Committee on MUTAGENICITY

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

Statement GXX or 2019/XX

Guidance Statement on the use of QSAR models to predict genotoxicity

Questions for the Committee:

- Are there any other aspects which should be included within the statement (such as ease of use)?
- Although the acceptability of a prediction is based on individual expert judgement, do the member's wish to include broad examples of predictions that are likely to be considered acceptable and unacceptable?
- This guidance statement describes the OECD QSAR principles in detail. Is such detail necessary, or should it link to the OECD report?

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

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). A staged approach to testing was recommended, consisting of stages 0 (preliminary considerations including physico-chemical properties), 1 (*in vitro* genotoxicity tests) and 2 (*in vivo* genotoxicity tests). QSARs are Stage 0 of the COM guidance. QSAR models and their predictions cannot replace the need to undertake the *in vitro* and *in vivo* genotoxicity tests required to derive conclusions on mutagenic hazard.

2. This initial assessment of potential genotoxicity can be based on the publically available QSAR models. The statement presented here provides guidance on the use of such models.

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

4. QSAR models may be knowledge- or statistical-based or a hybrid of the two approaches. Knowledge-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 knowledge-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 knowledge-based and statistical-based QSARs, for example, by providing identifying a mechanism of action with a statistical analysis of the data.

5. 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 as a weight-of-evidence approach.

The models used should be a combination of knowledge- 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 knowledge-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 of these QSARs are sufficient to conclude that a pharmaceutical impurity is of no mutagenic concern.

6. 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. These models were previously identified in report MUT/2018/02.

OECD QSAR principles

7. 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).

8. Toxtree, TOPKAT, DEREK Nexus, Danish QSAR Database, SARAH Nexus, Case Ultra, VEGA, OECD QSAR Toolbox, Leadscope Model Applier and ToxRead were assessed in terms of the OECD principles.

Principle 1 - A defined endpoint

9. 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 describe a number of 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.

Application of Principle 1 to each (Q)SAR

10. Toxtree is a collection of several modules that provide qualitative estimates for different endpoints. Three modules are of relevance when assessing potential genotoxicity namely *in vitro* mutagenicity (Ames test), carcinogenicity (genotoxicity and non-genotoxicity) and mutagenicity rule base and structural alerts for the *in vivo* micronucleus assay in rodents. Toxtree generally does not provide a further definition of the endpoint. However, under the carcinogenicity (genotoxicity and non-genotoxicity) and mutagenicity rule base, if a structure is identified as either containing aromatic amines or αβ-unsaturated aliphatic aldehydes, specific QSARs are available for predicting genotoxicity in *Salmonella typhimurium* TA100 (with S9) (Ideaconsult Ltd, 2015).

11. DEREK Nexus includes *in vitro* and *in vivo* predictions for genotoxicity endpoints including chromosomal damage, photo-induced chromosomal damage, mutagenicity, photo-induced mutagenicity, non-specific genotoxicity and photo-induced non-specific genotoxicity (Lhasa Ltd, year unknown).

12. The Danish QSAR database contains models for both *in vitro* and *in vivo* genotoxicity. The *in vitro* predictions consist of Ames assays (defined as reverse mutation test, direct acting Ames mutagens without S9, base pair Ames mutagens, frame shift Ames mutagens and potent Ames mutagens, reversions ≥ 10 times controls), chromosomal aberration assays in Chinese hamster overy (CHO) and Chinese hamster lung (CHL) cells, hypoxanthine-guanine phosphoribosyltransferase (HGPRT) assays in CHO cells, unscheduled DNA synthesis (UDS) in rat, Syrian hamster embryo cell transformation. The *in vivo* predictions consist of a sex-linked recessive lethal test in *Drosophila*, micronucleus test in mouse erythrocytes, dominant lethal mutations in rodents and sister chromatid exchange in mouse bone marrow cells (DTU Food, 2016).

13. Case Ultra contains several mutagenicity and genotoxicity models that are licensed in ‘bundles’. These bundles contain several models for predicting genotoxicity in *Salmonella typhimurium* and *Escherichia coli.* Strain-specific bundles contain strain-specific models for *Salmonella typhimurium*, with and without S9. (MultiCASE Inc, 2017).

14. Leadscope Model Applier contains QSAR models for *Salmonella* mutagenicity, *Escherichia coli* mutagenicity, mouse lymphoma assay, *in vitro* chromosome aberrations and *in vivo* micronucleus assay without further definition of the endpoints in the publically available material (Leadscope Inc., 2012b).

