EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection
The European Reference Laboratory for Alternative Methods to Animal Testing (EURL-ECVAM)

ECVAM SCIENTIFIC ADVISORY COMMITTEE (ESAC)

ESAC Working Group Peer Review Consensus Report

on

the ECVAM-coordinated follow-up study to assess the predictive capacity of the already validated Neutral Red Uptake cytotoxicity assay for acute oral toxicity testing

Title page information	
File name	ESAC-WG_REPORT_FINAL-CLEAN.doc
Abbreviated title of ESAC request	Follow-up study to assess the PC of the Neutral Red Uptake cytotoxicity assay for acute oral toxicity testing
Relating to ESAC REQUEST Nr.	2011-02
Request discussed through	ESAC34 plenary March 2011

TABLE OF CONTENTS

ESAC Working Group

This report was prepared by the "ESAC Working Group 3T3 NRU" (ESAC WG) with support from the ESAC Secretariat. The ESAC WG was charged with conducting preparing a detailed scientific peer review of the ECVAM follow-up study concerning the predictive capacity of the already validated 3T3 NRU assay for identifying non-classified substances (acute oral toxicity) using the 2000 mg/kg b.w. threshold of the EU CLP regulation.

The ESAC WG had been set up by the ESAC during its meeting on March 2011. Basis for the scientific review was the ECVAM request to ESAC concerning the scientific review (ESAC request 2011-02, see Annex 5).

The ESAC WG conducted the peer review from June 2011 to February 2012. This report was endorsed by the ESAC WG on 29. February 2012 and represents the consensus view of the ESAC WG.

This ESAC WG peer review consensus report was endorsed by the ESAC on 20. March 2012.

The ESAC WG had the following members:

- Dr. Neil CARMICHAEL (ESAC member, Chair of ESAC WG)
- Prof. Lucio COSTA (ESAC member)
- Dr. Ian DEWHURST (external expert)
- Dr. Eugene ELMORE (external expert, proposed by NICEATM)
- Prof. Annette KOPP-SCHNEIDER (external expert, proposed by ECVAM)
- Prof. Kristin SCHIRMER (ESAC member)

ESAC Secretariat:

- Dr. Claudius GRIESINGER (EC-ECVAM, ESAC Coordination/Scientific Secretariat)
- Dr. Anita TUOMAINEN (EC-ECVAM, specific support to ESAC Secretariat)

NOTE ON THIS REPORTING TEMPLATE

The template follows the ECVAM modular approach and allows at the same time for the description of the analysis and conclusions concerning more specific questions. The template was approved by the ESAC through written procedure on 29 October 2010.

The template can be used for various types of validation studies (e.g. prospective full studies, retrospective studies, performance-based studies and prevalidation studies).

Depending on the study type and the objective of the study, not all sections may be applicable. However, for reasons of consistency and to clearly identify which information requirements have not been sufficiently addressed by a specific study, this template is uniformly used for the evaluation of validation studies.

- Explanatory notes to the paragraph titles (in green) have been added on 17 November 2010. These notes provide guidance on the type of information / analysis expected under each section. Depending on the purpose and scope of the study to be reviewed, some of the aspects mentioned in the explanatory notes may not be applicable or only be applicable to some extent. Moreover, the explanatory notes are not intended to represent an exhaustive list of possible issues to be addressed under the respective heading, but are thought to provide some guidance with respect to the considerations typically expected.
- For all of the template's numbered sections the summary view of ESAC WG is given in bold followed by more detailed comments ("general observations" and "specific observations").

ABBREVIATIONS USED IN THE DOCUMENT

BLR	Between-laboratory reproducibility
ECVAM	European Centre for the Validation of Alternative Methods
ESAC	ECVAM Scientific Advisory Committee
ESAC WG	ESAC Working Group
FN	False Negative
FP	False Positive
GCCP	Good Cell Culture Practice
GLP	Good Laboratory Practice
PC	Positive Control
TN	True Negative
TP	True Positive
SOP	Standard Operating Procedure (used here equivalent to 'protocol')
VC	Vehicle Control
VMT	Validation Management Team
VSR	Validation Study Report
WLR	Within-laboratory reproducibility

Executive summary

Following a request from ECVAM to ESAC concerning the scientific peer review of an ECVAM-coordinated study on the predictive capacity of the 3T3 NRU assay for acute toxicity testing, an ESAC Working Group (ESAC WG) was set up to review the Validation Study Report (VSR). The follow-up study had addressed the test method's capacity to identify specifically negatives (i.e. substances not requiring classification for acute oral toxicity) on the basis of the 2 000 mg/kg b.w. threshold implemented through the EU Classification, Labelling and Packaging (CLP) regulation. The mandate of the ESAC and its WG included an assessment of the study design, study conduct and the conclusions drawn by the Validation Management Team, including the possible future use of the 3T3 NRU assay within an integrated testing strategy to screen for negatives amongst substances requiring data on acute oral toxicity. The ESAC WG met once in person at ECVAM in Ispra from 12-14 September 2011. The ESAC WG reviewed the ECVAM follow-up study report (referred in the following as Validation Study Report = VSR) and took also other existing information into consideration such as the previous NICEATM/ECVAM validation study (NIH, 2006) as well as an analysis of the possible prevalence of acute oral toxicants amongst industrial chemicals performed by ECVAM in 2009 (Bulgheroni et al., 2009).

Background to this follow-up study was that the previous joint NICEATM/ECVAM validation study had established good reproducibility of the 3T3 NRU assay already but found that it was unable to predict the different GHS (Globally Harmonised System) categories currently used for classification and labelling of acute oral toxicity. The previous NICEATM/ECVAM study had concluded that the test could be used, in a weight-of-evidence approach, to determine the starting dose for the in vivo experiment (OECD, 2010).

In contrast, the follow-up study explored the usefulness of previously assessed prediction models to predict the absence of acute oral toxicity potential (i.e. LD50 >= 2000 mg/kg b.w.). Importantly, according to the study objective, only negative test results are to be considered. In contrast, taking into account the results of the NICEATM/ECVAM study and the rather high sensitivity of the assay (thus yielding a high false positive rate), positive test outcomes are not to be considered but followed up by appropriate other means of testing.

The study under review had evaluated three different prediction models: two based on a regression analysis had already been used during the NICEATM/ECVAM study, while a ROC-based model had been developed for the purpose of the study. Moreover, the study assessed three different protocols: the original protocol used in the NICEATM/ECVAM study (assessed in the HSL laboratory), a shortened version of this protocol (assessed in the IIVS laboratory) and a version adapted for use on an automated platform (JRC laboratory).

The conclusions of the ESAC WG are as follows:

(1) the scientific work presented was considered of good quality and the conclusions largely justified by the data obtained and overall plausible in the context of existing information. However, some minor weaknesses in study design were observed such as (a) exclusion of CMR substances (=carcinogenic, mutagenic, reproductive toxicants) may have led to a rather low number of highly toxic substances in the set and which excluded a priori substances outside the known applicability domain of the test method, although the testing of such substances might have added valuable information; (b) the derivation of one single reference

LD50 values and thus reference classification in cases chemicals had several reference values. Single reference values were obtained through averaging, but should have rather been determined by using the median. Moreover, obviously outlying reference values were not excluded through expert judgement. However, the ESAC WG acknowledges that the reference values and thus the overall outcome would not have significantly changed had another approach been chosen.

(2) The ESAC WG concluded, in agreement with the VSR, that the assay has a high sensitivity, irrespective of the protocol used, allowing to reliably identify negatives, i.e. test substances non-classified in the rat in vivo reference test. As suggested in the VSR, the test may reduce the need for traditional testing by up to ca 40%. This number is based on an assumed prevalence of non-classified substances of 87% (based on an analysis of the EU New Chemicals Database: Bulgheroni et al., 2009). The ESAC WG holds however that this figure may be considerably lower (depending on the real prevalence currently not known) and is moreover based on a possibly restricted applicability domain relating to acute toxicants that trigger adverse effects through basal cytotoxicity, while more organ-specific mechanisms (e.g. specific voltage- or ligand-activated channels etc.) are not modelled in the 3T3 assay.

(3) In order to ensure appropriate future use of the test, the ESAC WG recommends that suitable exclusion criteria be developed allowing to decide whether substances can be tested in the assay. Such exclusion criteria should be based on the known physico-chemical properties, the structure and the possible structure-activity relationships taking the intended purpose of the substance into account, i.e. was it purposefully developed to interfere with a biological mechanism (e.g. pesticides, biocides, pharmaceuticals etc.) or is there no such indication.

(4) As all three different protocols (the original one, the abridged protocol as well as the automated one) had obtained comparable performance (reproducibility and predictive capacity), the ESAC WG concludes that the 3T3 assay is amenable to both, simplification and automated testing, making it a candidate assay for economical and fast high-throughput screening approaches.

(5) In agreement with the VSR, the ESAC WG concludes that the test, due to its limitations regarding specific mechanisms of action and its inability to resolve the various categories of acute toxicants (as already shown in the joint NICEATM/ECVAM study) cannot be used as a stand-alone replacement test but may be a useful component of an integrated testing strategy for acute oral toxicity assessment.

(6) Finally the ESAC WG recommends to consider amending the current SOP, such as to include the improved solubility protocol used by the IIVS laboratory which may allow to use higher testing concentrations.

1. Data collection

1.1 Information / data sources used

NOTE: (Pre)validation studies typically make use of existing data, e.g. either as reference data (prospective studies) or as reference data and testing data as well (retrospective study). Have other data been used during the studies that were not generated during the study? If yes, for which purpose (e.g. reference data etc.)? What were that data sources?

Several sources of existing data have been used:

1. Data sources for selection of chemicals to be included in the study

Chemical selection is described as a top-down approach. From a starting pool, chemicals were successively eliminated according to predefined criteria for chemical selection (see sections 3.1.1 to 3.1.3 of the VSR).

For classified chemicals, the starting pool of chemicals was the list from Annex I Dir 67/548/EEC. The pool initially contained 1020 chemicals and, after application of several selection criteria, had been reduced to 30 chemicals (Table 1 of VSR, page 17).

For unclassified chemicals the starting pool of chemicals (n=691) was the list of chemicals found in the ORATS database and the Registry of Cytotoxicity (RC) with LD_{50} > 2 000 mg/kg bw. This set was reduced to 26 chemicals following application of the selection criteria (Table 1 of VSR, page 18).

2. Data sources for in vivo reference LD_{50} values

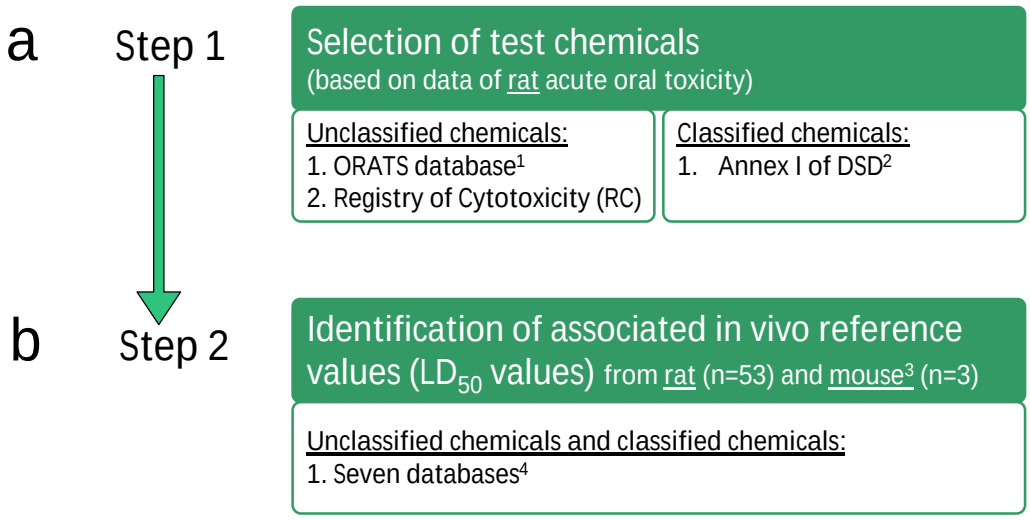
Once suitable chemicals had been selected, the associated reference data (in vivo LD_{50} data) were identified. The study uses in vivo LD_{50} reference values to determine the classification of test items according to a binary categorisation (non-classified vs classified). Notably only one of the categories 'non-classified' is suggested to be taken as a useful test outcome.

In the VSR, the authors reference only two exemplary internet data bases as source for LD_{50} data collection These are ChemIDplus and HSCB (both sites are link to the Toxnet page which and this link has been mentioned in the VSR. In addition, the VSR report states that original references were used to support reference data collection (including checking for repetitions in citations).

For the majority of chemicals rat reference data were used (n=53). Only in three cases, when rat data did not show finite LD_{50} values, mouse LD_{50} values were used.

The ESAC WG is of the opinion that this procedure was not sufficiently clearly explained in the VSR and that, furthermore, the 3 (out of 56) chemicals for which rat LD_{50} values of sufficient quality were lacking, should have been rather excluded from the validation testing set.

Figure 1: Schematic depiction of the selection of test chemicals and associated reference values. In a first step, chemicals with specific desired properties were extracted from 3 sets of databases which differed between unclassified and classified chemicals. In a second step, the associated LD_{50} values were identified through searching seven web-based databases for reported LD_{50} values associated with the chemicals identified in step 1.



1) ORATS = Online European Risk Assessment Tracking System
2) DSD = the so-called "Dangerous Substance Directive", Directive 67/548/EEC
3) Mouse data were used only if the rat in vivo values were not finite numbers.
4) The databases were: (1) ChemIDplus; (2) IUCLID (=-International Uniform Chemical Information Database); (3) RTECS (=Registry of Toxic Effects for Chemical Substances); (4) Merck index; (5) Sax' Dangerous Properties of Industrial Materials; (6) HSDB (=Hazardous Substances Databank); (7) EU Risk Assessment Report.
Moreover, in the case of 5 chemicals, unidentified sources were used for the determination (by averaging) of the LD50 reference value (4 of the five are unclassified, 1 is classified). For further 15 chemicals unidentified sources contributed to the determination of the reference LD50 value.

3. Data sources for deriving threshold values from a Receiver Operating Characteristic (ROC) analysis

The data sets used for establishment of prediction models are clearly defined. Two of the three prediction models were taken from the previous NICEATM/ECVAM validation study (these were the millimole and the weight regression prediction models).

These prediction models had already been used in the context of the NICEATM/ECAVM validation study that aimed at analysing whether LD50 values could be predicted with accuracy as to allow sufficiently correct classification of substances according to the five GHS hazard classes of acute toxicity. In total 72 substances had been used in the previous NICEATM/ECVAM study and the prediction models had been applied to these substances.

The third prediction model used in the current study was based on a ROC analysis using the 540 chemicals in the Registry of Cytotoxicity.

The main prerequisite for constructing prediction models is the availability of IC_{50} as well as LD_{50} information of suitable quality in order to allow to correlate one with the other and test to which extent LD_{50} can be extrapolated from IC_{50} values.

1.2 Search strategy

NOTE: How was the search for existing data planned, organised and executed? Has a search strategy been described and consistently applied?

Data sources for in vivo reference, i.e. LD_{50} estimation

No search strategy is described for identifying suitable databases. However, no search strategy is necessary in that case since, for the endpoint of acute oral toxicity, a wealth of databases is available.

However, no explanation is provided as to why a specific set of databases have been selected. Only two web-based databases were mentioned as examples in the VSR (paragraph 4.1) and it is unclear whether and if so, which other sources were used. As an outcome of the review, this has been clarified in the VSR and amended accordingly. Consequently, the data sources used for in vivo references were not transparent. Checking of a random sample of chemicals from Table 4 reveals that probably other sources were used in addition to the ChemIDplus, HSCB (and Toxnet) websites. Table 4 has now, as a result of the ongoing ESAC WG review, been amended in view of identifying for each test item the data sources (i.e. web-based databases) used.

