Ensuring Confidence in Predictions: A Scheme to Assess the Scientific Validity of *In Silico* Models

Mark Hewitt[a,b], Claire M Ellison[b], Mark TD Cronin[b*], Manuel Pastor[c], Thomas Steger-Hartmann[d], Jordi Munoz-Muriendas[e], Francois Pognan[f], Judith C Madden[b]

[a]School of Pharmacy, Faculty of Science and Engineering, University of Wolverhampton, City Campus, Wulfruna Street, WV1 1SB, England.
[b]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, England.
[c]Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, IMIM (Hospital del Mar Medical Research Institute), Dr. Aiguader 88, E-08003 Barcelona, Spain.

[d]Bayer HealthCare. Bayer Pharma AG, Investigational Toxicology, Müllerstraße 178, 13352 Berlin, Germany.

[e]Chemical Sciences, Computational Chemistry. GlaxoSmithKline, Stevenage, SG1 2NY, England.

[f]Biochemical & Cellular Toxicology, Discovery Investigative Safety – PreClinical Safety, Novartis Pharma AG, Werk Klybeck, Postfach, CH-4002 Basel, Switzerland.

m.hewitt@wlv.ac.uk; c.m.ellison@ljmu.ac.uk; m.t.cronin@ljmu.ac.uk (*corresponding author); manuel.pastor@upf.edu; thomas.steger-hartmann@bayer.com; jordi.4.munoz-muriedas@gsk.com; francois.pognan@novartis.com; j.madden@ljmu.ac.uk

*Tel: +44 (0)151 231 2402

Abstract:

The use of *in silico* tools within the drug development process to predict a wide range of properties including absorption, distribution, metabolism, elimination and toxicity has become increasingly important due to changes in legislation and both ethical and economic drivers to reduce animal testing. Whilst *in silico* tools have been used for decades there remains reluctance to accept predictions based on these methods particularly in regulatory settings. This apprehension arises in part due to lack of confidence in the reliability, robustness and applicability of the models. To address this issue we propose a scheme for the verification of *in silico* models that enables end users and modellers to assess the scientific validity of models in accordance with the principles of good computer modelling practice. We report here the implementation of the scheme within the Innovative Medicines Initiative project "eTOX" (electronic toxicity) and its application to the *in silico* models developed within the frame of this project.

Keywords: (6 – 10)

Model reliability; peer-verification; good computer modelling practice; validation

Table of Contents

1. Introduction

The need to develop and utilise reliable *in silico* models to predict the behaviour of chemicals has never been greater. *In silico* models can be used to predict desired biological activity (such as pharmacological effects), toxicity (to human and environmental species) and absorption, distribution, metabolism and elimination (ADME) profiles [1-4]. This makes *in silico* approaches very appealing as computational screening approaches during early stages of drug discovery. These methods are also applicable to the prediction of properties relevant to formulation development and handling of chemicals (e.g. prediction of vapour pressure, solubility, melting point etc.) meaning they also play a role in lead optimisation and related stages. Models are employed by pharmaceutical, cosmetic, agrochemical and fine chemical industries, all of which have financial and ethical reasons for using alternatives to animal testing wherever possible. Model development and use are also promoted by regulatory agencies with an obligation to reduce animal testing. Legislation such as the REACH regulation [5], aimed at ensuring safety of employees and consumers and the 7th amendment to the cosmetics directive [6] requiring cessation of animal testing for cosmetic products, have further driven the need for alternative methods.

Whilst the importance of developing and using *in silico* models has been ascertained, there remain barriers to acceptance of predictions based on such models, not least amongst the regulatory authorities. Lack of confidence in the validity of a given *in silico* model (and hence predictions based thereon) is a major reason for rejection of such methods. Unlike *in vitro* alternatives, which have a distinct protocol for validation by organisations such as the European Centre for the Validation of Alternative Methods (ECVAM) (https://eurl-ecvam.jrc.ec.europa.eu/), *in silico* models have not yet been verified in such a formalised manner. Given the degree of diversity seen in available *in silico* models (model architecture, statistical analyses used, dataset size and composition, etc.) developing a universal approach is difficult.

Another significant factor that impedes greater acceptance of models is not that the model itself lacks validity but the level of detail by which the model is recorded is insufficient to allow judgement of model quality; this again means the model cannot be used with confidence. Increased acceptance and uptake of *in silico* modelling approaches will only be possible where confidence in the applicability and usefulness of a model to provide a given prediction can be assured. In a recent scientific report on modern methodologies and tools for human hazard assessment of chemicals the European Food Safety Authority (EFSA) highlighted the need for validation of predictive models as an important step in their utilisation for chemical risk assessment [7]. Herein, we discuss factors that should be considered when assessing the validity of an *in silico* model, appropriate recording of model details and a pragmatic scheme that can be applied for model verification.

1.1 Recent advances and the current stage of *in silico* model verification

The importance of developing high quality models is well recognised and, despite the lack of formal verification schemes, there have been numerous publications relating to best practice

and common errors (and the resulting problems these cause) in model development. Cronin and Schultz [8] reported 'essential' and 'desirable' features for toxicological QSARs. These included factors relating to model development, similar to those later identified within the OECD Principles for the Validation of QSARs (*vide infra*) [9], as well as factors relating to model use e.g. avoiding extrapolation beyond the original domain of the QSAR and appreciating the expected predictivity of the model, when taking into consideration the biological data on which the model was founded. Their paper discusses reliable modelling practices addressing the three components of a QSAR individually i.e. (i) the biological data and their quality, (ii) the physico-chemical descriptors used and (iii) the statistical algorithm applied. The paper reinforces the importance of external validation as a means to confirm utility of the model. Dearden et al [10] tabulated and discussed (using examples from the literature) 21 types of common error in the development and reporting of quantitative structure-activity and structure-property relationships indicating which OECD Principle(s) were violated by the errors. Recommendations were provided as to how to avoid such errors and improve model development and reporting. Stouch et al [11] carried out validation studies on four externally derived models using in-house data. Although rationally developed, the models performed poorly using the in-house data. Reasons for the apparent 'failures' of the models are discussed by the authors. Problems, such as models being developed using inappropriate / highly variable data or data covering a very narrow range were identified. The authors make recommendations as to how to improve on the development, reporting and use of models emphasising the importance of fully describing the data used and the model itself so that the applicability of the model to new compounds can be assessed.

In 2009, Judson [12] proposed a series of rules to be used in establishing good computer modelling practice (GCMP), analogous to the good laboratory practice (GLP) rules that are applied to ensure high quality of experimental procedures. In addition to highlighting the problems of inconsistent interpretation of the OECD Principles, Judson identified the additional problems associated with model reproducibility e.g. inadequate information on software settings, lack of availability of training or test sets, changes to versions of software or availability of programs and lack of suitable documentation for subsequent auditing. Thirteen illustrative rules were proposed for GCMP along with 11 illustrative rules to support auditing of GCMP. The proposed rules cover a range of issues for example, confirming chemical and biological data, providing sufficiently detailed information to allow other modellers to repeat tests performed (e.g. parameter and option settings, program versions etc), reporting of anomalies, maintaining records of changes to models etc. Problems in inadequate recording of modelling procedures were also identified by Kristam et al [13] who proposed 12 hypothesis specification requirements that would enable replication of pharmacophore models published in the literature.

The need for developing high quality models and for detailed reporting of the modelling process is clear, however progress in developing validation strategies for *in silico* models has been relatively slow. It is more than a decade since the OECD published its principles for the validation of Quantitative Structure Activity Relationships (QSARs) for regulatory purposes [9]. These principles state that a model should be associated with (1) a defined endpoint; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness-of-fit, robustness and predictivity and; (5) a mechanistic interpretation, if possible. Although QSARs are stipulated explicitly, the Principles are equally

applicable to other *in silico* predictive methods. Whilst these principles are based on sound scientific philosophy, more guidance is needed on how to apply them in practice to model development and assessment. Additional regulatory guidance on the information requirements and chemical safety assessment utilising predictive models has been released by the European Chemicals Agency (ECHA) [14,15]. Thousands of *in silico* models have already been published within the scientific literature with model development on an increasing trajectory due to the necessity to replace animal tests. Furthermore, many models developed within the industrial environment remain unpublished, despite being used routinely within that setting. What would greatly benefit model developers and users is a robust method whereby the scientific validity of models and their suitability for a given purpose could be ascertained, such that an end user could have greater confidence in their results.

