

Unclassified

English - Or. English

21 November 2023

**ENVIRONMENT DIRECTORATE
CHEMICALS AND BIOTECHNOLOGY COMMITTEE**

Cancels & replaces the same document of 17 August 2023

**(Q)SAR Assessment Framework: Guidance for the regulatory assessment of
(Quantitative) Structure – Activity Relationship models, predictions, and results based
on multiple predictions**

**Series on Testing and Assessment
No. 386**

Two Annex documents are available with the following cotes:

- ENV/CBC/MONO(2023)32/ANN1
- ENV/CBC/MONO(2023)32/ANN2

The checklist (EXCEL version) is available at the following link:
<https://www.oecd.org/chemicalsafety/testing/QAF-Checklist.xlsx>

JT03532229

OECD Environment, Health and Safety Publications
SERIES ON TESTING AND ASSESSMENT
NO. 386

(Q)SAR Assessment Framework: Guidance for the regulatory assessment of
(Quantitative) Structure – Activity Relationship models, predictions, and results
based on multiple predictions

IOMC

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

A cooperative agreement among **FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD**

Environment Directorate
ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT
Paris 2023

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 38 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in twelve different series: **Testing and Assessment; Good Laboratory Practice and Compliance Monitoring; Pesticides; Biocides; Risk Management; Harmonisation of Regulatory Oversight in Biotechnology; Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; Safety of Manufactured Nanomaterials;** and **Adverse Outcome Pathways.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

- **Also published in the Series on Testing and Assessment: [link](#)**

**For this and many other Environment,
Health and Safety publications, consult the OECD's
World Wide Web site (www.oecd.org/chemicalsafety/)**

or contact:

**OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France**

E-mail: ehscont@oecd.org

© OECD 2023

Applications for permission to reproduce or translate all or part of this material should be made to: Head of Publications Service, RIGHTS@oecd.org, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France

OECD Environment, Health and Safety Publications

Foreword

In November 2004, the 37th OECD's Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology (Joint Meeting) agreed on the “OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models”.

In 2007, the same OECD Working Party published the “Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models” for the use of (Q)SARs in regulatory applications (OECD, 2007).

Since then, the possibility to use (Q)SARs was included within various regulations related to chemical substances. This has allowed different stakeholders to gain practical experience in assessing (Q)SARs for different regulatory purposes.

This experience has highlighted that not all predictions produced by a valid model are acceptable for all regulatory purposes. When a (Q)SAR prediction or a result generated from multiple predictions is used for a given regulatory purpose, it needs to be verified in the context of the specific application. While there was agreement on the principles for the assessment of models, there was a need to establish a set of commonly agreed principles for the regulatory assessment of (Q)SAR predictions, and results from multiple predictions.

In late 2020, a project led by the Istituto Superiore di Sanità (ISS) to develop a (Q)SAR Assessment Framework (QAF) was proposed to the OECD Working Party on Hazard Assessment (WPHA) and was added to the work plan in early 2021. The project includes the principles for assessing (Q)SAR models and predictions in the form of a checklist and criteria for evaluating the element in the checklist. In addition, the European Chemical Agency (ECHA) joined ISS as a project co-lead.

A request for the nomination of experts was sent to delegates of the Working Party on Hazard Assessment (WPHA) and the group of more than 40 experts was convened to provide input and review on the QAF. The QAF Expert Group met through a series of teleconferences in 2021 - 2023 and smaller subgroups contributed to the drafting. A face-to-face meeting of the QAF Expert Group was convened in November 2022 to help finalise the draft document.

With this document, the OECD QAF expert group has established OECD principles for the assessment of (Q)SAR predictions and results based on multiple predictions and agreed on checklists to perform the assessment of models, predictions, and results from multiple predictions in practice.

This document is published under the responsibility of the Chemicals and Biotechnology Committee of the OECD.

Table of contents

Foreword	6
Executive summary	9
Visual Abstracts	11
1 Assessment of (Q)SAR Models (Model Checklist)	13
1.1 Defined endpoint	13
1.2 Unambiguous algorithm	14
1.3 A defined domain of applicability	15
1.4 Appropriate measures of goodness-of-fit, robustness and predictivity	16
1.5 Mechanistic interpretation	17
1.6 Outcome of the assessment of the model	17
1.7 Conclusions on the assessment of the model	17
2 Assessment of (Q)SAR Predictions (Prediction Checklist)	19
2.1 Correct input(s) to the model	19
2.2 Substance within the applicability domain	20
2.3 Reliability of the prediction(s)	21
2.4 Outcome fit for the regulatory purpose	24
2.5. Conclusion on the assessment of an individual prediction	25
3 Assessment of a (Q)SAR Result derived from multiple predictions (Result Checklist)	27
3.1 When to use the Result Checklist	27
3.2 Uncertainty and outcome of the (Q)SAR Result	28
4 Final considerations	29
Annex I –(Q)SAR model reporting format (QMRF) v.2.1	30
Annex II – (Q)SAR prediction reporting format (QPRF) v.2.0	31
Glossary of selected terms	32

FIGURES

Figure 1. (Q)SAR Assessment Framework (QAF) Result based on an individual prediction	11
Figure 2. (Q)SAR Assessment Framework (QAF) Result based on multiple predictions	11

Executive summary

The aim of the (Quantitative) Structure-Activity Relationship ((Q)SAR) Assessment Framework (QAF) is to develop a systematic and harmonised framework for the regulatory assessment of (Q)SAR models, predictions, and results based on multiple predictions. The proposed assessment is meant to be applicable irrespective of the modelling technique used to build the model, the predicted endpoint, and the intended regulatory purpose. The primary audience of this document is regulatory authorities and their stakeholders. In addition, any other (Q)SAR user is encouraged to refer to the QAF when using (Q)SARs for regulatory purposes.

The assessment of (Q)SARs for regulatory purposes should not be limited to checking the validity of the model used because even a valid model can produce unacceptable predictions in certain conditions. Therefore, individual predictions and results from multiple predictions need dedicated assessments. To this purpose, the QAF is based on the OECD principles for the model validation (OECD, 2007¹, referred as OECD (Q)SAR Model Principles in the rest of the document), and newly defines the principles for the assessment of (Q)SAR predictions and results based on multiple predictions. Four principles have been established (referred as OECD (Q)SAR Prediction Principles in the rest of the document) related to 1) the correct input, 2) the fit of the substance within the applicability domain of the model, 3) the reliability of the prediction, and 4) the outcome's fitness for the identified purpose. To streamline the evaluation, each principle has been subdivided in elements that should be considered in the assessment (Assessment Elements, AEs). AEs are included in three checklists (Model Checklist, Prediction Checklist, and Result Checklist, provided as a separate document) that can be used to evaluate the acceptability of the use of (Q)SARs in practice. Each AE can either be fulfilled or not, not documented, or not applicable. The Checklists also provide further details and examples for each AE.

