Committee on MUTAGENICITY

Committee on Mutagenicity of Chemicals in Food, Consumer Products and the Environment (COM)

Statement GXX or 2020/XX

Guidance Statement on the use of QSAR models to predict genotoxicity

https://www.gov.uk/government/groups/committee-on-carcinogenicity-of-chemicals-in-food-consumer-products-and-the-environment-coc

COM Secretariat

c/o Public Health England

Centre for Radiation, Chemical and Environmental Hazards

Chilton, Didcot, Oxfordshire OX11 0RQ

Questions for the Committee:

- A draft statement was discussed in February 2019. The members asked for a more general statement including an evaluation of the OECD principles rather than an evaluation of specific QSAR models. Does this new draft address these issues?
- Are there any other published documents which should be considered?
- Does the statement reflect current views balanced with the short-term longevity of the recommendations?

Specific Questions

- In the previous statement emphasis was placed on freely available models - In February 2020 members commented that freely available models are not necessarily better models. Do member agree with the deletion in paragraph 3.
- Principle one relates to the evaluation of a specific endpoint, examples are given in paragraph 11. These examples were in the last draft - do members agree with these examples of defined endpoints? Or are there better examples?
- The term “knowledge-based” QSAR has been changed to “expert rule-based” - do members agree. This was suggested by experts in LHASA.

COM/2020/20

COMMITTEE ON MUTAGENICITY OF CHEMICALS IN FOOD, CONSUMER PRODUCTS AND THE ENVIRONMENT (COM)

Guidance statement on the use of QSAR models to predict genotoxicity

Introduction

1. A range of Quantitative Structure-Activity Relationship (QSAR) models have been developed to predict genotoxicity. The COM has previously agreed that where no genotoxicity data are available, the intrinsic chemical and toxicological properties of a chemical must be considered prior to developing a genotoxicity testing programme, as reported in "*Guidance On A Strategy For Genotoxicity Testing Of Chemical Substances*" (COM, 2011) and as updated in 2020 (REFERENCE). This guidance describes a staged approach to testing consisting of stages 0 (preliminary considerations including physico-chemical properties), 1 (*in vitro* genotoxicity tests) and 2 (*in vivo* genotoxicity tests).

2. QSARs are incorporated into Stage 0 of the COM guidance. Alternatives to animal testing and the usefulness of computational methods in the prediction of genotoxicity are areas of increasing research. QSAR models and their predictions currently cannot replace the need to undertake the *in vitro* and *in vivo* genotoxicity tests required to derive conclusions on mutagenic hazard except in specific regulatory settings.

3. This guidance statement will be updated periodically as the use of QSARs in regulatory frameworks evolves.

Assessment

~~1.~~4. Initial assessment of potential genotoxicity can be based on ~~publicly~~ available QSAR models. The statement presented here provides guidance on the use of such models.

~~2.~~5. It should be noted that data from a QSAR should not overrule test data from adequately designed and conducted genotoxicity tests.

3.6. QSAR models may be expert rule-based or statistical-based or a hybrid of the two approaches. Expert rule-based QSARs provide reasoning for predictions, such as a mechanism of action of a functional group, which are often supported with literature references and expert knowledge. However, the domain of applicability may not be clear and negative results may reflect insufficient knowledge of a mechanism of action within the database, rather than a lack of genotoxic activity for a chemical. Statistical-based QSARs use the statistical analysis of data to produce quantitative outputs. As such, they tend to have a higher accuracy of prediction than expert rule-based approaches. However, interpretation of the results is more difficult and there may not be a mechanistic rationale behind the predictions. Hybrid approaches combine the expert rule-based and statistical-based QSARs, for example, by identifying a mechanism of action with a statistical analysis of the data.

7. QSARs are predictive models, and as such are inherently uncertain. To compensate for this uncertainty, at least two QSAR models should be applied to predict the same endpoint for the same chemical in a weight-of-evidence approach. The models used should be a combination of expert rule-based and statistical-based approaches. For example, the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) M7 guideline "*Assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk*" provides a framework for assessing and controlling DNA reactive impurities in pharmaceutical products. In the absence of experimental data, the guideline requires the use of one expert rule-based and one statistical-based QSAR to predict bacterial mutagenicity. These QSARs are required to adhere to the Organization for Economic Cooperation and Development (OECD) principles for validating QSARs. Negative predictions with both QSAR models are sufficient to conclude that a pharmaceutical impurity is of no mutagenic concern. The guideline states that predictions should be reviewed with the use of expert knowledge which provides a rationale to support the conclusion.

8. The following QSAR models have been considered in comparison with OECD QSAR principles: Toxtree, TOPKAT, Derek Nexus, Danish QSAR Database, Sarah Nexus, Case Ultra, VEGA, OECD QSAR Toolbox, Leadscope Model Applier and ToxRead. The developers state that these models meet the OECD 5 principles but the user needs to evaluate the validity in relation to their data requirements. These models were previously reviewed in report MUT/2018/02 and allowed the members to reach their conclusions.

