Identification and Description of the Uncertainty, Variability, Bias and Influence in Quantitative Structure-Activity Relationships (QSARs) for Toxicity Prediction

Mark T.D. Cronin[1],* Andrea-Nicole Richarz[2], and Terry. W. Schultz[3]

[1]Liverpool John Moores University, School of Pharmacy and Biomolecular Sciences, Liverpool, England

[2]European Commission, Joint Research Centre (JRC), Ispra, Italy

[3]The University of Tennessee, College of Veterinary Medicine, Knoxville TN, USA

*Author for correspondence:

Mark Cronin:

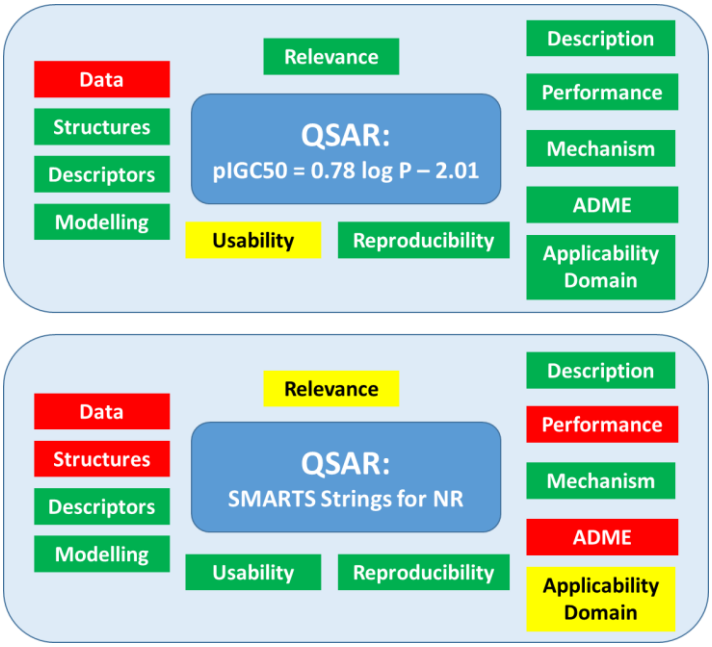
Address for Correspondence: School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, England.

Email: m.t.cronin@ljmu.ac.uk

Short title: Uncertainty in QSARs

Disclaimer: The views expressed are solely those of the authors and the contents of this manuscript do not necessarily represent the views or position of the European Commission.

Graphical Abstract



Summary of overarching assessment criteria for the two QSAR Case Studies showing areas of high (red), moderate (yellow) and low (green) uncertainty, variability, bias or influence.

Abstract

Improving regulatory confidence in, and acceptance of, a prediction of toxicity from a quantitative structure-activity relationship (QSAR) requires assessment of its uncertainty and determination of whether the uncertainty is acceptable. Thus, it is crucial to identify potential uncertainties fundamental to QSAR predictions. Based on expert review, sources of uncertainties, variabilities and biases, as well as areas of influence in QSARs for toxicity prediction were established. These were grouped into three thematic areas: uncertainties, variabilities, potential biases and influences associated with 1) the creation of the QSAR, 2) the description of the QSAR, and 3) the application of the QSAR, also showing barriers for their use. Each thematic area was divided into a total of 13 main areas of concern with 49 assessment criteria covering all aspects of QSAR development, documentation and use. Two case studies were undertaken on different types of QSARs that demonstrated the applicability of the assessment criteria to identify potential weaknesses in the use of a QSAR for a specific purpose such that they may be addressed and mitigation strategies can be proposed, as well as enabling an informed decision on the adequacy of the model in the considered context.

Keywords: QSAR; toxicity prediction; uncertainty; variability; bias; influence; barriers; assessment criteria

Abbreviations:

Absorption, Distribution, Metabolism and Excretion (ADME); Adverse Outcome Pathway (AOP); Animal Research: Reporting of *In Vivo* Experiments (ARRIVE); Artificial Intelligence (AI); classification and labelling (C&L); concentration causing 50% inhibition of growth (IGC50); Defined Approach (DA); Good Computer Modelling Practice (GCMP); Good Laboratory Practice (GLP); Integrated Approaches to Testing and Assessment (IATA); Intellectual Property (IP); logarithm of the octanol-water partition coefficient (log P); National Research Council (NRC); New Approach Methodology (NAM); Organisation for Economic Cooperation and Development (OECD); QSAR Model Reporting Format (QMRF); Quantitative Structure-Activity Relationship (QSAR); Read-Across Assessment Framework (RAAF); Registration, Evaluation, Authorisation and restriction of CHemicals (REACH); Standard Operating Procedure (SOP); SYstematic Review Center for Laboratory animal Experimentation (SYRCLE); Test Guideline (TG); United States Environmental Protection Agency (US EPA); Weight of Evidence (WoE); World Health Organisation International Programme on Chemical Safety (WHO IPCS)

Highlights

- Uncertainties, variabilities and potential areas of bias and influences of QSARs are identified
- The creation, description and application of QSARs is evaluated
- 13 types of uncertainty, variability, bias and influence of QSARs established
- 49 assessment criteria for QSARs are presented
- Application of the assessment criteria will improve the uptake and use of QSARs

Introduction

To understand the confidence that may be assigned to a prediction, there is a need to assess the underlying model and its suitability to make the prediction in question (Patterson and Whelan, 2017). With regard to the risk assessment of chemicals, many predictions can be made relating to hazard identification and potency as well as exposure assessment. For the prediction of toxicity and data gap filling in particular, the use of Quantitative Structure-Activity Relationship (QSAR) models is a well-established technique (Cronin and Yoon, 2019). QSARs attempt to formalise the relationship between toxicity and chemical structure and properties such that a model may make a prediction, from structure, when data are missing. For the purposes of this paper, the term "QSAR" is taken in its broadest possible sense to include relationships between chemical structure and properties and toxicity that have been formalised into some type of model – this may range in complexity from structural alerts to machine learning, categoric definition of activity and continuous potency, and be based on any type of descriptor or property. Being at the forefront of *in silico* toxicology for several decades, predictions from QSARs have found use in chemical regulations, such as for adaptation of information requirements in Annex XI of REACH (Spielmann et al., 2011). Despite this, in practice, QSARs are used mostly as supporting information or for screening purposes, although successfully integrated for example in the regulatory assessment practice of drug impurities, to predict mutagenicity according to the ICH M7 guidelines (ICH, 2017). However, for a wider uptake and successful regulatory use, an assessment and user-friendly communication of the confidence in the model and application is needed in order to enable the user or regulator to make an informed decision upon, and feel comfortable with, its use for the specific purpose in question (Worth, 2010).

Whilst it is not the purpose of this paper to give a full review of QSARs, it is accepted that they stand at the juncture of biology, chemistry and statistics. QSARs for toxicity require (preferably high quality) data to be modelled with regard to appropriate descriptors, parameters and/ or properties of a set of chemicals. In theory and practice, any set of high-quality toxicity data for a coherent set of compounds is applicable. Descriptors for the various molecular properties are selected either empirically using mechanistic understanding, for example based on Molecular Initiating or Key Events of Adverse Outcome Pathways (AOPs) (Cronin and Richarz, 2018), or by statistical methods. Over five decades of progress has resulted in a multitude of descriptors of molecular structure and properties that include empirical, quantum chemical, or non-empirical parameters. Whilst empirical descriptors may be measured or estimated and include physico-chemical properties, non-empirical descriptors are typically structural features developed from the knowledge of 2D structure. Statistical methods to develop QSARs are typically either correlative or use pattern recognition approaches. The most

common correlative method is regression analysis whereas pattern recognition techniques are varied, often complex and maybe multi-dimensional and non-linear (Cronin and Madden, 2010). Overall, it can be surmised that the QSAR modeller has an enormous number of techniques and approaches to use resulting in a wide diversity of QSAR models. These approaches must be used appropriately to develop models that are robust and fit for purpose; where the purpose is to support regulatory assessment a means of assigning confidence to a prediction is required.

The assignment of confidence to a prediction of toxicity requires the definition and assessment of the model and its suitability to make a prediction for a particular chemical (i.e., whether the chemical falls within the applicability domain) (Netzeva et al., 2005). The definition of a QSAR and discussions regarding the means of assessing their relative quality go back many decades and are well, and extensively, reviewed with regard to toxicity prediction elsewhere (cf. Cronin and Madden, 2010; Cronin et al., 2013). The assessment of the statistical fit of a QSAR and its overall performance are essential components to assess its quality and are ubiquitous in the science. As computational models, it was acknowledged early in the evolution of QSARs that statistical fit must be adequate for the intended purpose, but not too high as to imply over-fitting (Eriksson et al., 2003). Once computational techniques to perform statistical analysis became more widely available, more sophisticated statistical analyses were undertaken and there was a move from statistical fit to the assessment of predictivity of large external test sets, not necessarily considered in the model (Tropsha, 2010). Although an essential component of evaluating a QSAR, statistical veracity is only one part of the process to determine whether a prediction from a QSAR may be acceptable for a particular purpose.

The use of QSARs to make prediction of toxicity to support legislation goes back at least to the 1980s (Cronin and Yoon, 2019; Worth, 2010). However, the modern paradigm of the regulatory use of QSARs for toxicity prediction was defined by the “Setubal Workshop” in 2002 (Jaworska et al., 2002; Cronin et al., 2003a, 2003b) with a particular focus on preparing for the requirements of the European Union’s then upcoming Registration, Evaluation, Authorisation and restriction of CHemicals (REACH) legislation. The Setubal Workshop sparked interest in assessing the limitations and practical considerations in using QSARs in a regulatory setting (e.g., Tong et al., 2005; Tunkel et al., 2005; Worth, 2010), in particular, the issue of uncertainty (e.g., Sahlin et al., 2011; Ball et al., 2015). In order to assess and ensure the quality of a QSAR, aligned to the analysis of statistical performance, was the need for a better appreciation of the need for accurate representation of chemical structure, the intrinsic variability of the biological data and mechanistic basis on which a QSAR is based. This knowledge was cyrstallised into a set of six Principles for the “validation” of QSARs at the Setubal Workshop. These six principles were condensed to five when taken up as the OECD Principles for the

Validation of (Q)SARs for Regulatory Use (OECD, 2007; Worth, 2010). The Principles are often used as a framework to describe the content and performance of the model, for instance as applied using the QSAR Model Reporting Format (QMRF). In this context, the Principles, if applied appropriately by the model developer, may allow for formal validation of a model in terms of it being fit-for-purpose, or acceptable, in a regulatory context. The process of formal validation usually requires assessment against pre-defined validation criteria such that a non-expert in the field will have confidence in the use of the method. One of the shortcomings in the use of the Principles has been the failure to evaluate models fully against such criteria. As a result, to a certain extent they are a descriptive, rather than a diagnostic, means of defining and analysing a QSAR. Despite this, during the past 15 years the Principles have served the QSAR community (both users and developers) well, however they were not intended to be used for the breadth of QSAR approaches now available or be implemented with the context of 21st Century Toxicology – some examples of the types of uncertainties and how the science has developed are provided in Table 1. In addition, they do not necessarily allow or describe the ability to reproduce a QSAR or ensure its transparency (Patel et al., 2018; Piir et al., 2018), and have been applied in different ways and with different levels of detail by QSAR developers. As such, there is an opportunity to broaden the principles and incorporate newer thinking around toxicological problems. For instance, toxicology is moving to a more considered use of information with an emphasis on building weight of evidence (WoE) from individual lines of evidence. As part of this, there has been a growing emphasis on describing uncertainty(ies) associated with a model in an attempt to qualify, or even quantify, the areas where more information may be beneficial (Patterson and Whelan, 2017). In addition, there is a growing prominence of broader topics such as Good Computer Modelling Practice (GCMP) (Judson, 2009; Judson et al., 2015) as well as the understanding on how bias and variability affect into a model. In the broader context, the needs for all types of models in order to use them to make decisions have been defined with checklists presented to ensure model users and developers have considered the most significant factors for success (Calder et al., 2018). Thus, the assignment of the confidence in a model is dependent to a large extent on the identification of uncertainties, variability and biases and understanding, for QSARs, of how they may affect the prediction or the decision made based on the prediction, which may depend on the context considered. This information is necessary for the decision-maker to make an informed evaluation taking into account the potential limitations and for example an adequate risk-benefit-evaluation.