15. Within the OECD QSAR Toolbox, data are arranged in levels, with each sub-level offering more detail about the endpoint. The complete list of endpoints contained within the database is not immediately visible to the user; the QSAR development process is user-driven, and as such, specific endpoints are displayed based on the choices of the user. Therefore, a user can select to develop a QSAR based on a limited definition of an endpoint, such as mutagenicity in all available bacteria, or can set a highly specific definition of an endpoint, such as mutagenicity in a specific strain of a single bacteria with or without S9.

16. TOPKAT, SARAH Nexus, VEGA and ToxRead provide endpoint predictions for Ames mutagenicity without further definition of the endpoint.

Principle 2 - An unambiguous algorithm

17. 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.

Application of Principle 2 to each (Q)SAR

18. A number of the knowledge-based QSAR models provide genotoxicity predictions based on the presence of specified functional groups, which, for the purposes of this document, will be referred to as “structural alerts”. It should be noted that some models may apply their own nomencalture. With the exception of two specific QSAR calculations (for aromatic amines and αβ-unsaturated aliphatic aldehydes predictions), the Toxtree QSAR model applies a “decision tree” for the prediction of genotoxicity, with each “branch” of the tree representing a structural alert. As each structural alert will result in the QSAR model following one of several subsequent branches, a defined algorithm cannot be provided for this QSAR. However, the decision tree can be observed within the QSAR model, and in this sense, the QSAR model can be considered transparent and reproducible. The two specific QSAR calculations for aromatic amine and αβ-unsaturated aliphatic aldehyde genotoxicity predictions for genotoxicity in *Salmonella* typhimurium TA100 with S9, are based on unambiguous algorithms (Benigni *et al.*, 2008; Benigni *et al.*, 2009).

19. DEREK Nexus also identifies structural alerts in the prediction of genotoxicity, which are provided with the nomenclature “toxicophores” (Lhasa Ltd, year unknown). As such, it is not possible to produce a defined algorithm. However, the derivation of each alert is described within the QSAR model, and in this sense, the QSAR demonstrates transparency.

20. VEGA utilises a combination of structural alerts and statistical models. The structural alerts are defined within VEGA. However, the algorithms for the statistical models are not publically available (Mario Negri, 2017a; Mario Negri, 2017b; Mario Negri, 2017c).

21. SARAH Nexus is a statistical-based QSAR that is based on an unambiguous algorithm. Within the OECD QSAR toolbox, a QSAR prediction is generated via the user selecting a chemical profiler to collate an initial dataset, and subsequently reducing that dataset to a smaller number of chemicals to generate a prediction. An unambiguous algorithm is generated from this set of user-defined chemicals.

22. TOPKAT is a commercial model and its predictive algorithms are not reported in the public domain. Similarly, the Danish QSAR database, CASE Ultra and the Leadscope Model Applier contain several commercial models and their algorithms are either not reported within the QSAR model or are not available in the public domain. At the time of preparation of this document, ToxRead was still in development and undergoing beta testing. As such, the documentation of this model was incomplete, and no details on the algorithms contained within the model were available.

Principle 3 - A defined domain of applicability

23. 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).

Application of Principle 3 to each (Q)SAR

24. A number of the knowledge-based QSARs provide genotoxicity predictions based on structural alerts. These structural alerts may offer a domain in the sense that the chemical for which the prediction has been produced must share

characteristics with the dataset containing those structural alerts. However, the boundaries of that domain may be difficult to define. Therefore, determination of the applicability domains for such models is often subject to the quality of the documentation of the domains within the models.

25. The domain range within Toxtree is unclear, and cannot easily be determined from the data presented. For example, if a result of “No alerts for S. *typhimurium* mutagenicity” is obtained, it is not clear if this is a.) because there are no functional groups in the chemical that would trigger a structural alert, or b.) if there are no matching structural alerts within the QSAR model due to a limitations in its dataset, and the chemical would otherwise produce a positive result in an experimental test system (Benigni *et al.*, 2008; Benigni *et al.*, 2009).

26. DEREK Nexus has a clearer domain of applicability. This QSAR model indicates where a chemical does not trigger an alert. For example, the chromosome damage QSAR will generate a result of “nothing to report”. The mutagenicity QSAR will indicate if a chemical contains “unclassified” or “misclassified” structural characteristics. The user can then apply expert judgement to determine if this is a negative result, or if this is likely to be the result of an out of domain prediction (Lhasa Ltd, 2018).