The ESAC WG observes that the following issues related to the reference data search were not sufficiently clear in the VSR:

- Which other sources (apart from those mentioned in paragraph 4.1) were used.
- Whether all chosen databases were systematically interrogated for reference data (i.e. in all cases).
- Whether recourse to original references was systematic.
- How original references were found (i.e. search strategy for original publications of LD_{50} studies)

1.3 Selection criteria applied to the available information

NOTE: Have consistent evaluation/decision criteria been pre-defined and applied in order to select the data and has data selection been explained in a transparent manner?

1.3.1 Chemical selection

(a) Data sources for selection of chemicals to be included in the study

Data selection has been clearly defined.

For classified chemicals the elimination criteria were in order of application:

- Non-existence of EC number
- Classification as CMR, E, O, F, C, T, T+ (by alternative dose routes), T+ (by chronic exposure (R48,R33)
- Mixed CAS
- Pesticides
- Pharmaceuticals
- Chemicals used in previous studies
- Chemicals with inconsistent or missing LD_{50} values
- Chemicals not available in SIGMA
- Chemicals potentially difficult to handle (see paragraph 3.1.2 page 16 of VSR), however, no further specification is provided what this means)
- Insoluble metals
- Chemicals with uncertain classification

Observations by the ESAC WG regarding these criteria:

- It is not sufficiently clear what the rationale was for excluding CMRs and other specific toxicants from the testing set. This may limit the testing set and exclude categories that may be important to assess for acute toxicity during a validation study (e.g. mutagens, which are normally very toxic).
- Some exclusion criteria are not well explained, e.g. exclusion of chemicals "difficult to handle" (based on physico-chemical properties). Furthermore, in the discussion of the results it is stated that some chemicals were difficult. The ESAC WG is of the opinion, that it would possibly have been better to exclude such chemicals in the first place?

(b) For unclassified chemicals the elimination criteria were in order of application

- Chemicals not available in SIGMA
- Flammable and highly flammable chemicals
- If comparison of LD_{50} value between RC and Annex I showed large differences (No quantification was given concerning what was considered large)

Observations by the ESAC WG reg. these criteria:

- From the report it appears that different criteria were used for selecting classified and non-classified chemicals. This is probably due to the different information available in Annex I of the Dangerous Substance Directive as regards classified and non-classified substances.
- There was some degree of confusion regarding the selection criteria as the text in the body of the report and the information in the table were not entirely consistent.

- The source of physicochemical properties is not clear (e.g. Table 2). Where these data simply taken over from the databases?

1.3.2 Data sources for in vivo LD_{50} reference data

The criteria for inclusion of LD50 values from original citations which are stated are

(a) Unit of LD50 (mg/kg),

(b) Rat or mice data,

(c) Oral or gavage administration.

Reference LD_{50} values are given in Table 4, but without a list of respective references (i.e. databases and original reference whenever used). As a result, these references have been added and are included in an amended version of the VSR.

Table 4 shows mice and rat oral LD_{50} values. Checking the entries in Table 4 outlines that for classification into C/UC, only rat LD_{50} values were considered. In retrospect it appears that inclusion of mouse data (in cases where no finite rat LD_{50} values were available) did not provide any additional information that could not also have been derived from the non-finite rat data, i.e. the final classification of the chemical.

1.3.3 Data sources for training set for prediction model

The data for constructing the ROC-based prediction model were the RC set of chemicals (n=540). No selection criteria where hence applied.

2. Study objective and design

2.1 Clarity of the definition of the study objective

NOTE: Is the objective of the study clearly and comprehensibly defined?

The objective stated in the report (Section 1.4, page 9) "was to assess the ability of the 3T3 NRU test method to discriminate between classified (toxic/hazardous) (LD_{50}<2 000 mg/kg) and unclassified (LD_{50} > 2 000 mg/kg) chemicals according to the current EU CLP system for acute toxicity". The objective of the study is clear. This test having been shown in the previous NICEATM/ECVAM validation study to have poor predictivity for precise LD_{50} acute toxicity values, has been re-targeted at a much more limited objective i.e. the discrimination between no acute toxicity (by EU CLP definition) and all other classes of toxicity. Importantly, the high sensitivity of the test method is used to identify negatives with some degree of certainty. Thus, only negative test results (=suggested to be non acutely toxic) are proposed to be used as outcomes of the test. In contrast, positive test results are not considered as the a) lack sub-classification according to GHS categories and b) contain many false positives due to the high sensitivity of the test method.

2.2 Analysis of the scientific rationale provided

NOTE: Is the scientific rationale for the test method AND (consequently) for conducting the study clearly explained? How does the test method contribute scientifically to the scientific understanding / prediction of the specified health/environmental effect or aspects of it?

The scientific rationale for the test method, conducting the study and a sense of the strategy for the study were clearly explained (Sections 5.1 and 5.2) of the VSR.

The previous data (NICEATM/ECVAM) summarized in the report (Section 5.1) show that chemicals used in the previous study were selected to distinguish the five toxicity classes as well as the non-classified category according to the GHS but were not balanced between classified and unclassified chemicals of the EC CLP classification scheme for acute oral toxicity. Section 5.2 summarizes the purpose as to assess the capacity of the assay to provide a "yes/no" answer for the two categories. If the method can be shown to be predictive for unclassified chemicals, the information can be used as part of a tiered approach to minimize or eliminate the use of animals for testing unclassified chemicals. Testing in animals would only be required for all chemicals that were positive in the 3T3 NRU test (true positives or false positives). Section 1.2 reviews the historical data that would suggest that in vitro based cytotoxicity assays may predict human acute toxicity better than rodent acute toxicity studies. The WG observes that, in the VSR, there is no real attempt to explain why cytotoxicity should be expected to be a generally predictive surrogate for acute systemic toxicity. The WG further is of the view that the method contributes little to the understanding of organ-specific mechanisms of acute toxicity other than basal cytotoxicity, affecting – universally – all cells.

2.3 Analysis of the regulatory rationale provided

NOTE: Is a regulatory rationale specified, i.e. a specific application of the test method for purposes of generating data with respect to regulatory requirements as specified in legislation or internationally agreed guidelines etc.? If so, how does the study and its objective and design relate to this regulatory rationale?

The regulatory requirements for the EC CLP acute oral toxicity categories and the GHS acute oral toxicity categories were clearly explained (Section 1.1). The differences in classification between the EC CLP and GHS were clearly explained and the potential for prediction of each was considered during the analysis. The study targeted a balanced number of chemicals that were considered to be classified or unclassified in the EC CLP scheme based on their LD_{50}s. Within the unclassified group, the study also included chemicals that fall into category 5 and unclassified of the GHS to allow consideration of the response predictions for the 2 000 to <=5 000 mg/kg (category 5) and >5 000 mg/kg (unclassified) GHS classifications.

According to the VSR, this test may, if it proves reliable, lead to a reduction in animal use for classification purposes.

2.4 Appropriateness of the study design

NOTE: This includes an analysis of the selection of test items, the number of test items, the number of laboratories involved in the study, retesting in case of unqualified tests and other technical aspects of the study.

Background

The 3T3 NRU assay was previously validated as a result of the NICEATM/ECVAM validation study carried out between 2002 and 2005. Thus, reproducibility within and between laboratory, and transferability had already been assessed. The rationale for this study was to analyse the predictive capacity specifically for identifying unclassified chemicals (in contrast to the previous study which had assessed the PC for predicting all acute toxicity classes). For this purpose, according to the ECVAM modular approach, only one laboratory was needed. The laboratory chosen for this was the Health and Safety laboratory in the UK, which was awarded a contract by ECVAM for this purpose. Though this would have been sufficient, two additional laboratories were involved in this study. One was the IHCP at JRC, whose major objective was that of testing chemicals in an automated fashion. The second laboratory was a private, non-profit enterprise (Institute for In Vitro Sciences) in the USA, which carried out a simplified version of the protocol. Thus the study aimed at assessing, with one set of chemicals, (a) the possible use of the assay as validated for identifying substances not requiring classification, (b) whether the assay is amenable to automation and (c) to simplification.

The assay was intended to provide, in a rapid, relatively inexpensive, and animal free manner, a yes or no answer, based on cytotoxicity, and related to acute oral toxicity, and hence appropriate classification of chemicals. Cytotoxicity as a mean of determining acute toxicity has its intrinsic limitations, some of which are discussed on p. 46-47 and Section 11 of the report.

Transferability

The transferability of the protocol was confirmed by HSL using 9 chemicals (Section 8.1). Also, JRC conducted a 12 chemical analysis to show transferability using an automated approach (Section 8.2.)

SOP issues (solubility / volatility)

The solubility was assessed independently by each laboratory using the ICCVAM protocol (Annex F of VSR). A summary of the solubility results for all agents in each laboratory are presented in Table 6 of the report. Difficulties were reported including precipitation of the chemicals in the treatment medium or volatility. The solubility protocol specified an incubation of 5 to 60 minutes for chemicals that were insoluble. One laboratory included a 3 hr incubation at 37°C to increase solubility. When agents produced precipitates, IIVS sonicated and heated the agents to get homogenous suspensions at 400 to 800 µg/mL in culture medium. For all labs, some agents were tested as precipitates at one or more concentrations. Table 8 summarizes the test chemicals that formed precipitates in each laboratory. The report comments on the testing of precipitates in each laboratory and the possible resulting consequences, false positive and/or false negative agents. This issue was previously commented on in the Peer Report (June 2006) from the ICCVAM/NICEATM study report Appendix B, Section 1.3 Solubility Recommendations. Thus, the solubility problems experienced by the laboratories were similar to the previous 3T3 validation study.

Volatility was addressed by each laboratory. Variability was observed between the laboratories as to which chemicals were found to be volatile (Table 9). Notably, during normal testing conditions, volatility of substances should be known based on the physicochemical properties. However, during the validation study (due to blinding of the substances), it was not known by the participating laboratories whether or not individual substances were volatile. The issue of volatility as handled in this study is thus a consequence of the specific experimental conditions of a validation study.

Test items:

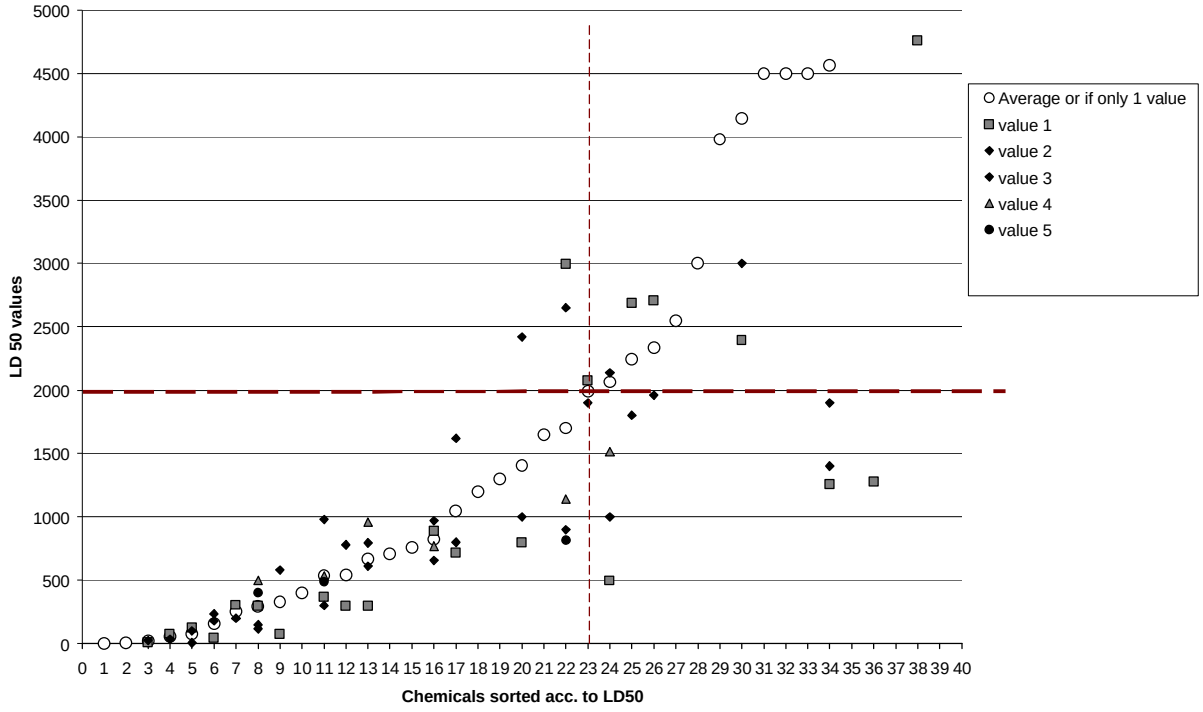
The choice of chemicals is discussed in detail in Section 5.1 of the report. A number of chemicals were excluded a priori (see section 1.3.1). These included any chemical previously used in the NICEATM/ECVAM validation study and in the ACuteTox integrated Project, pesticides, any pharmaceuticals. Two natural highly toxic alkaloids (brucine and aconitine) were however included. The rationale for this is was not sufficiently clear from the VSR.

The chemicals (total =56) were chosen so that ca 50% had oral LD_{50} > 2 000 mg/kg, and ca 50% had oral LD_{50} < 2 000 mg/kg, with a balance between solid and liquid. Through a series of evaluations, an initial number of more than 300 chemicals, led to 30 chemicals which were chosen as those "classified". Starting from more than 500 chemicals, a final number of 26 chemicals among those with LD_{50} > 2 000 mg/kg. The specific acute oral toxicities of the final 56 chemicals are discussed on Section 4.2 of the report.

The number of test chemicals is rather small to judge the False Negative Rate (i.e. 1 – sensitivity, with sensitivity being the True Positive Rate), probably the most important criterion. The numbers of compounds in the critical classes were: 16 in category 4 (the class of LD_{50} below 2 000 mg/Kg) and 11 in GHS class 5 the class immediately above the chosen cut off. The most critical distinction for this test is between these two classes. This number is further reduced due to censorship and other factors to give total numbers of compounds in the analysis from 44-54 (millimole regression model) and 40-54 (weight regression model) depending on the laboratory. The ESAC WG however acknowledges that the variability of the reference values at this cut-off is very high, thus poses a difficulty for an in vitro test that is based on the in vivo reference data to adequately distinguish substances at this very threshold. Given the inherent imprecision of rodent oral LD_{50}s, the precise prediction of in vitro LD_{50}s for chemicals that are predicted to be just below or just above the

classified/unclassified cut off is therefore difficult to achieve. It has long been recognised that rodent LD_{50} values, as reported by different studies, can routinely vary by a factor of more than 2 times even under similar test conditions[1]. In such a situation it is hard to see how a compound with an LD_{50} value of around 1 500 mg/Kg (classified) can be considered a wrong result when it is predicted to be above the cutoff of 2 000 mg/Kg.

Figure 2: Graph of the reported and averaged LD_{50} values for the test chemicals with averaged LD_{50} values from 0 to about 4 700 (n=38 of 56 substances).



<u>Derivation of reference values and classification (2 category system)</u>

While acknowledging that the heterogeneity of reported LD_{50} values poses a difficulty with regard to defining reliable reference data, the ESAC WG had concerns about the handling of multiple LD_{50} reference values for individual chemicals with regard to the derivation of one single LD_{50} value and the relevant class for the purpose of the 2 category classification system relevant to this study (non-classified vs classified).

In brief, when multiple LD_{50} values were available, values were simply averaged in order to derive one single LD_{50} value for the chemical in question and, from that, derive the categorisation of the chemical (non-classified/classified). The averaging approach is problematic as, for instance, a single outlying value is taken into account with exactly the same weight (through averaging) as a group of values that are within a close range. This is exemplified by substance Nr. 14 in Table 4 (2-

[1] One reasonable explanation for the observed variability of LD_{50} values which starts to increase from about >1000 mg/kg onwards is related to a simple issue of dosing volume: dosing about 1g of substance requires a very high vehicle volume which may lead to differences in distribution of the substance (in dilution) within the gastrointenstinal tract and, hence, a high variability of effects observed.