The Model Checklist consists of a list of AEs to evaluate a model according to the OECD (Q)SAR Model Principles (OECD, 2007). The Model Checklist should be used together with the other Checklists when assessing predictions and results based on multiple predictions. In this case, the use of an acceptable model can be considered the first step of the assessment. If the model is considered acceptable, then the assessment needs to further consider the other Checklists. When a model is considered not acceptable, then the assessment could be concluded without further considering predictions and results. Alternatively, the Model Checklist can be used as a standalone tool when e.g., (Q)SARs are used for screening databases without the possibility to assess predictions individually, or to keep a separate record for the assessment of a model that could be reused in future. The assessment of a model is specific for the regulatory purpose and should be repeated when assessing the same model for other purposes.

The Prediction and Result Checklists are used to evaluate individual (Q)SAR predictions and results based on multiple predictions, respectively. They consist of AEs based on the OECD (Q)SAR Prediction Principles, which have different weight depending on how critical they are for the assessment. Suggested default values for the weights are given in the Prediction and Result Checklists, but assessors can modify

¹OECD principles for the Validation, for Regulatory Purposes, of (Q)SAR Models, ENV/JM/MONO(2007)2 <https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf> (OECD, 2007)

the weights to fit their own regulatory framework or paradigm. Moreover, assessors can assign a semi-quantitative uncertainty value (low, medium, high) to each assessment element, following this guidance and the examples in the Checklists. Finally, the overall uncertainty of the prediction is determined by considering the uncertainty associated with each AE and its weight in the assessment. Based on the purpose of use and the level of uncertainty of the individual prediction, the assessor can conclude on the outcome of the assessment (i.e., if the prediction is acceptable for the intended purpose of use).

In case of results based on multiple predictions, in addition to the individual assessment of each prediction, the Result Checklist considers one additional AE to evaluate if the predictions are integrated correctly for the determination of the final result. The level of uncertainty of the final result is assigned by weighting the uncertainty of the individual predictions and the additional AE. Finally, also in this case the assessor concludes on the outcome of the assessment depending on the purpose of use of the result, which determines the level of acceptable uncertainty (i.e., if the result is acceptable for the intended purpose of use).

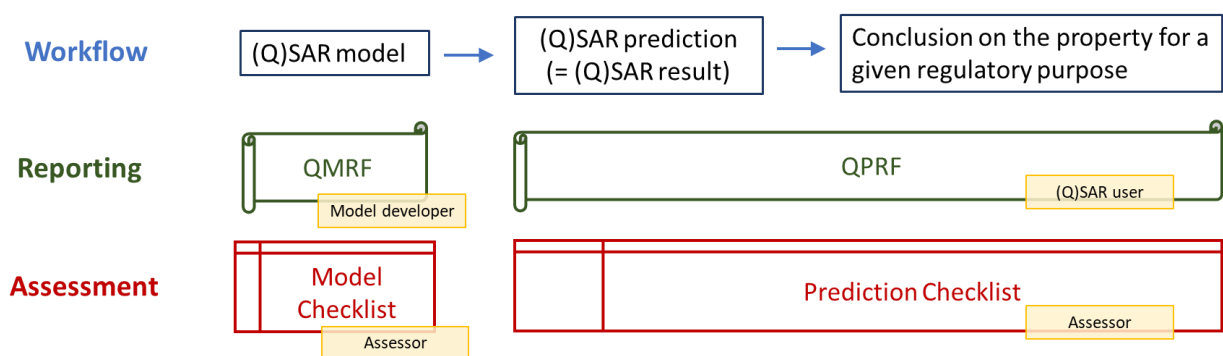
Complementing the QAF, updates for the (Q)SAR model reporting format (QMRF) and (Q)SAR prediction reporting format (QPRF) have been developed. While the update of the QMRF only concerns the description of the expected information in each field, without changes in the field names and order, the QPRF was more extensively updated, to reflect the newly established OECD (Q)SAR Prediction Principles. The updated QMRF and QPRF templates are provided as Annexes of this document.

Furthermore, examples that illustrate the use of the Checklists have also been developed and will be provided as separate documents. The expert group recommends the further usage and application of the QAF principles under OECD IATA case studies project for the case studies including (Q)SAR approaches.

Finally, some experts in the group have identified the need for further guidance on how to measure external predictivity of (Q)SAR models and a discussion was initiated towards the end of the QAF project. However, the topic was not elaborated in the final version of the QAF because there is no scientific consensus on how to measure external predictivity, and thus, this topic would require extensive review of the literature which is outside of the scope of this project. Further, model validation is a task for model developers and as such beyond the scope of the QAF, which aims primarily at providing instructions for regulators assessing pre-existing models.

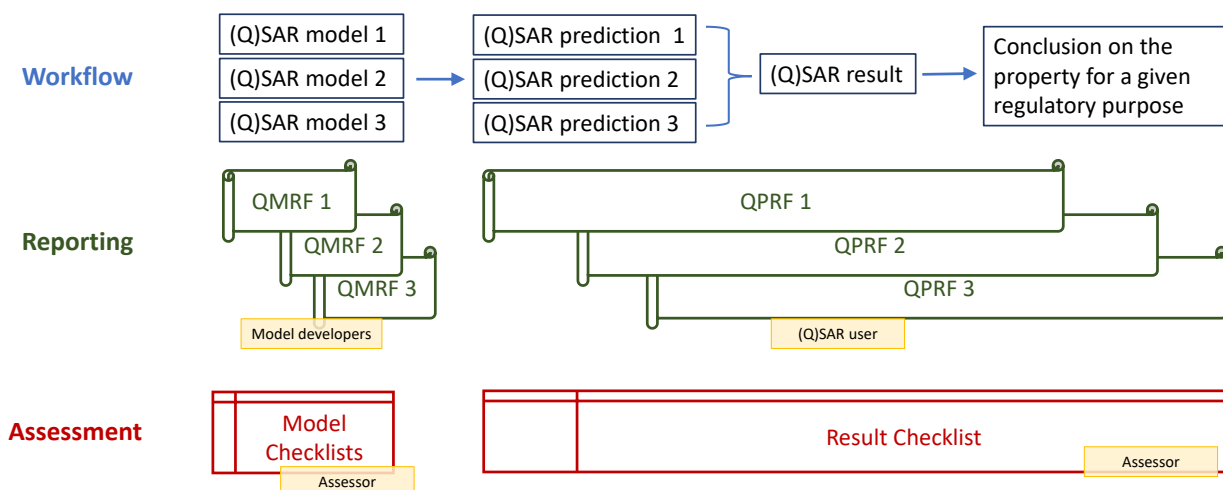
Visual Abstracts

Figure 1. (Q)SAR Assessment Framework (QAF) Result based on an individual prediction



Note: Workflow of (Q)SAR information for a result based on an individual prediction according to the OECD (Q)SAR assessment framework (QAF). The information about the model is reported in the (Q)SAR Model Reporting Format (QMRF) document prepared by the model developer and assessed by regulators using the QAF Model Checklist. The information about the (Q)SAR prediction is reported in the (Q)SAR Prediction Reporting Format (QPRF) document by the (Q)SAR user and assessed by regulators using the QAF Prediction Checklist. Checklists could also be pre-compiled by the (Q)SAR user to facilitate the work of the assessor.