The inherent quality of the data upon which a model is built is arguably the most important characteristic of any model. Data quality here refers to the accuracy and completeness of the information on the chemicals studied as well as the adequacy and reliability of the experimental data [16,17]. Ensuring the accuracy of chemical structures is not a simple task. It is relatively easy to check the basic structure of a compound is correct via a comparison to high quality, online resources such as ChemSpider (http://www.chemspider.com). However, if the original data source contains an error (for example chirality, ionisation, etc.) this can be difficult to identify. In addition, tools such as ToxRTool [18] can be used to help assess the quality of both *in vivo* and *in vitro* experimental data. ToxRTool provides a series of questions relating to experimental information from which a Klimisch reliability score is generated - Klimisch scores being the most commonly used methods to assess experimental data quality [19]. Although data quality is of paramount importance in QSAR development a detailed discussion of quality is beyond the scope of this article; excellent reviews of data quality and assessment schemes are already available in the literature [19-22].

A key factor identified from the literature is the requirement for detailed recording of the data, model and supporting documentation to enable the validity of the model and its applicability for a given purpose to be ascertained. A checklist-style reporting format has previously been developed by the European Commission's Joint Research Centre (JRC, Ispra), known as the (Q)SAR Model Reporting Format (QMRF) [23]. This provides a template for recording key information about QSAR models and associated validation studies. The format was designed with adherence to the OECD Principles in mind. Therefore, this template provides a useful starting point for recording relevant model details which can be used to aid the assessment of the overall validity of the model.

Note that the term "assessment of scientific validity" is used herein, this is distinct from formal "model validation" processes which may be undertaken, for example, by a regulatory authority. Such formal validation would require a more laborious analysis to be undertaken and the outcome to be agreed by stakeholders. Here we are concerned with proposing a scheme that can be implemented by model users / developers to demonstrate that a published model has undergone a rigorous assessment procedure. An alternative phraseology would be "peer-verification" of models. Verification is considered here as the process by which models are assessed for their compliance with a set of standard criteria developed in consultation with model developers and users. The standard criteria reflect good computer modelling practice and are independent of the relevance of the model for a

particular endpoint. The objective is to develop a “standard operating procedure” (which may be confirmed using a checklist approach) to confirm that adequate information has been recorded about a given model to enable it to be accurately reproduced and judgements to be made on the statistical reliability of the model. This allows a user to determine the model’s suitability for predicting behaviour of other chemicals (e.g. whether or not a new chemical would fall within the applicability domain of the model). This would enable users of models to have more confidence in predictions based on the models and a greater appreciation of which models are (not) suitable for a given purpose.

The process of “peer-verification” was considered as being key for increasing user confidence within the European Union Innovative Medicines Initiative eTOX project, which is described in detail below [24,25]. Whilst the work reported here was carried out as part of the eTOX project, the assessment scheme developed was designed to be universally applicable. Extending beyond regulatory acceptance, it is also anticipated that the verification scheme could have a role in developing rules for peer-reviewing models submitted for publication in scientific literature. Currently, the quality of predictive models submitted for publication is variable, with the editor/reviewer having to assess whether a given model is valid and suitable for publication. The development of a verification scheme, as detailed here, could be used, not only by model developers and regulatory bodies, but also by the aforementioned editors and reviewers to inform on the scientific validity of a given model.

The aim of this paper is to provide a scheme, for model developer, users and the wider scientific community, which can assess the validity of *in silico* models. This paper provides a harmonised method for peer-verification of models, ensuring where possible the OECD Principles are met and good computer modelling practice is followed to maximise confidence in predictions based on the models. Emphasis is placed on accurate and detailed documentation to accompany the models, as insufficient information is acknowledged to be a significant barrier to acceptance.

1.2 eTOX project overview

The eTOX (“electronic toxicity”) project is a joint collaboration between the European Commission and the European Federation of Pharmaceutical Industries and Associations (EFPIA) as a Joint Technology Initiative under FP7 (http://www.etoxproject.eu/). The eTOX project started in January 2010 for a duration of 5 years. The international project consortium consists of 13 pharmaceutical companies, 7 academic institutions and 6 SMEs with a total budget of €13.9M. The aims of the project are to develop a drug safety database from the pharmaceutical industry legacy toxicology reports and public toxicology data, build innovative *in silico* strategies and novel software tools to better predict the toxicological profiles of small molecules in early stages of the drug development pipeline. It is anticipated that wider uptake of a model verification process will increase confidence in, and therefore greater acceptance of, the *in silico* models built within the project.

A major output of eTOX is an online prediction platform known as eTOXsys, which contains not only a searchable version of the database, but also access to all the models developed

within the project (currently access is restricted to project partners). At the time of writing the database contains over 1700 structures linked with over 6000 repeat dose studies, and the models available cover a range of endpoints including cardiotoxicity, hepatotoxicity, phospholipidosis and many others. These models are developed from the available public data and, in some cases, with confidential data supplied under agreement with industrial partners. These data have been extracted from legacy toxicological reports and developed into a large database to advance *in silico* drug-induced toxicity prediction [25, 26].

Given the project extracts, collates and models both public and industry toxicological data, a critical component of the eTOX project has been to research and construct a strategy to assess the scientific validity of the models it develops. This strategy is a significant output of the project in itself, since no comprehensive verification scheme has yet been published.

2. Development of the Assessment Scheme

Several factors were considered to be of key importance in developing the assessment scheme to be used for model verification: (i) carrying out an assessment of a model had to be a realistic task both in terms of the required expertise of the individual and the time needed to conduct such an assessment; (ii) the assessment criteria had to be presented in a format which would be compatible with a wide range of operating systems and software and; (iii) the verification process had to be transparent, scientifically justifiable and the results readily accessible to end users.

The OECD Principles currently represent the state-of-the-art in terms of model validation and the QMRF is a well-recognised method for recording key information concerning models. Therefore the OECD Principles and the QMRF documents were used as the starting point for further development of the model assessment scheme described herein. The developed scheme also drew on the rules proposed by Judson [12] for GCMP to ensure that the models themselves were scientifically valid in addition to being correctly recorded. Modifications were made to the existing QMRF in order to:

- Extend the scope of the criteria beyond those appropriate only to QSAR models, to yield a set of criteria applicable across a wider range of *in silico* modelling methods
- Provide additional guidance to the user on how to assess each individual criterion.

In addition to the development of the criteria themselves, a key element of the current work is that it proposes a peer-review style model verification protocol, where external model verifiers perform robust model assessments using the aforementioned criteria. To help support this, additional functionalities such as dataset/model repositories, verification datasets and detailed guidance documents are introduced and discussed.

2.1 Peer-Verification

A novel component of this approach is the introduction of a peer-review aspect to model verification. This is not to say that an external verifier is responsible for all aspects a model's successful verification. We propose that model builders themselves prepare and submit all required documentation and supporting data, such that an external verification can be readily carried out. The role of the model verifier is then to check the submission for accuracy, completeness and reproducibility.

There are several possibilities regarding who could take responsibility for model verification. Researchers in the modelling community could undertake verification of each other's models in the spirit of cooperation, so that greater confidence is built in published models. Users within industry, who have an interest in the applicability of models to their chemical space of interest, could also be involved in verification. Within larger projects, for example European project consortia the task of model verification can be distributed amongst partners. There is also the possibility of international organisations such as the Joint Research Centre, QSAR DataBank (www.qsardb.org) [27] or the OECD playing a role in verification, perhaps being appointed by regulatory bodies themselves to further increase acceptance.

In the initial stages of developing the concept of peer-verification it was important to have the input of both modellers and end-users who could comment on the process as it was being established, and advice on improvements in an iterative manner. The peer-verification process was developed here in collaboration with partners on the EU eTOX project. There were several advantages to this: many models were under development within the project enabling a range of model types to be put forward for the pilot verification process; one partner could act as a coordinator liaising between model developers and verifiers; a large project team was already working together in model development; end users could have direct input into model verification identifying features that were most important to them; an "honest broker" (Lhasa Limited, Leeds) was available to perform peer-verification on models that were built using confidential data. Part of the peer-verification process requires investigation of the data used to build the model, where confidential data are used in model building alternative solutions have to be considered. In this case a trusted third party could access the data; this approach may be adopted for other models where the model builder can identify a trusted third party as verifier or the builder would need to provide assurance that adequate checks have been made concerning the data.

In silico models that were available within the eTOX project were put forward for verification using a proposed model verification schema. An overview of the process is given in Figure 1.

INSERT FIGURE 1 HERE

The criteria used for assessment represent an amalgamation of the OECD Principles, the (Q)MRF documentation requirements and elements of GCMP, as identified by Judson [12].