Phenoxyethanol) for which three of the reported values were between one and two thousand mg/kg, while the fourth value was at 13 700. The calculated average was thus 4565. It appears questionable whether this average is biologically meaningful. This approach hence led to questionable classifications in case of two chemicals:

- Chemical Nr. 4 (1,2 Dichlorobenzene) and
- Chemical Nr. 14 (2-Phenoxyethanol)

Based on the values reported and taking outlying values into account, both chemicals considered unclassified by the VSR, should have been considered classified instead (see more detailed discussion in section 4.1 reference data and table 1). Briefly, one chemical was predicted by the 3T3 NRU assay as unclassified (Nr. 4) leading to a possibly unjustified True Negative (TN) test outcome, while the other (Nr. 14) was predicted as classified and was thus considered by the study as a False Positive (FP). Taking the adjusted classification (based on considering outliers) into account, substance Nr. 4 would now be considered a False Negative (FN), while substance Nr. 14 would be considered a True Positive (TP).

In conclusion, although this approach was considered not ideal, the ESAC WG acknowledges that an alternative approach (i.e. deleting outliers by expert judgement or using the median rather than the average) would not have changed the outcome of the study.

The ESAC WG suggests that the following two approaches may have been better:

- Best study approach: This would depend on obtaining the original studies and determining whether or not any of the reported values (i.e. studies) could be excluded on the basis of methodological considerations. This would require defining, a priori, criteria of study quality, i.e. of what defines a well-conducted study and what defines a less well-conducted study. Such criteria could have been defined specifically in view of this endpoint. Therefore, the Klimisch scores, being rather general, may not have been suited to identify optimal studies. Issues such as the physical form of dosing (not considered by the Klimisch system) may have been of relevance for defining suitable criteria.
- Preference given to reference value search, followed by application of chemical selection criteria: Instead of selecting first a range of chemicals followed by the identification of the associated LD_{50} value(s), one could, alternatively, have started identifying chemicals with low LD_{50} reference value variability and from such a set construct a suitable range of test items in view of chemical class etc.

2.5 Appropriateness of the statistical evaluation

NOTE: Are the statistical methods used for evaluating the study data appropriate. Is there a sufficient justification for the use of the methods chosen? Was the statistician independent from the test method submitter/developer?

a) General observations

The aim of the study was the application of existing classifiers (i.e. prediction models), these were the millimole and weight regression models developed during the previous NICEATM/ECVAM study,

on a new set of empirical testing data based on test items that were specifically selected for the objective of analysing the capacity of the test to identify negatives.

Due to differences reg. solubility and other handling issues, the number of chemicals tested varied between laboratories. Of the 56 chemicals of the validation set, 50 were tested in HSL, 53 in JRC and 56 in IIVS. After application of these prediction models to the data, some data points were censored and excluded leading to a data matrix (for millimole regression) of 44 in HSL, 51 in JRC and 54 in IIVS. For the weight the number of data points were 40 for HSL, 46 for JRC and 47 for IIVS (table 34, page 155 of VSR). Finally, the predictive capacity of the assay was determined, on the basis of these data matrices, by analysing the sensitivity, specificity, accuracy, negative and positive predictive values of the assay in relation to the reference LD_{50} values derived as described under 2.4.

Moreover, the study attempted to construct a new classifier (i.e. prediction model) from historical data sets (i.e. the RC, n=540 chemicals) based on a ROC analysis and the application of this classifier to the newly generated data. The presentation and explanation of the ROC analysis in the VSR however was incomplete and it could thus not be reviewed how exactly the threshold values had been derived. The prediction model derived from the ROC analysis could thus not be analysed by the ESAC WG.

The study design is to be complimented: training sets for development of prediction model and test set for evaluation of the prediction model are completely distinct. Test set chemicals were evaluated in blinded fashion.

However, calculated sensitivities, specificities etc. should have been reported with 95%-confidence Intervals (exact confidence intervals as the number of chemicals is not very large), which is missing in the VSR. Furthermore, it is unclear whether the staff performing the statistical analysis was sufficiently independent from the Validation Management Group. Upon request of the ESAC WG, ECVAM clarified that the statistician was a temporary employee of IHCP.

b) Derivation of IC_{50} values / Dose-response analysis

Statistical analysis involves dose-response analysis for the test compounds. Dose-response analysis is described in Table 5, p. 48. Full versions of the analysis specification should be obtained in Annex A to C but are however not fully satisfactory.

HSL laboratory

Although some details of data analysis can be found In Annex A for the HSL laboratory, more details ideally should be given on, for example:

- Which Hill function exactly had been used (i.e. number of parameters)
- How the IC_{50} was exactly derived from the model fit using the Hill function

JRC laboratory

No details about data analysis can be found in Annex B for JRC.

IIVS laboratory

Details of data analysis can be found in Annex C for IIVS, although the method used for IC_{50} estimation is rather crude as it constitutes a simple interpolation between 2 concentration values surrounding the putative IC_{50} values. Other measured values are thus not used for analysing the data/deriving the IC_{50} value.

c) Censoring of IC_{50} values

Dose-response analysis can lead to right and left censored estimates of IC_{50}. Dealing with both types of censored observations was very good and appropriate.

d) Development and application of prediction models

Two methods were used for the development of prediction models which are in principal different: (1) a regression approach (already performed during the NICEATM/ECVAM study) and (2) a ROC approach.

Ad (1) Regression approach

The regression approach was carried out as weight regression and as millimole regression. The regression approach (performed in the previous NICEATM/ECVAM validation study) is based on the historic evaluation of the reported LD_{50} values of 282 chemicals listed in the original Registry of Cytotoxicity (RC) (notably there are, in the meantime, more chemicals in this dataset, n=540).

The weight regression approach is used to derive a cutoff value for the IC_{50} (µg/ml)

On the basis of the above formula, the ESAC WG concludes that chemicals with IC_{50} > 2 709 ug/ml would be predicted as unclassified. However, importantly this value has not been used in the study but was arrived at post-hoc during the ESAC WG review.

The millimole regression model is different as it depends on the molecular weight of the chemical, i.e. every chemical has its own specific threshold (due to the transformation mg/kg • mmol taking the mol. Weight into account).

Ad (2) ROC approach

The ROC approach is based on the evaluation of 540 chemicals from the RC database. This approach is much more flexible as it allows the choice of, e.g., sensitivity, one wants to allow for and the respective threshold is determined accordingly. From a statistical point of view, the ROC approach is to be preferred, as the ROC curve analysis is more in line with typical statistical evaluations used for the construction of classifiers. However, the ROC analysis performed during the study was insufficiently transparent in the VSR. In particular, the thresholds 'a=maximizes specificity' and 'c=maximizes sensitivity' were incorrect (i.e. neither maximising sensitivity nor specificity, respectively).

3. Test definition (Module 1)

3.1 Quality and completeness of the overall test definition

NOTE: This included an analysis of the description of the test system, the protocol, test acceptance criteria etc.

The test system was clearly defined and in sufficient detail to facilitate understanding of the test method, and the criteria for a response was clearly stated and the test acceptance criteria were applied uniformly.

However, the following issues were raised by the ESAC WG:

- The actual dissolved concentrations were not measured/analytically confirmed in case of difficulties with solubility, leaving, in some cases, uncertainty with regard to the actual concentration employed when treating the cells.
- While a specific non-'breathable' (i.e. gas impermeable) foil is listed as necessary equipment in the SOPs (Annex A, section VII.A.2 (v)) in case of volatility, the specifications of this foil are not detailed in the SOP, only the name of a possible supplier is given. This information should however been provided in case the supplier of this particular foil does not supply the foil any more.

3.2 Quality of the background provided concerning the purpose of the test method

NOTE: What is the overall purpose of the test method (scientific use, regulatory application, guidelines, etc.)

The background for the intended purpose is well described in the VSR: The NICEATM/ECVAM validation study provided evidence that the overall accuracy of the 3T3 NRU assay for predicting the acute oral toxicity of chemicals was low (about 30% overall). However, the assay appeared to be better at predicting chemicals with LD_{50} >2 000 mg/kg. Since the majority (87%) of chemicals in the EU New Chemicals database are unclassified (i.e. have LD_{50} > 2 000 mg/kg), it was thought that this assay may be utilized as a broad “screener” of chemicals for being below or above this value. The NICEATM/ECVAM validation study had 45 and 22 chemicals in each group, respectively, while in the study presented in the report the number of chemicals was still low, but more balanced (30 and 26). The actual protocol utilized in the assay of cytotoxicity is described on p. 54 et al. and is quite straightforward, as it is based on the ability of the cells to take up Neutral red. The two separate labs carried out modified versions of the assay. In one case (IHCP, JRC) changes were minimal, with a robotic system carrying out basically the same passages. In the other lab, a reduced number of replicates were utilized, and chemical concentrations were fixed for all chemicals, without dose-ranging experiments, which were carried out in the main study of the report. Finally, the range of concentrations used in the in vitro testing appears appropriate.

3.3 Quality of the documentation and completeness of SOPs and prediction models

NOTE: Are the SOPs sufficiently detailed and complete? Are the prediction models sufficiently well explained to be applied in the correct manner?

The basic starting protocol for all laboratories was identical to that in the NICEATM/ECVAM Study. The prediction models are the same as the initial NICEATM/ECVAM study with the exception of the ROC-based prediction model which was newly constructed in the context of the present ECVAM study (see however comments of ESAC WG on this in section 2.5). The two original prediction models (weight and millimole regression) were explained in sufficient detail to ensure their uniform and correct application.

The ESAC WG has the following comments on the study protocol:

(a) It is understood that the studies conducted in the three laboratories were not executed according to GLP, although they may have been performed in a GLP-compliant environment. This is not considered as affecting the quality of the study by the ESAC WG.

(b) The criteria for repeat testing in each laboratory (in case of data not meeting the acceptance criteria) need to be specified in the protocols. It was clear that assays were repeated due to assay response.

(c) The plate sealer method (VII.E.6 p18) has two wrong references: the first Section VII.E.2.b should read VII.E.2.c, and the second Section VII.E.2.b should read VI.E.2.d.

(d) Although not specifically stated, the plate seals specified in the protocol were not gas permeable (Excel Scientific). Since the buffering of the medium depends on carbogen delivered from the incubator, there may be possible effects of basification/alkalinisation of the medium in case plate sealing foils are used for 48h incubation time.

(e) The protocol used by HSL (Annex A) was simply a copy of the basic protocol without any changes. If the basic protocol was used, HSL should have stated that the protocol shown was the one used and modifications that were made in the basic procedure stated in the protocol. Thus the protocol should include the modification of the solubility protocol to provide up to 3hr heating for insoluble chemicals (Section 6.1).

(f) With respect to the JRC protocol (Annex B), it also needs to be stated that the basic protocol for the study was the one recommended from the NICEATM/ECVAM Study (Annex A) with any modifications clearly detailed. The description of the automated processes including the equipment used was presented in sufficient detail; however, references to the handling of test chemicals from solubility to preparation of the stocks for use in preparing the plate used for the automated dispensing of test chemical to the treatment plates needs better description. Some of the questions that were not addressed also included: 1) Were the plates evaluated morphologically following chemical treatment, 2) How were points with low cell number or other evidence of toxicity handled, 3) How were precipitates determined and how did precipitates impact the data used. 4) Were the solutions in the dilution plate mixed prior to sampling for treatment?

(g) Automation frequently does not provide access to information that would be easily accessible in a manual assay. The JRC protocol did include some information on modifications

from the original protocol, but more detailed information would allow automated protocol to be transferred more readily.

(h) The IIVS protocol (Annex C) was found acceptable as written.

(i) The solubility protocol (Annex F) was the same at that for the NICEATM and ECVAM study.

4. Data quality

4.1 Overall quality of the evaluated data

NOTE: What is the quality of the data evaluated (testing data).

Reference data

There are some issues relating to the choice of reference in vivo LD_{50} values. These faults are, to some extent, intrinsic to the nature of LD_{50} values which are inherently imprecise (Rowan et al., 1983). As a result, it is difficult to have a "true" LD_{50} value. (See Table 4 of the VSR, and Section 2.4 including Figure 2.)

The choice was made to use an arithmetic mean of the multiple LD_{50} values found. This has resulted in some anomalies. For instance, 2-phenoxy ethanol (table 4 page 37) has three LD_{50} values below 2 000mg/kg in the rat and one of 13 700. The arithmetic mean is 4 565 (not classified). However, if the extreme value were excluded as an outlier, the arithmetic mean of the other three values would have been 1 520mg/kg (classified). In total there are two chemicals where averaging lead to inappropriate weighting of outlying values (see table 1) and which had implications on the predictive capacity as, when considering the new reference values, one TN became one FN, while one FP became a TP. Furthermore (p35) several of the chemicals were classified in the EU, but would not have been on the basis of the LD_{50} values found.

Table 1: Test outcomes of substance 4 and 14, when disregarding outlying values for the derivation of a reference LD_{50} value

Nr. of chemical	Chemical	A) Reference classification in VSR (values below, outliers in bold)	B) Reference classification of ESAC WG disregarding outliers	Prediction of 3T3 assay (HSL laboratory, millimole regression)	Test outcome based on reference value A)	Test outcome based on reference value B) [disregarding outlier]
4	1,2-Dichlorobenzene	UC 500 / 1000 / 1516 / 2138 / 5170	C (median of four similar values = 1258)	Negative	TN	FN
14	2-Phenoxyethanol	UC 1260 / 1400 / 1900 / 13'700	C (median of three similar values = 1400)	Positive	FP	TP

In vitro testing data

Data presented in the VSR clearly show the quality of the data found in each laboratory, including the variability between laboratories. Some of the generated in vitro data showed high levels of variability between experiments in the same laboratory and it was not obvious whether and if so which acceptance criteria had been applied.

There was a good concordance between the laboratories for most chemicals which indicates a level of reliability of the data. One of the highly toxic compounds, aconitine, gave very variable results which complicates the interpretation.

4.2 Sufficiency of the evaluated data in view of the study objective

NOTE: Are the data and their quality sufficient in view of the stated objective of the study?

In general, the data were of sufficient quality for the purpose of the study. All laboratories provided the required number of replicates for the chemicals that were successfully evaluated. Data that did not meet the criteria were excluded from consideration.

4.3 Quality of the reference data for evaluating relevance[2]

NOTE: What is the quality of the reference data used? Are the data and their quality sufficient in view of the study objective?

The quality of the reference data is difficult to assess as the criterion for inclusion appears to be based on them being included in published reference databases.

The presentation of the reference data is clear and helpful with regard to evaluating the intrinsic variability of reported LD_{50} values (in case of multiple studies per chemical). However, the values do not have references in the tables which might give some idea of quality (age of the reference for example).

It would have been beneficial if quality audit had been done on the published data.

The report indicates that the data of JRC automated methodology was good quality according to the Z-prime parameter which the ESAC WG is not familiar with.

[2] OECD guidance document Nr. 34 on validation defines relevance as follows: "Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of accuracy (concordance) of a test method."

5. Test materials

5.1 Sufficiency of the number of evaluated test items in view of the study objective

NOTE: Is the number of test items tested during the study sufficient in order to draw conclusions with respect to the objective of the study? If not, are there reasons for deviations and are these explained and justified?