Table 1. A non-exhaustive list of the types of uncertainties in QSARs, with examples for the level of uncertainty, and where the scientific knowledge and methods have been extended since the publication of the OECD Principles (OECD, 2007).

General aspects of the QSAR or its application	Specific considerations and / or parts of the QSAR	Examples of low / moderate / high uncertainty	Extension of scientific knowledge since the publication of the OECD Principles
Biological Data in the QSAR	Quality of data	Low – OECD TG, GLP, adherence to Animal Research: Reporting of *In Vivo* Experiments (ARRIVE) guidelines etc.; multiple concordant values etc. Moderate – Accepted guideline, no GLP, or GLP but non-guideline; few or only one value High – Method not known or stated; little concordance between multiple values	These issues were mostly considered in the original Principles
	Relevance of data for the endpoint of interest to its intended use	Low – Endpoint measured (which will be the endpoint predicted by the model) is directly related to the (regulatory) endpoint of interest Moderate – Endpoint measured not the (regulatory) endpoint of interest, but closely related High – Endpoint measured related to an activity which might contribute to insights into the endpoint of interest	More and different types of data are now utilised e.g. omics, high throughput etc. which may not have direct relevance to regulatory endpoint or apical effect but which may be useful for WoE in decision making
	Mechanisms of action	Low – Strong evidence of underlying mechanism(s) of action and the model being demonstrably-related to the mechanism(s); for example, in terms of a Molecular Initiating Event or Key Event Moderate – Some, or partial, knowledge of mechanism and relevance to the QSAR High – No mechanistic hypothesis or relevance	More mechanistic knowledge is collated, for example through Adverse Outcome Pathways (AOPs) than envisioned in the Principles

Descriptors in the QSAR	Experimentally measured properties e.g. log P	Low – OECD TG, GLP etc.; multiple concordant values etc. Moderate – Accepted guideline: no GLP, or GLP but non-guideline; few or only one value High – Method not known or stated; little concordance between multiple values	These issues were mostly considered in the original Principles
	Calculated physico-chemical properties	Low – Reliable and reproducible descriptors; multiple reproducible values with low error Moderate – Two or more values with reasonable concordance High – Calculated descriptors not reproducible, little concordance between multiple values	These issues were mostly considered in the original Principles
	Relevance of descriptors	Low – Clear association to mechanism of action Moderate – Probable association to mechanism of action High – No association to mechanism of action	Many more descriptors are currently applied including, for instance, molecular fingerprints etc.
Statistical (or Other) Method Used in the QSAR	Transparency and reproducibility of the model developed	Low – Clearly defined and reproducible, i.e. fully transparent, model Moderate – Poor or no definition and / or transparency but model is reproducible High – Model is not defined, transparent nor reproducible	There is a greater use of multivariate statistical methods, especially in machine learning, deep learning, artificial intelligence (AI) etc. since the publication of the original Principles
	Model performance	Low – Accurate model, good performance including low systematic error Moderate – Poor model performance or accuracy High – Accuracy/performance not known	More is known about the assessment of model performance since the publication of the original Principles e.g. different types of metrics, need to perform external validation etc.
Applicability of the QSAR	Applicability domain	Low – Relevant applicability domain clearly defined Moderate – Applicability domain not clearly defined or ambiguous High – Applicability domain not known	More is known about describing applicability domains and the different aspect that may be required, e.g.

			structural, properties, metabolism etc., since the original Principles
	Relevance of the QSAR to the prediction or assessment goal, i.e. how much uncertainty does the QSAR bring to the WoE, IATA, Defined Approach (DA) etc.	Low – The QSAR is well-used and considered unambiguous; related directly to the endpoint being predicted Moderate – The QSAR is novel and /or poorly transparent; or poorly, or not, related to the endpoint being predicted High – The QSAR is not transparent; poorly or not related to the endpoint being predicted; contradictory to other evidence in a WoE, IATA or DA	The widespread use of QSARs as part of strategies including WoE, IATA and DA, as opposed to being standalone, was not foreseen in the original Principles

There are various definitions of uncertainty with regard to toxicological assessment, it is beyond the scope or possibility of this paper to review all of these, or indeed to standardise them, but some key definitions include the following. The World Health Organisation International Programme on Chemical Safety (WHO IPCS) gave a general definition of uncertainty as "*imperfect knowledge concerning the present or future state of an organism, system, or (sub)population under consideration*" (WHO IPCS, 2004). Relevant to this paper, the WHO IPCS (2004) definition was refined specifically to hazard characterisation "*as lack of knowledge regarding the "true" value of a quantity, lack of knowledge regarding which of several alternative model representations best describes a system of interest, or lack of knowledge regarding which probability distribution function and its specification should represent a quantity of interest*" (WHO IPCS 2017). ECHA (2012) builds on the WHO IPCS definition stating "*Uncertainty can be caused by limitations in knowledge (e.g. limited availability of empirical information), as well as biases or imperfections in the instruments, models or techniques used. An example is an emission estimate that is based on a reasonable-worst case assumption.*". Whilst these (and other) formal definitions are not discounted, this paper has preferentially considered the possibly broader terminology proposed by EFSA (2018a) which defined uncertainty as "*all types of limitations in available knowledge that affect the range and probability of possible answers to an assessment question*" (EFSA, 2018a). The EFSA Guidance is based around identifying, assessing, describing and, in some cases, quantifying uncertainty. There have been a variety of approaches that have attempted to define uncertainty in toxicological QSAR including those based on epistemological and other analyses (Vallverdu, 2012) and various statistical approaches (e.g., Sahlin 2013; Sahlin et al., 2013; Sahlin, 2015). Aligned to the concept of uncertainty is its quantification and the fact that increasingly robust quantitative uncertainty analysis frameworks have been developed which have provided a unified framework for hazard characterisation (WHO IPCS, 2017; Chiu and Slob (2015). It is acknowledged that this study does not itself attempt full quantification, but that there is a growing need for it to enable decisions regarding the acceptability of a prediction to be made. Whilst the areas of uncertainties relating to read-across have been identified (Schultz et al., 2019), and a checklist of elements to consider as part of an expert review of QSARs has been developed (Myatt et al 2018), as yet there has been no coherent mapping or definition of uncertainty with regard to QSAR models.

As with uncertainty, there are various definitions of variability (cf. ECHA, 2012; EFSA, 2018b; NRC, 2009; US EPA, 2001 amongst many others). Again, this paper does not propose a formal definition of variability but has taken it to refer, in a broad sense, to an actual variation or heterogeneity that can be measured or assessed in some manner and may possibly be reduced with further information – i.e., in common with EFSA (2018b). This paper also considered areas of bias and influence. Bias can be

defined as the possibility of introducing systematic error in the results (e.g., for the purposes of this study, a prediction) resulting from methodological criteria (Higgins and Green, 2008) with a number of approaches of measuring it (cf. Hooijmans et al., 2014; Krauth et al., 2014). For this study influence implies any aspect (particularly not relating the algorithm or data behind a model) that made a particular model preferable to another e.g. a cognitive bias (cf. Arnott, 2006), thus meaning that other predictions could have been made which may have had higher confidence. In the context of bias and influence in toxicological QSAR, these concepts have been taken to mean any direct or indirect aspects and/or motivations relating to the development, use or interpretation of a model that could be subject to change in a different context or situation – in other words, areas where human decisions have affected the model or use of a model.

The assessment of uncertainties, variabilities, biases and areas of influence of QSARs will undoubtedly allow for a more didactic and context dependent evaluation of their use for toxicity prediction. Overall, this would allow to increase confidence in QSAR models and, as a consequence, further their uptake and use in practice. It is not the purpose of the proposed scheme in this paper to consider any uncertainties, variabilities, biases or influences together or to combine in some way to provide an overall score – this would require a more mathematical approach. However, it is intended that a review of a QSAR model according to the assessment criteria defined will allow a user of a model and a user of a prediction of a model to identify any aspect where confidence may be lacking and make an assessment as to whether this is acceptable for the decision to be made (e.g., is the level of confidence sufficient to make a specific regulatory decision).

The aim, therefore, of this paper was to identify and describe the areas of uncertainty, variability, bias and influence with regard to the prediction of toxicity by QSARs, namely the issues that affect the development, description and utilisation of (Q)SARs. A list of criteria was compiled for each of these three areas that can be applied systematically in order to identify key uncertainties with a QSAR model. These assessment criteria were applied to two published QSAR studies in order to illustrate their utility to identify areas of high uncertainty, variability, bias and influence, allowing the confidence in the model to be evaluated and for mitigation strategies to be formulated.

Methods

Identification of Assessment Criteria for the Uncertainties, Variabilities and Areas of Potential Bias and Influence in a QSAR for Toxicity Prediction

An expert review was undertaken of various types of QSAR analyses for the prediction of toxicity. The specific QSARs are not listed here but drew on the experience of the authors. The review included published and unpublished QSARs, as well as free-to-use and with-payment computational products. From the review based on the experience of the authors, types of uncertainties, variabilities, bias and influences on the development and use of a QSAR for toxicity prediction were compiled and organised logically into assessment criteria. The assessment criteria were intended to be objective, although it is noted that subjectivity will inevitably be part of the evaluation of a model when the criteria are applied. Whilst the classifications are only indicative and for regulatory use would need to be placed in the context of decision and protection goal, Table 2 provides generic definitions of what is termed low, moderate and high uncertainty. These are definitions for use in this study and it is acknowledged that specific definitions may be available in guidance documents (cf. EFSA (2018b), WHO IPCS (2017)) and relevant legislation.