It was a prerequisite that the verification criteria were delivered in an open format, accessible by all end users. It was agreed that Microsoft Excel would be used as most users have access to this software or the file can easily be reformatted as a simple text file (e.g. csv) for input into other tools if needed. Since it was anticipated that verification documentation

would be archived and remain accessible to end users upon completion, this file format also benefits from relatively small file sizes. A checklist style architecture was adopted for simplicity, with the aim to enable a rapid (tick box) verification to be performed where possible. The verifier's role is to ensure that all of the required information is present and appropriate. A front cover page (worksheet) was created to record details of the model, its developer and verifier. In addition, the main aspects of the verification process are summarised and fields are created to house any comments or questions arising following verification (Figure 2).

INSERT FIGURE 2 HERE

The checklist details all of the criteria to assess as well as a short text field to provide additional notes relating to each criterion. It was intended that the single file would contain the criteria, be completed by the model verifier and then be made available to show the verification status of a given model.

Model verification is underpinned by comprehensive model documentation. As discussed previously, one of the major limitations to model acceptance at present is the lack of sufficiently detailed model documentation. Basing the information required for model verification on the criteria required for the JRC's (Q)MRF template alleviates this problem, as it was specifically designed to collect all of the information needed to assess a model's adherence to the OECD Principles. The complete verification criteria, termed the "verification template", used here (developed in collaboration within eTOX project) are available as supplementary material. In an effort to make the verification criteria both easy to complete and also easy to review, it was decided to split the criteria into three sections ((i) – (iii) below), each covering a different aspect of validation.

2.1.1 Model Documentation

This section forms the backbone of the verification criteria and consists of a range of questions relating to model documentation. This is the most important aspect of the assessment since it is contains all of the modelling details and supporting statistics. This section covers aspects relating to:

- A model's identity and its developers
- The endpoint investigated
- Training/test set data (including source)
- Modelling algorithm and summary statistics
- External predictivity
- Mechanistic information
- Applicability domain
- Interpretation of prediction

There should be sufficient detail within this section, not only to assess the model in terms of the OECD Principles, but also to rebuild the model. Any uncertainties arising during model verification should be addressed with the model developer. In terms of scientific validity, the documentation represented here should contain sufficient details to identify any potential issues resulting from application of inappropriate statistical approaches, overtraining of models, limitations in domain of applicability, etc.

2.1.2 Data Used to Build the Model and Consistency of Model Output

The nature and quality of data used to build a model is a major determining factor in model acceptability. It was therefore considered prudent to include an assessment of the dataset(s) used to develop a model as part of the verification process. Missing or incorrect data are recognised as major problems within model development [8, 10]. Unless prohibited for reasons of confidentiality, dataset(s) used to build (and test) a model should be available to the verifier. The information (which can be provided in the form of an Excel spreadsheet) should include chemical structure data (e.g. chemical name, SMILES strings), known (experimentally determined) endpoint values and endpoint values as predicted by the model in question. Knowledge of what the predicted values should be is required for assessing model implementation.

A major consideration in data quality is confirmation of the accuracy of the chemical structures contained within the data files submitted for model verification. Inaccuracies in chemical structures are one of the most common sources of error within computational modelling [16, 17, 28]. If these structures are incorrect (in terms of their chemical composition, ionisation, chirality, etc.), the error is carried across into the predictions making them unreliable. Whilst checking all structures within a dataset may be impractical for large datasets, a representative sample of data (for example 5 or 10% of structures can be checked using appropriate databases (such as ChemSpider)). If a problem is identified when checking a small number of structures then the remainder of the dataset can be checked. Another option would be to utilise freely accessible tools developed and supported by public organisations, such as OCHEM (www.ochem.eu) [29]. OCHEM has the functionality to rapidly screen a dataset (chemical name and structure) against PubChem and additionally applies methods to check for structural correctness. Structures that are suspected of being incorrect are then highlighted to the user for further examination. The same features as implemented in OCHEM are available within QSPR Thesaurus database [30] which, whilst more focused on the estimation of environmental toxicity, is a useful tool for identifying hazard. This approach dramatically reduces the number of manual checks that need to be performed by the mode verifier, enabling the entire dataset to be assessed, increasing the quality of modelling datasets. Of course, since this is a web-based tool, it is not suitable for assessing the quality of confidential data.

Quality of experimental data can be assessed using criteria, such as Klimisch scores [19]. However assessment of experimental data is a complex task and a detailed discussion is beyond the scope of the current work. As mentioned previously, for models containing confidential data, an “honest broker” can be assigned. This trusted third party could access the data and perform the verification checks required. This approach may be adopted where

the model builder can identify a trusted third party as verifier or the builder would need to provide assurance that adequate checks have been made concerning the data.

2.1.3 Implementation of the Model

The final component of the assessment criteria relates to how a model is implemented. For example, if a model is made available as a web service or is recommended to be implemented using specific software then the infrastructure for implementation needs to be confirmed as being stable and capable of providing reproducible results for different users.

In terms of the eTOX project, each model is made available via a web-based platform. Where confidentiality is an issue (e.g. for industrial partners), the project also implemented its predictive capabilities within a virtual machine environment that could be installed locally behind their firewall. Therefore, in this case it is necessary to ensure that a model is functioning as intended within the eTOX platform. Three aspects were considered here for model implementation:

I. **Model stability** – Here, the model is tested to ensure the model functions and returns predictions in all cases. Any issues relating to descriptor calculation error, problem structure(s) or more general software problems are quickly detected utilising a predefined dataset containing a selection of “common” structures. In the case of the eTOX project these structures were designed to include commonly encountered drug types and associated moieties. A validation set is designed to represent the types of compounds an end user may routinely analyse using the model.
II. **Robustness to input files** – The second stage is to assess a model’s robustness to a dataset containing structurally diverse compounds covering a broad spectrum of structural domains and molecular weights. Any sensitivities or limitations of particular models/software would be exposed under this more rigorous analysis of a model’s applicability
III. **Consistency of output** – This final stage is designed to check that a model returns what is expected and is functioning as intended. There are two parts to this step:
    a. When using the model’s training or test set, do the predictions generated by the model match the expected output provided by the model developers?
    b. Does the model generate the same output when the same structure is given in multiple formats? For example, differing SMILES notations for the same compound or when comparing the output given by structure input files of different formats.

Outside of the eTOX project, the possibility exists for modellers to upload models onto publically accessible online resources, such as QsarDB (http://www.qsardb.org/) [27] or OpenTox (http://www.opentox.org/) [31]. In addition to the actual models, these resources are also able to store modelling datasets, model documentation, model predictions, etc. This functionality provides an ideal platform from which to conduct model verification studies

since it allows full access to the model and all of the related information. Furthermore, it may also be possible for these systems to automatically generate summary statistics based on the datasets provided, increasing consistency and negating the need for modeller/verifiers to calculate these values themselves.

For any predictive model implemented in any accessible system or tool, sustainability of that platform is a key factor to consider. In terms of eTOXsys (and any other web-based platform) it is initially developed and then maintained only for a specified period of time. Depending on subsequent usage, this period may be extended, but the life of a particular online system or tool is uncertain. If the system is deleted (or access restricted) the model(s) are therefore lost. This in no way is berating the importance of online resources and tools, but inevitably their usefulness diminishes with time with the release of new tools and data. A possible solution to ensure that models are not lost is again to upload them onto publically accessible resources including QsarDB [27] and OpenTox [31]. An additional resource suitable for this purpose is the Joint Research Centre's QSAR Model Database (http://qsardb.jrc.ec.europa.eu/qmrf/). Their upload to and storage on such repositories ensures that models are preserved and can be accessed for future use.

Given the multitude of differing modelling approaches now being utilised, the verification template was intentionally developed as a non-prescriptive, flexible document. It was anticipated from the outset that not all fields would be applicable in all cases and that certain fields were open to interpretation by different modellers and verifiers. For example, applicability domain definition may be determined in different ways depending on the modelling approach (e.g. QSAR model versus the use of structural alerts).

It was anticipated that iterative refinement of the peer-verification process would be necessary. Therefore, once a draft procedure for verification had been devised (as described above) a pilot study was carried out, within the eTOX project, to determine necessary amendments (e.g. what steps would be needed to improve the process, was additional guidance documentation necessary, etc.). For the pilot study a diverse set of ten models was selected; this contained models that employed different algorithms, endpoints, dataset characteristics and differing model developer backgrounds (industry, academic etc.). Ten appropriate model verifiers were assigned (based on their expertise in the area) and the results of this pilot study were used to refine the verification process. It must also be noted that the eTOX project is made up of EFPIA (European Federation of Pharmaceutical Industries and Associations) and public partners. Both of these user groups were represented in the choice of model verifiers. Each model verifier was provided with the information and documentation detailed by Table 1.