The number of 56 chemicals and their split between classified and unclassified by EU criteria 24/32, seems reasonable. However, the number of chemicals actually tested was between 50 and 56 according to the laboratory (see table G1 in Annex G). Due to censoring and excluding values that could not be dealt with, the number of data available for analysis was further reduced. Using the millimole regression model 44, 51 and 54 values (depending on laboratory) could be included in the analysis (pages 145-148 of VSR). The weight regression model led to the censoring/exclusion of more data points so that 40, 46 and 47 values (depending on laboratory) could be used for the analysis of predictive capacity. There were a number of issues regarding solubility of chemicals, which were unfortunate. This affected mostly the comparison among the three laboratories. However, even in the main study (HSL), three compounds were not tested because of insolubility. All three belong to the "unclassified" group, thus reducing the number of unclassified substances tested to 29 from 32. The number of Classified compounds was also reduced in the main study. Indeed, brucine and aconitine were not tested because they were "potent toxins" (p. 61) and malononitrile was not properly retrieved from the vial (p. 61). This reduced the number of classified chemicals to 21 in HSL, to 27 in JRC and 29 in IIVS. Thus the total number of chemicals that underwent testing in the main study was (according to laboratory):

- 50 (21 Classified [C] + 29 unclassified [NC]) in HSL
- 53 (24 C + 29 NC) in JRC
- 56 (24 C +32 NC) in IIVS

Nevertheless, a formal sample size calculation can be performed on the basis of a targeted sensitivity of ca. 80% and a targeted length of the 95%-Confidence Interval of, e.g., 0.3. This would lead to a number of 32 classified chemicals. The analogous calculation can be made for a targeted specificity of 80%. With 32 chemicals, the maximal width of the 95%-CI would be 0.36 (reached at a specificity of 50%).

In summary, the number of chemicals seems appropriate.

5.2 Representativeness of the test items with respect to applicability

NOTE: Analysis of how well the test items were selected in order to gain – through empirical testing during the study – insight into the applicability domain / limitations of the test method OR analysis to which extent the test items used during the study map an applicability domain already known.

It should be noted that compounds with specific modes of action (pharmaceuticals, pesticides etc) were excluded, which may have implications for the testing of unknown chemicals. See Section 10.

6. Within-laboratory reproducibility (Module 2)

6.1 Assessment of repeatability and reproducibility in the same laboratory

NOTE: How were repeatability and reproducibility assessed? Are the conclusions justified by the data as evaluated?

Within laboratory reproducibility was assessed by the analysis of the coefficient of variation (CV), and the concordance between the toxicity predicted. CV mean values were 28% (HSL), 19% (JRC). Concordance predictions were 98% (HSL), 94% (JRC), and 100% (IIVS) with the millimole regression model, and very similar with the weight regression. No issues were noted with this aspect off the report.

6.2 Conclusion on within-laboratory reproducibility as assessed by the study

NOTE: How was within-laboratory reproducibility assessed? Are the conclusions justified by the data as evaluated?

The reproducibility within laboratories was deemed to be acceptable. No issues were noted.

7. Transferability (Module 3)

7.1 Quality of design and analysis of the transfer phase

NOTE: Was the transfer phase appropriately planned, e.g. transfer instructions, training, minimum requirements, training SOP (if appropriate). Where evaluation / decision criteria defining a successful transfer established beforehand and consistently applied during the analysis?

The transfer phase was appropriately planned. The evaluation / decision criteria were established beforehand, and were consistently applied to the analysis. Nine test chemicals were utilized in this phase in case of the HSL laboratory, while 12 were used in case of the JRC laboratory. Issues are appropriately described on p. 120-125 of the report.

7.2 Conclusion on transferability to a second laboratory as assessed by the study

NOTE: Are the conclusions justified by the data generated? Have critical issues that may impact on transferability been identified?

The transferability of the assay was only evaluated in the two laboratories (HSL and JRC), as outlined on pages 122-125 of the VSR. The conclusions were justified and there are no critical issues that may impact on transferability.

8. Between-laboratory reproducibility (Module 4)

8.1 Assessment of reproducibility in different laboratories

NOTE: How was reproducibility between laboratories assessed?

Only one laboratory (HSL) was utilized in this follow-up validation study. The other two laboratories (JRC and IIVS) conducted studies which were similar, but not identical. However, concordance for toxicity predictions for all three laboratories was high (p. 127 et al.) Taking account of the simple requirement to split into >2 000 or <2 000 mg/kg bw the degree of variation is acceptable. It is certainly less than the general variation between predicted and observed LD_{50} values.

8.2 Conclusion on reproducibility as assessed by the study

NOTE: Are the conclusions justified by the data generated?

Overall, reproducibility within and between laboratories appears to be acceptable.

9. Predictive capacity (Module 5)

For a test method that should fulfil the purpose of reliably identifying negatives, a high sensitivity is required, while the specificity may be lower but should on the other hand not be too low in order to avoid too high a FP rate.

Sensitivity / True Positive Rate

Using the millimole regression, sensitivity (defined as the ability to correctly identify classified chemicals) was 94.4 (CI 72,7 – 99,9%), 91.7 (CI 73,0-99,0%) and 95.8 % (CI 78,9-99,9%) for HSL, JRC and IIVS, respectively. Thus, about 5-8% of classifiable chemicals would be misclassified as negatives. Taking the confidence interval into account up to 27,3% of the AP substances could be misclassified as Negatives.

Using the weight regression, sensitivity was 100% for all laboratories/protocols. The lower confidence limits were 85.2 (IIVS), 80,5% (HSL) and 84.6% (JRC). The upper confidence limit was 100% for all three laboratories/protocols. Furthermore, the weight regression prediction model led to a higher number of censored / excluded chemicals. This, notably, concerned the substances misclassified (as FN) by the millimole regression prediction model. Therefore, the improvement of sensitivity observed in the weight prediction model is due to the exclusion of data.

Negative Predictive Value (NPV)

The Negative Predictive Value was derived correctly as well as accuracy. However, predictive values are dependent on the composition of the test set (prevalence of the toxic vs non-toxic compounds in the test set) and therefore can not be used to extrapolate to other sets of compounds with different compositions.

Percentage of unclassified chemicals

On page 145 of the VSR a formula is provided to extrapolate to the proportion of chemicals with negative outcome in the test depending on sensitivity, specificity and prevalence. This formula is correct and useful to estimate the number of animal tests potentially avoided. This calculation is valuable provided that this estimate of 13% of toxic compounds is representative of the population of the chemicals which will be tested.

Specificity / True Negative Rate

Using the millimole regression, the specificity (defined as the ability to correctly identify unclassified chemicals) was quite low and similar in all three laboratories (42.3, 44.4, and 40.0 %, respectively). Thus many compounds of very low toxicity would be identified instead as chemicals of moderate-high toxicity, and submitted to further tests, presumably in vivo. The weight regression analysis provided qualitatively similar results, but with some important quantitative differences: specificity was only 12-17% vs 40-44.4%).

Conclusions

(1) Based on the data of this study, the millimole regression model appears to have a better balance between sensitivity and specificity. While the sensitivity of the weight appears better than that of the millimole, this is due to a greater number of censored/excluded chemicals. Moreover, the specificity of the weight regression model is considerably lower than that of the millimole (12-17% vs 40-44.4%).

Overall, as stated in the report, given that >85% of all chemicals in the EU database are in the unclassified category, one may predict that this assay may reduce the need for in vivo animal testing by about 30-40%.

This however, will depend on the appropriate use of the test method, taking its applicability and limitations into consideration. For example, a compound acting through a specific receptor (e.g. domoic acid) may not be detected, i.e. may be a false negative. Domoic acid is a natural compound of high acute toxicity in humans. Similarly, digoxin was a false negative in the NICEATM/ECVAM study (NIH, 2006). However, notably, two substances with presumably a specific mode of action (the alkaloids brucine and aconitine) were largely correctly predicted in the current study. This may indicate that such substances may in addition act through basal cytotoxicity.

10. Applicability domain (Module 6)

The report highlights the areas already acknowledged as limitations of this test method, most of which require a priori knowledge of the chemicals to be tested. With this in mind, it is clear, as also concluded in the VSR, that this method is unsuitable for testing unknown chemicals as a stand alone method.

However, it could be useful in an appropriately constructed ITS strategy (including an extensive list of exclusion criteria). This would require knowledge of physiochemical properties of the test materials and the structure for QSAR analysis. Some of the limitations need to be taken into account are described below.

In reviewing the applicability domain the ESAC WG could not comment on the applicability to classes of agents that were excluded from the validation study. For example:

Organ-specific modes of action leading to acute toxicity (e.g. specific proteins, receptors etc.)

The VSR excluded a number of classes of chemicals from the set to be analysed (pesticides, pharmaceuticals, mutagens, carcinogens and reproductive toxicants). This severely restricted the number of highly toxic chemicals considered and those having receptor mediated toxicity. While the VSR acknowledges such chemicals as being outside the applicability domain, it would be necessary to develop some exclusion criteria / SAR model to exclude such chemicals from the assay in view of the application of the test for chemicals of unknown acute toxicity. The ESAC WG would expect that toxins (e.g. alkaloids) and venoms should be excluded. For example of two alkaloids included one was not reliably detected.

Solubility

All cell culture methods have limitations with compounds of low solubility in culture media and 3T3 NRU is reported to be no exception (as acknowledged in the VSR, p. 169). The ESAC WG makes a specific recommendation that solubility is verified in the plates containing the test medium.

Metabolism

As noted in the VSR, chemicals from classes known to produce metabolites of higher acute oral toxicity could be under predicted by 3T3 NRU as there is no metabolic capability within the system. Similarly chemicals that are rapidly and extensively de-toxified would have their toxicity over-predicted. It is unclear how many of the compounds in the database of the validation study are subject to activation or de-toxification. It might therefore be necessary to develop some exclusion criteria / SAR model to exclude from the assay such chemicals likely to produce higher toxicity metabolites. [Other more extensive data sets (e.g. RC) indicate that this is unlikely to be an issue if appropriate evaluation criteria are chosen.]

Other limitations on the applicability domain could be dependent on the evaluation / selection criteria used and the balance between false positive and false negative predictions.

The report states that the test methods works for both industrial chemicals and cosmetic ingredients. However, the test data set does not include any hydrocarbon solvents, low boiling point liquids (< 100 C), and there is a limited range of cosmetic dyes and UV absorbing agents (sunscreens).

11. Performance standards (Module 7)

Not applicable.

12. Readiness for standardised use

12.1 Assessment of the readiness for regulatory purposes

NOTE: Is the test method ready for regulatory purposes? If yes, why? If no – what impediments currently exclude application for regulatory purposes?

On the basis of results of the study and in agreement with the conclusions of the VSR and the mandate given to the WG by ECVAM, it is concluded that the 3T3 NRU assay is not suitable as a stand-alone test for regulatory use in determining chemicals that are unclassified via the oral route using either the EC CLP criteria or GHS criteria. The reasons for this are

- a limited applicability domain / limitations with respect to specific mechanisms of action (see Section 10);
- the limited number of chemicals in the testing set (n=24) that would be classified as 'Toxic' (GHS categories 1,2, 3 and 4);
- the variable prediction of the toxicity of the only chemical for which there is currently clear indication that it acts via a specific mechanism of action other than basal cytotoxicity; with the associated questions about general applicability to receptor active toxicants. The latter is linked to the exclusion criteria that removed pesticides and pharmaceuticals from the present validation study;

The low specificity of the method resulting in many false positives will require further in vivo testing of chemicals of low toxicity therefore reducing the cost effectiveness of the method.

The different test methodologies evaluated have been shown to have high concordance which suggests that the method is robust, amenable to automation and simplification as well as transferable between laboratories.

12.2. Assessment of the readiness for other, non-regulatory, uses

The method may useful for in-house screening by companies who have a good knowledge of the chemistry that they are testing as a way of avoiding unnecessary animal use.

The automated protocol variant (JRC) could have particular value in screening for chemicals of very low acute oral toxicity (unclassified under CLP), however appropriate exclusion criteria concerning toxicity induced by specific mechanism of action (i.e. other than basal cytotoxicity) should be used.

12.3 Critical aspects impacting on standardised use

General limitations provided in the applicability domain require that ITS information be obtained prior to initiating the test and with the broad range of exclusion criteria the number of classes of chemicals appropriate for the test would be restricted.

12.4 Gap analysis

In terms of the data shown in the VSR, there are gaps which may be worth of further investigation:

- Testing a higher number of chemicals of high toxicity in order to better evaluate the "real world" risk of false negatives with high toxicity;
- Evaluating whether excluding pesticides and pharmaceuticals (and other specific toxicities) is justified as the data of the previous NICEATM/ECVAM study do not appear to support excluding them. However, it is noted that when analysing the 72 chemicals of the NICEATM/ECVAM study (containing pesticides, pharmaceuticals) with regard to the 2000 mg/kg cut-off value, only one chemical was a false negative (FN).
- Chemicals which may be susceptible for metabolism and with the capacity to give rise to toxic metabolites (would require modification of the protocol).

13. Other considerations

NOTE: Please address any other consideration you might have in relation to the proposed approach under this section.

- Limitations need to be communicated clearly in order to avoid testing of chemicals that are not within the applicability domain.
- In comparing the predictive data from the two protocol variations to that observed with the manual protocol, the data support their standardized use alongside the already validated protocol. The predictive data are essentially indistinguishable from the manual protocol data. In addition, the improvements in test chemical solubility handling by IIVS provide options for testing insoluble chemicals and allowed for testing of additional agents. Both the protocol variations should reduce assay performance costs by either automation or reduced test concentration requirements.
- Using ROC to make adjustments in the perceived assay performance for the sake of gaining higher accuracy does not seem justified. If the 3T3 test method is to be used to identify unclassified agents from further animal testing, then the issue of specificity becomes paramount. Identifying a severe acute toxicant as a negative does not seem to be acceptable. One agent, aconitine, exemplifies this concern, correctly predicted in one laboratory (JRC), falsely predicted in IIVS and not tested in HSL. However, it should be noted that IIVS had problems with the solubility of aconitine, which may have led to the classification obtained (page 150 of VSR). The remaining agents that were considered as false negatives fall into the EC CLP category 4 and would therefore raise fewer concerns, especially since the predicted LD_{50} for these agents was close the upper cut-off for category 4.
- Although not considered as part of this study, the ICCVAM/ECVAM initial study was similar in its predictive potential for the GHS unclassified chemicals.

14. Conclusions on the study

NOTE: This section should presents an overview over the study results and conclusions as described in the study reports (subsection 14.1), discuss to which extent the conclusions drawn in the study reports are justified by the study results on their own (subsection 14.2) and evaluate to which extent the conclusions are plausible with respect to other information (subsection 14.3).

14.1 Summary of the results and conclusions of the study

14.1.1 Test items

The VSR reported 3T3 NRU results for a test set of 56 chemicals (24 classified and 32 unclassified chemicals) with available acute oral LD_{50} values and which had not been tested in previous RC and NICEATM/ECVAM studies. The 3T3 NRU testing was performed by three laboratories (HLS, JRC, IIVS) using variations of the basic 3T3 NRU cytotoxicity assay. Cytotoxicity results were not obtained for all chemicals in all laboratories.

The range of chemicals in the test set was restricted by the prior removal from the core set of any compounds classified as mutagenic, carcinogenic or toxic to reproduction; insoluble metals / salts; and all chemicals used as pesticides or pharmaceuticals. This left a set of chemicals with only a small number of chemicals of moderate to high acute oral toxicity (8 out of 56 had acute oral LD_{50}s <300 mg/kg bw).

14.1.2 Summary of study results

The report results show a high degree of consistency between the three test laboratories demonstrating the test methodology is robust, stable and amenable to automation.

The IC_{50} results from the cytotoxicity assay did not provide a reliable prediction of the reported acute oral LD_{50} values (this was not the stated aim of this exercise). Despite this lack of direct correlation, using the evaluation criteria of the study it was possible to determine a subset of chemicals that were unclassified for acute oral toxicity using the EC CLP criteria (LD50 >2 000 mg/kg bw) with a low percentage of false negative results (compounds predicted as unclassified but with LD_{50}s <2 000 mg/kg bw).