Table 2. Generic definitions and putative relevance to regulatory decision making of the terms low, moderate and high uncertainty of QSARs for toxicity prediction as used in this study.

Uncertainty	Generic Definition	Relevance to Regulatory Decision Making
Low	The evidence and / or hypothesis on which a decision or result is based indicates it is highly probable that the finding is correct	High confidence in QSAR models and predictions; they may support decisions relating to risk assessment, classification and labelling (C&L) and prioritisation
Moderate	The evidence and / or hypothesis on which a decision or result is based indicates it is probable that the finding is correct	Moderate confidence in QSAR models and predictions; they may probably be used to support decisions relating to C&L and prioritisation
High	There is no or little evidence and / or hypothesis that the result is correct	Low confidence in QSAR models and predictions; they may be used to inform prioritisation as part of a weight of evidence

Application of the Uncertainty, Variability, Bias and Areas of Influence Assessment Criteria to Case Studies

The compiled assessment criteria for uncertainty, variability, bias and influences of a QSAR were designed to allow for the evaluation of QSARs. To illustrate their application, the assessment criteria were applied to two case studies as examples using a simplistic scale of high, moderate and low uncertainty, variability or bias and area of influence for each criterion. The case studies, two of the authors' studies, were chosen to represent different modelling approaches to address different endpoints with the intention of demonstrating how applications of the identified sources of uncertainties, variabilities, biases and areas of influence in QSARs may emerge.

Case Study 1. A regression-based QSAR utilising the logarithm of the octanol-water partition coefficient (log P) to describe the inhibition of growth (IGC50) of compounds considered to be acting by the non-polar narcosis mechanism of action to the ciliated protozoan *Tetrahymena pyriformis* (Ellison et al., 2008). The QSAR is as follows:

$$\text{Log } 1/\text{IGC50} = 0.78 \log P - 2.01$$

Case Study 2. An *in silico* workflow that is based firstly on a variety of predicted physico-chemical properties and descriptors and secondly on molecular fragments associated with toxicity which had been coded in SMARTS strings with the aim of identifying compounds that have the capacity to bind to nuclear receptors associated with hepatic steatosis (Mellor et al., 2016).

The findings of the application of assessment criteria were summarised using expert judgement. This applied defined examples of the scoring system for low, medium or high uncertainty, variability, bias or influence. Possible strategies to mitigate areas of high uncertainty, variability, bias or influence were proposed on a pragmatic basis that could be achieved with limited further testing or resources.

Results and Discussion

Toxicology is a science based on measurement, description and analysis of experimental findings. As a scientific discipline it inevitably relies on interpretation of evidence to make a decision. With the paucity of resources available for toxicological assessment, much has been placed into maximising the possible information that may be obtained without recourse to animal testing and / or by utilising existing data. Part of this process of maximising the value and utility of data has been to assign confidence to information by understanding the uncertainties within toxicological results, with a view to minimising the knowledge required in a cost- and resource-efficient manner (Patterson and Whelan, 2017). *In silico* models contribute to toxicological knowledge through chemistry- and property-based predictions (Cronin and Madden, 2010) increasing the understanding of complex causal interrelations. These form a special case in toxicology as an understanding of their uncertainties means that the predictions can be improved and may be considered to have greater confidence. For instance, the inclusion of New Approach Methodology (NAM) data has been shown to reduce uncertainty in read-across (Schultz and Cronin, 2017) and a similar approach could be envisioned for QSARs. Whilst the uncertainties in read-across (Schultz et al., 2019) and for mathematical modelling in Next Generation Risk Assessment (Gosling, 2019) have been defined and described this is not the case for QSARs specifically. In addition, ECHA (2017) has defined the Read-Across Assessment Framework (RAAF) as a means of evaluating a read-across argument for toxicity prediction based, in part at least, on an understanding of the quality of the justification and data and assessment of similarity. Whilst a one-to-one read-across (i.e. an analogue approach) can be considered a unique situation, the RAAF does also allow for read-across from many chemicals. When a reasonable number of chemicals are included in a category for read-across, especially for a quantitative endpoint, the distinction between read-across and QSAR becomes blurred. As such, this investigation has developed a comprehensive set of assessment criteria that describe uncertainties and potential areas of variability, bias and areas of influence in QSARs and demonstrated its applicability to published QSARs, whilst being mindful that these criteria could find use, in certain circumstances, for read-across when trend analysis or QSAR is applied. It is emphasised that the assessment criteria go beyond uncertainties to areas that influence the development and use of a QSAR (e.g. motivation, cost, etc.) but which may assist in the evaluation of quality, or practical applicability, of a model.

Following an expert review and evaluation of numerous *in silico* approaches, uncertainties, as well as areas of variability, bias and influence, associated with QSARs have been assigned to one of three "thematic" areas, namely that from the creation of the model, the description of the model and the

application of the model. Specifically, the three main thematic areas relating to uncertainties, variabilities and areas of potential bias and influence in a QSAR were:

1. **Model Creation:** the information on which the QSAR is based including the quality, consistency, variability, reporting and reliability of the toxicological data, chemical structures and properties / descriptors as well as the development and nature of the algorithm that forms the basis of the QSAR. The completeness of the data set is assessed as is how appropriate the modelling approach is.
2. **Description:** the reporting of the model, i.e. transparency, in terms of the type of an (unambiguous) algorithm, its performance and definition of applicability domains and mechanistic and toxicokinetic relevance.
3. **Application:** aspects of the usability, or barriers to it, of the model are considered, including the actual reproducibility based on the documentation; different influences on the choice of using a particular model; as well as the relevance and adequacy of a model for a particular purpose (if known), in particular for possible regulatory applications.

The assessment criteria for the three thematic areas are described in detail in Tables 3-5 respectively along with examples, comments and reference to whether the information may be retrieved from the QMRF. The first and second thematic areas (i.e., Model Creation (Table 3) and Description (Table 4)) draw heavily from, and extend, the OECD Principles for the Validation of QSARs. The third thematic area, Application (Table 5), goes beyond what is considered in the OECD Principles to draw on experience of the practical usage of a model – it is acknowledged at the outset that many of the issues considered here are not uncertainties but potential areas of variability, bias as well as influences on and barriers to the use of QSARs. In Tables 3 – 5 each criterion is classified as being an uncertainty, variability, bias or influence, with some criteria having two such classifications. These classifications of the criteria are broad and are unlikely to be definitive at this time. However, such criteria could form the foundation of formal validation, guiding an assessor through the assessment and documentation of the different issues described.

Table 3 summarises a total of 20 potential assessment criteria, in terms of areas of concern (i.e., uncertainties, variability, bias) that are associated with the raw data used to develop a QSAR and the modelling technique. The assessment criteria are sub-divided according the type of data and their role in the development of a QSAR. The issues associated with the accuracy of chemical structures are defined in Section 1.1 with an overriding requirement for correct definitive structures (Young et al., 2008; Ball et al., 2016). Seven areas of uncertainty, variability or bias have been identified for the biological / toxicological data on which a QSAR is developed (Section 1.2). Key amongst these are the

intrinsic quality and error associated with the data – well-established and newer schemes to evaluate such criteria are available (Klimisch, 1997; Przybylak et al., 2012; Moermond et al., 2016; Molander et al., 2015, Myatt et al, 2018; NTP, 2015), as well as the consistency and comparability of the data compiled to a data set. Also noted is the use of nominal or measured concentrations – with the realisation that whilst measured internal concentrations are preferred, the majority of historical toxicological data for modelling are based on nominal concentrations (Partosch et al., 2015). Five uncertainties relating to the molecular descriptors or properties on which a model is derived are accounted for in Section 1.3, the intention here is to encourage correct use and documentation of software, or description of experimental measurements, rather than being prescriptive of the type of descriptor. Sections 1.4 and 1.5 account for the uncertainties or potential bias that may be related to issues of the completeness and content of the data set and appropriateness of the modelling approach.

Table 3. Description of the assessment criteria relating to uncertainties, variability and bias that may be associated with the creation of QSAR models.

IID Number	Assessment Criteria for Individual Areas of Uncertainty, Variability or Bias	Example of Low Uncertainty, Variability or Bias	Example of Moderate Uncertainty, Variability or Bias	Example of High Uncertainty, Variability or Bias	Comment or Other Information	Type of Criterion	Information Potentially Retrievable from the QMRF
1.1 Definition of Chemical Structures							
1.1a	Accuracy of chemical structures	Structures unambiguously defined including any isomerism such as tautomerism and stereoisomerism	Structures well-defined with small numbers of ambiguities	Structures not defined	Definitive chemical structures are required including any possible tautomerism, stereochemistry etc.	Uncertainty about the active molecule if exact structure not known Variability of structures due to different forms of isomerism (leading to uncertainty about which structures and related properties/activities are included in the data for modelling and possible impact on the model)	Yes
1.1b	Assessment of significant impurities or mixtures	Impurities / mixtures defined and stated	Major impurities / and components in mixtures defined and stated; only low concentration components omitted	Impurities / mixtures not stated		Uncertainty about the active molecule(s), if not known which components present at the time of measurement Variability of composition for example for nanomaterials	Yes