Figure 2. (Q)SAR Assessment Framework (QAF) Result based on multiple predictions



Note: Workflow of (Q)SAR information for a result based on multiple predictions according to the OECD (Q)SAR assessment framework (QAF). The information about the models is reported in the (Q)SAR Model Reporting Format (QMRF) documents prepared by the model developers and assessed by regulators using the QAF Model Checklist. The information about the (Q)SAR predictions and result is reported in the (Q)SAR Prediction Reporting Format (QPRF) document by the (Q)SAR user, and assessed by regulators using the QAF Result Checklist.

Note: the QPRF is designed to report individual predictions and has limited fields to describe the integration of multiple predictions into an individual result. For this reason, the publication of an additional template to report (Q)SAR results based on multiple predictions is under consideration.

1 Assessment of (Q)SAR Models (Model Checklist)

This section of the document provides practical advice for completing the Model Checklist. It is based on and complements the more theoretical OECD guidance on model validation (OECD, 2007).

The OECD Principles for (Q)SAR validation state that “to facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

1. a defined endpoint
2. an unambiguous algorithm
3. a defined domain of applicability
4. appropriate measures of goodness-of-fit, robustness and predictivity
5. a mechanistic interpretation, if possible.”

The following chapters and the Model Checklist provide more details on these principles and how to verify that a (Q)SAR model is compliant with them. The assessment should be based on the information provided in the QSAR Model Reporting Format (QMRF). A map of the AEs in the Model Checklist to the QMRF fields is also provided in the Model Checklist to facilitate the retrieval of the information relevant for the assessment. Each OECD (Q)SAR Validation Principle is also further considered in one or more AEs of the Prediction and Result Checklists, as indicated at the end of each sub-chapter below.

1.1 Defined endpoint²

According to OECD (Q)SAR Model Principle 1 (OECD, 2007), a (Q)SAR should be associated with a “defined endpoint”, where endpoint refers to any physicochemical, biological, or environmental property that can be measured and therefore modelled. The intent of this principle is to ensure transparency in the endpoint being predicted by a given model, since an endpoint could be determined by different experimental protocols and under different experimental conditions.

The Model Checklist includes the following AEs to verify that the endpoint is clearly defined:

1. Clear scientific and regulatory purposes
2. Transparency of the underlying experimental data
3. Quality of the underlying experimental data

² As explained in the glossary, this document uses the term “property” to refer to the endpoint predicted by the model. However, the original term “endpoint” has been preserved in the title and first paragraph of this chapter to avoid confusion for the readers familiar with “OECD (Q)SAR Validation Principle 1: Defined Endpoint”

Clear scientific and regulatory purposes (AE 1.1 in the Model Checklist)

To have a clear scientific purpose, the predicted property has to be precisely described. To have a clear regulatory purpose, a model should address a specific regulatory requirement, which is often associated with a specific test method or test guideline, or it should provide supporting information to such requirement (e.g., mechanistic information). The description of the predicted property should be as detailed as possible by including all elements that have been considered (e.g., the unit of measurement, timescale, observations such as growth, mortality, etc.). The complexity of the predicted property influences the extent of documentation required (i.e., models predicting more complex properties such as developmental toxicity require more details in the definition of the property compared to models predicting simpler properties such as *in vitro* mutagenicity in Ames test).

Transparency of the underlying experimental data (AE 1.2 in the Model Checklist)

This AE concerns the transparency of the underlying experimental data and of the related data selection and curation procedure. The sources of the experimental data should be adequately reported, as well as information on experimental data selection criteria, data processing and information on chemical identifiers (including at least one identifier that codifies the chemical structure, such as InChi/InChIKey or (canonical) SMILES, and other commonly reported information such as CAS registry numbers) of tested substances. Potential biases in the data selection should also be investigated (e.g., systematic inclusion in the training set of data measured according to test guidelines not related to the predicted endpoint). The original studies (or an accessible reference) represent the highest level of transparency, but they are rarely available. On the contrary, the underlying studies may not be available at all for some models due to confidentiality or insufficient documentation. For many existing (Q)SAR models, the level of transparency is between these two extremes, with some but not all details available for the experimental studies used to build the models.

Authorities responsible for the assessment can decide the minimum acceptable level of transparency needed for specific purposes, with the understanding that for some models the available information might be limited for e.g., commercial reasons. In general, there should be sufficient information on the underlying data or on the data curation procedure to be able to assess data quality.

Quality of the underlying experimental data (AE 1.3 in the Model Checklist)

The (Q)SAR model should be built on data of sufficient quality. However, the individual assessment of the quality of each data point is often not feasible. In these cases, the quality of the underlying data can be assessed based on the description of the data curation procedure. For instance, assessors can verify how the relevant experimental parameters (e.g., sex, species, temperature, exposure period, protocol) that could affect the results of experimental studies have been considered when selecting data to build the model. Assessors may also consider whether all data points applied to develop and validate a model are generated by use of 1) the same assay protocol; and 2) the most updated assay protocol – and what are the consequences for the reliability. The quality of individual data should also be assessed to the extent possible.

OECD (Q)SAR Validation Principle 1 is further considered in the Prediction and Result Checklists under the element "Correspondence between predicted property and property required by the regulation".

1.2 Unambiguous algorithm

According to OECD (Q)SAR Model Principle 2 (OECD, 2007), a (Q)SAR model should be expressed in the form of an unambiguous algorithm (intended as unambiguous description of the algorithm). The intent

of this principle is to ensure transparency in the description of the model algorithm to allow an independent reproducibility of its predictions.

The Model Checklist includes the following AEs to verify the principle of an unambiguous algorithm:

1. Description of the algorithm and/or software
2. Inputs and other options
3. Model accessibility

Description of the algorithm and/or software (AE 2.1 in the Model Checklist)

The first element to be checked is the availability of a transparent description of the algorithm. The model equation, if applicable, including all descriptors and approach used for their selection, should be detailed. Furthermore, if applicable, a list of fragments/structure alerts (e.g., active, inactive, masks) and their description should be provided. The rationale that guided their identification could also be included. Calculated descriptors should be denoted with the software name and version used for their calculation. Furthermore, the version, developers' contact information and any available description of the software for the (Q)SAR model should also be provided. When an exact description of the algorithm is not publicly available (e.g., for commercial models), any available relevant information should still be assessed.

Inputs and other options (AE 2.2 in the Model Checklist)

Secondly, assessors should check if the documentation includes a description of inputs and settings of the model software. The allowed (or preferred) input formats for the chemical structure and its descriptors, including applicable pre-processing procedures (e.g., for salts and tautomers) should be documented. Further, customisable options/settings on the software should be reported and explained. Unless justified otherwise, the recommended input formats and options are expected to be the same as those used by model developers when developing the model and assessing its performance.

Model accessibility (AE 2.3 in the Model Checklist)

Finally, it should be checked if the model version under assessment is publicly accessible. A working link to access or download the model is expected in the QMRF documentation. When assessors have access to a different version of the model under assessment (e.g. a newer version), any differences in the outputs should be investigated.