Two prediction models were evaluated, one based on regression of mg/kg b.w. and the other on a millimole basis. The authors proposed the millimole model as giving the better balance between sensitivity and specificity. The millimole prediction model produced a high level of sensitivity (about 95%, CI 73-100%). Same model gave specificity of about 40% with CI 23-65%. The resultant false positives would trigger in vivo testing but had acute oral LD_{50} values above 2 000 mg/kg bw. Alternative evaluation criteria can be used to alter the balance between false negative and false positive results. The ROC analysis in view of constructing a third prediction model was provided in the VSR but was not able to be used. A new ROC analysis was performed which showed that this approach could be used to manipulate sensitivity and specificity to favour one or the other if required but was not considered at this stage to add any value to the study.

For the 24 classified substances results for all chemicals were obtained at least from one laboratory. With the regression model based on millimole, three compounds were false negatives at least in one laboratory. With the weight regression model 23 (out of 24) results were obtained and all classified compounds were identified.

The compounds which gave false negative (FN) were:

- 1,2,4-trichlorobenzene (JRC),
- aconitine (IIVS, not tested in HSL and borderline in JRC)
- benzyl benzoate (HSL and JRC).
- 1,2,4-trichlorobenzene (results were ambiguous in JRC but were clearly identified in other laboratories). Benzyl benzoate is not of concern as its toxicity is very close to the limit of 2 000 mg/kg b.w.

One substance (aconitine, with the highest acute oral toxicity in the test set) was not clearly identified as having significant toxicity in either of the laboratories where tested. It was not tested at HSL, gave false negative results in IIVS and a borderline result in JRC. Although this compound has a known mode of action which a priori is outside the applicability domain of this test, nevertheless the result raises concerns about the ability of the 3T3 NRU test to detect compounds of very high toxicity and unknown mode of action. It is however acknowledged that aconitine is an alkaloid with a typical structure for this class of substances. Thus, using appropriate considerations of chemical structure, such chemicals may be excluded from testing (i.e. development of suitable exclusion criteria).

14.2 Extent to which study conclusions are justified by the study results alone

The VSR concludes that "The study has shown that the test chemicals categorised as unclassified in vivo (LD_{50} > 2 000 mg/kg b.w.) are most likely categorised as unclassified also by the 3T3 NRU in vitro test method." This is incorrect as only 40% of chemicals unclassified in vivo are also unclassified in this test.

The ESAC WG assumes that the intention of this sentence was to say inversely and correctly: "The study has shown that the test chemicals tested as unclassified in vitro are in most cases indeed unclassified chemicals based on the in vivo test (LD_{50} > 2 000 mg/kg b.w.)."

Nevertheless, assuming the hypothesized distribution of unclassified chemicals (87%), the ESAC WG agrees that this could allow reduction in animal testing for acute oral toxicity up to 40%, subject to the applicability issues discussed above.

A note of caution is needed as the choice of the preferred prediction model was based on the actual test set of chemicals and may not be representative of a wider universe of chemicals.

The study has shown that the method is reproducible, stable and amenable to automation. Depending on which data analysis model is used, it is possible to produce a high level of conservatism, provided chemicals with specific modes of action and specific physicochemical properties are excluded. However, this high level of conservatism is obtained at the expense of a very high false positive rate.

14.3 Extent to which conclusions are plausible in the context of existing information

If additional databases (e.g RC) are included, concerns about the number of chemicals with 'Toxic' classifications and receptor mediated modes of action are reduced. The poor predictivity of absolute LD_{50} values is still evident within the extended database. Preliminary evaluation of pooled data from NICEATM/ECVAM validation study would appear to be consistent with the present study.

15. Recommendations

Note: This section should provide recommendations on the test method (e.g. further work, possible use) and their constituting elements (e.g. test system, prediction model, SOP).

15.1 General recommendations concerning the 3T3 NRU test method

a) Possible use of the 3T3 test method

- In agreement with the Validation Study Report, the 3T3 NRU test method should be considered only in the context of an ITS strategy or for non-regulatory screening.
- Also when used within such a testing strategy, rigorous application of exclusion criteria for test chemicals is required to ensure adequate level of safety. These criteria may be based on an integration (e.g. via an weight of evidence analysis) of existing information on the compounds. Existing information may come from
    - The intended purpose of the substance and whether or not it has been designed/screened out to interfere with specific biological targets/modes of action (i.e. pharmaceuticals, pesticides etc.)
    - Physicochemical properties of the substance that may indicate acute toxicity potential
    - Information from read-across, grouping, SARs, (Q)SARs, structural alerts pointing to acute toxicity potential
    - Existing human information (poison centres, epidemiology etc.)
    - Existing information from in vivo tests (including repeated-dose toxicity tests)

b) Possible further work in view of characterising the performance (applicability, limitations, predictive capacity for negatives)

- Further testing using an increased number of compounds belonging to the highly toxic acute toxicity classes (i.e. category 1 and category 2) should be considered.
- Consideration should be given to the fact that a high proportion chemicals of very low acute oral systemic toxicity turned out to be False Positives (FP).
- The analysis of the data from NICEATM/ECVAM, RC database and this study together would be better done in the following way:

  Comparison of different prediction models should be performed on the basis of the historical data (RC and NICEATM/ECVAM) by random splitting into a training and validation set. The models are developed on the training set and evaluated on the validation set. The best model is identified from this evaluation and then applied to the current study data.

  This approach is generally preferred to building and testing for prediction models.

15.2 Recommendations for improvement of the SOPs associated with the 3T3 NRU test method

- Solubility should be assessed in the final evaluation of test plates. Treatment conditions with evidence of insolubility should not be considered.
- Data analysis should be described more clearly in the SOPs.
- Recommend in addition to the SOP a cut-off value for volatility (because volatile compounds will partition between the head space and the medium and consequently reduce the effective concentration).

16. References

Rowan A (1983) Shortcomings of LD_{50}-values and acute toxicity testing in animals. Acta Pharmacologica et Toxicologica, 52: 52–64.

OECD (2010) Series on Testing and Assessment No. 129. Guidance Document on using cytotoxicity tests to estimate starting doses for acute oral systemic toxicity tests.

NIH (2006) Background Review Document (BRD): Validation of Neutral Red Uptake Test Methods NIH/In Vitro Cytotoxicity Test Methods for Estimating Acute Systemic Toxicity. Publication NO. 07-4518, November 2006. http://iccvam.niehs.nih.gov/methods/acutetox/inv_nru_brd.htm

Bulgheroni A, Kinsner-Ovaskainen A, Hoffmann S, Hartung T, Prieto P (2009) Estimation of acute oral toxicity using the No Observed Adverse Effect Level (NOAEL) from the 28 day repeated dose toxicity studies in rats. Regul Toxicol Pharmacol. 53(1):16-9.

17. Annexes

Annex 1 – Question / Answer session with validation management staff

ECVAM
SCIENTIFIC
ADVISORY
COMMITTEE
(ESAC)

ESAC Working Group 3T3 NRU

September 2011

This document summarises a question and answer session held during the ESAC WG meeting on 13. September 2011 at EC-JRC/ECVAM, Ispra. The ESAC WG had several questions on the study design that were answered by one member of the VMG. Moreover, section B) of this document summarises three more clarification of the VMG & ESAC Secretariat that were added to this document after the ESAC WG meeting.

A) Question/Answer session with VMG during ESAC WG meeting

1. ESAC WG: What is meant with using original references as support to the collection of LD50 reference data?

VMG: This was sub-contracted to Fraunhofer who went back to the original references when possible (not all of them were accessible) and they verified, when possible, that the numbers were correct. Methodology is described in Hoffmann et al., Regulatory Toxicology and Pharmacology 58 (2010) 395–407 (Acute oral toxicity: Variability, reliability, relevance and interspecies comparison of rodent LD50 data from literature surveyed for the ACuteTox project)

2. ESAC WG: Is there are justification for using mouse data for Nr. 49 tri ethylene glycol? Why were mouse data used? (they did not seem to make a difference)

VMG: This was due to the exclusion criteria when rat data were range or non finite numbers; In retrospect it is clear that the rat data (e.g. centre values of ranges) could have been taken for deriving classifications.

3. ESAC WG: Chemicals explored in previous studies (NICEATM/ECVAM and ACuteTox were excluded. Pharmaceuticals and pesticides were excluded. Unclear why Brucine and Acotinine were kept.

VMG: The exclusion criteria (incl. used in previous studies) left very few highly toxic compounds.

4. ESAC WG: What was the rationale for excluding CMRs etc from the testing set? Issues of safety of testing personnel? Could this not be managed in another way? May it not exclude important acute toxicants? (e.g. mutagens are normally very toxic)

VMG Health and safety issues.

5. ESAC WG: Why were different selection criteria used for classified vs non-classified chemicals (chemical selection, not ref. data selection)

VMG: One reason is that the starting lists came from the different sources. However, the general selection criteria were the same for both categories. It is correct however that three exclusion criteria were only listed for classified chemicals even though they were also considered for unclassified chemicals (not listed in the report for unclassified chemicals):

1. Insoluble metals were excluded
2. Chemicals potentially difficult to handle in the laboratory or in vitro (based on indication from the physico-chemical properties) were excluded.
3. Chemicals used in previous studies i.e. the NICEATM/ECVAM validation study and the EU ACuteTox project were excluded;

6. ESAC WG: why were chemicals T+ excluded if they were also toxic by dermal or inhalation route? To protect lab personnel? There seems no obvious scientific reason. (Page 14 criterion 5)

VMG: Only used to decrease the number of compounds.

7. ESAC WG: What is the cost of conducting the 3T3 test per chemical?

VMG: About 1000 euros.

8. ESAC WG: Where data excluded where chemicals had shown precipitates?

VMG: IIVS data were all used whereas the other two laboratories did not use the data.

B) Further clarifications of the VMG after the ESAC WG meeting

Added 16.9.2011

(1) Incorrect sentence in section 12 (conclusions)

The VMG confirmed that the sentence at the beginning of the second paragraph in section 12.0 was incorrect and that essentially, as suspected, the intention was to describe the inverse of what the sentence describes at present.

The sentence reads at present in the report:

"The study has shown that the test chemicals categorized as unclassified in vivo (LD50 > 2000 mg/gk b.w.) are most likely categorized as unclassified also by the 3T3 NRU in vitro test method".

Instead the sentence should rather read:

"The study has shown that the test chemicals categorized as unclassified by the 3T3 NRU in vitro test method are most likely categorized as unclassified also by in vivo (LD_{50} > 2000 mg/gk b.w.)"

(2) Exclusion criteria as explained in section 3.1.2 (on chemical selection)

With regard to the exclusion of chemicals that showed toxicity via other than the oral route (criterion 5 on page 14 VSR), the VMG would like to clarify that chemicals were only then excluded if they showed higher toxicity via the dermal or inhalation routes than via the oral route.

Selection criterion 5 on page 14 says:

"In case of chemicals labeled Xn [for the oral route], the compounds classified as T+ by other routes (dermal, inhalation) [thus higher toxicity in other routes] ...were excluded".

For example Aconitine and Brucine are both classified as T+ via the oral route AND T+ R26/28 (very toxic by inhalation and if swallowed) were not excluded.

However, it is noteworthy that there were other chemicals classified T+ by more than one route but excluded from the final list because they did not meet other criteria, for example:

- Hydrogen cyanide(CAS No. 74-90-8) was classified in Annex 1 as T+; R26/27/28 N; R50-53 but excluded because the chemical was not available in SIGMA
- 2-hydroxy-2-methylpropionitrile (CAS No. 75-86-5) was classified in Annex 1 as T+; R26/27/28 N; R50-53 but excluded because unstable, especially in water.
- 1-isopropyl-3-methylpyrazol-5-yl dimethylcarbamate (CAS No. 119-38-0) was classified in Annex 1 as T+; R27/28 but excluded because the chemical was not available in SIGMA

(3) Typing error in bullet point 6 (on chemical selection)

Finally, we have realized that there is a typing error in bullet point 6 (surplus T). It should read as follows:

“In case of chemicals labelled T, the compounds classified T+ by other routes (dermal, inhalation) and/or T effect after prolonged exposure (R48) and/or cumulative effect (R33) were excluded”

Annex 2 – Synoptic tables of testing data and predictions yielded (prepared by ESAC Secretariat for use during ESAC review).

The following three datasets of the ECVAM-coordinated study are shown:
A) Validation dataset: data from HSL using the validated protocol
B) Adjunct dataset 1: data from JRC using the automated version of the protocol
C) Adjunct dataset 2: data from IIVS using the shortened version of the protocol
In all cases, results obtained with all three prediction models are shown.

Moreover, a reanalysis of the testing data of the previous NICEATM/ECVAM study is shown, using the 2000 mg/kg cut-off.

VALIDATION DATA SET: Testing data of HSL using the validated protocol.

Predicted LD50, Predicted Toxicity (dichotomous), Prediction in relation to in vivo reference data

LEGEND

TN = true negative	not tested
TP = true positive	censored.excluded
FN = false negative	*in vivo* positive (classified)
FP = false positive	*in vivo* negative (unclassified)