						(inherent distribution of size etc.)	
1.2 Biological Data							
1.2a	Quality of individual studies in the data set	Standard test, OECD Test Guideline (TG), performed to e.g. Good Laboratory Practice (GLP) standard, adherence to ARRIVE Guidelines	A well-recognised and described test but not necessarily performed to GLP and/or OECD Test Guidelines	Quality not known / determinable or a not recognised non-standard test protocol applied	Evaluation of data quality may require analysis of the original study reports / publication. Criteria such as Klimisch (1997) or SciRAP (Molander et al., 2015) scores, or the SYRCLE risk of bias tool for animal studies (Hooijmans et al., 2014) may be useful. It is acknowledged that a non-standard/ non-guideline test may be high quality.	Uncertainty about the model if underlying data of questionable quality Bias relating to systematic errors in the studies leading to erroneous data	No
1.2b	Consistency of the data set including comparability of data	Consistent set in terms of assay, same laboratories	Minor inconsistencies in the data set, e.g., in test protocols or testing laboratories	Varied test guidelines / sources of data	Do the individual data points form a consistent set of data from which to develop a model in terms of the assays, measurement etc.	Uncertainty about quality, i.e. consistency, of the underlying data set and its impact on the model	No
1.2c	Checking of toxicological data	Source data / study reports checked against the original study reports	Most data checked, or in the case of a large data set a random sample checked and verified	No checking	Have the data been checked e.g. with reference to the source of the data	Uncertainty that may be introduced due to the quality/correctness of the underlying data set for the model	No
1.2d	Error associated with biological data	Error is known and stated and within what would be	Error is not known for some compounds	Unknown error	Is an estimate of the error associated with biological data e.g. the measurement	Uncertainty about measurement errors and thus	No

		normally associated with the test	in the data set or very significant		error and intrinsic variability of the test/ measured property provided? The requirement for knowledge of error is so that the model is not overfitted. Naturally modelling with data with low error is preferred.	quality of the data set and its impact on modelling Variability relating to the property to be measured which may vary among individuals	
1.2e	(if required) Units of concentration known, stated and appropriate for use	Appropriate units stated	Units stated but they may not be wholly appropriate	Not known or not stated	For measures of potency, typically molar units are required	Uncertainty about the model if based on not appropriate values	Yes
1.2f	(If appropriate) Use of nominal or measured concentrations	Measured experimental concentrations used taking account of degradation, uptake etc. over time course of the test	Experimental concentrations used, although without reference to time course (e.g., a single measurement)	Nominal concentrations used		Uncertainty about adequate data used for modelling	No
1.2g	Taking into account of the internal exposure	Internal exposure known	Internal exposure estimated	Internal exposure not known	The internal exposure is the actual concentration at the tissue or in the cell	Uncertainty about the actual concentration at the relevant site and impact on the data and modelling	No
1.3 Measurement and / or Estimation of Physico-Chemical Properties and Structural Descriptors							
1.3a	Measurement of physico-chemical	Measurement by appropriate OECD TG / according to	A well-recognised and described test but not necessarily	Experimental procedure not known, little		Uncertainty relates to underlying data quality and impact on the model	Yes

	properties and descriptors	GLP, multiple concordant values etc.	performed to GLP and/or OECD Test Guidelines	concordance between multiple values		Variability is anticipated for measured properties, such as, for example, for nanomaterials	
1.3b	Calculation of properties and 2-D descriptors	Reliable and reproducible descriptors, multiple reproducible values with low error	Two or more values with reasonable concordance	Calculated descriptors not reproducible, little concordance between multiple values		Uncertainty about impact of possible low descriptor quality on the model	Yes
1.3c	Calculation of 3-D descriptors, if utilised	Full structure optimisation and conformational analysis	Structure optimisation performed without conformational analysis	No optimisation of chemical structure		Uncertainty within the model if 3-D structure is not fully optimised Variability relating to variations in 3-D conformations which might have an impact on descriptors used for modelling	Yes
1.3d	Software utilised for descriptor calculation	Full details of software and any options / non-default	Software known, but use (options/ parameters) not fully described	Software not described	Full version number, parameters used etc. of software reported	Uncertainty about the quality of the descriptors and thus of the model	Yes
1.3e	Definition of molecular fragments	Correct fragments used, definition of the fragment and its domain defined	Fragments used, but significance unknown	Non-defined fragments		Uncertainty about the quality of the molecular fragments and impact on model	Yes
1.4 Compilation of the Data Set for QSAR Modelling							
1.4a	Data set is complete	No data gaps	Minor number of missing data (e.g., when experimental	Data gaps present	All data (e.g., structure, toxicological and physicochemical	Uncertainty about the model if developed based on incomplete dataset	Yes

			properties are missing or supplemented by calculated values)		properties) should be available for all chemicals used for modelling		
1.4b	Data set has appropriate variation in potency (quantitative) or balance of actives vs inactives (qualitative)	Good variation in potency (e.g., several log units) or balanced actives / inactives	Moderate range of potency or balance of actives / inactives	Limited range of potency or imbalanced data set		Uncertainty about the model based on unbalanced dataset Bias relating to the selection of the data set relating to the endpoint, application and intended use, and experience of the model developer	Yes
1.4c	Selection of training set data for modelling	Data selection reported to be without bias, removal of outliers has been explained. This may include mechanistic selection of compounds or according to specific chemistries.	Data selection assumed to be without bias (although not reported). No outlier removal apparent or reported	Unknown or unexplained bias in the selection of data to model	QSAR models are seldom developed with all available data; selection criteria should be acknowledged	Uncertainty about the model (and possible bias) if data selection procedure not known and different compilations of data sets are possible Bias from the model developer who may have applied implicit or existing/perceived knowledge to the selection of the training set	Yes
1.4d	Homogeneity of the chemical space of the training and test sets (related to Criterion 2.3a)	Chemicals are heterogeneous across chemical space (and e.g. descriptor space, mechanistic space)	Chemicals are moderately well distributed across chemical space with some areas of high density and some areas of low density	Chemicals are poorly distributed across chemical space or the distribution is not known	Heterogeneity and distribution can be demonstrated by for example, visualisation of density (cf. Jaworska et al., 2005; Hanser et al., 2016)	Uncertainty relating to the distribution across chemical space of the underlying data sets which may skew the model and give false predictions for the target chemical Bias from the chemical space of the data set which might be	No

						influenced by the experience of the model developer or the model's intended use or application	
1.4e	Suitable split into training and test sets, sets defined and utilised	As required, appropriate training and test sets	Only small test sets utilised, lack of resampling of data etc.	Split into training and tests lacking		Uncertainty about appropriate building and use of the training and test sets which may impact on the model	Yes
1.5 Modelling Approach							
1.5a	How appropriate is the modelling approach for the endpoint and to deal with the complexity / non-linearity of the data	Appropriate modelling approach for the endpoint	Modelling approach likely, but unproven, to be appropriate for the endpoint	Approach likely to be too complex or simplistic	This requires a pragmatic and subjective assessment, e.g. a data set based on one mechanism with a single overriding descriptor can be afforded a simpler modelling approach as to one that is more complex	Uncertainty about the model if the modelling approach chosen not appropriate Bias from different approaches to modelling which may result from personal knowledge, experience or prejudice	No

Table 4 describes 13 assessment criteria for areas of concern (i.e., uncertainties) associated with the documentation of the model. This is closest to the QMRF which describes many of the features accounted for in Table 4 (Worth, 2010). Indeed, it is anticipated that a QMRF could be a valuable source of reference to address these criteria. As well as criteria closely aligned to the OECD Principles for the Validation of QSARs, i.e. transparent description of the model, performance, applicability domain, and mechanistic relevance are summarised in Sections 2.1-2.4, Section 2.5 extends the analysis to include explicit consideration of toxicokinetics (which were not envisaged in the original Principles).

Table 4. Description of the assessment criteria relating to uncertainties that may be associated with the description of QSAR models.

ID Number	Assessment Criteria for Individual Areas of Uncertainty	Example of Low Uncertainty	Example of Moderate Uncertainty	Example of High Uncertainty	Comment or Other Information	Type of Criterion	Information Potentially Retrievable from the QMRF
2.1 Description of Model							
2.1a	Definition and description of model (related to Criterion 3.1a)	Model fully defined	A small number of aspects of the model non-defined or ambiguous	Model non-defined or ambiguous	All terms e.g. descriptors, statistical values, algorithms should be defined. The QMRF is a possible reporting format	Uncertainty about model if not completely defined or described: model cannot be retraced and evaluated	Yes
2.1.b	Underlying data set is fully described	All data are provided and described	The data set is described/provided partially (not all properties or structures are fully described)	Data set is neither provided nor described	All data (e.g., structure, toxicological and physico-chemical properties) should be reported	Uncertainty about model if underlying data cannot be retraced or verified	Yes

2.1c	Transparency of the model	Model is transparent in terms of the algorithm and can be interpreted and reproduced	Model is defined providing some aspect of transparency, but may not be reproducible. The algorithms of, e.g. neural networks, may be difficult to interpret even if transparent.	Non-transparent model	A transparent model can be fully understood, reproduced and coded by another user	Uncertainty about the model if not retraceable/ reproducible, cannot be evaluated	Yes
2.2 Statistical Performance							
2.2a	Statement of statistical fit, performance and predictivity	Full description of model performance	Some key measures of model performance missing	Limited or no description of model performance	The use of appropriate validation methods and / or external test sets should be demonstrated, different metrics may be required for different models	Uncertainty about model accuracy and quality of the prediction if no information about the model performance	Yes
2.2b	Interpretation of statistical fit etc with respect to biological measurement error and variability (see Criterion 1.2d)	Statistical performance is significant but not overfitted	Statistical performance moderate or possibly overfitted	No statistical significance or overfitted as compared to experimental error		Uncertainty about the model if performance is not adequate or overfitted	No
2.3 Applicability Domains							

2.3a	Chemical, structural and descriptor applicability domain of model	Fully defined in terms of relevant physico-chemical properties and structural descriptors	Domain defined but not in terms of all key aspects	Not defined	This can be defined by many methods but should be appropriate for the model	Uncertainty about the applicability of the model for the target chemical	Yes
2.3b	Mechanistic applicability domain of model	Fully defined in terms of relevant mechanism(s) of action	Partial definition of mechanistic domain	Not defined	Reference may be made here to relevant AOPs	Uncertainty about the applicability of the model for the target chemical	Yes
2.3c	Biological applicability domain of model	Fully defined including possible metabolism	Partial definition of biological domain/ metabolic domain	Not defined	This could refer to, for instance, consistent NAM data and profiles for the data set, e.g. from ToxCast	Uncertainty about the applicability of the model for the target chemical	Yes
2.4 Mechanistic Relevance and Interpretability							
2.4a	Mechanistic justification	Causative definition of mechanism of action or reference to AOP	Putative definition of mechanism of action or reference to AOP	No mechanistic basis	If multiple mechanisms are present in the dataset, this should be acknowledged and, where possible, defined	Uncertainty about the model due to unknown mechanistic justification	Yes

2.4b	Presence / availability of other and supporting information to support mechanistic interpretation or validity	Use of supporting NAMs, other data or other evidence relating to mechanistic basis	Partial supporting data e.g. for limited proportion of the data set	No supporting information	This would normally form part of the accompanying justification and documentation and may, for instance, confirm mechanism of action	Uncertainty about the model if insufficient evidence provided	No
2.4c	Relevance of descriptors to mechanism of action / AOP	Descriptors or properties clearly and causally related to mechanism	Partial or correlated relationship to mechanism	No mechanistic basis of descriptors		Uncertainty about model if relevance of descriptors used for modelling not known or interpretable	Yes
2.5 Adequate coverage of Absorption, Distribution, Metabolism and Excretion (ADME) effects							
2.5a	Metabolism and / or effect of significant metabolites have been considered	Role of metabolism in eliciting the toxicity is established	Metabolism is assumed but without experimental evidence	No reference to metabolism	This could include whether the test system is metabolically competent	Uncertainty about possible impact of metabolism or metabolites on the model and predicted effect	No
2.5b	Toxicokinetics have been addressed in the model	Model relates to toxicokinetic considerations that affect toxicity or potency	Model only partially relates to toxicokinetics	No reference to toxicokinetics		Uncertainty about the role and possible impact of toxicokinetics on the model	No

The criteria relevant to the practical application and use of a QSAR, in particular for possible regulatory uses, are addressed in Table 5. The 16 assessment criteria in Table 5 continue to move away from uncertainties, variability and bias to provide a greater understanding of areas of bias, influence and the appropriateness of a QSAR to address a specific problem as well as potential barriers to their use. Section 3.1 considers whether a model is sufficiently documented and the documentation allows the model to be reproduced for the required purpose. Such considerations have been shown to be very important in terms of use of a model and reproducibility not always given even when QMRFs were available (Patel et al., 2018; Piir et al., 2018). Section 3.2 considers the model from the user's perspective and requires consideration of the accessibility and utility – such as in terms of user-friendliness or costs – of a particular QSAR. In addition, the sustainability of a QSAR is addressed, as this is seen as a significant factor in the use of a model (Cronin et al., 2019). These criteria go beyond the model itself into understanding where it is practically usable and possible related concerns/barriers, e.g. intellectual property or ethical concerns. Finally, but most importantly especially for regulatory applications, the relevance and adequacy of a QSAR to make a prediction for a specific effect for the endpoint considered and the specific hazard/risk assessment context is addressed in Section 3.3.