OECD (Q)SAR Model Principle 2 is further considered in the Prediction and Result Checklists under the element "Reproducibility". Note that when the model is implemented in a software program that is accessible to the assessor, the reproducibility of the results should be possible even for cases when the description of the algorithm is not fully disclosed. Assessors may decide that this is acceptable for some regulatory purposes.

1.3 A defined domain of applicability

According to OECD (Q)SAR Model Principle 3 (OECD, 2007), a valid model is associated with a defined applicability domain (AD). The AD of a (Q)SAR model, as described in the Guidance Document (OECD, 2007), is *the response and chemical structure space in which the model makes predictions with a given reliability*. Elaborating on the AD definition given above, the AD should therefore consider the parametric, structural, mechanistic, metabolic and response space of the model. Nevertheless, the QAF does not prescribe a specific way to define the AD of a model because multiple valid methodologies can be used.

These are described in the Guidance Document (OECD, 2007), which can be consulted for further scientific aspects concerning the AD, while this paragraph focuses on practical aspects of the assessment within the QAF.

The Model Checklist includes one AE related to the applicability domain:

- Clear definition of the applicability domain and limitations of the model

Clear definition of the applicability domain and limitations of the model (AE 3.1 in the Model Checklist)

Assessors should verify that the definition proposed by model developers is sufficiently detailed to allow the assessment of how a given substance relates to the AD of the model in the Prediction and Result Checklists (e.g. is the substance within the AD of the model etc.?). To facilitate the assessment, developers of new models are encouraged to implement functionalities to automatically include in the prediction report information on how the input substance relates to the AD. In addition, the Prediction and Result Checklists include a separate reliability assessment that considers crucial aspects that may influence the reliability of the result, irrespective of whether they are included in the AD definition, as described in Chapter 2.3 below.

The model documentation may include a list of limitations of the model in addition to the AD definition, such as substance classes for which the use of the model is not recommended. The QAF is not prescriptive in terms of how to define such limitations, but it should be verified that the description of potential additional limitations is sufficiently detailed for assessing how a given input substance relates to them.

Principle 3 is further considered in the Prediction and Result Checklists under the elements "Substance within the applicability domain" and "Any other limitation of the model is considered".

1.4 Appropriate measures of goodness-of-fit, robustness and predictivity

According to Principle 4 (OECD, 2007), a (Q)SAR should be associated with "appropriate measures of goodness-of-fit, robustness and predictivity." This principle expresses the need to provide information on the goodness-of-fit and robustness of a model (as determined by internal validation) and the predictivity of a model (as determined by external validation). The AD and the performance of a model are connected. The performance should be measured within the applicability domain defined by its developers. An enhanced model performance can generally be obtained with a narrower AD. The Guidance Document (OECD, 2007) can be consulted for further scientific aspects concerning Principle 4.

The Model Checklist includes the following AEs to verify the appropriateness of measures of goodness-of-fit, robustness and predictivity of the model:

- Goodness-of-fit, robustness
- Predictivity

Goodness-of-fit, robustness and predictivity (AEs 4.1 and 4.2 in the Model Checklist)

Assessors should verify that the information available on the model's internal and external performance can be used as an indication of the expected accuracy of the model when predicting new substances. Size of the training, test and external sets, statistical methods and metrics used, values of the statistical metrics, and transparency of the procedure to measure performance are some of the important aspects to consider when assessing the performance of the model. Training, test, and external sets should be independent. The information describing how the QSAR model under scrutiny was selected, and its predictive performance estimated, has to be assessed in order to check if the model was correctly trained without accounting for the information to be used to estimate the model's external predictivity. For instance, the

use of information from an external test set of chemicals (originally devised for the exclusive assessment of external predictivity) during the development of the model (i.e., “data leakage” from test set for model calibration or selection), is very likely to result into an overestimation of the model's external predictivity.

OECD (Q)SAR Model Principle 4 is further considered in the Prediction and Result Checklists under the element “Overall performance of the model”.

1.5 Mechanistic interpretation

According to Principle 5 (OECD, 2007), a (Q)SAR “should be associated with a mechanistic interpretation, if possible”. Statistical methods used to describe relationships between chemical structure and activity are not intended to replace other knowledge from chemistry and toxicology when such knowledge exists. Assessors may require that the model documentation includes considerations on how the rationale behind a (Q)SAR model is consistent with or accounts for the knowledge related to the predicted property (such as known Adverse Outcome Pathways, AOPs, relevant for the predicted property), namely a mechanistic interpretation. Toxicokinetic considerations are also part of the mechanistic interpretation, if relevant for the property of interest.

The Model Checklist includes the following AE related to mechanistic interpretation:

- Plausibility of the mechanistic interpretation

Plausibility of the mechanistic interpretation (AE 5.1 in the Model Checklist)

For fragment-/alert- based models, the mechanistic interpretation can be based on the explanation of the chemical reactivity or molecular interaction caused or inhibited by the fragments associated with the alerts. For equation-based models, the mechanistic interpretation can be based on the physicochemical interpretation of each descriptor and its association with a mode or mechanism of action. An indication whether the mechanistic basis of the model was determined *a priori* (e.g., by pre-selecting descriptors or fragments to fit a specific mechanism of action) or *a posteriori* (e.g., after the modelling, by interpretation of the final set of training structures and/or descriptors) is also expected.

OECD (Q)SAR Model Principle 5 is further considered in the Prediction and Result Checklists under the element “Mechanistic and/or metabolic considerations”.

1.6 Outcome of the assessment of the model

The assessment of each AE and of the model in general depends on (and is specific for) a given regulatory purpose.

A model should be considered as acceptable when the outcome of each AE is “fulfilled”, or, alternatively, when sufficient information about the model is available for assessing the AEs related to the model in the Prediction and Result Checklists. For some regulatory purposes and with a valid justification, models that do not fulfil all AEs can also be accepted.

1.7 Conclusions on the assessment of the model

The compilation of the Model Checklist is the first step in the assessment of predictions and results from multiple predictions. It also supports the evaluation of AEs related to the validity of the model in the Prediction and Result Checklists. Once completed, the Model Checklist can be reused, i.e. assessors do not need to re-evaluate the model each time that a new prediction is submitted. Furthermore, the Model

Checklist can serve as a standalone tool to assess models used for screening of databases for which predictions are not meant to be assessed individually.

Finally, the Model checklist includes a mapping between the Model Checklist and the QMRF. This mapping, together with the outcome of the assessment of the Model Checklist, when disclosed, can serve as feedback to model developers for further improvement of their models and related documentation.

2 Assessment of (Q)SAR Predictions (Prediction Checklist)

In order to assess (Q)SAR predictions and results from multiple predictions, principles for their regulatory assessment needed to be established, in addition to the (Q)SAR Model Principles. The QAF establishes four principles for the assessment of (Q)SAR predictions and results from multiple predictions for regulatory purposes:

1. the model input(s) should be correct;
2. the substance should be within the applicability domain of the model;
3. the prediction(s) should be reliable;
4. the outcome should be fit for the regulatory purpose.