	LABORATORY:	**PREDICTION MODELS APPLIED TO DATA**							
	HSL (VAL'ED PROTOCOL)	**A) MILLIMOLE REGRESSION**			***B) WEIGHT REGRESSION***			***C) ROC Analysis***	
Nr.	**Chemical**	**Predicted LD50**	**Toxicity Prediction (0 = toxic, 1 = non-toxic)**	**Test outcome (EU CLP)**	**Predicted LD50**	**Toxicity Prediction (0 = toxic, 1 = non-toxic)**	**Test outcome (EU CLP)**	**Toxicity Prediction (0 = toxic, 1 = non-toxic)**	**Test outcome (EU CLP)**
1.	(4-Ammonio-*m*-tolyl)ethyl(2-hydroxyethyl)ammonium sulphate	205.14 ± 28.09	0	TP	192.48 ± 22.36	0	TP	0	TP
2.	1,2,4-Trichlorobenzene	674.26 ± 87.36	0	TP	661.90 ± 72.50	0	TP	0	TP
3.	1,2-Benzenedicarboxylic acid	> 2 563.6	1	TN	censored excluded	-	-	1	TN
4.	1,2-Dichlorobenzene	> 2 130.9	1	TN	censored excluded	-	-	1	TN
5.	1-Naphthylamine	292.50 ± 86.50	0	TP	363.96 ± 90.79	0	TP	0	TP
6.	1-Phenyl-3-pyrazolidone	474.81 ± 33.94	0	TP	518.92 ± 31.54	0	TP	0	TP
7.	2-(2-Butoxyethoxy)ethanol	2 038.7 ± 230.89	1	TN	1 783.1 ± 170.68	0	**FP**	1	TN
8.	2,2',6,6'-Tetrabromo-4,4'-isopropylidenediphenol	780.99 ± 20.05	0	**FP**	445.18 ± 9.69	0	**FP**	0	**FP**
9.	2,4,6-Tris(dimethylaminomethyl)phenol	censored excluded	-	-	censored excluded	-	-	censored excluded	-
10.	2,6-Diethylaniline	> 2 149.0	1	TN	censored excluded	-	-	1	TN
11.	2-Butoxyethyl acetate	1 580.1 ± 48.99	0	**FP**	1 446.1 ± 37.98	0	**FP**	1	TN
12.	2-Chloro-4-nitroaniline	413.12 ± 151.75	0	**FP**	445.10 ± 141.14	0	**FP**	0	**FP**
13.	2-Ethylhexyl acrylate	censored excluded	-	-	censored excluded	-	-	1	TN
14.	2-Phenoxyethanol	936.65 ± 21.99	0	**FP**	996.15 ± 19.84	0	**FP**	0	**FP**
15.	4'-Tert-butyl-2',6'-dimethyl-3',5'-dinitroacetophenone	724.82 ± 197.97	0	**FP**	557.82 ± 128.83	0	**FP**	0	**FP**
16.	Acetophenone	522.58 ± 34.26	0	TP	649.15 ± 36.09	0	TP	0	TP
17.	Aconitine	not tested	-	-	not tested	-	-	not tested	-
18.	Ammonium chloride	527.23 ± 44.69	0	TP	960.78 ± 68.81	0	TP	0	TP
19.	Barium chloride	censored excluded	-	-	censored excluded	-	-	1	**FN**
20.	Benzaldehyde	censored excluded	-	-	censored excluded	-	-	1	**FN**
21.	Benzyl benzoate	> 2 618.5	1	**FN**	censored excluded	-	-	1	**FN**
22.	Brucine	not tested	-	-	not tested	-	-	not tested	-
23.	Caprylic acid	957.49 ± 98.64	0	**FP**	994.01 ± 86.80	0	**FP**	0	**FP**
24.	Copper sulphate	480.51 ± 28.08	0	TP	528.25 ± 26.11	0	TP	0	TP
25.	Diallyl phthalate	722.29 ± 115.06	0	TP	606.56 ± 82.60	0	TP	0	TP
26.	Diepoxide 126	778.49 ± 68.49	0	**FP**	639.43 ± 47.88	0	**FP**	0	**FP**
27.	Di-"isodecyl" phthalate	> 27 213	1	TN	> 9 907.9	1	TN	1	TN
28.	Diisopropanolamine	1 104.4 ± 25.27	0	**FP**	1 165.5 ± 22.62	0	**FP**	1	TN
29.	Dimethyldioctadecylammonium chloride	338.67 ± 175.28	0	**FP**	209.19 ± 92.18	0	**FP**	0	**FP**
30.	Edetic acid	1 262.8 ± 189.83	0	**FP**	897.82 ± 114.11	0	**FP**	0	**FP**
31.	Ethoxyquin	209.87 ± 49.56	0	TP	225.55 ± 45.78	0	TP	0	TP
32.	Ethyl acetoacetate	1 315.8 ± 243.36	0	**FP**	1 365.1 ± 212.70	0	**FP**	1	TN
33.	Ethyl chloroacetate	207.52 ± 3.73	0	TP	294.07 ± 4.49	0	TP	0	TP
34.	Glycerol triacetate	3 364.1 ± 194.81	1	TN	2 368.5 ± 116.55	1	TN	1	TN
35.	Maleic acid	790.19 ± 134.61	0	TP	935.76 ± 136.35	0	TP	0	TP
36.	Malononitrile	not tested	-	-	not tested	-	-	not tested	-
37.	Methenamine	470.34 ± 39.41	0	**FP**	551.69 ± 39.14	0	**FP**	0	**FP**
38.	*N*-isopropyl-*N*'-phenyl-*p*-phenylenediamine	85.44 ± 30.20	0	TP	102.96 ± 31.61	0	TP	0	TP
39.	Octyl 3,4,5-trihydroxybenzoate	35.29 ± 5.33	0	**FP**	44.03 ± 5.68	0	**FP**	0	**FP**
40.	*P*-benzoquinone; quinone	85.36 ± 8.85	0	TP	146.97 ± 12.88	0	TP	0	TP
41.	Phthalic anhydride	1 274.7 ± 23.18	0	**FP**	1 251.4 ± 19.27	0	**FP**	1	TN
42.	Potassium sulfate	3 975.4 ± 144.25	1	TN	3 036.5 ± 93.23	1	TN	1	TN
43.	Resorcinol	429.46 ± 14.37	0	TP	573.05 ± 16.23	0	TP	0	TP
44.	Sodium cyanate	506.60 ± 13.75	0	TP	846.81 ± 19.46	0	TP	0	TP
45.	Sodium salt of chloroacetic acid	535.32 ± 20.50	0	TP	672.47 ± 21.79	0	TP	0	TP
46.	Sorbitan monolaurate	censored excluded	-	-	censored excluded	-	-	censored excluded	-
47.	Tetramethylthiuram monosulphide	42.49 ± 7.99	0	TP	59.50 ± 9.46	0	TP	0	TP
48.	Triethanolamine	2 137.1 ± 91.29	1	TN	1 932.0 ± 69.91	0	**FP**	1	TN
49.	Triethylene glycol dimethacrylate	1 019.6 ± 359.04	0	**FP**	753.18 ± 223.91	0	**FP**	0	**FP**
50.	Tripotassium citrate	2 965.3 ± 224.30	1	TN	1 810.9 ± 116.13	0	**FP**	1	TN
51.	Tris(nonylphenyl) phosphite	censored excluded	-	-	censored excluded	-	-	censored excluded	-
52.	Trizinc bis(orthophosphate)	not tested	-	-	not tested	-	-	not tested	-
53.	Tween 20	2 994.9 ± 419.94	1	TN	943.39 ± 111.69	0	**FP**	0	**FP**
54.	Urea	3 278.6 ± 143.16	1	TN	4 279.2 ± 158.60	1	TN	1	TN
55.	Zinc distearate	not tested	-	-	not tested	-	-	not tested	-
56.	Zinc oxide	not tested	-	-	not tested	-	-	not tested	-

	MILLIMOLE		**WEIGHT**		**ROC**	
TN	11		TN	4	TN	15
FP	15		FP	19	FP	12
TP	17		TP	17	TP	17
FN	1		FN	0	FN	3
SUM	44		SUM	40	SUM	47
% of subst.	78.6		% of subst.	71.4	% of subst.	83.9
Specificity	42.3		**Specificity**	17.4	**Specificity**	55.6
Sensitivity	94.4		**Sensitivity**	100.0	**Sensitivity**	85.0
Accuracy	63.6		**Accuracy**	52.5	**Accuracy**	68.1
NPV	91.7		**NPV**	100.0	**NPV**	83.3
PPV	53.1		**PPV**	47.2	**PPV**	58.6

ADJUNCT DATA SET 1: Testing data of JRC using a non-validated protocol adapted for automated use

Predicted LD50, Predicted Toxicity (dichotomous), Prediction in relation to in vivo reference data

LEGEND

TN = true negative	not tested
TP = true positive	censored.excluded
FN = false negative	*in vivo* positive (classified)
FP = false positive	*in vivo* negative (unclassified)

	LABORATORY:	PREDICTION MODELS APPLIED TO DATA							
	JRC	A) MILLIMOLE REGRESSION			B) WEIGHT REGRESSION			C) ROC Analysis	
Nr.	Chemical	Predicted LD50	Toxicity Prediction (0 = toxic, 1 = non-toxic)	Test outcome (EU CLP)	Predicted LD50	Toxicity Prediction (0 = toxic, 1 = non-toxic)	Test outcome (EU CLP)	Toxicity Prediction (0 = toxic, 1 = non-toxic)	Test outcome (EU CLP)
1.	(4-Ammonio-*m*-tolyl)ethyl(2-hydroxyethyl)ammonium sulphate	163.76 ± 25.89	0	TP	158.98 ± 21.26	0	TP	0	TP
2.	1,2,4-Trichlorobenzene	> 2 398.05	1	**FN**	censored excluded	-	-	1	**FN**
3.	1,2-Benzenedicarboxylic acid	> 2 563.6	1	TN	censored excluded	-	-	1	TN
4.	1,2-Dichlorobenzene	censored excluded	-	-	censored excluded	-	-	1	TN
5.	1-Naphthylamine	125.21 ± 14.99	0	TP	177.88 ± 17.97	0	TP	0	TP
6.	1-Phenyl-3-pyrazolidone	260.02 ± 21.76	0	TP	311.51 ± 22.16	0	TP	0	TP
7.	2-(2-Butoxyethoxy)ethanol	2 262.7 ± 383.34	1	TN	1 946.3 ± 282.05	0	**FP**	1	TN
8.	2,2',6,6'-Tetrabromo-4,4'-isopropylidenediphenol	796.17 ± 35.66	0	**FP**	452.48 ± 17.18	0	**FP**	0	**FP**
9.	2,4,6-Tris(dimethylaminomethyl)phenol	832.34 ± 29.45	0	TP	660.83 ± 19.79	0	TP	0	TP
10.	2,6-Diethylaniline	> 2 149.0	1	TN	censored excluded	-	-	1	TN
11.	2-Butoxyethyl acetate	2 435.8 ± 104.69	1	TN	2 086.7 ± 76.15	1	TN	1	TN
12.	2-Chloro-4-nitroaniline	358.72 ± 22.73	0	**FP**	397.32 ± 21.40	0	**FP**	0	**FP**
13.	2-Ethylhexyl acrylate	> 2 419.0	1	TN	censored excluded	-	-	1	TN
14.	2-Phenoxyethanol	1 321.3 ± 23.65	0	**FP**	1 333.4 ± 20.23	0	**FP**	1	TN
15.	4'-Tert-butyl-2',6'-dimethyl-3',5'-dinitroacetophenone	653.12 ± 92.27	0	**FP**	511.87 ± 61.49	0	**FP**	0	**FP**
16.	Acetophenone	1 792.5 ± 64.65	0	TP	1 845.1 ± 56.43	0	TP	1	**FN**
17.	Aconitine	1 944.3 ± 371.23	0	TP	887.24 ± 145.17	0	TP	0	TP
18.	Ammonium chloride	667.01 ± 87.15	0	TP	1 172.2 ± 129.25	0	TP	1	**FN**
19.	Barium chloride	1 871.3	0	TP	1 471.8	0	TP	1	**FN**
20.	Benzaldehyde	1 211.0 ± 26.34	0	TP	1 403.9 ± 25.89	0	TP	1	**FN**
21.	Benzyl benzoate	> 2 618.5	1	**FN**	censored excluded	-	-	1	**FN**
22.	Brucine	718.67 ± 24.46	0	TP	483.32 ± 13.96	0	TP	0	TP
23.	Caprylic acid	1 289.5 ± 177.96	0	**FP**	1 278.7 ± 150.67	0	**FP**	1	TN
24.	Copper sulphate	466.73 ± 11.77	0	TP	515.45 ± 11.02	0	TP	0	TP
25.	Diallyl phthalate	723.00 ± 13.16	0	TP	607.75 ± 9.38	0	TP	0	TP
26.	Diepoxide 126	620.72 ± 14.89	0	**FP**	527.94 ± 10.73	0	**FP**	0	**FP**
27.	Di-"isodecyl" phthalate	2 955.1	1	TN	1 505.6	0	**FP**	1	TN
28.	Diisopropanolamine	1 222.5 ± 93.27	0	**FP**	1 270.0 ± 82.07	0	**FP**	1	TN
29.	Dimethyldioctadecylammonium chloride	450.34 ± 24.97	0	**FP**	269.34 ± 12.68	0	**FP**	0	**FP**
30.	Edetic acid	1 252.8 ± 168.09	0	**FP**	891.99 ± 101.21	0	**FP**	0	**FP**
31.	Ethoxyquin	225.25 ± 23.88	0	TP	239.98 ± 21.59	0	TP	0	TP
32.	Ethyl acetoacetate	1 849.8 ± 23.03	0	**FP**	1 824.4 ± 19.24	0	**FP**	1	TN
33.	Ethyl chloroacetate	312.83 ± 30.95	0	TP	416.23 ± 34.77	0	TP	0	TP
34.	Glycerol triacetate	3 315.4 ± 797.76	1	TN	2 333.7 ± 480.36	1	TN	1	TN
35.	Maleic acid	1 229.0 ± 54.26	0	TP	1 362.2 ± 51.07	0	TP	1	**FN**
36.	Malononitrile	176.89 ± 15.35	0	TP	344.45 ± 25.38	0	TP	0	TP
37.	Methenamine	733.81 ± 75.91	0	**FP**	804.10 ± 70.39	0	**FP**	0	**FP**
38.	*N*-isopropyl-*N*'-phenyl-*p*-phenylenediamine	90.09 ± 13.72	0	TP	108.21 ± 14.06	0	TP	0	TP
39.	Octyl 3,4,5-trihydroxybenzoate	46.46 ± 7.36	0	**FP**	55.58 ± 7.45	0	**FP**	0	**FP**
40.	*P*-benzoquinone; quinone	78.16 ± 6.17	0	TP	136.43 ± 9.12	0	TP	0	TP
41.	Phthalic anhydride	1 263.6 ± 73.18	0	**FP**	1 241.9 ± 60.94	0	**FP**	1	TN
42.	Potassium sulfate	3 685.4 ± 122.43	1	TN	2 847.8 ± 80.08	1	TN	1	TN
43.	Resorcinol	490.08 ± 19.25	0	TP	640.88 ± 21.30	0	TP	0	TP
44.	Sodium cyanate	555.64 ± 45.48	0	TP	915.54 ± 63.71	0	TP	0	TP
45.	Sodium salt of chloroacetic acid	555.71 ± 57.64	0	TP	693.81 ± 61.31	0	TP	0	TP
46.	Sorbitan monolaurate	889.62 ± 63.77	0	**FP**	615.86 ± 37.36	0	**FP**	0	**FP**
47.	Tetramethylthiuram monosulphide	66.11 ± 13.27	0	TP	86.53 ± 14.68	0	TP	0	TP
48.	Triethanolamine	2 139.9 ± 112.44	1	TN	1 934.1 ± 86.32	0	**FP**	1	TN
49.	Triethylene glycol dimethacrylate	591.84 ± 30.68	0	**FP**	477.42 ± 20.94	0	**FP**	0	**FP**
50.	Tripotassium citrate	2 752.1 ± 100.57	1	TN	1 700.3 ± 52.72	0	**FP**	1	TN
51.	Tris(nonylphenyl) phosphite	censored excluded	-	-	censored excluded	-	-	censored excluded	-
52.	Trizinc bis(orthophosphate)	not tested	-	-	not tested	-	-	not tested	-
53.	Tween 20	2 404.2 ± 418.99	1	TN	782.80 ± 115.08	0	**FP**	0	**FP**
54.	Urea	3 064.8 ± 30.65	1	TN	4 041.9 ± 34.26	1	TN	1	TN
55.	Zinc distearate	not tested	-	-	not tested	-	-	not tested	-
56.	Zinc oxide	not tested	-	-	not tested	-	-	not tested	-

	MILLIMOLE		WEIGHT		ROC	
TN	12		TN	4	TN	17
FP	15		FP	20	FP	11
TP	22		TP	22	TP	17
FN	2		FN	0	FN	7
SUM	51		SUM	46	SUM	52
% of subst.	91.1		% of subst.	82.1	% of subst.	92.9
Specificity	44.4		**Specificity**	16.7	**Specificity**	60.7
Sensitivity	91.7		**Sensitivity**	100.0	**Sensitivity**	70.8
Accuracy	66.7		**Accuracy**	56.5	**Accuracy**	65.4
NPV	85.7		**NPV**	100.0	**NPV**	70.8
PPV	59.5		**PPV**	52.4	**PPV**	60.7

ADJUNCT DATA SET 2: Testing data of IIVS using a non-validated abbreviated protocol (no range finding)

Predicted LD50, Predicted Toxicity (dichotomous), Prediction in relation to in vivo reference data

LEGEND	
TN = true negative	not tested
TP = true positive	censored.excluded
FN = false negative	*in vivo* positive (classified)
FP = false positive	*in vivo* negative (unclassified)