Many of the assessment criteria in Table 5 move away from what is traditionally considered in the evaluation of a QSAR, however they relate to the applicability of a particular model to provide data for one or more problem(s). As with any of the assessment criteria, the assessor may omit specific criteria should they be deemed not relevant or out of scope for the model or purpose. Assessment criteria 3.3d and 3.3e (human health vs environmental effects) are likely to be mutually exclusive for most use case scenarios. In addition, there is consideration of aspects of inherent (and cognitive) bias and hence this is moving away from classic areas of uncertainty to other areas that should be considered to gain a full understanding of where bias and other areas of influence could enter in a model. For instance, these may be used to identify issues such as models being deliberately developed to be sensitive or specific – both of which are legitimate modelling approaches but require to be identified for use. Assessment Criteria 3.2b and 3.2d move away from the model itself into issues that may bias model development or application e.g. a model may be biased towards certain chemistries or solving a particular problem. Identification of areas of bias does not preclude any use of a model but it allows the user to understand how it may be used and potential strengths and weaknesses.

Table 5. Description of the assessment criteria relating to uncertainties, and areas of bias and influence that may be associated with the application of QSAR models.

ID Number	Assessment Criteria for Individual Areas of Uncertainty, Bias and Influence	Example of Low Uncertainty, Bias and Influence	Example of Moderate Uncertainty, Bias and Influence	Example of High Uncertainty, Bias and Influence	Comment or Other Information	Type of Criterion	Information Potentially Retrievable from the QMRF
3.1 Documentation and Reproducibility							
3.1a	Reproducibility of the model or QSAR (related to Criterion 2.1a)	Full documentation, availability of data and details of software do allow to repeat the QSAR *de novo*	Some aspects of the model, software or data are not available, meaning there is difficulty in reproducing the model	QSAR cannot be reproduced	To determine reproducibility, the model is assumed to be transparent (see Criterion 2.1c)	Uncertainty about the model if it cannot be reproduced	No
3.1.b	Reproducibility of the QSAR prediction	Application of the model to the same chemical always gives the same prediction result (using the same descriptors)	Model does not give reproducible predictions without careful control of descriptors	Model does not give reproducible predictions	To obtain reproducible predictions, all parameters (descriptors) need to be available and controllable	Uncertainty will be increased if predictions are not reproducible	No
3.2 Usability							

3.2a	Implementation of the model	Fully implemented into software	Model has the potential to be implemented but this has not been undertaken	No implementation possible	A usable model to obtain a prediction from a chemical structure is preferred	Influence as the ease of use of a model will influence if and how a user applies it Bias in the choice of the models implemented in the software may be biased by the sector for which the software is intended for	No
3.2b	Software accessibility	Software is publicly and freely available	Software may be obtained on specific license	Software is not publicly available, e.g. in-house software		Influence relating to software accessibility of a model will influence if and how a user applies it	No
3.2c	Software transparency	Software algorithm transparent	Design of software and implementation of model transparent, but not the code	Closed software which cannot be examined		Uncertainty about model if details not known, lower confidence in a model associated with software perceived as a "black box" Influence as a user may not wish to use	No

						non-transparent software	
3.2d	Relative cost	Free / cheap to use compared to a standard test	Moderate cost and / or need for licencing	Cost equivalent or greater than a standard test		Influence where a user may have budget constraints that pre-determine which models can / cannot be used	No
3.2e	Sustainability	Strong sustainability plan	Model available from a reliable source e.g. governmental web-site, but without published sustainability planning	No obvious sustainability e.g. a model on a freely available web-site, which might disappear at any time		Influence as a user may prefer the use of a model with assurance that it will be available in the future	No
3.2f	Maintenance and support	Strong maintenance plan and good product support	Limited possibility for maintenance and / or support	No obvious maintenance or support, or technical implementation is not updated	Software that is not maintained and updated could become redundant	Influence as a user may prefer a well maintained and supported product	No
3.2g	Intellectual Property (IP)	No IP limitations, i.e. open access with e.g. appropriate Creative Commons license	IP, but available via license	IP restrictive to use of the model		Influence as a user may preferentially wish to use (or develop further) a model without IP restrictions	No
3.2h	Ownership	Model ownership defined and contact provided	Ownership known, but no contact information or availability	Ownership not known		Influence as a user may prefer a model with known ownership (e.g. to obtain more	Yes

						information, or to trust the source as perceived as being reliable)	
3.2i	Ethics	No ethical concerns	Minor ethical concerns e.g. use of animal data without fully reported ethics	Ethical concerns about development or use of model	This may relate to the data e.g. animal tests or data protection issues	Influence as a user may prefer to use a model with known ethical pedigree	No
3.3 Applicability, Relevance and Adequacy							
3.3a	Heterogeneity and density of chemical space of the model for the chemical for which a prediction is required (related to Criteria 1.4d)	Prediction being made in an area of the applicability domain that has a high density of potentially similar compounds	Prediction being made in an area of the applicability domain that has a moderate density of potentially similar compounds	Prediction being made in an area of the applicability domain that has a low density of potentially similar compounds or density not known		Uncertainty about the applicability of the model for considered target chemical	No
3.3b	Relevance of the predicted endpoint for the regulatory risk assessment purpose/protection goal	Predicted endpoint related directly to the (regulatory) endpoint of interest and relevant for overall assessment purpose	Predicted endpoint generally relevant for the regulatory endpoint considered, might contribute to insights, but does not provide sufficient evidence	Not relevant to considered endpoint or stated purpose		Uncertainty about the relevance of the model for the risk assessment, purpose or regulatory endpoint	Yes

			to fulfil overall stated risk assessment purpose				
3.3c	Adequacy of the model to make a prediction for the stated (regulatory) purpose	Prediction is adequate for the stated purpose	Prediction is adequate for the considered endpoint generally, but not for the particular regulatory purpose (for example classification and labelling)	Not adequate for stated purpose	Adequacy is a function of the information and confidence required for a prediction in a certain set of circumstance. There are many potential uses for the same model and the acceptability and adequacy of a prediction might be different for them.	Uncertainty about the adequacy of the model for a specific purpose	No
3.3d	Extrapolation and relevance to humans	Relevance for humans generally agreed or demonstrated based on same biological pathways etc.	Potentially relevant to humans e.g. predictive of effects to a non-human mammalian species	Not relevant		Uncertainty about the relevance of a model to human health	Yes
3.3e	Extrapolation and relevance to environmental biota	Relevance to environmental biota generally agreed or demonstrated based	Partially relevant to environmental biota	Not relevant to environmental biota		Uncertainty about the relevance of a model to environmental biota	Yes

		on same biological pathways etc.					

The list of assessment criteria can be used for the evaluation of a QSAR model in a systematic way, in order to identify key uncertainties. The practical application of the assessment criteria in Tables 3-5 requires a thorough analysis of the QSAR model but also a use case scenario or problem formulation. As noted, some of the information can be obtained from a well-developed QMRF (see Tables 3-5). Other information requires expert judgement and detailed assessment of the QSAR model. In order to put the assessment criteria in the context of the OECD Principles for the Validation of QSARs the relevant criteria have been mapped onto the individual Principles, as shown in Table 6. This analysis also demonstrates that the assessment criteria cover many other aspects of QSAR models which are also summarised in Table 6. The assessment criteria allow for a greater understanding of where uncertainties in particular affect the application of QSARs for a particular purpose and what factors might also have an influence, thus extending the concept of an assessment or even formal validation of a QSAR to whether individual predictions may be appropriate for a specific purpose and how they could be improved.

Table 6. Mapping of the assessment criteria in Tables 3-5 onto the OECD Principles of the Validation of QSARs or other topics not addressed specifically or wholly by the OECD Principles.

OECD Principles	Assessment Criteria in Tables 3-5
1) a defined endpoint	1.2e, 3.3b, 3.3c, 3.3d, 3.3e
2) an unambiguous algorithm	2.1a, 2.1b, 2.1c
3) a defined domain of applicability	2.3a, 2.3b, 2.3c
4) appropriate measures of goodness-of–fit, robustness and predictivity	1.4e, 2.2a
5) a mechanistic interpretation, if possible	2.4a
Other Issues Not Necessarily Captured by the OECD Principles	
Data quality and curation	1.1a, 1.1b, 1.2a, 1.2b, 1.2c, 1.2d, 1.2f, 1.4b, 2.2b
Model / descriptors	1.3a, 1.3b, 1.3c, 1.3d, 1.3e, 1.4a
ADME and toxicokinetics considerations	1.2g, 2.5a, 2.5b
Bias in QSAR development	1.4c, 1.5a, 3.3a
Details of the documentation of the model	1.4d, 2.4b, 2.4c, 3.1a
Practical use of the QSAR and relevance / adequacy for regulatory application	3.1b, 3.2a, 3.2b, 3.2c, 3.2d, 3.2e, 3.2f, 3.2g, 3.2h, 3.2i

Two case studies applying the assessment criteria to QSARs representing different model types have been performed. The results of the evaluation with the assessment criteria are reported in the Supplementary Information. The assessment criteria were applied rapidly in the case studies as would be the case for a quick routine assessment of the usability of a model in a broader risk evaluation

exercise. Each criterion can be scored with a scoring system of "high, moderate or low" uncertainty, variability, bias or area of influence being used in this analysis. According to user requirements, the scoring system may be annotated to explain any particular decision. However, the scoring scheme is intended to be user-defined and whatever scoring scheme is employed, it should be flexible and adaptable. It is not intended to give an overall score of uncertainty, variability, bias or influence in this way but to highlight potential areas where further consideration may be required.