4.9. QSAR models for the Ames test are satisfactory but found to be less than reliable for other genotoxicity endpoints such as chromosomal aberrations. The application of QSARs is heavily reliant on expert judgement and even with significant advances in models and other computational methods this is still the case (EFSA 2019).

OECD QSAR principles

~~5.~~10. The OECD has published principles for validating QSARs:

- Principle 1 - A defined endpoint;
- Principle 2 - An unambiguous algorithm;
- Principle 3 - A defined domain of applicability;
- Principle 4 – An appropriate measure of goodness-of fit, robustness and predictivity; and
- Principle 5 - A mechanistic interpretation (if possible).

~~6.~~11. QSAR models are being developed and improved at a fast pace and the user needs to evaluate the reliability of the predictions in relation to their specific data requirements. The OECD QSAR principles are a good framework for this evaluation.

Principle 1 - A defined endpoint

~~7.~~12. The endpoint to be predicted by the QSAR should be fully documented by providing details on the specific effect within a specific organ/tissue under specific conditions, such as duration of exposure. (OECD, 2007). Therefore, the endpoint should be fully described within the QSAR. As an example, "*in vitro* cytogenicity study in mammalian cells or *in vitro* micronucleus study" is regarded as a regulatory endpoint under Annex VIII of the Registration, Evaluation, Authorisation and restriction of CHemicals (REACH) Regulations. However, as such a description could relate to several different assays, it cannot be regarded as a defined endpoint within the context of a valid QSAR. In contrast, "*in vitro* chromosomal aberration in Chinese hamster lung fibroblasts without S9" would be considered a fully defined endpoint. It may not always be possible to define endpoints to this level of detail using some QSAR models, as many cite an endpoint of "Ames mutagenicity", without defining the strain of bacteria or metabolic status. However, this would not necessarily indicate that a QSAR prediction is invalid as a prediction based on a dataset of studies conducted according to OECD 471 may provide useful predictions for bacterial mutagenicity, even if the specific strain is not clear. Therefore, expert judgement is required to determine a sufficient level of detail for an acceptable QSAR prediction.

Principle 2 - An unambiguous algorithm

~~8.~~13. The function of Principle 2 is to ensure that a QSAR model prediction is transparent and can be independently reproduced. However, such transparency may not be available in commercially developed QSAR models (OECD, 2007). In such cases, a prediction may be reproduced by another individual using the same commercial QSAR model, but they would not be able to explain the basis of the prediction.

Principle 3 - A defined domain of applicability

~~9.~~14. There will be limitations within QSAR models with regards to the types of chemical structures, physico-chemical properties and mechanisms of action for which a reliable prediction can be generated (OECD, 2007). These limitations represent the domain of applicability, and must be described to provide reassurance of the reliability of the prediction. There is typically a trade-off between constraining the domain of applicability of a QSAR and the applicability of that QSAR for use with multiple chemicals. The more constrained the domain of applicability, the fewer chemicals for which reliable predictions can be generated. The less constrained the domain of applicability, the wider the range of chemicals for which predictions can be generated, but the reliability of those predictions will decrease (OECD, 2007).

Principle 4 - Appropriate measure of goodness-of-fit, robustness and predictivity

~~10.~~15. Principle 4 is a set of principles by which the prediction is statistically measured to assess its reliability. “Measures of goodness-of-fit and robustness” test the internal performance of the QSAR model and “measures of predictivity” test the external performance of the QSAR model (OECD, 2007). These statistical measures should be considered in combination with the applicability domain of the QSAR model. There is no “absolute” cut-off by which a QSAR model is considered acceptable or unacceptable. Therefore, expert judgement is required to determine the acceptablity of the QSAR prediction.

Principle 5 - A mechanistic interpretation (if possible)

~~11.~~16. The statistical measures of a QSAR are intended to demonstrate an association between chemical structure and activity, but a mechanistic interpretation is intended to demonstrate a causal relationship between the knowledge of the chemistry and toxicology of a chemical structure and its activity. Therefore, the provision of a mechanistic interpretation can aid in the interpretation of the results of a QSAR model, adding transparency to the model and confidence in the result.

Reporting QSAR models and predictions

~~12.~~17. QSARs are typically reported using two formats, the QSAR Model Reporting Format (QMRF) and the QSAR Prediction Reporting Format (QPRF).