	LABORATORY:	PREDICTION MODELS APPLIED TO DATA							
	IIVS	A) MILLIMOLE REGRESSION			B) WEIGHT REGRESSION			C) ROC Analysis	
Nr.	Chemical	Predicted LD50	Toxicity Prediction (0 = toxic, 1 = non-toxic)	Test outcome (EU CLP)	Predicted LD50	Toxicity Prediction (0 = toxic, 1 = non-toxic)	Test outcome (EU CLP)	Toxicity Prediction (0 = toxic, 1 = non-toxic)	Test outcome (EU CLP)
1.	(4-Ammonio-*m*-tolyl)ethyl(2-hydroxyethyl)ammonium sulphate	< 692.32	0	TP	< 539.95	0	TP	0	TP
2.	1,2,4-Trichlorobenzene	759.99	0	TP	732.71	0	TP	0	TP
3.	1,2-Benzenedicarboxylic acid	> 3 833.0	1	TN	censored excluded	-	-	1	TN
4.	1,2-Dichlorobenzene	> 2 130.9	1	TN	censored excluded	-	-	1	TN
5.	1-Naphthylamine	< 185.98	0	TP	< 248.89	0	TP	0	TP
6.	1-Phenyl-3-pyrazolidone	295.73	0	TP	346.82	0	TP	0	TP
7.	2-(2-Butoxyethoxy)ethanol	2 312.8	1	TN	1 984.9	0	**FP**	1	TN
8.	2,2',6,6'-Tetrabromo-4,4'-isopropylidenediphenol	656.02	0	**FP**	384.03	0	**FP**	0	**FP**
9.	2,4,6-Tris(dimethylaminomethyl)phenol	1 105.9	0	TP	840.80	0	TP	0	TP
10.	2,6-Diethylaniline	1 474.0	0	**FP**	1 405.2	0	**FP**	1	TN
11.	2-Butoxyethyl acetate	> 2 236.3	1	TN	censored excluded	-	-	1	TN
12.	2-Chloro-4-nitroaniline	317.03	0	**FP**	357.53	0	**FP**	0	**FP**
13.	2-Ethylhexyl acrylate	> 2 419.0	1	TN	censored excluded	-	-	1	TN
14.	2-Phenoxyethanol	955.32	0	**FP**	1 012.5	0	**FP**	0	**FP**
15.	4'-Tert-butyl-2',6'-dimethyl-3',5'-dinitroacetophenone	censored excluded	-	-	censored excluded	-	-	censored excluded	-
16.	Acetophenone	710.04	0	TP	841.81	0	TP	0	TP
17.	Aconitine	> 2 186.5	1	**FN**	censored excluded	-	-	censored excluded	-
18.	Ammonium chloride	499.44	0	TP	917.98	0	TP	0	TP
19.	Barium chloride	1 084.6	0	TP	928.11	0	TP	0	TP
20.	Benzaldehyde	458.42	0	TP	616.38	0	TP	0	TP
21.	Benzyl benzoate	1 574.3	0	TP	1 261.2	0	TP	1	**FN**
22.	Brucine	609.76	0	TP	420.35	0	TP	0	TP
23.	Caprylic acid	1 072.6	0	**FP**	1 094.9	0	**FP**	1	TN
24.	Copper sulphate	< 163.93	0	TP	< 212.40	0	TP	0	TP
25.	Diallyl phthalate	733.75	0	TP	615.41	0	TP	0	TP
26.	Diepoxide 126	780.41	0	**FP**	640.68	0	**FP**	0	**FP**
27.	Di-"isodecyl" phthalate	> 3 975.	1	TN	censored excluded	-	-	1	TN
28.	Diisopropanolamine	947.78	0	**FP**	1 023.9	0	**FP**	0	**FP**
29.	Dimethyldioctadecylammonium chloride	460.28	0	**FP**	274.40	0	**FP**	0	**FP**
30.	Edetic acid	1 041.4	0	**FP**	763.26	0	**FP**	0	**FP**
31.	Ethoxyquin	291.24	0	TP	298.49	0	TP	0	TP
32.	Ethyl acetoacetate	1 084.6	0	**FP**	1 160.5	0	**FP**	1	TN
33.	Ethyl chloroacetate	273.41	0	TP	371.48	0	TP	0	TP
34.	Glycerol triacetate	3 799.0	1	TN	2 625.8	1	TN	1	TN
35.	Maleic acid	925.16	0	TP	1 070.9	0	TP	1	**FN**
36.	Malononitrile	< 300.55	0	TP	< 539.95	0	TP	0	TP
37.	Methenamine	711.73	0	**FP**	783.42	0	**FP**	0	**FP**
38.	*N*-isopropyl-*N*'-phenyl-*p*-phenylenediamine	< 240.44	0	TP	< 248.89	0	TP	0	TP
39.	Octyl 3,4,5-trihydroxybenzoate	< 272.19	0	**FP**	< 248.89	0	**FP**	0	**FP**
40.	*P*-benzoquinone; quinone	< 144.17	0	TP	< 229.27	0	TP	0	TP
41.	Phthalic anhydride	censored excluded	-	-	censored excluded	-	-	censored excluded	-
42.	Potassium sulfate	3 216.1	1	TN	2 537.3	1	TN	1	TN
43.	Resorcinol	< 400.31	0	TP	< 539.95	0	TP	0	TP
44.	Sodium cyanate	588.16	0	TP	961.00	0	TP	0	TP
45.	Sodium salt of chloroacetic acid	428.60	0	TP	557.02	0	TP	0	TP
46.	Sorbitan monolaurate	913.69	0	**FP**	630.08	0	**FP**	0	**FP**
47.	Tetramethylthiuram monosulphide	< 190.38	0	TP	< 212.40	0	TP	0	TP
48.	Triethanolamine	1 890.8	0	**FP**	1 741.7	0	**FP**	1	TN
49.	Triethylene glycol dimethacrylate	822.23	0	**FP**	630.28	0	**FP**	0	**FP**
50.	Tripotassium citrate	2 872.8	1	TN	1 763.3	0	**FP**	1	TN
51.	Tris(nonylphenyl) phosphite	> 4 596.1	1	TN	censored excluded	-	-	1	TN
52.	Trizinc bis(orthophosphate)	831.31	0	**FP**	552.39	0	**FP**	0	**FP**
53.	Tween 20	2 544.1	1	TN	822.25	0	**FP**	0	**FP**
54.	Urea	2 096.3	1	TN	2 929.6	1	TN	1	TN
55.	Zinc distearate	1 639.2	0	**FP**	776.19	0	**FP**	0	**FP**
56.	Zinc oxide	< 112.35	0	**FP**	< 212.40	0	**FP**	0	**FP**

	MILLIMOLE		WEIGHT		ROC	
TN	12		TN	3	TN	15
FP	18		FP	21	FP	15
TP	23		TP	23	TP	21
FN	1		FN	0	FN	2
SUM	54		SUM	47	SUM	53
% of subst.	96.4		% of subst.	83.9	% of subst.	94.6
Specificity	40.0		**Specificity**	12.5	**Specificity**	50.0
Sensitivity	95.8		**Sensitivity**	100.0	**Sensitivity**	91.3
Accuracy	64.8		**Accuracy**	55.3	**Accuracy**	67.9
NPV	92.3		**NPV**	100.0	**NPV**	88.2
PPV	56.1		**PPV**	52.3	**PPV**	58.3

Analysis of the testing data of the NICEATM/ECVAM validation study using the 2000 mg/kg cutoff criteria

TN = true negative
TP = true positive
FN = false negative
FP = false positive
Data not included in Table 6-7 in BRD.
Marked as LD50 > 5000 mg/kg in Table L2-1.

Nr.	Chemical	Predicted LD50	Toxicity Prediction (0 = toxic, 1 = non-toxic)	Test outcome (EU CLP)	In vivo ref data
1	ACETAMINOPHEN	380.56	0	**FP**	2163
2	ACETONITRILE	1731.27	0	**FP**	3598
3	ACETYLSALICYLIC ACID	1345.69	0	TP	1506
4	AMINOPTERIN	15.10	0	TP	7
5	5-AMINOSALICYLIC ACID	1824.65	0	**FP**	3429
6	AMITRIPTYLINE HCL	247.83	0	TP	348
7	ARSENIC III TRIOXIDE	109.08	0	TP	25
8	ATROPINE SULFATE	1099.11	0	TP	819
9	BORIC ACID	1148.61	0	**FP**	3426
10	BUSULFAN	620.12	0	TP	12
11	CADMIUM II CHLORIDE	58.29	0	TP	135
12	CAFFEINE	729.78	0	TP	310
13	CARBAMAZEPINE	686.30	0	**FP**	2805
14	CARBON TETRACHLORIDE				3783
15	CHLORAL HYDRATE	722.62	0	TP	638
16	CHLORAMPHENICOL	901.50	0	**FP**	3491
17	CITRIC ACID	1498.15	0	**FP**	5929
18	COLCHICINE				15
19	CUPRIC SULFATE PENTAHY	477.66	0	TP	474
20	CYCLOHEXIMIDE	47.40	0	TP	2
21	DIBUTYL PHTHALATE	545.93	0	**FP**	8892
22	DICHLORVOS	305.15	0	TP	59
23	DIETHYL PHTHALATE	674.43	0	**FP**	9311
24	DIGOXIN	2600.15	1	**FN**	28
25	DIMETHYLFORMAMIDE	1990.03	0	**FP**	5309
26	DIQUAT DIBROMIDE MONOH	284.38	0	TP	160
27	DISULFOTON	834.74	0	TP	5
28	ENDOSULFAN	273.70	0	TP	28
29	EPINEPHRINE BITARTRATE				4
30	ETHANOL	1693.24	0	**FP**	11324
31	ETHYLENE GLYCOL	3565.84	1	TN	7161
32	FENPROPATHRIN	452.50	0	TP	76
33	GIBBERELLIC ACID	5682.90	1	TN	6040
34	GLUTETHIMIDE	824.43	0	TP	600
35	GLYCEROL	4476.80	1	TN	19770
36	HALOPERIDOL	257.74	0	TP	330
37	HEXACHLOROPHENE	228.19	0	TP	82
38	LACTIC ACID	1765.14	0	**FP**	3639
39	LINDANE	785.26	0	TP	100
40	LITHIUM I CARBONATE	753.93	0	TP	590
41	MEPROBAMATE	1333.95	0	TP	1387
42	MERCURY II CHLORIDE	180.47	0	TP	40
43	METHANOL				8710
44	NICOTINE	961.78	0	TP	70
45	PARAQUAT	348.85	0	TP	93
46	PARATHION	494.39	0	TP	6
47	PHENOBARBITAL	1442.40	0	TP	224
48	PHENOL	337.23	0	TP	548
49	PHENYLTHIOUREA	476.81	0	TP	3
50	PHYSOSTIGMINE	406.92	0	TP	5
51	POTASSIUM I CHLORIDE	1699.16	0	**FP**	2799
52	POTASSIUM CYANIDE	206.26	0	TP	7
53	PROCAINAMIDE HCL	1404.01	0	TP	1950
54	2-PROPANOL	1517.53	0	**FP**	5105
55	PROPRANOLOL	322.32	0	TP	466
56	PROPYLPARABEN				6332
57	SODIUM ARSENITE	56.77	0	TP	44
58	SODIUM CHLORIDE	1680.37	0	**FP**	4046
59	SODIUM DICHROMATE DIHY	80.03	0	TP	51
60	SODIUM I FLUORIDE	230.28	0	TP	127
61	SODIUM HYPOCHLORITE	1015.54	0	**FP**	10328
62	SODIUM OXALATE	321.01	0	TP	633
63	SODIUM SELENATE	347.03	0	TP	3
64	STRYCHNINE	1005.04	0	TP	6
65	THALLIUM I SULFATE	296.92	0	TP	25
66	TRICHLOROACETIC ACID	1445.20	0	**FP**	5229
67	1,1,1-TRICHLOROETHANE	4711.55	1	TN	12078
68	TRIETHYLENEMELAMINE	46.66	0	TP	4
69	TRIPHENYLTIN HYDROXIDE	19.29	0	TP	329
70	VALPROIC ACID	1356.74	0	TP	995
71	VERAPAMIL HCL	642.86	0	TP	111
72	XYLENE	1028.26	0	**FP**	4667

TN	4
FP	18
TP	44
FN	1
SUM	67
Specificity	18.18
Sensitivity	97.78
Accuracy	71.64
NPV	80.00
PPV	70.97

Annex 3 – ESAC Request concerning the review of the follow-up study

ESAC Request 2011-02

ECVAM Scientific Advisory Committee

(ESAC)

ECVAM REQUEST FOR ESAC ADVICE

on

the ECVAM-coordinated follow-up study to assess the predictive capacity of the already validated Neutral Red Uptake cytotoxicity assay for acute oral toxicity testing.

INSTRUCTIONS FOR IVM/ST STAFF:

Blue text: to be filled in by the ECVAM Scientific Officer completing the draft request in collaboration with ESAC Secretariat.

Green text: to be filled in by the ESAC Secretariat.

Title page information	
Abbreviated title of ESAC request	Follow-up study to assess the PC of the Neutral Red Uptake cytotoxicity assay for acute oral toxicity testing
ESAC REQUEST Nr.	2011-02
Template used for preparing request	EP 2.01
Date of finalising request	2011-03-07
Date of submitting request to ESAC	2011-03-09
Request discussed through	Plenary discussion at ESAC 34, 22-23 March 2011
ESAC WG report / opinion expected at (date)	ESAC 35, 4-5 October 2011
File name of this request	ER2011-02 3T3 NRU follow-up MANDATE.doc

1. TYPE OF REQUEST

Request Type		Identify request ("YES")
R1 ESAC Peer Review of a Prevalidation Study or Validation Study		YES
If R1)applies please specify further:		
• Prevalidation Study		
• Prospective Validation Study		YES This study, finished in 2010, was planned and conducted as a follow-up to the previous full prospective validation study of the 3T3 NRU cytotoxicity assay conducted by NICEATM in collaboration with ECVAM and finalised in 2005. The study was designed to complement the information on predictive capacity of the 3T3 NRU assay for the specific purpose of identifying substances that do not need to be labelled for acute oral toxicity according to the EU CLP regulation (i.e. substances with LD_{50} doses above the limit dose of 2000 mg/kg body weight). Importantly, being a follow-up and complement exercise according to ECVAM's modular approach, the study deviates to some extent from the typical design of a full prospective validation study (e.g. no transferability/reproducibility assessment as these have been addressed in the previous study).
• Retrospective Validation Study		
• Validation Study based on Performance Standards		
R2 Scientific Advice on a test method submitted to ECVAM for validation (e.g. the test method's biological relevance etc.)		NO
R3 Other Scientific Advice (e.g. on test methods, their use; on technical issues such as cell culturing, stem cells, definition of performance standards etc.)		NO

2. TITLE OF STUDY OR PROJECT FOR WHICH SCIENTIFIC ADVICE OF THE ESAC IS REQUESTED

Follow-up study on the predictive capacity of the 3T3 Neutral Red Uptake cytotoxicity assay to correctly identify substances not classified for acute oral toxicity under the EU CLP system (LD_{50} > 2000 mg/kg).

3. BRIEF DESCRIPTION OF THE STUDY OR PROJECT

3.1 Summary of the follow-up study

This follow-up study was conducted as a complement to the previous full prospective validation study of the 3T3 Neutral Red Uptake (NRU) cytotoxicity assay (=the "3T3 NRU assay") conducted by NICEATM/ICCVAM in collaboration with ECVAM. The test exploits the correlation between the systemic toxicity (i.e. acute oral toxicity) of substances and their cytotoxicity exerted on 3T3 cells. Cytotoxicity is measured as reduction of uptake of the vital dye 'neutral red', which accumulates in lysosomes of healthy cells.

As a follow-up, the study deviates to some extent from the manner in which a full prospective validation exercise is typically conducted. The study was designed to specifically assess whether the 3T3 NRU assay is able to discriminate classified chemicals from non-classified ones (i.e. those beyond the limit dose of 2000mg/kg according to the EU CLP regulation implementing UN GHS). Thus, the study was intended to provide additional information on predictive capacity of the 3T3 NRU assay for this specific purpose, without addressing reproducibility/transferability of the protocol which had been previously demonstrated in the NICEATM/ECVAM validation study.

To assess the capacity of the 3T3 NRU to correctly identify chemicals not requiring classification, 56 test items (a sufficient number to analyse dichotomous classifications) with good in vivo reference data were selected and tested in one laboratory using the already validated protocol. In addition, the same chemicals were tested in two more laboratories using slight modifications of the protocol. These variations were 1) an abbreviated version of the validated protocol, and 2) a protocol modified for use on an automated platform. This additional testing was intended to provide information on the extent to which the original protocol is amenable to simplification (protocol variant 1), and automation (protocol variant 2).