Table S1 gives the findings for a simple log P regression-based QSAR and Table S2 for the *in silico* workflow for nuclear receptor binding. As well as assigning uncertainties, variabilities and areas of biases and influences Tables S1 and S2 also include a brief comment and explanation which will be especially important to explain a decision. The outcome of the analysis detailed in Table S1 for Case Study 1 is summarised in Table 7. This shows that the greatest uncertainties relate to the use of a non-standard test with historical protocols although existing knowledge is presented that may reduce uncertainty (Cronin et al., 1991; Hewitt et al., 2011). Likewise, the analysis in Table S2 for Case Study 2 is summarised in Table 8 showing uncertainty with biological data quality and its relevance. Such analyses also provide an opportunity to identify mitigation strategies that can help to reduce (if possible) uncertainty. The intention of the mitigation strategies is to allow for resources to be directed to improve the use and relevance of QSAR models.

Table 7. Summary analysis of the analysis of the criteria for uncertainty, variability, bias and influence (as reported in Table S1), as a means to highlight possible areas for improvement, for Case Study 1: QSAR for the inhibition of growth of non-polar narcotics to *Tetrahymena pyriformis.*

Number of Criteria with **low** uncertainty, variability, bias, influence	Number of Criteria with **moderate** uncertainty, variability, bias, influence	Number of Criteria with **high** uncertainty, variability, bias, influence	Number of Criteria Not Considered
36	3	2	8
Main Areas of Uncertainty	**Relating to Criteria**	**Mitigation Strategy**	
Biological data quality	1.2a, 1.2f, 1.2g	Biological data are historical but well characterised (Hewitt et al., 2011). They are shown to relate strongly to e.g. fish toxicity (Cronin et al., 1991). It is unlikely that internal concentrations will be calculated and this must be borne in mind if predictions are used.	
No training / test set	1.4e	Model could be redeveloped with a training / test set	
Model not implemented	3.2a	Model could be implemented in appropriate software	

Table 8. Summary analysis of the analysis of the criteria for uncertainty, variability, bias and influence (as reported in Table S2), as a means to highlight possible areas for improvement, for Case Study 2: *In silico* workflow for nuclear receptor binding leading to hepatic steatosis.

Number of Criteria with **low** uncertainty, variability, bias, influence	Number of Criteria with **moderate** uncertainty, variability, bias, influence	Number of Criteria with **high** uncertainty, variability, bias, influence	Number of Criteria Not Considered
30	7	9	3
Main Areas of Uncertainty	**Relating to Criteria**	**Mitigation Strategy**	
Biological data quality	1.1b, 1.2a, 1.2b, 1.2c, 1.2d, 1.2e, 1.2f, 1.2g	Biological data are historical but well characterised. Further analysis of the original source data / publications could assist in understanding and confirming data quality. It is unlikely that internal concentrations will be calculated and this must be borne in mind if predictions are used.	
No training / test set	1.4e	Model could be redeveloped with a training / test set	
Reporting of statistical performance	2.2a, 2.2b	Re-analysis of the models could be performed to determine statistical performance	
Relevance of the endpoint modelled	2.3c, 3.3d	More information could be provided, e.g. via a relevant AOP, to demonstrate relevance	
ADME and TK consideration	2.5a, 2.5b	A greater consideration of metabolic stability and relevance of toxicokinetics could be provided	
Lack of maintenance and support	3.2f	A maintenance and support plan could be initiated	

With the application of the assessment criteria reported in Tables 3-5 it is important to recognise the purpose of the analysis. The primary purpose is to identify uncertainties or areas of variability and / or bias and influence in a QSAR. The intent of the identification of these areas is two-fold. Firstly, to alert the user to potential limitations in the use of a QSAR for a specific purpose, the limitations identified can then be considered on an individual basis in the context of the use of the prediction. For example, in Case Study 1, analysis of the first thematic area (i.e., Model Creation) indicates that assessment criteria 1.2a and 1.4d each have moderate uncertainty as a result of the data being derived from a non-standard test and data were not split into training and test sets, respectively. However,

since the *Tetrahymena* population growth inhibition testing was performed to a rigorous Standard Operating Procedure (SOP) and the QSAR was developed with a tightly defined applicability domain, this uncertainty is likely to be acceptable. Similarly, for Case Study 1, assessment criteria 1.1f and 1.1g have high uncertainty as the potency data were based on nominal concentrations and internal exposure is not known, respectively. If the predictions from such a QSAR are used in a weight-of-evidence to support aquatic toxicity predictions, these uncertainties are likely be deemed unimportant; however, such uncertainty may be significant for different types of use. The second intent of the identification of uncertainties is to establish how QSARs may be improved through the inclusion of further information or data such that key areas of concern can be improved (e.g., uncertainties reduced). Other criteria listed in Table 5 in particular allow for the highlighting of potential problems or barriers to the actual application of the models and for the raising of awareness of the issues which relate to the uptake of models in practice. For the use of QSARS in regulatory applications, the relevant criteria for the type of risk assessment and legislative context considered are particularly important to evaluate.

With regard to the application of the assessment criteria in Tables 3-5, the proposal is to use these criteria to assess uncertainty and other issues in terms of a low – moderate – high semi-quantitative scheme. As noted above, this may provide a means to extend the general concepts and discussion of validation of QSARs (OECD, 2007) and provides specific criteria where pre-defined measures for each are applied. The exact application of these criteria, setting of possible pre-defined measures of "acceptability" and inculcation into regulatory frameworks, however, goes beyond the scope of this paper or the capabilities of the authors alone – this is a topic that should be addressed with open dialogue in the QSAR modelling and regulatory risk assessment communities to facilitate international regulatory uptake. There is no reason not to suppose that other scoring schemes could be applied (e.g. as discussed by Schultz et al (2015)) or that uncertainty could be classified as simply acceptable or not acceptable. It would seem that a classification scheme is, however, preferable and that a three-step scale from low to high is pragmatic. In addition, this type of classification scheme would lend itself to a traffic light-style of presentation (as in Tables S1 and S2), or even a summary presentation (as illustrated in the Graphical Abstract). In part, this type of assessment will inform about the adequacy, or otherwise, of a model for a particular purpose, in the same manner as QSAR reporting formats currently used, and may assist in utilisation of QSAR predictions, for instance in accordance with guidance proposed ECHA (2011 – Section R4.3.2). However, future effort should be placed into the development of quantitative uncertainty analysis to determine whether a prediction from a QSAR is acceptable or not for a specific regulatory purpose – especially in light of the unified frameworks currently available for hazard characterisation (WHO IPCS, 2014; Chui and Slob, 2015). In addition,

such an overview of the model, as proposed by the assessment criteria, would allow the user or regulator to quickly grasp any potential areas of concern, which need to be looked into further, and thus contribute to support making an informed decision on the adequateness of the model for the intended purpose – keeping in mind that the relevance of the model has to be verified depending on the endpoint and hazard/risk assessment goal in question.

Conclusions

The uncertainties and areas of variability, biases and influence of QSAR models have been identified and a set of assessment criteria to evaluate QSAR models is presented. In total 49 assessment criteria have been defined to account for aspects of the data behind the model and the model development approach, the description of the model and its application. The criteria allow for the identification of areas of high uncertainty or other issues that may be of concern for the confidence in a particular prediction or use of a model and also point towards issues which might influence or prevent the update of models in practice. Application of the assessment criteria could allow for mitigation strategies to be proposed. The assessment criteria are not intended as a means to rank models in terms of their suitability, and should not be seen, or utilised in that way. However, the criteria are intended to make users of a model aware of the potential areas of uncertainty and / or variability and bias and to provide them an opportunity to reduce them, as well as to make an informed decision on their impact and on the adequateness of use of the model for the intended purpose. In addition, the criteria may provide a stepping stone towards the validation of models.

Acknowledgements

The detailed, knowledgeable and extremely helpful comments from the anonymous reviewers assisted greatly in the improvement of this manuscript. The authors thank them for their considerable efforts and contribution.

References

Arnott, D., 2006. Cognitive biases and decision support systems development: a design science approach. Info. Systems J. 16, 55-78.

Ball, N., Bartels, M., Budinsky, R., Klapacz, J., Hays, S., Kirman, C., Patlewicz, G., 2014. The challenge of using read-across within the EU REACH regulatory framework; how much uncertainty is too much?

Dipropylene glycol methyl ether acetate, an exemplary case study. Reg. Toxicol. Pharmacol. 68, 212-221.

Ball, N., Cronin, M.T.D., Shen, J., Blackburn, K., Booth, E.D., Bouhifd, M., Donley, E., Egnash, L., Hastings, C., Juberg, D.R., Kleensang, A., Kleinstreuer, N., Kroese, E.D., Lee, A.C., Luechtefeld, T., Maertens, A., Marty, S., Naciff, J.M., Palmer, J., Pamies, D., Penman, M., Richarz, A.-N., Russo, D.,P.. Stuard, S.B., Patlewicz, G., van Ravenzwaay, B., Wu, S., Zhu, H., Hartung, T., 2016. Toward Good Read-Across Practice (GRAP) guidance. ALTEX 33, 149-166.

Calder, M., Craig, C., Culley, D., de Cani, R., Donnelly, C.A., Douglas, R., Edmonds, B., Gascoigne, J., Gilbert, N., Hargrove, C., Hinds, D., Lane, D.C., Mitchell, D., Pavey, G., Robertson, D., Rosewell, B., Sherwin, S., Walport, M., Wilson, A., 2018. Computational modelling for decision-making: where, why, what, who and how. R. Soc. Open Sci. 5: 172096. http://dx.doi.org/10.1098/rsos.172096

Chiu, W.A., Slob, W., 2015. A unified probabilistic framework for dose–response assessment of human health effects. Environ. Health Persp. 123, 1241–1254.

Cronin, M.T.D., Dearden, J.C., Dobbs, A.J., 1991. QSAR studies of comparative toxicity in aquatic organisms. Sci. Tot. Environ. 109/110, 431-439.

Cronin, M.T.D., Jaworska, J.S., Walker, J.D., Comber, M.H.I, Watts, C.D., Worth, A.P., 2003. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ. Heal. Persp. 111, 1391-1401.

Cronin, M.T.D., Madden, J.C., (eds), 2010. *In Silico* Toxicology: Principles and Applications. The Royal Society of Chemistry, Cambridge, England.

Cronin, M.T.D., Madden, J.C., Enoch, S.J., Roberts, D.W., 2013. Chemical Toxicity Prediction: Category Formation and Read-Across. The Royal Society of Chemistry, Cambridge, England.

