Applying criteria for model evaluation to TKTD models

Tjalling Jager[1]

[1]DEBtox Research, De Bilt, The Netherlands (www.debtox.nl)
E-mail contact: tjalling@debtox.nl

1. Introduction

Environmental risk assessment (ERA) is the process of evaluating the impact of chemicals on the environment (or parts thereof). Estimation of exposure concentrations relies heavily on mechanistic fate models, but on the effects side, the focus lies on toxicity testing and descriptive statistical treatment of data. Mechanistic effect models are gaining increasing interest in a regulatory context. However, if these effect models are to be applied for ERA, it is important that their 'quality' is established. Recently, several papers (see [1]) and an EFSA opinion [2] discussed quality criteria, documentation and evaluation of effect models in the context of ERA. These discussions on 'model evaluation' are written from the perspective of population and community modelling. However, effect models also exist at the individual level, generally falling into the category of toxicokinetic-toxicodynamic (TKTD) models. In contrast to the higher-level models, TKTD models are almost always completely parameterised by fitting them to a data set. In fact, one of their explicit aims is to replace the descriptive hypothesis testing and dose-response fitting as data-analysis tools [3]. Furthermore, the development of these models generally does not fit nicely into an orderly modelling cycle. The best-established TKTD models have been, and are being, developed, applied and tested by many different groups, over at least half a century, applying them to very different problems, coming up with model adaptations, and borrowing elements from each other. Rather than a discreet cycle, development of TKTD models is more like a complex interconnected web.

These aspects have consequences for the application of frameworks for model evaluation. To illustrate, I will use the General Unified Threshold model of Survival (GUTS, [4]). GUTS is a TKTD model; it links a TK model to a TD model to translate an external concentration (which might be time varying) to effects on a life-history trait (survival) over time. The TK component of GUTS is a one-compartment model with first-order kinetics; the TD component applies a damage module and a simple stochastic representation of death. The GUTS framework integrates virtually all of the previously published models for the endpoint survival. Specific models can thus be seen as special cases of GUTS, generated by fixing parameters to specific values.

2. Issues with model evaluation

What is the model? Frameworks for model evaluation assume that it is clear what is meant by 'the model' that is to be evaluated. However, for TKTD models, this is often far from clear. In the case of GUTS, is the GUTS framework the model, or a special case such as GUTS-SIC-SD (Fig. 1)? Or is the model the implementation of a special case into software, such as Matlab? Or is the model the specific case of GUTS *after* it has been calibrated to a specific species and toxicant (or even a specific data set)? Each of the common elements of model evaluation applies only to some definition of 'the model'. For example, we can only perform meaningful sensitivity and uncertainty analysis once the model has been calibrated.

Can we reconstruct model development? The TRACE tool [1] emphasises the importance of documenting every step in the development of a model. Dynamic models for survival have been developed, tested and applied for at least half a century by many different (unconnected) groups. They incorporate TK models with an even longer and more convoluted history. GUTS brings this long development together into a unified framework. Hundreds of papers have appeared on survival modelling, and countless software implementations exist or have existed. It is clearly impossible to reconstruct this entire history, and also hardly informative. It is far more useful to have a clear description of GUTS, with clarification of the underlying assumptions and limitations (an extensive e-book on GUTS is currently being developed).

What is the meaning of sensitivity and uncertainty analysis? Sensitivity and uncertainty analysis feature prominently in model evaluation frameworks. Firstly, it is important to realise that there is no point in performing such analyses for GUTS or GUTS-SIC-SD in general; one would first need to parameterise a special case for a certain species and chemical, and define an exposure scenario. Secondly, GUTS is fully parameterised by fitting the model on a data set. Therefore, all of the relevant information on uncertainty and sensitivity is contained in an appropriate statistical treatment of the model fit: confidence intervals on parameters, correlations between parameters, and intervals on predictions. There will be little added value by performing classical sensitivity or uncertainty analysis, and such analyses will not inform us on the quality of the conceptual model or its implementation; it only reflects a specific parameterisation (Fig. 1).