The same principles are included in the Prediction and Result Checklists. This chapter and the Prediction and Result Checklists provide more details on these principles and how to verify that an individual prediction or a result from multiple predictions is compliant with them. The assessment should be based on the information provided in the QMRF and QPRF. This chapter describes the AEs common to the Prediction and Result Checklists, while the next chapter focuses on the additional elements to consider in the Result Checklist.

2.1 Correct input(s) to the model

(Q)SAR models require one or more inputs to generate a prediction. Depending on the model, the input may be limited to information on the structure of the substance or also include some of its descriptors. Some models also have customisable settings in the software.

The input is correct when it is clearly and completely described, is representative of the substance being analysed, and uses reliable parameters (e.g. values for descriptors to be used in the prediction). The documentation needed to ensure that an input is correct depends on the complexity of the model (or of the software implementing it) and of the substance under analysis. In general, the input should be prepared by carefully following the instructions of the model or software developers, if available.

The Prediction Checklist includes the following AEs to verify that the input is correct:

- Clear and complete description of the input and model settings
- Input representative of the substance under analysis
- Reliable input (parameters)

Clear and complete description of the input and model settings (AE 1.1 in the Prediction and Result Checklists)

The first element to check is the description of the input and ensure that it is unequivocal and complete. In the simplest case, the model takes information on the structure (e.g., SMILES) as the sole input and does not have other editable options accompanying the structural input. In this case, the description of the exact structural information and the model/software version that were used to obtain the prediction are sufficient. For more complex cases, the requirement is to provide all information, including three-dimensional information on the chemical structure, customisable options (“settings”) and parameters of the software application (e.g., manual input of values of the descriptors and their source) that are needed as input to the model.

Input representative of the substance under analysis (AE 1.2 in the Prediction and Result Checklists)

Secondly, it is important to check that the input is representative of the substance under analysis and thus relevant for its assessment. When the substance consists of a single well-defined constituent, checking the agreement between the substance name, structure and numerical identifiers is sufficient. For three-dimensional models, information on the rationale for the selection of the conformation used as input is expected. For substances with complex compositions, a (Q)SAR result can be derived from multiple predictions that cover the constituents and impurities. In fact, one of the advantages of (Q)SARs is that more constituents and metabolites can be predicted to investigate their contribution to the overall toxicity of the substance with limited additional costs.

In addition, some models may require that inputs undergo structural curation before they can be used for a prediction. This is often the case for e.g., salts, ionisable structures, or structures subject to tautomerism. In these cases, different approaches exist. The choice of the approach should be decided on a case-by-case basis and special attention should be paid to how the pre-processing was performed by the model developers for the training set substances, and recommendations of the regulatory framework of interest, if relevant.

Reliable input (parameters) (AE 1.3 in the Prediction and Result Checklists)

Finally, for models that utilise direct input beyond the chemical structure, such as a physicochemical descriptor(s), the source of that descriptor value, whether experimentally measured or itself predicted by a model, needs to be evaluated for reliability before it is used to predict another property. The same approach applied by model developers during model development and assessment of performance of the model should be applied, unless properly justified. In case the (Q)SAR model relies on many physicochemical descriptors, and it is unfeasible to evaluate the reliability of each input, the focus should be on the most influential descriptor(s).

2.2 Substance within the applicability domain

The second principle requires that the model is applicable to the substance under analysis. The assessment of the applicability of the model to the substance relies on the verification of how the substance under analysis relates to the applicability domain (AD) and to the limitations of the model as defined by the model developers. If there are aspects potentially influencing the reliability of the prediction that have not been considered by the model developers when defining the applicability domain, these can be evaluated when assessing the next principle “Reliable prediction(s)”. Applicability domain informs the reliability of the prediction. In this document, AD and reliability are evaluated separately to streamline the assessment procedure.

The Prediction Checklist includes the following AEs to verify that the substance is within the applicability domain of the model:

- Substance within the applicability domain
- Any other limitation of the model is considered

Substance within applicability domain, and any other limitation is considered (AEs 2.1 and 2.2 in the Prediction and Result Checklists)

In general, the AD definition should include criteria for assessing whether the substance under analysis falls within the AD. In some cases, model developers also provide a description of the known limitations of the model and of cases when the model should not be used (e.g., a list of chemical classes or descriptor ranges for which the model is known to make inaccurate predictions). Such descriptions may be provided separately from the given definition of AD but are of similar importance. Many recent software tools provide an automatic assessment of the applicability of the model to the substance along with the prediction result. In this case, it is important that the automatic assessment can be verified independently.

In the preferable scenario, the substance falls within the applicability domain of the model and model specific limitations do not apply to the substance.

On the contrary, in some cases it is clear that the substance is outside the AD of the model, or model specific limitations apply. In these cases, the prediction should not be considered further, (unless a valid justification is provided, e.g., it is not technically possible to perform an experimental test, but a numerical value is still needed). An example is a model for which the applicability domain is defined as range of physicochemical properties, and in addition the description of the model clarifies that the model has been developed to predict organic chemicals within certain chemical classes. In this case, predictions of organic substances outside these classes should not be considered further, irrespective of whether the physicochemical values are within AD.

Between these two clear-cut scenarios there are intermediate cases. An example of such cases are substances for which their relation to the applicability domain cannot be unequivocally established but at the same time do not meet the exclusion criteria for which the model should not be used. For these substances the accuracy of the model prediction may be unknown but not necessarily unacceptable for all regulatory purposes. Another example are predictions for chemicals that are located near the boundaries of the applicability domain and therefore may be associated with higher uncertainty. For these cases, the assessment of the reliability of the prediction, as described in the next section, is decisive for its overall acceptance for a given regulatory purpose.

2.3 Reliability of the prediction(s)

In addition to the AD, several other aspects can be considered when assessing the reliability of a prediction. If these aspects are already included in the AD definition, the assessment does not need to be repeated.

The Prediction Checklist includes the following AEs to verify that a prediction is reliable:

- Reproducibility
- Overall performance of the model
- Fit within the physicochemical, structural and response spaces of the training set of the model
- Performance of the model for similar substances
- Mechanistic and/or metabolic considerations
- Consistency of information

Reproducibility (AE 3.1 in the Prediction and Result Checklists)

First, it should be verified that it is possible to reproduce the prediction (i.e., to obtain the same result) using the documented input and model. When assessors have access to the same model and version used to generate the prediction, they can repeat the prediction to confirm that the same prediction is obtained. If a different prediction is obtained without a valid explanation (e.g. different model version), is symptomatic of lower reliability.

Overall performance of the model (AE 3.2 in the Prediction and Result Checklists)

Second, the overall model performance should be considered. It represents the baseline for the reliability assessment and is expected in the model documentation (c.f. QMRF). As an example, a similar standard error for models predicting continuous outcomes (or the accuracy of models predicting categorical outcomes) can be expected for the prediction of the substance under analysis. For this reason, a prediction generated by a model with better performance should be considered more reliable than a prediction generated by a model with lower performance.

