modelling in a UK chemical risk assessment framework - Use
5. Benchmark dose modelling in a UK chemical risk assessment framework - Advantages of BMD modelling
6. Current challenges to the use of benchmark dose modelling in regulatory toxicology
7. Benchmark dose modelling in a UK chemical risk assessment framework - COT’s discussion
8. Benchmark dose modelling in a UK chemical risk assessment framework - Conclusions
9. Annex A - Introduction and Background
10. Annex A- Benchmark dose modelling
11. Annex A - Selected Previous Publications
12. Annex A - COT previous discussions
13. Annex A - NOAEL approach vs BMD approach
14. Annex A - Modelling the data
15. Annex A - Fitting the model to the data
16. Annex A - Bayesian vs frequentist approach
17. Annex A Case Study (FSA Computational Fellow)
18. Annex A - User experience
19. Annex A - Conclusions
20. Annex A - Questions on which the views of the Committee are sought
21. Annex A - List of Abbreviations
22. Annex A - Technical terms

23. Annex A - References

71. The objective of model fitting is to best describe the dose-response relationship of a given data set. The process typically involves searching for parameter values in the model that lead to a function or curve that describes the data well, using some statistical criterion that defines a good fit (FAO/WHO, 2020).

Constraining or not constraining the models

72. As curve fitting typically optimise a model's "best fit" to a given set of data without knowledge of the biological dose-response, this may lead to model fits that describe the data well but contain parameter values which are "biologically improbable". It has been argued that setting parameter bounds adds information which may improve the accuracy of the model and mitigate the likelihood of biologically implausible responses (FAO/WHO, 2020).

73. Some constraints serve a practical necessity e.g., constraining the probabilities of an effect in a dichotomous model to no greater than one (US EPA, 2012). The EPA and EFSA agree on other general restrictions such as biological measures generally being positive, and that dose-responses will be generally monotonic (i.e., a higher dose of a given substance will have an equal or greater effect than a lower dose). Much existing practice constrains models to avoid non-monotonic curves (EFSA, 2017; US EPA, 2012).

74. Other model restraints are controversial, and guidance from the EPA and EFSA diverge (Haber et al., 2018). An example is constraining models that are steeply supralinear. In some models, such as the Weibull model, where the dose is raised to a power of a given parameter, the slope of the dose-response curve can become very steep at low doses if the power parameter is estimated at values lower than 1. Thus, the US EPA recommends that the modeler should consider constraining power parameters to be 1 or greater (this is the default in the BMDS software). While EFSA (2017) acknowledge this concern exists, they point to work by Slob and Setzer (2014) that demonstrates that this constraint is largely based on a false argument and is contradicted by real dose-response data (EFSA, 2017; Slob and Setzer, 2014). They recommend against constraining the model in this way, as it could produce artificially high BMDLs (EFSA, 2017).

75. The US EPA encourages the use of constrained models as a frontline approach, to avoid biologically unreasonable dose-response curves. They

recommend unconstrained models only be used if an acceptable fit is not achievable using constrained models (US EPA, 2012). Similarly JECFA/JMPR guidance accepts that constraints may be needed “when it is deemed biologically appropriate” and also highlights that parameter constraints are less necessary when using model averaging or Bayesian methods in general (FAO/WHO, 2020).

Convergence

76. The goal of the fitting process is to find values for the model parameters so that the resulting fitted model describe the data most optimally. The practical matter of determining the “best” parameters for model fit typically involve a BMD software starting with an initial “guess” for the parameter values. Then, this guess is iteratively updated, producing a sequence of estimates that (usually) converge. Many models will converge to the right estimates for most datasets from just any reasonable set of initial parameter values; however, some models, and some datasets, may require multiple guesses at values before the model or models converge (US EPA, 2012).

77. After fitting all models, the first step is to evaluate model convergence. If the model did not converge to a single maximum likelihood, it is possible that there may be more than one set of parameter estimates that would result in similar log-likelihood values (Haber et al., 2018). The EFSA guidance states that convergence does not guarantee a reliable BMD confidence interval, and a message of non-convergence does not necessarily indicate that the model should be rejected. EFSA state that simulations have shown that non-convergence may have little impact on the BMD confidence interval but recommend that in instances where convergence is not achieved that a BMD specialist should be consulted. They note that a lack of convergence could be because the data are not informative, or the model may be over-parameterised (EFSA, 2017).

Evaluating the model fit

78. The JECFA (2020) guidance provide a list of commonly applied methods for evaluating if a given dose-response model fits a data set well. These methods include examination of the visual fit, bootstrap statistics to evaluate goodness-of-fit (frequentist model averaging), and appropriate Bayesian methods if applicable. For individual models, JECFA state that users can compare models using the AIC or BIC (Bayesian information criterion) and evaluate them using analysis of deviance and Pearson $\chi 2$ goodness-of-fit tests. They note that no one technique is recommended for every case, and stress that the model fit criteria should be

justified and documented (FAO/WHO, 2020).