INSERT TABLE 1 HERE

To highlight the diversity of the modelling approaches covered with the pilot study and the applicability of the devised verification protocol and documentation, the ten pilot study models developed within eTOX are summarised by Table 2.

INSERT TABLE 2 HERE

Those involved in the pilot study were invited to comment on the operation of the peer-verification scheme. Following feedback received, an improved scheme for model assessment was devised. The outcome of the pilot study and the resulting new schema are discussed below.

3. Results of pilot study

The aim of this work was to devise an assessment scheme to enable peer-verification of *in silico* models, such that end users could have greater confidence in accepting predictions based on such models. A pilot study using a draft scheme was carried out within the eTOX project, the results of which were used to inform the next stage of development of the peer-verification process resulting in a schema with broad applicability to a range of *in silico* models.

The pilot study was extremely informative as it gave an insight into how the verification process fared in a real-world situation. It highlighted a range of issues, some predictable and others which were unanticipated; this information was critical to the refinement of the verification process. During the initial pilot study none of the ten models selected were successfully verified. The reasons for this partly relate to problems within the verification process (which will be discussed in more detail below) but also to the fact that the modellers were not aware of the verification criteria when they began developing models. Thus, critical information was missing in certain cases because specific steps were not taken during model development. The second reason why models could not be verified during the process related to the inexperience of the verifier. As this was a new process for all parties involved, certain aspects of the process were not communicated in sufficient detail. Misunderstandings led to misinterpretation of the criteria and hence models failed in areas where all the information was available but the verifier did not know where/how to access the relevant information. These issues were identified within the feedback provided by the verifiers which forms the basis for the results of this analysis and the focus of the remainder of the paper; performance of individual models and their verification is not discussed further.

3.1 Feedback on the verification process

Overall, verifiers stated that they were satisfied in general with the model verification process and the tasks involved. However, they reported confusion regarding some areas of the documentation which led to them being unable to complete some of the tasks.

3.1.1 Data used to build the model

Assessing the dataset(s) used to develop and test predictive models is a key element in the verification process. The main problems found when assessing data quality in this study related to the format of the provided data. There was no consistency between the modellers; each supplied the data in their favoured format (InChi, SMILES, SDF etc.) and also supplied varying levels of detail (e.g. which compounds were used as test or training data, their predictions from the model, their experimental results etc). Therefore, an obvious solution would be supply modellers with a standard template with which to submit the required data.

The second factor of assessing data quality was to check the validity of the structures provided. This can be an extremely difficult and time consuming task if all the structures within a supplied dataset are to be verified. Therefore, only a sample of the structures was required to be verified. However, the feedback from the verifiers still highlighted this as the most time consuming task of the whole process. Methods implemented to complete the task included performing manual checks on a small percentage of the dataset (e.g. 5%), comparing the InChi-keys generated from the provided structures versus those obtained from ChemSpider (using CAS numbers) and crosschecking the SDfile against UNICHEM Cross Reference Code. Verifiers also commented in a number of cases on the lack of stereochemical definition in structures. This would become important when different stereoisomers display different biological activities. If stereoisomerism information was lacking, a model would be insensitive to any differences between two stereoisomers.

The final factor which should be examined is the accuracy of the activity/endpoint data. This aspect of data quality was not assessed in the pilot study because it was outside the scope of the verification work. It is important to assess experimental data quality at the modelling rather than verification stage as models should not be built using poor-quality data. Modellers then need to describe how they have assessed the quality of their training (and test) data in the model documentation. The verifiers would then simply need to check that this information is available. An example of high quality data suitable for modelling is the central eTOX database. Here, the data underwent a strict routine of tests and checks before being entered into the central database. As discussed by Przybylak et al [22] completing data evaluation at the data curation stage is easier as modellers do not need to look back through historic reports or publications to find the data which is relevant. However, it should be noted that although the data was checked by the donating partner for correctness against the original study report and GLP status are provided; no global description of quality was assigned to the data. Also, the models used in the pilot study had not been produced using the main eTOX database.

3.1.2 Model Documentation

This is an area where all of the pilot study models failed the verification process. The issues found with the model documentation were often minor and consisted of typing errors and discrepancies between the data supplied by the modeller and the formal documentation in terms of reported statistics, number of compounds etc. There were also reports where the information provided were too vague to be useful to the end user and would benefit from additional clarification.

The largest problem with the model documentation was the presence of multiple missing entries where no information had been provided (i.e. the fields had simply been left blank). In many cases the fields were not applicable to the model under investigation but the verifier could not know if this was the case or if the field had been missed. For example, when considering structural alert models with qualitative endpoints (e.g. positive/negative), certain quantitative goodness of fit criteria (R^2, RMSE, etc.) are not applicable.

Some reporting errors were expected but it was surprising that all of the models in the pilot study failed in this area. This highlights the importance of fully involving the modellers in the verification process. If they were aware of the specific elements which were going to be analysed during verification, they may have prepared their documents differently and been more successful during verification. Subsequent improvements to this process are discussed later.

The applicability domain and reliability of each model should be described in the model documentation so that the model is compliant with the fourth OECD validation principle. This information should show the user where a model can be used to make reliable predictions and act as a guide/warning to end users. However, it was found that most models within the pilot study did not have any assessment of the applicability domain. This was not totally unexpected since the version of eTOXsys in use at this time did not support models providing applicability domain information (such information needed to be added as part of a models eTOXvault entry). However, it was surprising to see that most models lacked even basic applicability domain information. This is a major problem and needs to be rectified for all models. The exceptions to this were the structural alert models for which the applicability domain is inherently described: the presence of an alert within a compound dictates that the compound is within the applicability domain of the model. It must be remembered that the applicability domain is a complicated issue with many different approaches one can take in its definition. Even for the aforementioned structural alerts models, defining the applicability domain can be a complex task [32]. Within eTOX, the new release of eTOXsys is now able to accept applicability domain information, aiding in its characterisation. At the same time, work is ongoing within the project to develop tools able to assess a models applicability domain (e.g. ADAN [33]).

3.1.3 Implementation of the Models

Model implementation during the course of the eTOX project is via a bespoke web-based platform eTOXsys. Upon completion of the project, a fully deployable, self-contained system is envisaged. This online system used to run the models and return predictions in the pilot study was found to be stable under these test conditions. This is a significant achievement in itself considering the complex architecture of the system and that it relies on input from many independent researchers (Fig. 3). The only two points raised with regard to the stability of the eTOX system were the enforced maximum size of the input files, which was smaller than some of the test sets, and also problems with SDFiles created in specific software. The first problem was easily rectified by the developers simply increasing the maximum file size allowed by the system. The second issue related to a specific bond type created by third party software (namely the aromatic type ‘4’ bond created by the ChemAxon, MarvinBeans suite of tools (http://www.chemaxon.com/products/marvin)). Therefore, the solution to this was to either not use ChemAxon for generating SDFiles or to the use the “Convert to

Kekulé" function with Marvin. Whilst the issues identified were clearly specific to the eTOXsys platform, the study highlights that whichever system/software is used for model implementation, it needs to be assessed as part of the model verification process.

INSERT FIGURE 3 HERE

Although stable, some models were unable to make predictions in all cases due to their inability to generate required molecular descriptors. It must be said that this was encountered rarely (for a single compound), but it did result in all compounds within a dataset receiving an error rather than a prediction. It is understood that more complex compounds may, sometimes, cause a failure in the calculation of certain parameters. However, in this study it was only the training, test and verification datasets that were being considered which rarely contained any exotic compounds. Within the pilot study, one exception to this was a compound containing co-ordinate bonds which lead to multiple prediction failures. Such failures here suggest possible issues in relation to model robustness and applicability that should be made clear to the user.

In addition to missing predictions, it was also noted that the predictions generated by eTOXsys did not, in all cases, match those provided by the model builder. Such differences could be a result of the model verifier altering the input file in some way. This could include splitting the input file to overcome the input file size restriction or as a result of a verifier transforming the input file format (e.g. from SMILES to InChI). In addition, there could be problems resulting from how the model was implemented within eTOXsys. It is also possibly that erroneous predictions were provided by the model builder but this issue is easily resolved where model builders and verifiers are working together.

In addition to points raised above, three further issues were also raised by the verifiers. One of these was specific to the project but 2 others have a wider applicability and are hence discussed here:

I. How do we know that the verifiers are performing the process correctly?
II. The verification protocol examines model validity, but not whether the model is fit for a given purpose

The first point was easy to answer as the entire verification process was co-ordinated, in this case, by the verification co-ordinators who could check all incoming verification reports for accuracy and consistency. They were also there to act as an intermediary to enable communication between the modeller and verifier where required. Outside of projects where it is possible to have a verification co-ordinator, it is envisaged that a peer-review system similar to that used by scientific journals could be used. Modellers and verifiers would work together to ensure that verification is completed in a satisfactory manner.