Fit within the physicochemical, structural and response spaces of the training set of the model (AE 3.3 in the Prediction and Result Checklists)

The substances in the training set have defined values for the descriptors (such as physicochemical descriptors, molecular descriptors, etc.) and for the property of interest (response values). These values can be used to define descriptor and response spaces based on ranges or distributions. The assessors shall compare descriptor and response values of the substance under analysis with the ranges defined by the training set.

Likewise, the substances in the training set can be used to define a structural space in terms of functional groups or structural fragments. When all structural characteristics of the substance are known by the model, the prediction is considered more reliable. Structural characteristics not known by the model may have an impact on the reliability of the prediction. This aspect is of particular importance for predictions indicating a lack of biological activity by fragment/alert based models, where the lack of effect may be due to lack of knowledge by the model, which might have not been trained to predict the effect of certain fragments.

The reliability of the prediction is lower when the substance under analysis falls in regions of the training set spaces that are scarcely populated or associated with lower model performance. Special attention should be paid in such cases.

Performance of the model for similar substances (AE 3.4 in the Prediction and Result Checklists)

An assessment of local performance of the model requires the identification of similar substances with reliable experimental data for the property of interest. If not executed automatically by the software, this work should be performed by (Q)SAR users, and then verified by the assessors. Similar substances can be defined by considering multiple aspects, such as structural, physicochemical, and/or mechanistic similarity. Similar substances may be identified manually by the users using expert judgment, with or without the help of computational tools for analogue identification such as the OECD QSAR Toolbox³. Some (Q)SAR models automatically provide with the prediction a list of substances similar to the input, their experimental values, and the accuracy of their predictions.

³ More information about this software, which is freely available and co-developed by OECD and the European Chemicals Agency (ECHA), is available at www.qsartoolbox.org

However, molecular similarity indices alone are not necessarily sufficient to define similarity, as some molecular fragments may be very important for one property but not for another property, and in some cases co-occurrence of multiple fragments can be very important (e.g. cross-linking agents).

If reliable experimental data are available for substances similar to the one under analysis, the performance of the model when predicting these can be used to better characterize the reliability of the prediction. The support is higher when the similar substances used for this assessment are not part of the training set of the model, as this case mimics more objectively the real-life application of the model when predicting new substances. The more similar the substances are to the substance under analysis, especially in terms of the descriptors, fragments, or other properties most relevant to the prediction, the more informative is the accuracy of their predictions to estimate the accuracy for the prediction of substance under analysis. In some cases, the reliability improvement is applicable even when the substance under analysis or similar substances are formally outside the AD of the model. The substance under analysis may be out of AD due to lack of experimental data in its chemical space at the time of model development. If new data for similar substances shows good performance of the model in that chemical space, then there is a concomitant improvement in the reliability of the prediction of the substance under analysis.

For fragment/alert- based models, the performance of the identified positive/negative alerts (expressed as, e.g., Cooper statistics (OECD, 2007)) can also provide information on the local performance of the model.

In absence of data for similar substances or when their identification is not possible or feasible, then the outcome for this AE should be marked “not applicable/assessed”.

Mechanistic and/or metabolic considerations (AE 3.5 in the Prediction and Result Checklists)

This assessment element covers toxicokinetic and toxicodynamic considerations relevant for the property under analysis. A discussion on how considerations on mechanistic aspects (such as Adverse Outcome Pathways, AOPs) and ADME³ properties (such as (bio)transformations, i.e., metabolism and other biotic or abiotic transformations such as hydrolysis, autooxidation, and photolysis) fit the prediction can contribute to the reliability assessment and should therefore be provided when possible. If the structure is known or predicted to have a certain mechanism of action relevant for the property under analysis, or the formation of certain metabolites or other transformation products is expected, then their relation to the prediction and how the model takes them into account are expected to be described. One example is the prediction of a property for which modes of action are known, such as skin sensitisation due to protein binding. If the model provides a “non-sensitiser” prediction, but the substance under analysis is known or predicted to bind to proteins, then additional explanations are necessary to support the reliability of the negative prediction. Similarly, if the substance under analysis is known to undergo metabolism leading to the formation of hazardous metabolites, a prediction of lack of hazard would need additional justification to discuss how the model prediction considers the effects of metabolism and its products.

Consistency of information (AE 3.6 in the Prediction and Result Checklists)

Finally, it is often the case that predictions from more than one model are used to predict the property of interest, and/or additional information (measured or calculated values) for the same or related property is available. The consistency between (Q)SAR predictions and/or with other reliable information, if available, needs to be considered. Predictions consistent (i.e., in agreement) with each other are considered more reliable when generated by independent models (in terms of training sets, modelling techniques and/or descriptors/alerts used). Contradicting information, in the absence of some explanation, tends to decrease the reliability of the prediction. The Result Checklist should be used when a (Q)SAR result is based on

³ Absorption, Distribution, Metabolism, Elimination

multiple predictions (see Chapter 3). Further weight of evidence considerations are out of the scope of this document.

The reliability assessment is anticipated to be the most complex part of the assessment. It requires an in-depth analysis and expert judgment, even when facilitated by the comprehensive reports of modern (Q)SAR software. However, when a prediction-specific reliability assessment is thoroughly performed, it may offer more insight than the more generic information on the applicability domain of the model.

2.4 Outcome fit for the regulatory purpose

This principle describes the assessment of the usefulness of the (Q)SAR prediction to answer a specific regulatory question. Even if very reliable, a prediction or a result from multiple predictions can be used for a specific regulatory purpose only if it is fit for that purpose. The elements to consider when evaluating the fitness for purpose depend on the regulatory framework. The final decision on the fitness for the purpose (and on the acceptability in general) lies with the authority responsible for the regulatory framework.

The Prediction Checklist includes the following AEs to verify that an outcome is fit for the regulatory purpose:

- Compliance with additional requirements
- Correspondence between predicted property and property required by the regulation
- Decidability within the specific framework

Compliance with additional requirements (AE 4.1 in the Prediction and Result Checklists)

A regulatory framework may directly indicate specific criteria for the acceptable use of (Q)SARs. If the criteria include elements not covered in this document, then the compliance of the prediction with these additional criteria is expected for considering the use as acceptable. As an example, in case of regulations that address substances with complex compositions or mixtures, a crucial element to assess is the consideration of the whole composition of the substance or components of the mixture following the criteria specified by the regulation (e.g. by taking into account all components present at a concentration above a given threshold, using a concentration or dose addition approach or selecting a reasonable worst case). In some cases, the consideration of potential antagonistic and synergic effects of the different components may also be required. Another example are regulations that require the use in combination of two different types of models to produce a result (such as the ICH guideline M7 assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk⁴, which requires the use of one expert rule-based and one statistical-based model). In this case, the use of an individual prediction will not be fit for purpose.