27. The Danish QSAR database reports that it is defined by domain boundaries (DTU Food, 2016). However, they are not documented within the model. SARAH Nexus, VEGA and the OECD QSAR Toolbox report defined domain boundaries for each QSAR prediction. Based on the publically available information for the Leadscope Model Applier, the The Non-human Genetic Toxicity Model (one of the available modules within the QSAR model) (Leadscope Inc., 2012b) does provide a defined domain range. However, it is unclear whether the second module, the Genetox Expert Alerts Suite (Leadscope Inc., 2012a), has a defined domain.

28. TOPKAT and Case Ultra are commercial models and their domains of applicability are not reported. At the time of preparation of this document, ToxRead was still in development and undergoing beta testing. As such, the documentation of this model was incomplete, and no details on the domain range of the model were available.

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

29. 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.

Application of Principle 4 to each (Q)SAR

30. The majority of the QSAR models report some degree of measure by which their reliability can be assessed. Toxtree reports the overall sensitivity, specificity and accuracy of each of the modules, and details the true positive rate for each structural alert within its dataset (Benigni *et al.*, 2008; Benigni *et al.*, 2009). The Danish QSAR database reports on the overall sensitivity, specificity and concordance (DTU Food, 2016). VEGA details the overall sensitivity, specificity and accuracy for both the training set and the test set for two of the in-built mutagenicity models (CAESAR and SarPy/IRFMN), but only provides sensitivity, specificity and accuracy for the training set of two other mutagenicity models (ISS and KNN/Read-Across) (Mario Negri, 2017c; Mario Negri, 2017d; Mario Negri, 2017e; Mario Negri, 2017f).

31. The OECD QSAR Toolbox can report different measures, depending on the type of prediction the user selects to develop a QSAR. Using the “Read-across” approach, as would be typical for a qualitative assessment such as mutagenicity, the Toolbox provides statistics in the form of concordance between the values in the dataset and a p-value for the prediction confidence.

32. Leadscope Model Applier is a commercial model, however, a white-paper is publically available that provides a partial summary of the statistical principles by which the Leadscope Model Applier is measured. The Leadscope Model Applier reports the overall true positive, true negative, false positive, false negative, concordance, sensitivity, specificity, positive predictivity and negative predictivity of the Genetox Expert Alerts Suite (Leadscope Inc., 2016).

33. At the time of preparation of this document, ToxRead was still in development and undergoing beta testing. As such, the documentation of this model was incomplete. However, the documentation states that Fisher Test p-values are applied within the model to identify rules that have less statistical significance (Anon, year unknown).

34. TOPKAT, DEREK Nexus, SARAH Nexus and CASE Ultra are commercial models containing proprietary data, and therefore, their measures of goodness-of-fit, robustness and predictivity are not reported in the public domain.

Principle 5 - A mechanistic interpretation (if possible)

35. 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.

Application of Principle 5 to each (Q)SAR

36. The rule system within Toxtree is based on the identification of "structural alerts", whereby the presence of a particular functional group will trigger a particular alert. Within this context, there is an association between the chemical structure and its activity. However, Toxtree does not offer a mechanistic interpretation of those alerts to demonstrate a plausible biological mode of action. For example, the rule SA2_Ames, is explained within the QSAR model as "*methyl, ethyl, propyl, butyl or benzyl esters of sulphonic or phosphonic acid. P(=O)(O)(O)R or S(=O)(O)(O)R where R is not S or O The alkyl chains can have halogen substituents*" (Ideaconsult Ltd, 2015). It is not possible to readily relate this information to a genotoxic mechanism without further information that is not contained within the QSAR model.

37. DEREK Nexus utilises a combination of structural alerts and "reasoning rules", which appear to apply mechanistic considerations within the prediction. For example if a chemical operates via a rodent-specific mode of action, even if a structural alert is triggered, a result of "Impossible" will be provided for a bacterial prediction. Data on the mechanistic elements of SARAH Nexus are not reported in the public domain. However, the Lhasa website states that QSAR predictions "*Both Derek Nexus and Sarah Nexus have been designed independently to meet the OECD validation principles*" (Lhasa Ltd, 2017).

38. VEGA utilises a combination of structural alerts and statistical models. The structural alerts are defined within the QSAR. However, the model does not offer a mechanistic interpretation of the alerts to demonstrate a plausible biological mode of action, and no details are provided with respect to the mechanistic interpretation of the statistical model (Mario Negri, 2017a; Mario Negri, 2017b; Mario Negri, 2017c).

39. The OCED QSAR Toolbox provides mechanistic interpretations for each chemical profiler. A detailed "scheme" is accessible within the programme that provides a referenced description of the mechanistic interpretation of the profiler. As the application of these profilers is driven by the user, significant expertise is required to ensure the appropriate application of profilers to ensure that their application is biologically plausible.