The protocol described here intentionally only examines model validity and not whether the model is fit-for-purpose. This is because whether a model is fit for a specific purpose can only be determined by the end user (i.e. a model may be appropriate for one purpose but not for another – this does not reflect on the intrinsic quality of the model). In terms of eTOXsys

it is envisaged that it will be used by a variety of end users and therefore it is impossible for one or a group of parties to state which models are the “best”. This philosophy also applies outside of the eTOX project where model developers and end users may be working from very different perspectives. Information should be available to end users to help them to make the decision for themselves (i.e. description of endpoint, data used to build model, statistics etc.). A global fit-for-purpose assessment is not appropriate.

3.2 Lessons Learnt

The feedback provided from the pilot verification study clearly highlighted areas for improvement. The most important issue was the need for clear guidance for the verifiers. This study used a variety of verifiers from different institutions which led to variability in the way the verifications were completed and hence the information that was reported in the verification template. This problem could be easily rectified by providing clear, concise guidance documents to verifiers. The documents would state how to perform each stage of the verification and what is (not) acceptable in the fields of the model documentation. They would also detail which data should be used for testing and whether verifiers need to reproduce any reported test statistics.

Another lesson learnt was the amount of missing information the verifiers found in the model documentation. It became apparent that modellers need to know what is required for verification at an early stage so that they can ensure this information is available. It should not be presumed that they will automatically provide all the required information. For example, in this study, along with modellers not entering ‘N/A’ for fields not relevant to their models, they often omitted the following information:

- Definition of the applicability domain
- Mechanistic basis of the model
- Description of the types of compounds used to build the model (e.g. pharmaceuticals, small industrial chemicals)

In addition to the specific guidance required for verifiers, it became clear that further guidance on assessing data quality was required. This is in terms of checking the structures in both test and training set data, and whether or not to assess the quality of the experimental data. The method for checking structural integrity was chosen by each specific verifier, but this again led to variability in the results and also left some verifiers unsure how thorough to be with the task (e.g. is stereochemistry important for all models?). Therefore, it would be useful to provide a standard set of instructions on the specifics of assessing structural accuracy. It may also reduce the time taken to perform this task which was a concern to some of the verifiers.

Assessing experimental data quality is another important yet time consuming task, as previously discussed. Therefore it is important, as stated above, to provide verifiers with enough information on the topic to allow them to fully understand what is (not) required. Although experimental data quality was not examined in this study (due to the constraints of

the project) it is something which would be required if this verification scheme is to be used on a larger scale. As the initial use of the data is by the modellers, it would seem logical that the data quality is assessed at this stage, rather than during verification. A model builder can comment on how they have arrived at the decision that particular data are suitable for their purpose (e.g. experiments were performed according to GLP, database is highly curated). Therefore, one can envisage the data quality assessment being included in the model documentation. A standardised system would need to be agreed but something akin to the Klimisch [19] scheme could be implemented by modellers when they build their models and the verifiers would then only need to check whether the information has been made available in the model documentation.

4. Development of a Revised Assessment Scheme for Peer-Verification of Models

The lessons learnt during the pilot study have enabled us to make improvements to the verification scheme devised and to suggest a practical final scheme for the verification of predictive computer models (Figures 4-5). It is expected that the co-ordination aspect would be completed by a verification co-ordinator (as for the eTOX project) or split between the modeller and verifier. The scheme is implemented through four documents: 1) Model documentation along with field descriptions; 2) Data template for submission of data to verification; 3) Verification template for verifiers; 4) Guidance document on how to complete verification (Standard Operating Procedure). All of the documents are available as supplementary information and are also described below.

INSERT FIGURE 4 HERE

INSERT FIGURE 5 HERE

4.1.1 Model documentation along with field descriptions

The new model documentation retains the sections described in section 2 but field descriptions were made easily available to the modellers to allow them to better understand which fields were relevant for their models. This enables them to complete the documentation more thoroughly and indicate non-relevant fields with ‘N/A’.

4.1.2 Data template for submission of data to verification

The data template retained all of the fields described in section 2 but now stipulates that only SMILES can be used to submit compounds to overcome conversion problems encountered with InChi codes. This also negated the requirement for modellers to submit compounds as

SDfiles as there are a number of software packages (freely) available to perform a SMILES/SDF conversion.

4.1.3 Verification template for verifiers

The three main sections of the verification template remain but the order has been changed to make the process more efficient and to ensure verifiers do not need to repeat tasks when information is required in the different sections. The order of the template is now:

I. Assessment of data
II. Assessment of model implementation
III. Assessment of model documentation

Additional information has also been added to the template to provide guidance on what the verifier should be examining for each statement. For example, the 'accuracy of structures' statement is accompanied by the following information "*Datasets should be checked for agreement between chemical name, CAS, structure, etc. Any disagreement should be reported. A suggestion is to rank compounds via endpoint value and assess every 20th compound (5%), depending upon dataset size. The methodology implemented to assess structure accuracy is dependent on that available to the individual model verifier. As a minimum requirement, the structures within the dataset (or subset) should be compared to those within high quality online chemical databases, e.g. ChemSpider. Other suitable approaches may also be used.*"

The verification checklist remains as a useful tool for enabling the verifier to review all of the verification statements and assess whether a model should be considered 'verified'. This allows for the verifier to use their discretion where a model may have failed some insignificant criteria but they still feel the model should be given the verified status. For example if one or two of the statistics on model performance are missing but enough information is provided to give the user an understanding of model performance, then the verifier might perceive it to be appropriate to state that adequate performance statistics are provided and sign off the model as verified.

4.1.4 Guidance document on how to complete verification

This document can be used alongside the verification template to provide verifiers with all the information they require to complete the process. The verification process is described in four stages which relate to the three sections of the verification template. Additional actions which are (not) required for each stage are also described here. For example, it is noted that the verifiers need to convert the SMILES provided to SDFiles to run them through the models, and that it is not necessary to run all the test and training data through the model during stage 1 as this stage simply requires the presence of the data to be noted. Although some of these instructions seem obvious they have been included to ensure that assessments are consistent across numerous verifiers.

5. Conclusions

The requirement for predictive *in silico* models to be verified for use, particularly in regulatory settings, is clear. Work published over the past 10 years has shown how important this task is, but as yet there has been little practical guidance on how this might actually be achieved.

The work presented here has provided a clear scheme for model verification which can be applied across a variety of model types (e.g. structural alerts, QSAR, molecular interactions). The pilot study showed clear areas of improvement which have now been implemented within the eTOX project. The pilot study allowed initial ideas about the verification process to be tested and refined without impacting heavily on the on-going work of the project. It became obvious that it is not sufficient to supply verifiers with a single document to be completed. A standard operating procedure needs to be implemented to give modellers and verifiers a clear understanding of what is required. A graphical representation of the SOP is shown in fig. 4-5.

Overall testing ideas through the pilot study allowed a clear and concise scheme for model verification to be developed. These modifications are currently being tested within the eTOX project as the scheme described above is being used throughout the project to verify all models which have been produced. Once the verified models are available to users and there is a practical use-case, it is hoped that value of this work in a regulatory setting will become apparent.

In summary, the method for peer-verification of *in silico* models, as presented here, may be applied in a broad range of modelling scenarios and applied to other areas such as the evaluation of models submitted for publication. It is intended that using such a scheme to assess the scientific validity of models, will increase acceptance of the models and engender greater confidence in predictive methods.

6. Acknowledgements

This work has been funded in part by the eTOX project, grant agreement number 115002 under the Innovative Medicines Initiative Joint Undertaking (IMI-JU). The invaluable contribution, support and constructive feedback provided by the eTOX consortium is gratefully acknowledged.

7. References

[1] T.J. Hou, J. Wang, W. Zhang, W. Wang and X. Xu, Recent advances in computational prediction of drug absorption and permeability in drug discovery, Current Medicinal Chemistry, 13, (2006), 2653-2667.