Cronin, M.T.D., Madden, J.C., Yang, C., Worth, A.P., 2019, Unlocking the potential of *in silico* chemical safety assessment – A report on a cross-sector symposium on current opportunities and future challenges. Comput. Toxicol. 10, 38-43.

Cronin, M.T.D., Richarz, A.N., 2017. Relationship between Adverse Outcome Pathways and chemistry-cased in silico models to predict toxicity. Appl. in Vitro Toxicol. 3, 286-297.

Cronin, M.T.D., Walker, J.D., Jaworska, Comber, M.H.I, Watts, C.D., Worth, A.P., 2003. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. Environ. Heal. Persp. 111, 1376-1390.

Cronin, M.T.D., Yoon, M., 2019. Computational Methods to Predict Toxicity. In: Balls, M., Combes, R., Worth, A., (eds) The History of Alternative Test Methods in Toxicology. Academic Press. pp. 287-300.

ECHA (European Chemicals Agency), 2011. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.4: Evaluation of Available Information. https://echa.europa.eu/documents/10162/13643/information_requirements_r4_en.pdf

ECHA (European Chemicals Agency), 2012. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.19: Uncertainty Analysis. https://echa.europa.eu/documents/10162/13632/information_requirements_r19_en.pdf

ECHA (European Chemicals Agency), 2017. Read-across Assessment Framework (RAAF). https://echa.europa.eu/documents/10162/13628/raaf_en.pd

EFSA (European Food Safety Authority) Scientific Committee, Benford, D., et al, 2018a. Guidance on uncertainty analysis in scientific assessments. EFSA J. 16, 5123, pp. 39 https://doi.org/10.2903/j.efsa.2018.5123.

EFSA (European Food Safety Authority) Scientific Committee, Benford, D., et al, 2018b. Scientific Opinion on the principles and methods behind EFSA's Guidance on Uncertainty Analysis in Scientific Assessment. EFSA J. 16 , 5122 p. 235 https://doi.org/10.2903/j.efsa.2018.5122.

Ellison, C.M., Cronin, M.T.D., Madden, J.C., Schultz, T.W. 2008. Definition of the structural domain of the baseline non-polar narcosis model for *Tetrahymena pyriformis*. SAR QSAR Environ. Res. 19, 751–783.

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ. Heal. Persp. 111, 1361-1375.

Gosling, J.P., 2019. The importance of mathematical modelling in chemical risk assessment and the associated quantification of uncertainty. Comput. Toxicol. 10, 44-50.

Hanser, T., Barber, C., Marchaland, J.F., Werner, S., 2016, Applicability domain: towards a more formal definition, SAR QSAR Environ. Res. 27, 865-881.

Hewitt, M., Cronin, M.T.D., Rowe, P.H., Schultz, T.W., 2011. Repeatability analysis of the *Tetrahymena pyriformis* population growth impairment assay. SAR QSAR Environ. Res. 22, 621–637.

Higgins, J.P., Green, S., 2008. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: John Wiley & Sons Ltd.

Hooijmans, C.R., Rovers, M.M., de Vries, R.B.M., Leenaars M., Ritskes-Hoitinga M., Langendam, M.W., 2014. SYRCLE's risk of bias tool for animal studies. BMC Med. Res. Methodol. 14, 43.

Jaworska, J.S., Comber, M., Auer, C., van Leeuwen, C.J., 2003. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. Environ. Heal. Persp. 111, 1358-1360.

Jaworska, J., Jeliazkova, N., Aldenberg, T., 2005. QSAR applicability domain estimation by projection of the training set descriptor space: a review. ATLA. 33, 445-459.

ICH M7, 2017. (R1). Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Multidisciplinary/M7/M7_R1_Addendum_Step_4_31Mar2017.pdf

Judson, P.N., 2009. Towards establishing good practice in the use of computer prediction. Qual. Assur. J. 12, 120-125.

Judson, P.N., Barber, C., Canipa, S.J., Poignant, G., Williams, R., 2015. Establishing Good Computer Modelling Practice (GCMP) in the prediction of chemical toxicity. Mol. Inform. 34, 276-283.

Klimisch, H.J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Reg. Toxicol. Pharmacol. 25, 1–5.

Krauth, D., Woodruff, T.J., Bero, L., 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: A systematic review. Environ. Health Persp. 121, 985-992.

Mellor, C.L., Steinmetz, F.P., Cronin, M.T.D., 2016. Using molecular initiating events to develop a structural alert based screening workflow for nuclear receptor ligands associated with hepatic steatosis. Chem. Res. Toxicol. 29, 203-212.

Moermond, C.T.A., Kase, R., Korkaric, M., Ågerstrand, M., 2016. CRED: Criteria for reporting and evaluating ecotoxicity data. Environ. Toxicol. Chem. 35, 1297–1309.

Molander, L., Ågerstrand, M., Beronius, A., Hanberg, A, Rudén, C., 2015. Science in Risk Assessment and Policy (SciRAP): An online resource for evaluating and reporting *in vivo* (eco)toxicity studies. Hum. Ecol. Risk Assess. 21, 753-762.

Myatt, G.J., Ahlberg, E., Akahori, Y., et al., 2018. *In silico* toxicology protocols. Reg. Toxicol. Pharmacol. 96, 1-17.

Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton D.T., 2005. Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. The report and recommendations of ECVAM Workshop 52. ATLA 33, 155-173.

NRC (National Research Council), 2009. Science and Decisions: Advancing Risk Assessment. The National Academies Press, Washington, DC. https://doi.org/10.17226/12209.

NTP (National Toxicology Program), 2015. Handbook for Conducting a Literature-Based Health Assessment using OHAT Approach for Systematic Review and Evidence Integration. https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf

OECD (Organisation for Economic Cooperation and Development), 2007. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships. ENV/JM/MONO(2007)2. OECD, Paris, pp. 154.

Partosch, F., Mielke, H., Stahlmann, R., Kleuser, B., Barlow, S., Gundert-Remy, U., 2015. Internal threshold of toxicological concern values: enabling route-to-route extrapolation. Arch. Toxicol. 89, 941-948.

Patel, M., Chilton, M.L., Sartini, A., Gibson, L., Barber, C., Covey-Crump, L., Przybylak, K.R., Cronin, M.T.D., Madden, J.C. 2018. Assessment and reproducibility of Quantitative Structure–Activity Relationship models by the nonexpert. J. Chem. Inf. Mod. 58, 673–682.

Patterson, E.A., Whelan, M.P., 2017. A framework to establish credibility of computational models in biology. Prog. Biophys. Mol. Biol. 129, 13-19.

Piir, G., Kahn, I., García-Sosa, A.T., Sild, S., Ahte, P., Maran, U., 2018. Best practices for QSAR model reporting: Physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints. Environ. Health Persp. 126, 126001-1 – 126001-20.

Przybylak, K.R., Madden, J.C., Cronin, M.T.D., Hewitt, M., 2012. Assessing toxicological data quality: basic principles, existing schemes and current limitations. SAR QSAR Environ. Res. 23, 435-459.

Sahlin, U., 2013. Uncertainty in QSAR predictions. ATLA 41, 111-125.

Sahlin, U., 2015. Assessment of uncertainty in chemical models by Bayesian probabilities: Why, when, how? J. Comp.-Aided Mol. Des. 29, 583-594.

Sahlin, U., Filipsson, M., Oberg, T., 2011. A risk assessment perspective of current practice in characterizing uncertainties in QSAR regression predictions. Mol. Inform. 30, 551-564.

Sahlin, U., Golsteijn, L., Iqbal, M.S., Peijnenburg, W., 2013. Arguments for considering uncertainty in QSAR predictions in hazard and risk assessments. ATLA 41, 91-110.

Schultz, T.W., Amcoff, P., Berggren, E., Gautier, F., Klaric, M., Knight, D.J., Mahony, C., Schwarz, M., White, A., Cronin M.T.D., 2015. A strategy for structuring and reporting a read-across prediction of toxicity. Regul. Toxicol. Pharmacol. 72, 586-601.

Schultz, T.W., Cronin, M.T.D., 2017. Lessons learned from read-across case studies for repeated-dose toxicity. Reg. Toxicol. Pharmacol. 88, 185-191.

Schultz, T.W., Richarz, A.-N., Cronin, M.T.D., 2019. Assessing uncertainty in read-across: Questions to evaluate toxicity predictions based on knowledge gained from case studies. Comput. Toxicol. 9, 1-11.

Spielmann, H., Sauer, U.G., Mekenyan, O., 2011. A critical evaluation of the 2011 ECHA reports on compliance with the REACH and CLP Regulations and on the use of alternatives to testing on animals for compliance with the REACH Regulation. ATLA 39, 481-493.

Tong, W., Hong, H., Xie, Q., Shi, L., Fang, H., Perkins, R., 2005. Assessing QSAR limitations - A regulatory perspective. Curr. Comput.-Aided Drug Des. 1, 195-205.

Tropsha, A., 2010. Best practices for QSAR model development, validation, and exploitation. Mol. Inform. 29, 476-488.

Tunkel, J., Mayo, K., Austin, C., Hickerson, A., Howard, P., 2005. Practical considerations on the use of predictive models for regulatory purposes. Environ. Sci. Technol. 39, 2188-2199.

US EPA (United States Environmental Protection Agency), 2001. Risk Assessment Guidance for Superfund: Volume III - Part A, Process for Conducting Probabilistic Risk Assessment. US EPA, Washington DC. https://www.epa.gov/sites/production/files/2015-09/documents/rags3adt_complete.pdf

Vallverdu, J., 2012. An Epistemological Analysis of QSPR/QSAR models. In: Castro, E.A., Hagji, A.K. (eds) Advanced Methods and Applications in Chemoinformatics: Research Progress and New Applications. IGI Global, Hershey PA, USA, pp. 318-332.

WHO IPCS (World Health Organization International Programme on Chemical Safety) (2004). IPCS Risk Assessment Terminology. World Health Organisation, Geneva, Switzerland. http://www.inchem.org/documents/harmproj/harmproj/harmproj1.pdf

WHO IPCS (World Health Organization International Programme on Chemical Safety) (2017). Harmonization Project Document 11. Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterisation. Second edition. World Health Organisation, Geneva, Switzerland. http://apps.who.int/iris/bitstream/10665/259858/1/9789241513548-eng.pdf

Worth, A.P., 2010. The role of QSAR methodology in the regulatory assessment of chemicals. In Puzyn, T., Lesczynski, J., Cronin, M.T.D. (Eds.). Recent Advances in QSAR Studies: Methods and Applications. Springer, Dordrecht, The Netherlands, pp. 367-382.

Young, D., Martin, T., Venkatapathy, R., Harten, P., 2008. Are the chemical structures in your QSAR correct? QSAR Comb. Sci. 27, 1337-1345.