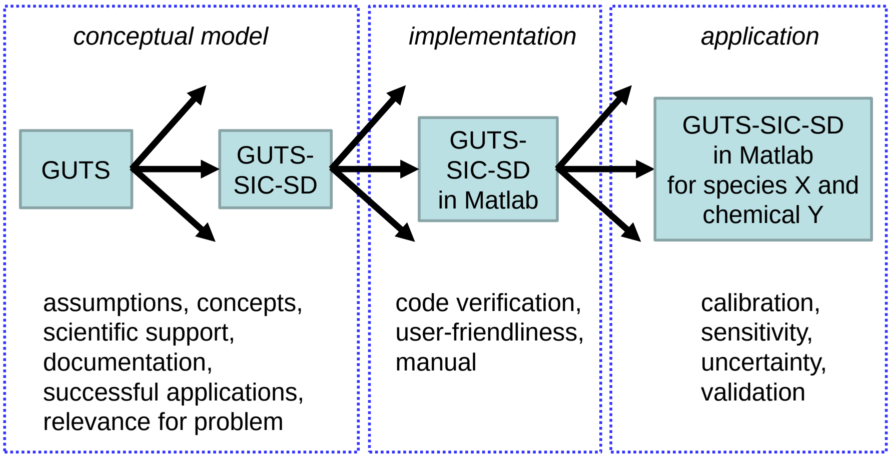
conceptual model
implementation
application
GUTS
GUTS-SIC-SD
GUTS-SIC-SD in Matlab
GUTS-SIC-SD in Matlab for species X and chemical Y
assumptions, concepts, scientific support, documentation, successful applications, relevance for problem
code verification, user-friendliness, manual
calibration, sensitivity, uncertainty, validation

Figure 1. Different definitions of 'model' as exemplified with the GUTS special case for scaled-internal concentration (SIC) and stochastic death (SD). GUTS has a number of special cases, and a range of software implementations. The number of model applications is huge, especially if we include earlier models that can now be viewed as special cases of GUTS.

What is the meaning of validation? Another crucial element in model evaluation is 'validation', which is generally intepreted as a comparison between model predictions and independent observations. What is it that we like to validate for a TKTD model like GUTS? If we use GUTS to analyse a data set for acute toxicity, we might use a model parameter (e.g., the threshold for effects) or a model prediction (e.g., the LC50 after 4 days) in ERA. How do we validate a model for such use when we cannot independently measure an LC50 or a threshold for effects? A good fit to the data does mean something: the model only has a limited behavioural repertoire, so a good fit is a support for the model. However, it is still a calibration. If we repeat the toxicity test, and compare the model predictions to these new results, what are we validating? The model, the model calibration, or the reproducibility of the toxicity test? If we use GUTS to make predictions for untested scenarios, validation would be more straightforward. For example, we can calibrate model using a test with constant exposure and predict survival under a pulsed scenario (or vice versa). However, if we obtain a successful validation exercise for one species and one chemical in one extrapolation, what does it say about the quality of the GUTS framework in general? And, what does it say about the validity of other extrapolations, such as to survival over a whole year, or effects on another species?

3. Conclusions and recommendations

Frameworks for evaluation of effect models are geared towards models at the population level and higher; TKTD models require a different outlook. Different aspects of model evaluation only apply to different definitions of 'the model' (see Fig. 1), and need a different interpretation for TKTD models. Using a case study on propiconazole in *Gammarus pulex*, I will demonstrate how sensitivity, uncertainty and validation can be approached for TKTD models like GUTS. Proper treatment of calibration, quantification of sensitivities and uncertainties, and a clear documentation, are all important aspects of a model analysis but say nothing about the quality of the model concept. We therefore have to explicitly separate the evaluation of the conceptual model from that of the software implementation, from that of the model application to a specific data set. The quality of the conceptual model is of primary importance. For any model, user-friendly software and documentation can be developed, including options for propagation of uncertainty. For any model, we can set up dedicated validation experiments. However, a model based on flawed assumptions will always remain a bad model. It is inefficient to place the burden of model evaluation on risk assessors; ERA is best served by a set of agreed models or modules. Finally, it should be stressed that the methods that are currently used in ERA (toxicity tests, dose-response analysis, assessment factors, etc.) are also models. They should be evaluated using the same criteria if we endeavour a fair comparison to mechanistic effect models.

4. References

[1] Grimm V, Augusiak J, Focks A, Frank BM, Gabsi F, Johnston ASA, Liu C, Martin BT, Meli M, Radchuk V, Thorbek P, Railsback SF. 2014. Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. Ecol Modell 280:129-139.

[2] EFSA. 2014. Scientific Opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. EFSA journal 12:3589.

[3] Jager T, Heugens EHW, Kooijman SALM. 2006. Making sense of ecotoxicological test results: towards application of process-based models. Ecotoxicology 15:305-314.

[4] Jager T, Albert C, Preuss TG, Ashauer R. 2011. General Unified Threshold model of Survival - a toxicokinetic-toxicodynamic framework for ecotoxicology. Environ Sci Technol 45:2529-2540.