[2] T.J. Hou and J. Wang, Structure – ADME relationship: still a long way to go? Expert Opinion on Drug Metabolism and Toxicology, 4, (2008), 759-771

[3] T.J. Hou, Y. Li, W. Zhang, and J. Wang, Recent developments of in silico predictions of intestinal absorption and oral bioavailability, Combinatorial Chemistry & High Throughput Screening, 12, (2009), 497-506;

[4] J.Y. Zhu, J. Wang, H. Yu, Y. Li and T.J. Hou, Recent developments of in silico predictions of oral bioavailability, Combinatorial Chemistry & High Throughput Screening, 14, (2011), 362-375

[5] European Union, Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/ EC and repealing Council Regulation (EEC) No. 793/93 and Commission Regulation (EC) No. 1488/94as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/ EEC, 93/105/EC and 2000/21/EC, Official Journal, L 396 (2006), pp. 1–850.

[6] Cosmetic Products Regulation (EC) No 1223/2009.

[7] http://www.efsa.europa.eu/en/efsajournal/doc/3638.pdf (Accessed July 2014) Scientific Report of EFSA, Modern Methodologies and tools for human hazard assessment of chemicals, EFSA journal, 12,(2014).

[8] M.T.D. Cronin and T.W. Schultz, Pitfalls in QSAR, Journal of Theoretical Chemistry (Theochem), (2003), 622:39–51.

[9] Organisation for Economic Cooperation and Development. Report from the expert group on (quantitative) structure-activity relationships ((Q)SARs) on the principles for the validation of (Q)SARs. (ENV/JM/MONO(2004)24). 2004.

[10] J.C. Dearden, M.T.D. Cronin, and K.L.E. Kaiser, How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR), SAR and QSAR in Environmental Research, 20, (2009), 241-266.

[11] T.R. Stouch, J.R. Kenyon, S.R. Johnson, X-Q Chen, A. Doweyko1 and Y Li, *In silico* ADME/Tox: why models fail, Journal of Computer-Aided Molecular Design, 17, (2003), 83–92.

[12] P.H. Judson, Towards establishing good practice in the use of computer prediction, The Quality Assurance Journal, 12, (2010), 120-125.

[13] R. Kristam, V.J. Gillet, R.A. Lewis, D. Thorner, Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst, Journal of Chemical Information and Modelling, 45, (2005), 461-476.

[14] ECHA. http://echa.europa.eu/web/guest/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment (Accessed August 2014).

[15] A.P. Worth, The role of QSAR methodology in the regulatory assessment of chemicals. In Recent Advances in QSAR Studies, T. Puzyn, J. Leszczynski and M.T. Cronin (Eds), (2010), Springer, pp. 367-382.

[16] D. Young, T. Martin, R. Venkatapathy, and P. Harten, Are the chemical structures in your QSAR correct?, QSAR and Combinatorial Science, 23, (2008), 1337-1345.

[17] B. Beck, and T. Geppert, Industrial applications of in silico ADMET, Journal of Molecular Modeling, 20, (2014), 2322-2336.

[18] K. Schneider, M. Schwarz, I. Burkholder, A. Kopp-Schneider, L. Edler, A. Kinsner-Ovaskainan, T. Hartung, and S. Hoffmann. "ToxRTool", a new tool to assess the reliability of toxicological data, Toxicology Letters, 189, (2009), 138-144.

[19] H.J. Klimisch, M. Andreae, and U. Tillmann, A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data, Regulatory Toxicology and Pharmacology, 25, (1997), 1-5.

[20] M. Nendza, T. Aldenberg, E. Benfenati, R. Benigni, M. Cronin, S. Escher, A. Fernandez, S. Gabbert, F. Giralt, M. Hewitt, M. Hrovat, S. Jeram, D. Kroese, J. Madden, I. Mangelsdorf, R. Rallo, A. Roncaglioni, E. Rorije, H. Segner, B. Simon-Hettich, and T. Vemeire, Data quality assessment for in silico methods: a survey of approaches and needs, in In Silico Toxicology: Principles and Applications, M.T.D. Cronin and J.C. Madden, eds., The Royal Society of Chemistry, Cambridge, (2010), pp. 59–118.

[21] J.C. Madden, Sources of chemical information, toxicity data and assessment of their quality, In Chemical Toxicity Prediction: Category Formation and Read-Across, M.T.D. Cronin, J.C. Madden, S.J.Enoch and D.W. Roberts (Eds), The Royal Society of Chemistry, Cambridge, pp. 98-126.

[22] K.R. Przybylak, J.C. Madden, M.T.D. Cronin, and M. Hewitt, Assessing toxicological data quality: basic principles, existing schemes and current limitations, SAR and QSAR in Environmental Research, 23, (2012), 435-459.

[23] http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/QRF (Accessed July 2014)

[24] http://www.etoxproject.eu/ (Accessed August 2014)

[25] K. Briggs, M. Cases, D.J. Heard, M. Pastor, F. Pognan, F. Sanz, C.H. Schwab, T. Steger-Hartmann, A. Sutter, D.K. Watson and J.D. Wichard, Inroads to predict *in vivo* toxicology – an Introduction to the eTOX Project, International Journal of Molecular Sciences, 13, (2012), 3820-3846.

[26] M. Cases, K. Briggs, T. Steger-Hartmann, F. Pognan, P. Marc, T. Kleinoder, C.H. Schwab, M. Pastor, J. Wichard, F. Sanz, The eTOX data-sharing project to advance *in silico* drug-induced toxicity prediction, International Journal of Molecular Sciences, 15, (2014), 21136-21154.

[27] V. Ruusmann, S. Sild and U. Maran, QSAR DataBank - an approach for the digital organization and archiving of QSAR model information. Journal of Cheminformatics, 14, (2014), 6-25.

[28] D. Fourches, E. Muratov, and A. Tropsha, Trust, but verify: on the importance of chemical structure curation in chemoinformatics and QSAR modeling research, Journal of Chemical Information and Modeling, 50, (2010), 1189-1204.

[29] I. Sushko, S. Novotarskyi, R. Körner, A.K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V.V. Prokopenko, V.Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I.I. Baskin, V.A. Palyulin, E.V. Radchenko, W.J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko and I.V. Tetko, Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. Journal of Computer-Aided Molecular Design, 25, (2011), 533-554.

[30] S. Brandmaier, W. Peijnenburg, M.K. Durjava, B. Kolar, P. Gramatica, E. Papa, B. Bhhatarai, S. Kovarich, S. Cassani, P.P. Roy, M. Rahmberg, T. Öberg, N. Jeliazkova, L. Golsteijn, M. Comber, L. Charochkina, S. Novotarskyi, I. Sushko, A. Abdelaziz, E. D'Onofrio, P. Kunwar, F. Ruggiu and I.V. Tetko, The QSPR-THESAURUS: the online platform of the CADASTER project. Alternatives to Laboratory Animals (ATLA), 42, (2014), 13-24.

[31] B. Hardy, N. Douglas, C. Helma, M. Rautenberg, N. Jeliazkova, V. Jeliazkov, I. Nikolova, R. Benigni, O. Tcheremenskaia, S. Kramer, T. Girschick, F. Buchwald, J. Wicker, A. Karwath, M. Gütlein, A. Maunz, H. Sarimveis, G. Melagraki, A. Afantitis, P. Sopasakis, D. Gallagher, V. Poroikov, D. Filimonov, A. Zakharov, A. Lagunin, T. Gloriozova, S. Novikov, N. Skvortsova, D. Druzhilovsky, S. Chawla, I. Ghosh, S. Ray, H. Patel and S. Escher, Collaborative development of predictive toxicology applications, Journal of Cheminformatics, 31, (2010), 7.

[32] C.M. Ellison, R. Sherhod, M.T. Cronin, S.J. Enoch, J.C. Madden, and P.N. Judson, Assessment of methods to define the applicability domain of structural alert models, Journal of Chemical Information and Modelling, 23, (2011), 975-985.

[33] P. Carrió, M. Pinto, G. Ecker, F. Sanz, and M. Pastor, Applicability Domain Analysis (ADAN): A robust method for assessing the reliability of drug property predictions, Journal of Chemical Information and Modeling, 54, (2014), 1500-1511.

Figure Legends:

Figure 1. General overview of the proposed model verification process.

Figure 2. The verification checklist to be completed by verifiers.

Figure 3. Basic architecture of how models are implemented in eTOXsys.

Figure 4. Overall process of model development and verification.