Supplementary Information

Identification and Description of the Uncertainty, Variability, Bias and Influence in Quantitative Structure-Activity Relationships (QSARs) for Toxicity Prediction

Mark T.D. Cronin[1], Andrea-Nicole Richarz[2], and Terry. W. Schultz[3]

[1]Liverpool John Moores University, School of Pharmacy and Biomolecular Sciences, Liverpool, England

[2]European Commission, Joint Research Centre (JRC), Ispra, Italy

[3]The University of Tennessee, College of Veterinary Medicine, Knoxville TN, USA

Table S1. Case Study 1: Application of assessment criteria in Tables 3-5 to a QSAR for non-polar narcosis to *Tetrahymena pyriformis*

Low Uncertainty, Variability, Bias or Influence	
Moderate Uncertainty, Variability, Bias or Influence	
High Uncertainty, Variability, Bias or Influence	

ID	Area of Uncertainty, Variability, Bias or Influence	Assignment of Uncertainty, Variability, Bias or Influence	Reason
1. Model Creation			
1.1 Definition of Chemical Structures			
1.1a	Accuracy of chemical structure		Structures unambiguously defined including any isomerism
1.1b	Assessment of significant impurities or mixtures		Impurities / mixtures defined and stated
1.2 Biological Data			
1.2a	Quality of individual studies in the data set		Non-standard test, although performed to a rigorous SOP
1.2b	Consistency of the data set including comparability of data		Consistent set in terms of assay, same laboratory
1.2c	Checking of toxicological data		Source data / study reports checked
1.2d	Error associated with biological data		Error is known and stated. There is a publication on the experimental variability.
1.2e	(if required) Units of concentration known, stated and appropriate for use		Appropriate units stated
1.2f	(If appropriate) Nominal or measured concentrations		Nominal concentrations used
1.2g	Internal exposure known		internal exposure is not known
1.3 Measurement and / or Estimation of Physico-Chemical Properties and Structural Descriptors			

1.3a	Measurement of physico-chemical properties	N/A	Measured properties not used
1.3b	Calculation of properties and 2-D descriptors		Well characterised software providing unambiguous properties
1.3c	Calculation of 3-D descriptors	N/A	3-D descriptors not utilised
1.3d	Software utilised		Full details of software provided
1.3e	Definition of molecular fragments	N/A	Molecular fragments not used in model (but used to define the applicability)
1.4 Creation of the Data Set for QSAR Modelling			
1.4a	Data set is complete		Full data set provided
1.4b	Data set has appropriate variation in potency (quantitative) or balance of actives vs inactives (qualitative)		Good variation in potency (e.g. several log units)
1.4c	Selection of training set data		Training set selected without bias
1.4d	Training set homogeneity		Training set is homogeneous
1.4e	Suitable training and test sets defined and utilised		No splitting into training and test set, although QSAR developed as much for the applicability domain as the model
1.5 Modelling Approach			
1.5a	How appropriate is the modelling approach for the endpoint and to deal with the complexity / non-linearity of the data		Regression analysis is an appropriate modelling approach for the endpoint
2. Description of the QSAR Model			
2.1 Description of Model			
2.1a	Documentation and reporting		Model fully defined
2.1b	Data set is complete and described		No data gaps
2.1c	Transparency of the model		Model is transparent in terms of the algorithm

2.2 Statistical Performance			
2.2a	Statement of statistical fit, performance and predictivity		Full description of model performance
2.2b	Interpretation of statistical fit etc with respect to biological error (see Criterion 1.2d)		Statistical performance is significant but not overfitted
2.3 Applicability Domains			
2.3a	Chemical applicability domain of model		Fully defined in terms of relevant physico-chemical properties and structure
2.3b	Mechanistic applicability domain of model		Fully defined in terms of relevant mechanism(s) of action
2.3c	Biological applicability domain of model		Fully defined including possible metabolism
2.4 Mechanistic Relevance, Interpretability and Transparency			
2.4a	Mechanistic justification		Definition of non-polar narcosis mechanism of action
2.4b	Presence / availability of other and supporting information		Use of evidence relating to mechanistic basis
2.4c	Relevance to descriptors to mechanism of action / AOP		Descriptors or properties clearly related to mechanism
2.5 Adequate coverage of ADME effects			
2.5a	Metabolism and / or effect of significant metabolites have been considered		Role of metabolism in eliciting the toxicity is established
2.5b	Toxicokinetics have been addressed in the model		Model relates to toxicokinetic considerations that affect toxicity or potency
3. Application of the QSAR Model			

3.1 Documentation and Reproducibility			
3.1a	Reproducibility of the models		Model transparent and fully documented
3.1b	Reproducibility of the prediction	N/A	No predictions made
3.2 Usability			
3.2a	Implementation of the model		Not implemented into software but it could be possible
3.2b	Software accessibility	N/A	No software utilised
3.2c	Software transparency	N/A	No software utilised
3.2d	Relative cost		Cheap to use compared to a standard test
3.2e	Sustainability		Published QSAR
3.2f	Maintenance and support	N/A	No software provided
3.2g	Intellectual Property		No IP considerations i.e. open access
3.2h	Ownership		Model in public domain
3.21	Ethics		No ethical concerns
3.3 Relevance			
3.3a	Heterogeneity and density of chemical space		Well populated and distributed chemical space
3.3b	Relevance of the predicted endpoint for the regulatory risk assessment purpose/protection goal		Fit for stated purpose. Likely to provide an estimate that could support e.g. hazard identification.
3.3c	Adequacy		Adequate for stated purpose. Likely to provide an estimate that could support e.g. hazard identification.
3.3d	Extrapolation and relevance to humans	N/A	Not for humans
3.3.e	Extrapolation and relevance to environmental biota		Relevant to environmental biota

Table S2. Case Study 2: Application of assessment criteria in Tables 3-5 to an *in silico* workflow to identify nuclear receptor binders that may be related to hepatic steatosis.

Low Uncertainty, Variability, Bias or Influence	
Moderate Uncertainty, Variability, Bias or Influence	
High Uncertainty, Variability, Bias or Influence	

ID	Area of Uncertainty, Variability, Bias or Influence	Assignment of Uncertainty, Variability, Bias or Influence	Reason
1. Model Creation			
1.1 Definition of Chemical Structures			
1.1a	Accuracy of chemical structure		Structures unambiguously defined including any isomerism
1.1b	Assessment of significant impurities or mixtures		Impurities / mixtures not known
1.2 Biological Data			
1.2a	Quality of individual studies in the data set		Non-standard test, quality of individual studies not known
1.2b	Consistency of the data set including comparability of data		Same / similar assays and endpoints but performed in different laboratories
1.2c	Checking of toxicological data		Source data not checked but used as provided in ChEMBL
1.2d	Error associated with biological data		Unknown error
1.2e	(if required) Units of concentration known, stated and appropriate for use		ChEMBL units applied (not molar)
1.2f	(If appropriate) Nominal or measured concentrations		Nominal concentrations used
1.2g	Internal exposure known		internal exposure is not known

1.3 Measurement and / or Estimation of Physico-Chemical Properties and Structural Descriptors			
1.3a	Measurement of physico-chemical properties	N/A	Measured properties not used
1.3b	Calculation of properties and 2-D descriptors		Well characterised software providing unambiguous properties
1.3c	Calculation of 3-D descriptors	N/A	3-D descriptors not utilised
1.3d	Software utilised		Full details of software provided
1.3e	Definition of molecular fragments		Molecular fragments well defined
1.4 Creation of the Data Set for QSAR Modelling			
1.4a	Data set is complete		Full data set provided
1.4b	Data set has appropriate variation in potency (quantitative) or balance of actives vs inactives (qualitative)		Good variation in potency (e.g. several log units)
1.4c	Selection of training set data		Training set selected without bias
1.4d	Training set homogeneity		Training set is homogeneous
1.4e	Suitable training and test sets defined and utilised		No splitting into training and test set, although QSAR developed as much for the applicability domain as the model
1.5 Modelling Approach			
1.5a	How appropriate is the modelling approach for the endpoint and to deal with the complexity / non-linearity of the data		Regression analysis is an appropriate modelling approach for the endpoint
2. Description of the QSAR Model			
2.1 Description of Model			
2.1a	Documentation and reporting		Model fully defined
2.1b	Data set is complete and described		No data gaps
2.1c	Transparency of the model		Model is transparent in terms of the algorithm

2.2 Statistical Performance			
2.2a	Statement of statistical fit, performance and predictivity		Limited evaluation of statistical performance
2.2b	Interpretation of statistical fit etc with respect to biological error (see Criterion 1.2d)		Statistical performance is significant, possibly over-predictive
2.3 Applicability Domains			
2.3a	Chemical applicability domain of model		Fully defined in terms of relevant physico-chemical properties and structure
2.3b	Mechanistic applicability domain of model		Fully defined in terms of relevant mechanism(s) of action
2.3c	Biological applicability domain of model		Binding data may not have biological relevance
2.4 Mechanistic Relevance, Interpretability and Transparency			
2.4a	Mechanistic justification		Definition of nuclear receptor binding
2.4b	Presence / availability of other and supporting information		Use of evidence relating to mechanistic basis
2.4c	Relevance to descriptors to mechanism of action / AOP		Descriptors or properties clearly related to mechanism
2.5 Adequate coverage of ADME effects			
2.5a	Metabolism and / or effect of significant metabolites have been considered		Role of metabolism not addressed
2.5b	Toxicokinetics have been addressed in the model		Model does not relate to toxicokinetic considerations
3. Application of the QSAR Model			
3.1 Documentation and Reproducibility			

3.1a	Reproducibility of the models		Model transparent and fully documented
3.1b	Reproducibility of the prediction	N/A	No predictions made
3.2 Usability			
3.2a	Implementation of the model		Model implemented ina a KNIME Workflow
3.2b	Software accessibility		KNIME and relevant nodes freely available; as is workflow
3.2c	Software transparency		KNIME and relevant nodes freely transparent; as is workflow
3.2d	Relative cost		Cheap to use compared to a standard test
3.2e	Sustainability		Published QSAR
3.2f	Maintenance and support		Free software with no maintenance and support
3.2g	Intellectual Property		No IP considerations i.e. open access
3.2h	Ownership		Model in public domain
3.2i	Ethics		No ethical concerns due to use of an *in vitro* assay
3.3 Relevance			
3.3a	Heterogeneity and density of chemical space		Well populated and distributed chemical space
3.3b	Relevance of the predicted endpoint for the regulatory risk assessment purpose/protection goal		Fit for stated purpose. Likely to provide an estimate that could support e.g. hazard identification
3.3c	Adequacy		Adequate for stated purpose. Likely to provide an estimate that could support e.g. hazard identification.
3.3d	Extrapolation and relevance to humans		Possible relevance for humans
3.3e	Extrapolation and relevance to environmental biota		Relevant to environmental biota