Correspondence between predicted property and property required by the regulation (AE 4.2 in the Prediction and Result Checklists)

It is important that the property predicted by the (Q)SAR model matches the property required by the regulation. The property predicted by a (Q)SAR model depends on the experimental data applied as training set and the data curation performed on the data prior to (Q)SAR modelling (e.g. inclusion/exclusion criteria, transparency of manual evaluations and removal of outliers, thresholds applied for defining positives and negatives, etc.). As an example, specific bacterial strains and presence of metabolic activation may need to be explicitly considered by the model if required by the regulation when evaluating

⁴ https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-m7r1-assessment-control-dna-reactive-mutagenic-impurities-pharmaceuticals-limit_en.pdf

in vitro mutagenicity in a bacterial reverse mutation assay. If the regulation refers to a specific test guideline, the model should include the experimental results obtained following the specified test guideline in its training set. However, this may not always be necessary, e.g. models for Ames mutagenicity may include historical data not performed by use of all currently required strains, but positive predictions from the model may still be adequate. Another example is degradation half-life, where predictions from models trained with data on primary degradation half-life are not fit for the purpose of assessing ultimate degradation half-life.

Decidability within the specific framework (AE 4.3 in the Prediction and Result Checklists)

A prediction is fit for purpose when it allows to take a regulatory decision in the framework of use. In presence of a regulatory threshold, the prediction will be decidable when it is comparable to the threshold (i.e., expressed in the same or convertible unit) and provides a sufficient level of confidence that the prediction lies on one side of the threshold or is close enough that a concern cannot be ruled out. As an example, if a model predicts a water solubility of 0.09 mg/L with high uncertainty and the regulatory decision depends on whether the substance has a solubility above or below 0.10 mg/L, then the prediction will not be decidable. The level of confidence required, as well as the way to establish it, will depend on the intended regulatory purpose.

2.5. Conclusion on the assessment of an individual prediction

The conclusion of the Prediction Checklist consists of an uncertainty value for the assessment and an outcome based on this uncertainty.

2.5.1. Uncertainty

When using the Prediction Checklist, assessors need to decide if each AE is fulfilled or not. In most cases, this decision will include a certain level of uncertainty⁴.

The uncertainty of each applicable AE can be described using semi-quantitative values (low, medium, or high), while AEs that are not applicable/assessed are not given an uncertainty score. For AEs that are fulfilled, the explanation of how to assign the uncertainty level for each AE is provided in a separate table of the Prediction Checklist. In addition, a default high uncertainty is assigned to AEs that are not fulfilled or not documented, unless a valid justification is provided. Then, the uncertainty associated with each AE can be used to decide on the overall uncertainty of the prediction.

The overall Prediction Checklist uncertainty is considered “Low” when:

- the prediction fulfils all elements with low uncertainty, OR
- most elements have low uncertainty and the elements with greater uncertainty (including elements not documented or not fulfilled, if any) are not considered of high weight in the overall assessment, and a convincing justification of this consideration is provided.

The overall Prediction Checklist uncertainty is considered “Medium” when:

- the prediction fulfils all or most elements with medium uncertainty. Elements with greater uncertainty (if any) are not considered of high weight in the overall assessment, OR

⁴ The uncertainty refers to the assessment itself, irrespective of whether the prediction is expressed as a point value, a range, or a probability estimate. For a definition of uncertainty, see the Glossary.

- the prediction fulfils most elements with low uncertainty, but some elements of high weight in the overall assessment have medium uncertainty.

The overall Prediction Checklist uncertainty is considered “High” when:

- the prediction fulfils all or most elements with high uncertainty, OR
- the prediction fulfils most elements with low or medium uncertainty, but some elements of high weight in the overall assessment have high uncertainty.

2.5.2. Outcome

The outcome of the assessment of the prediction is based on an integration of the outcome and uncertainty of the AEs in the Prediction Checklist. In general, AEs with high weight are the most critical for the assessment, and an acceptable prediction should fulfil all of them with low or medium uncertainty. AEs with lower weight are also important, but predictions that fulfil some of them with high uncertainty or do not fulfil them can also be acceptable for some applications. When applying the logic described in Chapter 2.5.1., this approach corresponds to considering predictions with Low or Medium uncertainty to be acceptable.

This document provides general advice on the assessment and suggests the “weight” that can be used for each AE, but each authority could establish (and communicate) different requirements for acceptable results for their applications.

For an individual prediction, the assessment is concluded at this stage.

3

Assessment of a (Q)SAR Result derived from multiple predictions (Result Checklist)

3.1 When to use the Result Checklist

The Result Checklist should be used when assessing a result derived from multiple predictions for the same or related properties. Cases that consider multiple predictions include:

- a. Predictions from different models for the same structure;
- b. Predictions from the same models for different structures (such as the multiple constituents of a substance or for the substance under analysis and its metabolites or transformation products);
- c. A combination of the above.

First, each prediction needs to be evaluated using the individual prediction checklists within the Result Checklist. For complex cases (point c), it is advised to start by addressing multiple predictions associated with the same structure, and then consider the predictions for different structures.

A checklist for each prediction (within the Result Checklist) needs to be completed. The assessment of each prediction can be carried out independently, except for the AEs listed below. For these AEs, the outcome needs to consider the information from all predictions:

- Input representative of the substance under analysis;
- Mechanistic and/or metabolic considerations;
- Consistency of information;
- All AEs referring to the principle “Outcome is fit for the regulatory purpose”.

This situation can be exemplified with a case where two constituents of the same substance are predicted individually. Two checklists should be compiled within the Result Checklist, one for each prediction. When evaluating the AE on the representativeness of the input, both constituents shall be considered, and the same outcome can be recorded in the checklists for each prediction within the Result Checklist.

Then, there is one additional AE to consider when evaluating a result from multiple predictions:

- Correct determination of the final result from individual predictions

Assessors should verify that the final result, derived from the individual predicted values, has been correctly determined. Depending on the type of endpoint and regulatory requirement, the final result can be determined by majority consensus, worst case, average value, or more complex techniques. In this calculation, different weights may be given to predictions depending on e.g., their reliability. In any case, assessors should verify that the determination is documented and adequately justified.

3.2 Uncertainty and outcome of the (Q)SAR Result

3.2.1. Uncertainty

Finally, there is an additional step aimed at weighing each prediction in order to reach a conclusion, i.e., to decide on the acceptability of the final result. The logic for the estimation of the uncertainty of the result derived from multiple predictions is as follows:

- When the final result is generated by integrating consistent (i.e., in agreement) predictions for the same structure from different independent models, then the uncertainty of the final result can be equal or lower than the uncertainty of the individual predictions. The uncertainty of a result from multiple predictions can be lower when the AEs with higher uncertainty are different for the predictions, and they can therefore support each other on their elements of uncertainty. This is not the case when higher uncertainty comes from the same AEs;
- When the final result is generated by integrating inconsistent predictions for the same structure from independent models, then the uncertainty of the final result will be equal or higher than the uncertainty of the individual predictions. Normally, results derived from inconsistent predictions will be acceptable only if one or more predictions in agreement outweigh the inconsistencies due to better fit to the applicability domain, higher reliability, or lower uncertainty. A justification on the rationale of the integration is expected from the (Q)SAR user, while the (Q)SAR assessor can comment on this aspect under the AE “Consistency of information”;
- When the final result depends on predictions for different structures (e.g., different constituents of the same substance, and/or a parent and its metabolites, etc.), then its uncertainty should take into account the uncertainty of the individual predictions and the composition of the substance under analysis. As an example, when predicting water solubility, the uncertainty of the prediction for the main constituent could have a higher influence on the uncertainty of the final result compared to the uncertainty of the prediction for a minor impurity. In this example, if the uncertainty of the prediction for the main constituent is “Medium”, and the uncertainty for the prediction of the minor impurity is “High”, then the uncertainty of the final result could be “Medium”. While if the uncertainty of the prediction for the main constituent is “High”, while for the minor impurity is “Medium”, then the uncertainty of the final result could be “High”. For other properties such as mutagenicity, where a mutagenic impurity can make the whole substance mutagenic, the concentration of a constituent or impurity in the composition will have a smaller role or no role in weighing the individual predictions to decide on the uncertainty of the overall result.