~~13.~~18. A QMRF is a reporting framework that summarises the key information related to a QSAR model, including the results of any validation studies. The QMRF is intended to provide users of the QSAR model detail related to the source of the model (including information on the model developer), the type of model and its development, validation and application. It also includes some information on the application of the

OECD principles within the QSAR model. The Joint Research Centre of the European Commission hosts a database of QMRFs[1], for genotoxicity endpoints including those produced for Case Ultra, Derek Nexus, Sarah Nexus and Toxtree, and some models, such as the OECD QSAR Toolbox and VEGA include QMRFs for some endpoints within their installation packages.

~~14.~~19. A QPRF is a standardised format for the reporting the results of a QSAR prediction to allow assessment of its adequacy. It provides detailed substance identification information and demonstrates the compliance of the QSAR model and the prediction with OECD principles. It is often a requirement for regulatory submission of a QSAR prediction. The Joint Research Centre of the European Commission has published a template QPRF with guidance on the completion of each data field[2].

Overall discussion and conclusions

20. QSAR models and their predictions cannot usually replace the need to undertake *in vitro* and *in vivo* genotoxicity tests required to derive conclusions on mutagenic hazard. However, QSAR approaches for the prediction of genotoxic activity can be a valuable tool to aid in the initial evaluation of genotoxic hazard and where relevant allow the development of a testing strategy. QSAR prediction of Ames results and gene mutations in bacteria are very robust, most models accurately predict this endpoint but the predictions of other genotoxicity endpoints are not as reliable.

~~15.~~21. Significant expert judgement is needed when using QSARs to ensure that the models are appropriate for the intended purpose and the predictions are robust and reliable. Adherence of a QSAR to OECD principles should be considered as part of an assessment of any prediction, and adherence to these principles should be documented in a QPRF.

~~16.~~22. The use of two or more different QSAR models, combining expert rule-based and statistical-based QSARs, may be used to generate predictions for an endpoint in order to provide adequate data as a weight-of-evidence approach. A single QSAR prediction, in the absence of any other data, should be considered with caution. QSARs are Stage 0 of the COM guidance; *in vitro* genotoxicity testing and *in vivo* genotoxicity testing are stages 1 and 2, respectively. The core tests in Stage 1 include bacterial gene mutation and mammalian cell micronucleus assays, as well as non-core tests including chromosomal aberration, mouse lymphoma, HPRT, *in vitro* assay for human reconstructed skin and the *in vitro* alkaline comet assay. Stage 2 details the core assays including rodent bone marrow and peripheral blood micronucleus assays or bone marrow chromosomal aberration assays, the transgenic rodent mutation assay and the rodent comet assay. Stage 2 also details the rat liver UDS assay. *In*

[1] https://qsardb.jrc.ec.europa.eu/qmrf/protocol?pagesize=250
[2] https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/qsar_tools/qrf/QPRF_version_1%201_DEREK_SS.pdf

vitro or *in vivo* genotoxicity tests should be attributed a much higher weight of evidence than (Q)SAR predictions, although all information should be assessed on a case-by-case basis.

References

COM (2011) *Guidance On A Strategy For Genotoxicity Testing Of Chemical Substances [Online]* Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/315793/testing_chemicals_for_genotixicity.pdf

OECD (2007) Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models. Environment Directorate. Joint Meeting on the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology. Organisation for Economic Co-operation and Development. ENV/JM/MONO(2007)2.

EFSA (2019), Romualdo Benigni et al, Evaluation of the applicability of existing (Q)SAR models for predicting the genotoxicity of pesticides and similarity analysis related with genotoxicity of pesticides for facilitating of grouping and read across. https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/sp.efsa.2019.EN-1598

Definition of terms

Training sets and test sets

Training sets represent the input data used to establish the model. Ideally, a 'test set' of data is also used as an external validation technique to check the predictability and applicability of the model. However, such approaches are not always possible. As a result, training sets are often divided into two reduced data sets, with one of the reduced training sets serving as the input data to establish the model, and the second reduced set serving as the external validation.

Sensitivity

Sensitivity represents the true positive rate, i.e. for those chemicals which are known to be positive in the experimental genotoxicity assay, the model correctly predicts a positive result for that same assay.

Specificity

Specificity represents the true negative rate, i.e. the proportion of chemicals that the model predicts to be negative that have also been experimentally determined to be negative in the genotoxicity assay.

Concordance

Concordance represents the amount of 'agreement' between two measures; these measures are typically the model that is applied within the QSAR and a 'gold standard' measure, which is the best approach for measuring the same endpoint. This gold standard may be an experimental assay or it may represent an alternative model.

Accuracy

Accuracy represents the precision of the software and is a ratio between the correctly predicted true positives and the true negatives.

Positive predictivity

Positive predictivity is the probability of a positive outcome from the model to be correctly positive, i.e.

$$\frac{True\ positive}{(True\ positive + False\ positive)}$$

Negative predictivity

Negative predictivity is the probability of a negative outcome from the model to be correctly negative, i.e.

$$\frac{True\ negative}{(True\ negative + False\ negative)}$$