40. The Danish QSAR database and ToxRead do not detail the mechanisms by which chemicals are predicted to display genotoxicity.

41. TOPKAT, CASE Ultra and the Leadscope Model Applier are commercial models containing proprietary data, and therefore, their mechanistic interpretations are not reported in the public domain.

Reporting QSAR models and predictions

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

43. 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], and some models, such as the OECD QSAR Toolbox and VEGA include QMRFs for some endpoints within their installation packages.

44. 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

45. QSAR models and their predictions cannot replace the need to undertake the *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. 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.

46. The use of two or more different QSAR models, combining knowledge-based and statistical-based QSARs, should be used to generate predictions for an endpoint in order to provide adequate data as a weight-of-evidence approach. A single QSAR

[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

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 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

Anon (year unknown) *Guide to ToxRead (0.9 BETA) [Online]* Available from: http://www.toxread.eu/downloads/ToxRead-0.9-beta-Guide.pdf

Benigni, R., Bossa, C., Jeliazkova, N., Netzeva, T. and Worth, A. (2008) The Benigni / Bossa rulebase for mutagenicity and carcinogenicity – a module of Toxtree. Joint Research Centre Scientific and Technical Reports. European Commission.

Benigni, R., Bossa, C., Tcheremenskaia, O. and Worth, A. (2009) Development of structural alerts for the in vivo micronucleus assay in rodents. Joint Research Centre Scientific and Technical Reports. European Commission.

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

DTU Food (2016) User Manual for the Danish (Q)SAR Database. National Food Institute, Technical University of Denmark.

Ideaconsult Ltd (2015) Toxtree (Estimation of Toxic Hazard - A Decision Tree Approach) v2.6.13. Available for download from http://toxtree.sourceforge.net.

Leadscope Inc. (2012a) *Genetox Expert Alerts Suite [Online]* Available from: http://www.leadscope.com/genetox_expert_alerts/

Leadscope Inc. (2012b) *Non-human Genetic Toxicity Suite [Online]* Available from: http://www.leadscope.com/genetic_toxicity_suite/

Leadscope Inc. (2016) An expert alert system to predict the mutagenic potential of impurities to support the ICH M7 guideline. Leadscope Genetox Expert Alerts Version 3.0 White Paper. May 2016. Available from http://www.leadscope.com/white_papers/LeadscopeAlertsWhitePaperV3FINAL-051816.pdf.

Lhasa Ltd (2017) *Sarah Nexus. Statistical-based software for the prediction of mutagenicity [Online]* Available from: https://www.lhasalimited.org/products/sarah-nexus.htm

Lhasa Ltd (2018) *Nexus 2.2: An Overview of Key New Features - Derek and Meteor Nexus [Online]* Available from: https://www.lhasalimited.org/publications/nexus-22-an-overview-of-key-new-features-derek-and-meteor-nexus/4625

Lhasa Ltd year unknown) *RE: Derek Nexus 5.0 Program Training.* Type to PRESENTATION PROVIDED TO PHE.

Mario Negri (2017a) VEGA. Istituto di Ricerche Farmacologiche, Laboratory of Environmental Chemistry and Toxicology, Milan.
Mario Negri (2017b) Guide to Mutagenicity Consensus version 1.0.2. VEGA user guide. Istituto di Ricerche Farmacologiche, Laboratory of Environmental Chemistry and Toxicology, Milan.
Mario Negri (2017c) Guide to Mutagenicity Classification Model version 2.1.13. VEGA user guide. Istituto di Ricerche Farmacologiche, Laboratory of Environmental Chemistry and Toxicology, Milan.
Mario Negri (2017d) Guide to Mutagenicity SarPy/IRFMN Model version 1.0.7. VEGA user guide. Istituto di Ricerche Farmacologiche, Laboratory of Environmental Chemistry and Toxicology, Milan.
Mario Negri (2017e) Guide to Mutagenicity ISS Model version 1.0.2. VEGA user guide. Istituto di Ricerche Farmacologiche, Laboratory of Environmental Chemistry and Toxicology, Milan.
Mario Negri (2017f) Guide to Mutagenicity Read-Across version 1.0.0. VEGA user guide. Istituto di Ricerche Farmacologiche, Laboratory of Environmental Chemistry and Toxicology, Milan.
MultiCASE Inc (2017) *CASE Ultra Models [Online]* Available from: http://www.multicase.com/case-ultra-models
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.

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.

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