79. The EPA list criteria on which the quality of a given model can be assessed but stop short of prescribing the choice of the model. Instead, they provide a series of steps to determine the best model or suite of models in each case (Haber et al., 2018; US EPA, 2012). These steps are summarised briefly here:

80. Assessing the goodness-of-fit: The EPA recommend using a value of $\alpha = 0.1$ (or $\alpha = 0.05$ / 0.01 if appropriate) to compute the critical value for goodness-of-fit, along with a visual inspection of the model fit.

81. They recommend rejecting models that do not sufficiently describe the low-dose portion of the dose-response, by a combination of examining the scaled residuals and visual fit of the relevant model or models.

82. Any models which pass the criteria are assumed to meet the recommended default statistical criteria for adequacy and visual fit, and theoretically could be used for determining the BMDL (US EPA, 2012).

83. If the BMDL estimates from the qualifying models are sufficiently “close” (i.e. there is no strong influence of any one individual model), then the guidance recommends that the model with the lowest AIC (Akaike, 1973) can be used to calculate the BMDL for the RP. If two or more models share the lowest AIC, the average BMDLs from these models may be used (US EPA, 2012).

84. If the BMDL estimates are disparate enough to be considered “not sufficiently close”, (i.e., some model dependence can be assumed), the EPA acknowledge that expert user judgment is needed to determine if the uncertainty is too great to rely on the results. They suggest that if the range of BMDLs is judged reasonable, but there is no obvious biological or statistical basis to choose one over another, the lowest BMDL may be selected as a conservative estimate (US EPA, 2012).

85. EFSA (2017) also recommend using the AIC in the selection of the models for frequentist approaches (EFSA, 2017). The AIC is calculated as:

$$AIC = -2 \log(L) + 2p$$

86. With log(L) being the log-likelihood of the model, and p being the number of parameters. The first term, -2 log(L), decrease as the model gets closer to the measured data, while the second term 2p acts to penalise the number of

parameters in the model. Thus, a model with a relatively low AIC may be considered as providing a good fit without using too many parameters (EFSA, 2017).

87. Based on work from Burnham and Anderson (2004), EFSA recommend that models resulting in AICs differing by less than two units may be regarded as describing the data equally well (Burnham and Anderson, 2004; EFSA, 2017). EFSA note that this cutoff between good and poor models is relatively arbitrary and acknowledge that in specific cases, a user may decide to use a larger value than 2, e.g. when using a value of 2 would lead to the selection of just one model being selected (EFSA, 2017).

88. Further, EFSA notes that the AIC criterion can be used to investigate if there is, in fact, a dose-related trend in the data. For a model to show statistical evidence of a dose-related trend, EFSA proposes that a model's AIC be lower than the AIC (null model) - 2. Similarly, the AIC criterion can also be used to compare the fit of any model with that of the full model. If the model with the minimal AIC is greater than two units larger than that of the full model, (AIC(min model) > AIC (full model) + 2), this could indicate an inappropriate dose-response model (e.g. it may contain insufficient numbers of parameters), or a misspecification of the distributional part of the model (e.g. litter effects are ignored), or to other non-random errors in the data (EFSA, 2017).

Model Averaging

89. A notable divergence between EPA and EFSA guidance, since 2017, is their approach to model uncertainty. Haber and colleagues note, in their review of BMD modelling, that there is a growing recognition that methods which attempts to choose a "best" model (and use the associated BMDL) do not reflect the true model uncertainty (Haber et al., 2018).

90. The EFSA (2017) guidance proposes that the goal of BMD analysis is not to identify the best fitting model and get an estimate of the (true) BMD. Rather, the goal should be to find a range of plausible values of the (true) BMD as described by a range of models, given the data available. In practice, this involves considering all models that offer plausible descriptions of the data - even models resulting in slightly poorer fits. EFSA note "After all, it could well be that the second (or third, ...) best-fitting model is closer to the true dose-response than the best-fitting model". This so-called 'model uncertainty' is the basis for their recommendation for BMD confidence intervals to be used and are based on the results from various models, instead of a single 'best' model (EFSA, 2017).

91. Model averaging has been proposed as an appropriate method to address model uncertainty in DRMs (EFSA, 2017; FAO/WHO, 2020). Model averaging permits an estimate of the dose-response relationship and associated statistics, such as the BMD and confidence intervals, using a weighted average of all model fits (Burnham and Anderson, 2004; Kang et al., 2000; Wheeler and Bailer, 2009, 2008, 2007). Individual model results are combined using weights, with higher weights accorded to models that fit the data better (EFSA, 2017; FAO/WHO, 2020).

92. The EPA guidance (2012) acknowledges the utility of model averaging to estimate levels of uncertainly in the model fits. However, they note this is a more complex undertaking; resulting model fits may give divergent results and more difficult interpretations. They recommend, instead of model averaging, users select a single well-fitting and plausible model (US EPA, 2012). They note that using the uncertainly of the model fits to derive the average BMD and associated confidence intervals also has disadvantages, including the fact the 95% lower bound (on the average BMD) is not, in fact, the lower bound described in the various individual estimates, but a lower bound of the average of the particular BMDLs under consideration (i.e., statistical properties of the individual estimates are lost) (US EPA, 2012).