3.2.2. Outcome

The uncertainty of the result can then be used to determine the outcome of the assessment. Consistently with the approach for individual predictions described in 2.5.2, results with Low or Medium uncertainty could be considered as acceptable. The decision of the outcome for the final result completes the assessment.

4 Final considerations

This document aims at providing a comprehensive list of the elements needed for assessing (Q)SAR models, predictions, and results from multiple predictions. Depending on the specificities of the case and context of use, not all AEs may be applicable or required. Each regulatory authority should decide and communicate what elements need to be systematically considered for the assessment in a specific regulatory framework and context (e.g., the reliability requirements may be lower when the (Q)SARs are used for the purpose of screening a database compared to when they are used for assessing an individual substance). Thresholds to judge whether an AE is fulfilled may also be introduced by some authorities but are not included in this document for at least two reasons: 1. lack of generally agreed thresholds in the scientific community and 2. acceptable thresholds may vary depending on a multitude of factors including the property of interest and the regulatory purpose or context of use.

In general, the use of valid (Q)SAR models shall be expected. This can be verified using the Model Checklist.

A correct input should always be required, irrespective of the context of use. If the input is not correct, the (Q)SARs will not be useful for further regulatory considerations.

The relation of the substance with the AD of the model is also important. Ideally, the substance should clearly fall within the AD of the model. If this is not the case, other reliability aspects may still lead to the acceptance of the use of the prediction or result for a given purpose.

Finally, even if otherwise correct and reliable, a prediction or result can only be used if it is fit for purpose. Clear regulatory requirements for the use of (Q)SAR results facilitate the assessment of this principle.

When integrating multiple predictions into one result, a description of and a justification for the approach used to derive the overall result is essential.

This document includes a major update to the QPRF format. The need for a more comprehensive update of the QMRF has been identified too, and is expected to start after the publication of this document. Furthermore, the development of a new template to report results based on multiple predictions is under consideration.

The QAF Checklists offer more explanations and examples useful to perform the assessment in practice. After the publication of this document, more regulation-specific guidance documents or case studies may be prepared to clarify prediction- and regulation-specific requirements.

Annex I –(Q)SAR model reporting format (QMRF) v.2.1

Please refer to the separate document for full Annex I
ENV/CBC/MONO(2023)32/ANN1

An editable format is available at

<https://www.oecd.org/chemicalsafety/risk-assessment/qsar-assessment-framework-annex-1-qsar-model-reporting-format.docx>

Annex II – (Q)SAR prediction reporting format (QPRF) v.2.0

Please refer to the separate document for full Annex II
ENV/CBC/MONO(2023)32/ANN2

An editable format is available at

<https://www.oecd.org/chemicalsafety/risk-assessment/qsar-assessment-framework-annex-2-qsar-prediction-reporting-format.docx>

Glossary of selected terms

Assessment element (AE): a critical aspect to consider when assessing (Q)SAR models, predictions and overall results meet. AEs are associated with the OECD (Q)SAR principles for models and results.

Assessor: the person evaluating the acceptability of a model and/or prediction for the intended regulatory purpose.

Conclusion: the uncertainty and outcome of the assessment for an individual (Q)SAR prediction or a (Q)SAR result.

Model checklist: a separate document to facilitate the assessment of a (Q)SAR models according to QAF principles. It includes a list of assessment elements to consider, columns to record the outcome of the assessment, practical advice, and examples.

Outcome of the assessment: the decision if the (Q)SAR prediction or result is acceptable for the intended regulatory purpose.

Prediction checklist: a separate document to facilitate the assessment of a (Q)SAR prediction according to QAF principles. It includes a list of assessment elements to consider, columns to record the outcome of the assessment, practical advice, and examples.

Property: a physicochemical, toxicological, ecotoxicological, or fate property; chemical reactivity or biological interaction. In this document, the term “property” is preferred to “endpoint” because of the different understanding of the meaning of the term endpoint depending on the audience.

(Q)SAR model: a model that predicts the property of a substance using as input information on the structure.

(Q)SAR prediction: an individual output (i.e., the predicted value of a property) of a (Q)SAR model. It can be a continuous or a categorical (two or more categories) output.

(Q)SAR result: the assessment of a property of a substance based on multiple (Q)SAR predictions.

Regulatory framework: the specific guidelines, rules or mechanisms used to fulfil regulatory requirements.

Regulatory purpose/use/purpose: the specific application within a regulatory framework for which the (Q)SAR result is used.

Result checklist: a separate document to facilitate the assessment of a (Q)SAR result based on multiple predictions according to QAF principles. It includes a list of assessment elements to consider, columns to record the outcome of the assessment, practical advice, and examples.

Substance: the chemical substance under analysis. A substance can be formed by one or more constituents, and hence by associated to one or more structures.

Uncertainty: according to EFSA Guidance on Uncertainty Analysis in Scientific Assessments (2018)⁵ “a general term referring to all types of limitations in available knowledge that affect the range and probability

⁵ EFSA (European Food Safety Authority) Scientific Committee, Benford, D, Halldorsson, T, Jeger, MJ, Knutsen, HK, More, S, Naegeli, H, Noteborn, H, Ockleford, C, Ricci, A, Rychen, G, Schlatter, JR, Silano, V, Solecki, R, Turck, D,

of possible answers to an assessment question". In the result checklist, uncertainty can be assigned a low, medium, or high value and it can refer to individual assessment elements, (Q)SAR predictions, or (Q)SAR results.

Internal and external validation of a (Q)SAR model: the statistical procedure to evaluate the performance of the model, based on the use of a training and (independent) test set.

Weight of the assessment element: the importance of an assessment element in the overall assessment of a (Q)SAR prediction. It depends on the regulatory purpose and can have a low, medium, or high value.

Younes, M, Craig, P, Hart, A, Von Goetz, N, Koutsoumanis, K, Mortensen, A, Ossendorp, B, Martino, L, Merten, C, Mosbach-Schulz, O and Hardy, A, 2018. Guidance on Uncertainty Analysis in Scientific Assessments. EFSA Journal 2018;16(1):5123, 39 pp. <https://doi.org/10.2903/j.efsa.2018.5123>