Figure 5. Specific requirements of the verifier

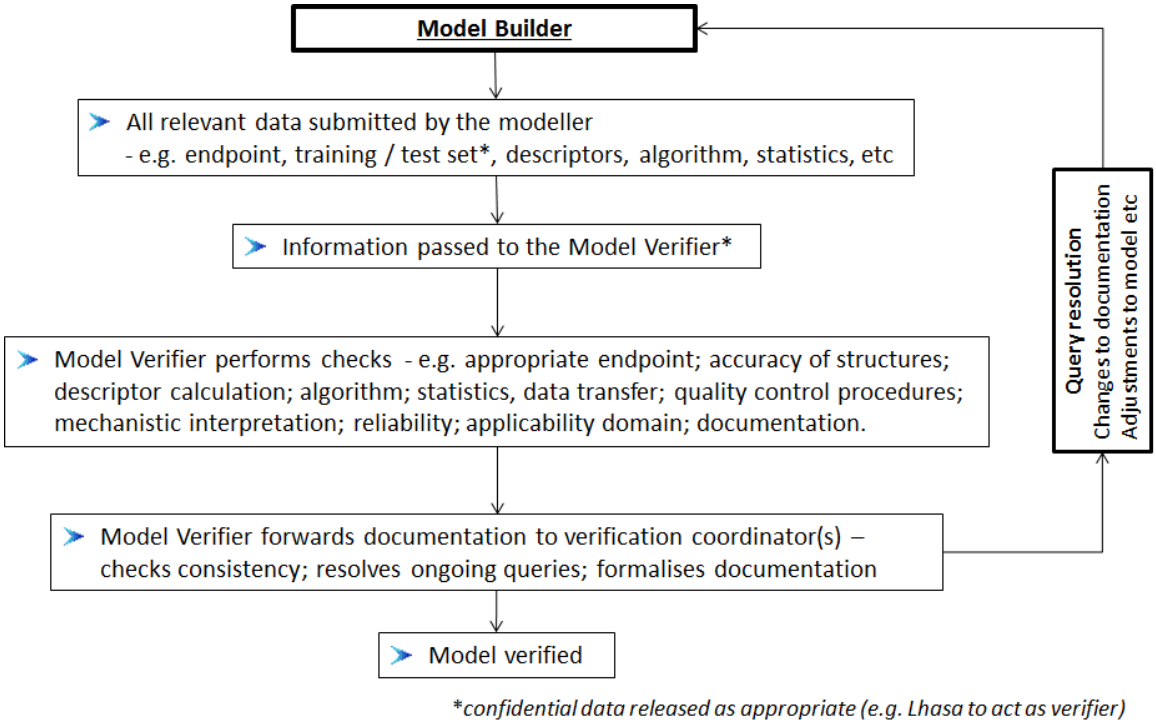
Model Builder
All relevant data submitted by the modeller
- e.g. endpoint, training / test set*, descriptors, algorithm, statistics, etc
Information passed to the Model Verifier*
Model Verifier performs checks - e.g. appropriate endpoint; accuracy of structures; descriptor calculation; algorithm; statistics, data transfer; quality control procedures; mechanistic interpretation; reliability; applicability domain; documentation.
Model Verifier forwards documentation to verification coordinator(s) – checks consistency; resolves ongoing queries; formalises documentation
Model verified
Query resolution
Changes to documentation
Adjustments to model etc
*confidential data released as appropriate (e.g. Lhasa to act as verifier)

Figure 1

Model Verifier Details	
Name	
Institution	
Date of verification	

Model being verified	
Model Name	
Model ID	
Model Developer (institution)	

Model summary information / documentation		Present/completed (YES [Y] or NO [N])	Verifier comment
Component	**Information Source(s)**		
Presence of eTOXvault entry	eTOXvault		
Model summary details present	eTOXvault		
Modelling partner details present	eTOXvault		
Executive summary present	eTOXsys		
Model training set supplied	Dataset(s) provided		
Test set(s) supplied	Dataset(s) provided		
Adequacy for purpose			
Interpretation statement adequate	Executive summary		
Endpoint information fit for purpose	eTOXvault		
Supporting statistics available	eTOXvault		
Reliability index given	eTOXvault		
Applicability domain assessment provided	eTOXvault		
Accurate structures provided	Dataset(s) provided		
Model implementation			
Model implemented in eTOXsys	eTOXsys		
eTOXsys implementation stable	eTOXsys		
Model robust to input files	eTOXsys		
Output consistent	eTOXsys		

Verification queries	
Please detail any queries below:	

Model Verification Status	

Figure 2

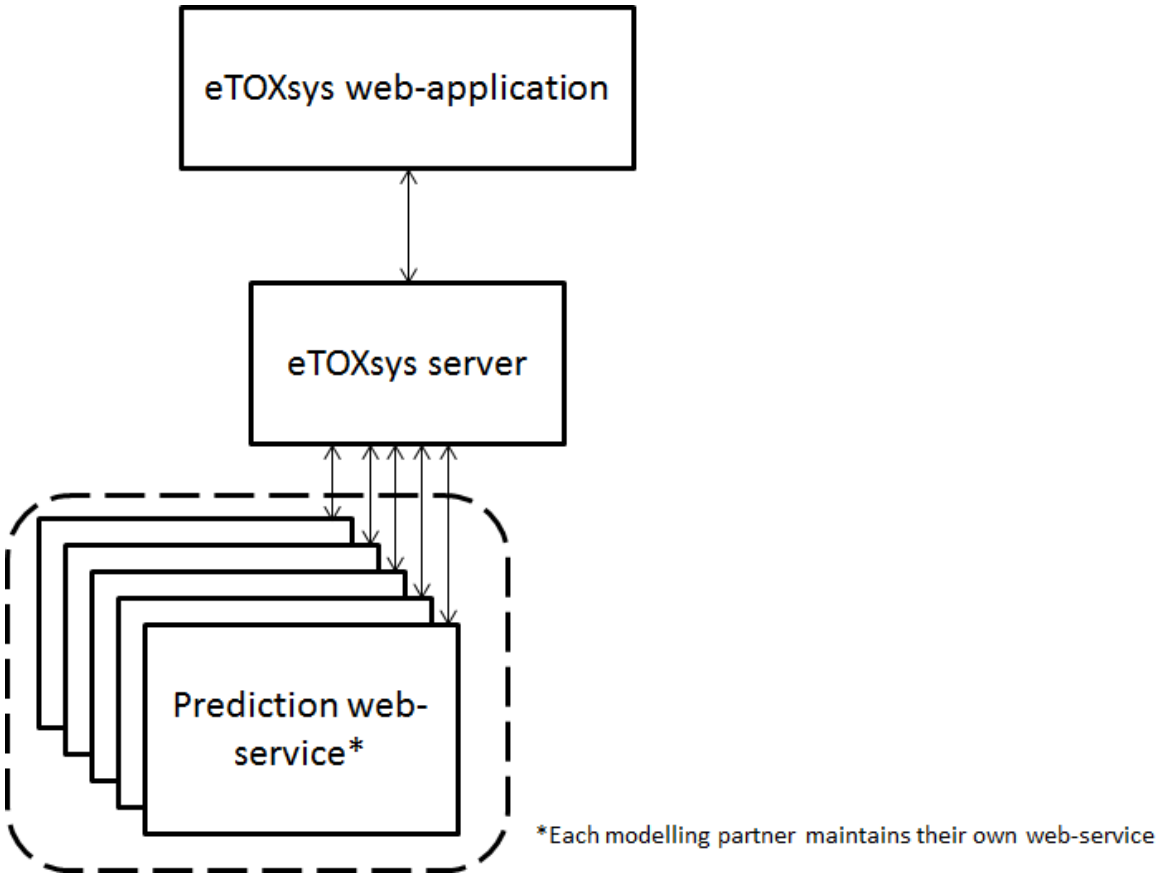
eTOXsys web-application
eTOXsys server
Prediction web-service*
*Each modelling partner maintains their own web-service

Figure 3

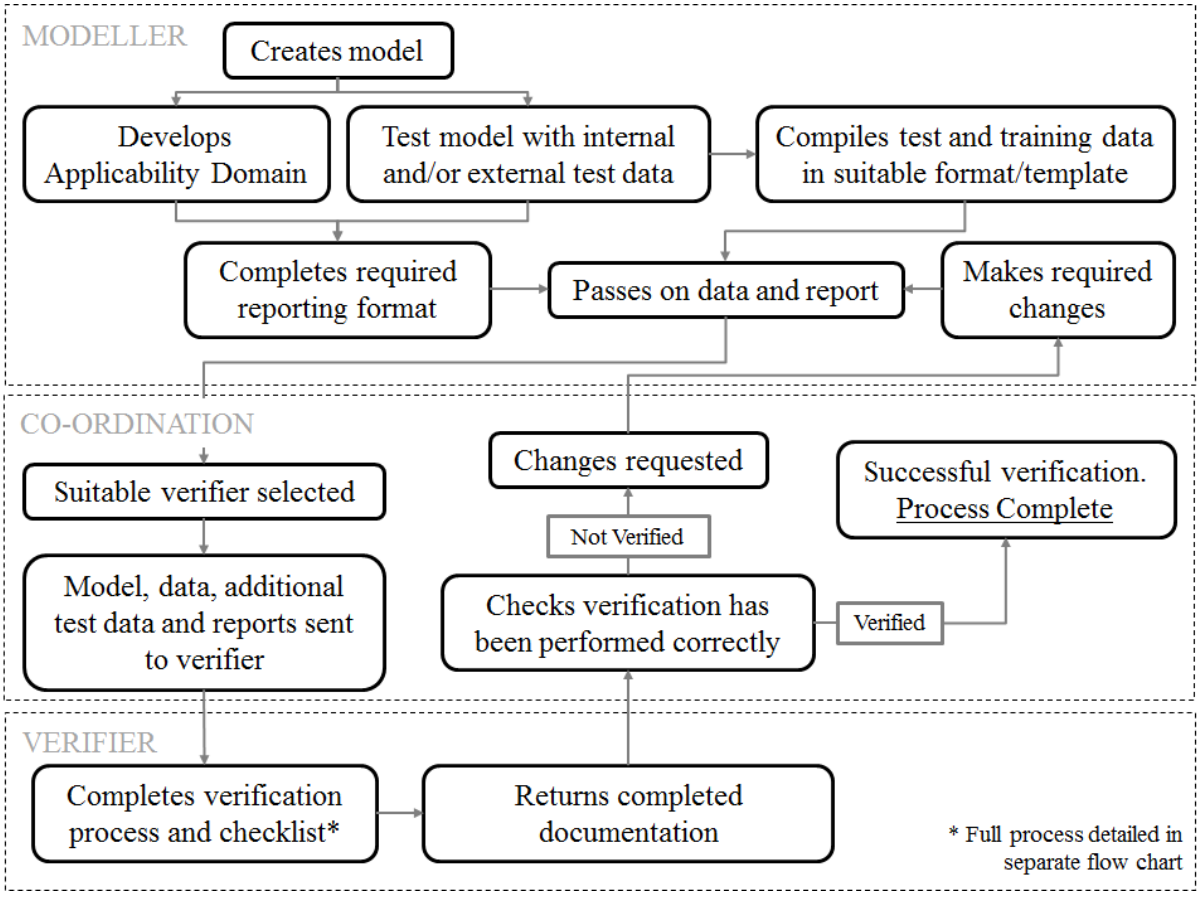
MODELLER
Creates model
Develops Applicability Domain
Test model with internal and/or external test data
Compiles test and training data in suitable format/template
Completes required reporting format
Passes on data and report
Makes required changes
CO-ORDINATION
Suitable verifier selected
Changes requested
Successful verification. Process Complete
Not Verified
Model, data, additional test data and reports sent to verifier
Checks verification has been performed correctly
Verified
VERIFIER
Completes verification process and checklist*
Returns completed documentation
* Full process detailed in separate flow chart

Figure 4

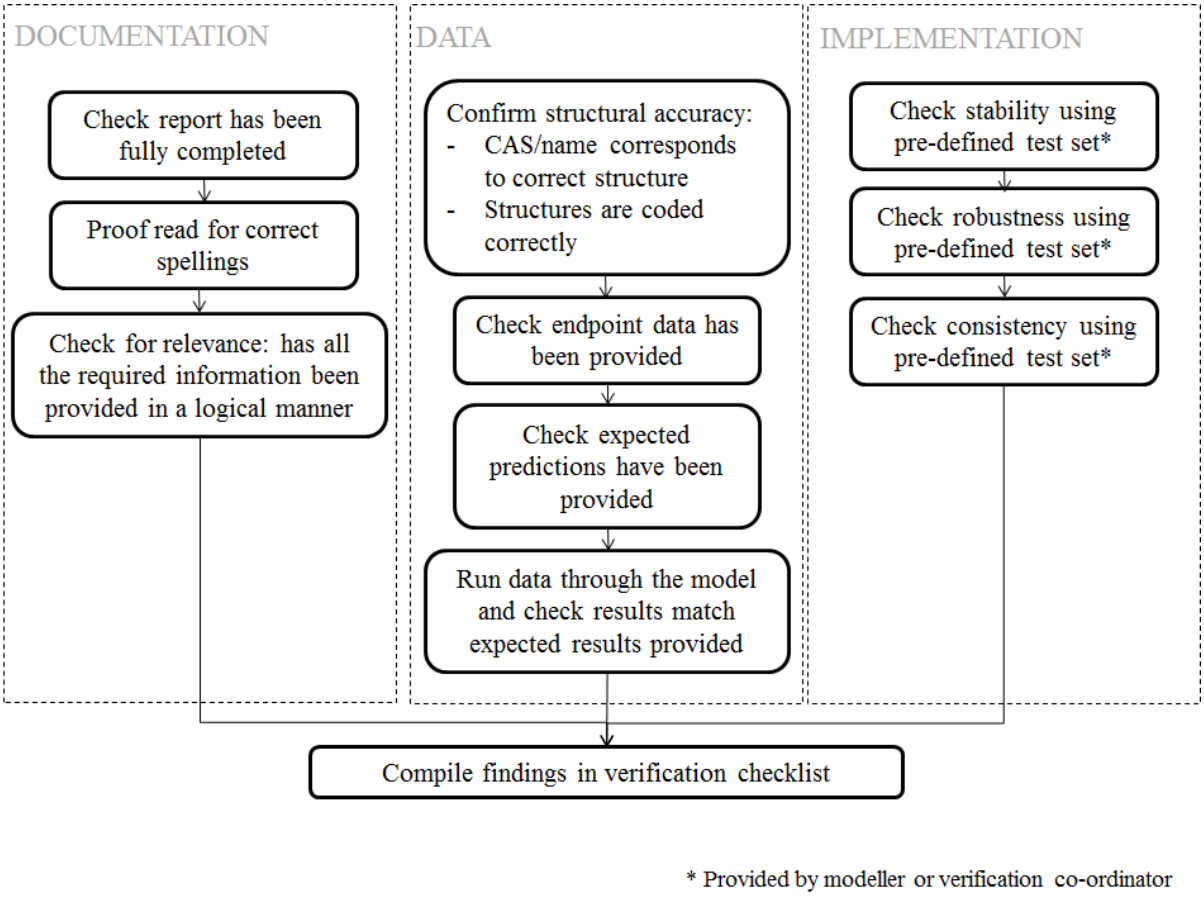
DOCUMENTATION
Check report has been fully completed
Proof read for correct spellings
Check for relevance: has all the required information been provided in a logical manner
DATA
Confirm structural accuracy:
- CAS/name corresponds to correct structure
- Structures are coded correctly
Check endpoint data has been provided
Check expected predictions have been provided
Run data through the model and check results match expected results provided
IMPLEMENTATION
Check stability using pre-defined test set*
Check robustness using pre-defined test set*
Check consistency using pre-defined test set*
Compile findings in verification checklist
* Provided by modeller or verification co-ordinator

Figure 5

Table 1. Summary of the resources and brief provided to each model verifier participating the verification pilot study.

Provided to the verifier	Verifiers Brief

An email containing:
• The identity of the model to be verified.
  ○ Model name and eTOXsys ID
  ○ Model developer
• The model verification template
• The relevant model training/test sets
• The expected model predictions
	Each verifier was asked to use the verification template provided to assess the model they were assigned.
	All verifiers had been involved in developing the draft template to some degree and had seen this document beforehand.
	Each modeller was given two months in which to perform the verification task.

Table 2. Summary of the model types participating in the eTOX verification pilot study.

Model ID	Model Type	Endpoint	Model Developer	Model Verifier
1	kNN	CYP 3A4 affinity	Chemotargets	Novartis
2	Decision tree and 3D QSAR	Drug-induced phospholipidosis	Fundació IMIM	Bayer HealthCare
3	Ligand-based pharmacophore	hERG inhibition	Inte:Ligand	Dutch Technical University
4	Ligand-based pharmacophore	Acetylcholinesterase inhibition	Inte:Ligand	GlaxoSmithKline
5	Expert system	Hepatotoxicity	Lhasa Limited	Lhasa Limited
6	Structural alerts	Drug-induced phospholipidosis	Liverpool John Moores Univesity	Molecular Networks
7	Classification SVM	Plasma protein binding	Lead Molecular Design	Vrije Universiteit Amsterdam
8	Consensus kNN	Human total clearance	Molecular Networks	Fundació IMIM
9	Consensus kNN	Drug induced phospholipidosis	Molecular Networks	Chemotargerts
10	Decision tree	Predominant CYP isoform (3A4, 2D6 and 2C9)	Molecular Networks	Liverpool John Moores University