The testing data of the validated protocol show that the 3T3 NRU identifies true positives with a sensitivity of 94%. Since the 3T3 NRU is able to correctly identify most positives, negative test results in the 3T3 NRU are very likely to represent either true negatives (non-classified chemicals) or false positives. In contrast the rate of false negatives is low. This is reflected by the high NPV (negative predictive value) of 92%. Therefore, the 3T3 NRU may be appropriate for identifying negatives as a first screening step in a tiered testing approach involving subsequent in vivo testing to 1) further categorise chemicals with positive results, 2) to identify false positive results of the 3T3 NRU (low specificity of 42%) and 3) to test, in specific cases where there is additional weight of evidence information, negative substances for confirmation. The two protocol variants gave similar predictive values suggesting that also these variants of the validated 3T3 NRU protocol may be used for the

screening of non-classified chemicals according to EU CLP within a tiered testing strategy.

3.2 Detailed background

Several international projects have studied the possibility of using in vitro methods to predict acute oral toxicity.

The first of these studies was the Multicentre Evaluation of In Vitro Cytotoxicity (MEIC) programme. It showed that in vitro methods used in the study predicted human acute oral lethality better than did mouse and rat in vivo LD_{50} data.

In a second study based on information of the Registry of Cytotoxicity (RC, a database for rodent acute oral LD_{50} values and in vitro IC_{50} values), over 70% of the substances tested in vitro were able to predict the rodent acute oral lethality.

Third, the international NICEATM/ECVAM validation study (the In Vitro Basal Cytotoxicity Validation Study finished in 2005) used a human-derived cell model (primary normal human epidermal keratinocytes) and a mouse cell model (BALB/c 3T3 mouse fibroblasts) to evaluate the usefulness and limitations of the in vitro basal cytotoxicity test methods based on measuring cell viability trough neutral red uptake (NRU) for predicting starting doses for systemic (i.e. in vivo) acute oral toxicity test methods. In addition, this validation study assessed the accuracy of the two basal cytotoxicity test methods to estimate rodent oral LD_{50} values across the five categories of the Globally Harmonized System (GHS) for acute oral toxicity as well as unclassified toxicities. The study concluded that the two NRU test methods could be used in a weight-of-evidence approach to determine the starting dose for acute oral in vivo toxicity protocols. The validation study also showed that the overall accuracy of the 3T3 NRU test method for correctly predicting each of the GSH acute oral toxicity classification categories was low (around 30%), however, substances falling in the GHS 4 category (i.e. $300 < LD_{50} \bullet 2.000$ mg/kg) were predicted better, with 81% accuracy.

Taken together, the results of MEIC, the RC, and the NICEATM/ECVAM international validation study have all shown a correlation of around 60-70% between in vitro cytotoxicity data and oral rodent LD_{50} values. These studies indicated that the in vitro methods are able to predict low systemic toxicity with much greater precision than high systemic toxicity, suggesting the potential usefulness of these methods for identifying chemicals not requiring classification.

3.3 Purpose of the study

This follow-up study was initiated in 2008 by ECVAM and was finalised in October 2010. The aim of this study was to further explore, on the basis of the previous validation study, whether the predictive capacity (e.g. sensitivity, specificity, concordance) of the 3T3/NRU cytotoxicity assay is sufficient to correctly distinguish chemicals not requiring classification for acute oral toxicity according to provisions of the EU CLP regulation (i.e. $LD_{50} > 2000$ mg/kg b.w.) from those that require classification ($LD_{50} \bullet 2000$ mg/kg b.w.). The scientific and regulatory rationale embedded in study's objective was to assess whether the 3T3 NRU assay could be used as the first step of a tiered approach to identify unclassified chemicals so that subsequent testing in vivo would focus on confirmatory testing to classify positives according to the 4 classified classes of EU CLP and identify substances with positive test results in the 3T3 NRU that are actual negative (=3T3 NRU false positives).

The study used the test method protocol validated in the NICEATM/ECVAM validation study. In

addition, two protocol modifications were assessed: one version of the 3T3/NRU protocol adapted to an automated platform and an abbreviated version of the validated protocol that was targeted at resolving acute oral toxicities around the 2000 mg/kg cut-off value. The aim of this additional testing was to assess whether a simplified version and a version adapted for automated testing would generate similar data on the basis of the 56 test chemicals selected and to assess, therefore, to which extent these variants of the validated protocol may be used for purposes of identifying negatives (LD_{50} > 2000 mg/kg b.w.).

3.4 Organisation of the study

The study was coordinated and managed by a Validation Management Team composed of two ECVAM staff members. Although testing was performed in three laboratories, the core validation exercise (aiming at more detailed information on predictive capacity) concerned only laboratory Nr. 1 which worked with the validated protocol. Laboratories 2 and 3 produced additional data on the basis of two protocol variants supporting a comparative analysis of protocol performance. The laboratories were:

1) Health and Safety Laboratory (HSL), UK (under ECVAM sponsored contract) using the already validated manual test method protocol
2) JRC (IHCP), Italy using the automated version of the test method protocol
3) IIVS, US (sponsored by IIVS and PETA, People for the Ethical Treatment of Animals) using the abbreviated test method protocol

A set of 56 coded industrial chemicals (including cosmetic ingredients) were tested using each test method protocol. The chemicals were purchased from Sigma-Aldrich (Italy) and coded by ECVAM. The distribution of chemicals and respective material safety data sheets were done by Sigma-Aldrich Germany (for the two European laboratories) and Sigma-Italy for the laboratory in the US. The data from blind testing were de-coded and analysed independently by ECVAM.

3.5 Results and conclusions

The results of all three protocol variants show that the 3T3/NRU assay has high sensitivity (92-96%) and high negative predictive value (86-92%). This indicates that compounds identified as negatives by the method (40% - 44%) will most likely be correctly categorised as unclassified (LD50 > 2000 mg/kg b.w.). Therefore, if the above proposed tiered strategy is applied, negatives may not be required to be tested in subsequent confirmatory in vivo testing. Positives of the 3T3 NRU however would require confirmatory in vivo testing on the basis of a starting dose approach as validated in the NICEATM-ECVAM validation study.

A recent analysis of the New Chemicals Database showed that over 85% of new industrial chemicals do not require classification for acute oral toxicity according to EU CLP (LD50 > 2000 mg/kg b.w.). With the 3T3/NRU method, which was demonstrated of being able to correctly identify about 42% of all true negatives, a testing strategy could be developed, limiting animal testing to only those substances identified as "classified" by the 3T3/NRU assay.

References

Clemedson et al. 1996. MEIC evaluation of acute systemic toxicity. Alt. Lab. Anim.24: 249-272

Ekwall B, Clemedson C, Craaford B, Ekwall B, Hallander S, Walum E, Bondesson I. 1998a. MEIC

evaluation of acute systemic toxicity. Part V. Rodent and human toxicity data for the 50 reference chemicals. Alt. Lab. Anim. 26 (suppl 2):571-616.

Ekwall B, Clemedson C, Craaford B, Ekwall B, Hallander S, Walum E, Bondesson I. 1998b. MEIC evaluation of acute systemic toxicity. Part VI. The prediction of human toxicity by rodent LD50 values and results from 61 in vitro methods. Altern Lab Anim 26 (suppl 2):617- 658.

Ekwall B, Clemedson C, Ekwall B, Ring P, Romert L. 1999. EDIT: A new international multicentre programme to develop and evaluate batteries on in vitro tests for acute and chronic systemic toxicity. Alt. Lab. Anim. 27:339-349

Ekwall B, Ekwall B, Sjostrom M. 2000. MEIC evaluation of acute systemic toxicity: Part VIII. Multivariate partial least squares evaluation, including the selection of a battery cell line tests with a good prediction of human acute lethal peak blood concentrations for 50 chemicals. Alt. Lab. Anim. 28 (suppl 1):201-34.

Halle W. 2003. The Registry of Cytotoxicity: Toxicity Tesing in Cell Cultures to predict Acute Toxicity (LD50) and to reduce Tesing in Animals. Alt. Lab. Anim. 31: 89-198.

Anon, 2006. Background Review Document (BRD): Validation of Neutral Red Uptake Test Methods NIH /In Vitro Cytotoxicity Test Methods for Estimating Acute Oral Systemic Toxicity. Publication No. 07-4518, November 2006, http://iccvam.niehs.nih.gov/methods/acutetox/inv_nru_brd.htm.

EU 2008. Regulation (EC) No 1272/2008 on classification, labelling and packaging of substances and mixtures, amending and repealing directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006 (REACH).

Bulgheroni, A., Kinsner-Ovaskainen, A., Hoffmann, S., Hartung, T., Prieto, P., 2009. Estimation of acute oral toxicity using the No Adverse Effect Level (NOAEL) from the 28-day repeated dose toxicity studies in rats. Regulatory Toxicology and Regulatory Toxicology and Pharmacology 53, 16–19.

Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., Hoffmann, S., Roi, A.J., Prieto, P., Sabbioni, E., Scott, L., Worth, A., Zuang, V., 2004. A modular approach to the ECVAM principles on test validity. Alt. Lab. Anim. 32:467–472.

4. OBJECTIVES, QUESTIONS, TIMELINES

4.1 OBJECTIVE

Objective Why does ECVAM require advice on the current issue?	The opinion of ESAC should support ECVAM with respect to the development of further recommendations regarding the ability of the 3T3/NRU test method to correctly identify substances not requiring classification for acute oral toxicity under the EU CLP system (LD50 > 2000 mg/kg b.w.) and the use of the test method in a tiered testing approach for acute oral toxicity testing.

4.2 QUESTION(S) TO BE ADDRESSED

Questions What are the questions and issues that should be addressed in view of achieving the objective of the advice?	1) DESIGN & CONDUCT OF STUDY: The ESAC is requested to review whether the validation study was conducted appropriately in view of the objective of the study, i.e. to assess the ability of the 3T3/NRU test method to correctly identify substances not requiring classification for acute oral toxicity under the EU CLP system (LD50 > 2000 mg/kg b.w.). In particular the following issues should be addressed: (a) Clarity of the definition of the study objective. (b) Appropriateness of the study design in view of study objective, inter alia: o Were the criteria for chemical selection appropriate? o Is the toxicity range of the selected chemicals appropriate for the purpose of the study (i.e. analysis of the ability to distinguish at the 2000mg/Kg b.w. threshold)? o In case of gaps (chemical class etc.) – are these justified? o Is the number of chemicals sufficient? o Is the number of laboratories sufficient? (c) Appropriateness of the study execution (e.g. were there pre-defined acceptance criteria, were these respected? How were exceptions / deviations handled, e.g. censoring of values, retesting etc?). (d) Appropriateness of the statistical analysis used for analysing predictive capacity. 2) CONCLUSIONS OF STUDY: The ESAC is requested to assess whether the conclusions, as presented in the Validation Study Report (VSR), are substantiated by the information generated during validation and are plausible with respect to existing information and current views (e.g. literature). In particular: (a) Do the data on the basis of these chemicals provide new information on applicability and possible limitations (in addition to the original information available upon completion of the original ICCAM/ECVAM study)?

	(b) Are the conclusions on predictive capacity justified and plausible with respect to existing information (c) Is the information on the two protocol variants (abbreviated and automated version) sufficient in view of supporting their standardized use alongside the already validated protocol? (d) Are there possible gaps between study design and study conclusions which remain to be addressed in view of the suggested conclusions / use (see also point 3)? 3) SUGGESTED USE OF THE TEST METHOD: The ESAC is requested to review the suggested use of the validated method within a strategy to identify only unclassified chemicals (LD50 > 2000 mg/kg b.w.) as proposed by the Validation Management Team. (a) Is the suggested use of the test method, based on the information generated in the Validation Study, plausible and scientifically justified? (b) Is there additional information required (i.e. are there gaps) to be able to conclude on the plausibility of the suggested use?

4.3 TIMELINES

Timelines concerning this request When does ECVAM require the advice?	Timeline	Indication
	Draft/finalised ESAC Opinion required by:	ESAC 35, 4-5 October 2011
	Request to be presented to ESAC by written procedure (e.g. due to urgency) prior to the next ESAC	NO
	Request to be presented to ESAC at ESAC plenary meeting	ESAC 34, 22-23 March 2011

5. ECVAM PROPOSALS ON HOW TO ADDRESS THE REQUEST WITHIN ESAC

5.1 ECVAM PROPOSAL REGARDING REQUEST-RELATED STRUCTURES REQUIRED

Specific structures required within ESAC to address the request Does the advice require an ESAC working group, an ESAC rapporteur etc.?	Structure(s) required	Required according to ECVAM? (YES/NO)
	S1 ESAC Rapporteur	NO
	S2 ESAC Working Group	YES Proposals from (a) ECVAM, (b) ESAC members and (c) ICATM partner organisations are listed in a separate document
	S3 Invited Experts	NO
	Ad S3: If yes – list names and affiliations of suggested experts to be invited and specify whether these are member of the EEP	
	If other than above (S1-S3):	

5.2 DELIVERABLES AS PROPOSED BY ECVAM

Deliverables What deliverables (other than the ESAC opinion) are required for addressing the request?	Title of deliverable other than ESAC opinion	Required? (YES/NO)
	D1 ESAC Rapporteur Report and draft opinion	NO
	D2 ESAC Peer Review Report and draft opinion	YES
	If other than above (D1-D2):	

6. LIST OF DOCUMENTS TO BE MADE AVAILABLE TO THE ESAC

Count	Description of document	Available (YES/NO)	File name
1	Final Study Report	yes	Final 3T3 NRU study report_March 2011
2	Study protocol HSL	yes	Annex 1_Study protocol of HSL
3	Study protocol JRC	yes	Annex 2_Study protocol of JRC
4	Study protocol IIVS	yes	Annex 3_Study protocol of IIVS
5	Solubility protocol	yes	Annex 4_ Solubility protocol
6	Seidle et al._2010_Cross sector review drivers and available 3Rs approaches acute toxicity testing	yes	2010_Seidel et al._ Toxicological Sciences
7	Creton et al._2010_Acute toxicity testing of chemicals—Opportunities to avoid redundant testing and use alternative approaches	yes	2010_Creton et al., Critical Reviews in Toxicology
8	Bulgheroni et al._2009_Estimation of acute oral toxicity using NOAEL	yes	2009_Bulgheroni et al._Toxicology In Vitro

7. TERMS OF REFERENCE OF THE ESAC WORKING GROUP

7.1 ESTABLISHMENT OF THE ESAC WORKING GROUP

During its 34th meeting on 22./23. March the ESAC plenary unanimously decided to establish an ESAC Working Group charged with the detailed scientific review of the ECVAM follow-up study on the predictive capacity of the validated 3T3 NRU assay for acute toxicity testing.

7.2 TITLE OF THE ESAC WORKING GROUP

Full title:
ESAC Working Group for the detailed scientific peer review of the ECVAM follow-up study of the 3T3 NRU assay for acute toxicity testing

Abbreviated title:
ESAC WG 3T3 NRU

7.3 MANDATE OF THE ESAC WG

The EWG is requested to conduct a scientific review of the ECVAM-conducted follow-up study concerning the predictive capacity of the 3T3 NRU assay. The review needs to address the questions put forward to ESAC by ECVAM and the more detailed questions developed by the ESAC members of the ESAC WG in collaboration with the ESAC Chair, Vice Chair and Secretariat.

The review should focus on the appropriateness of design and conduct of the study in view of the study objective and should provide an appraisal to which extent the conclusions drawn in the Validation Study Report are substantiated by the information generated during the study and how the information generated relates to the scientific background available.

7.4 DELIVERABLE OF THE ESAC WG

The ESAC WG is requested to deliver to the chair of the ESAC and the ESAC Secretariat a detailed ESAC Working Group Report outlining its analyses and conclusions. A reporting template has been appended (Appendix 1) intended to facilitate the drafting of the report.

The conclusions drawn in the report should be based preferably on consensus. If no consensus can be achieved, the report should clearly outline the differences in the appraisals and provide appropriate scientific justifications.

7.5 PROPOSED TIMELINES OF THE ESAC WG

The Secretariat has proposed timelines which should be agreed upon during the first Teleconference (Item 1 in the table):

Item	Proposed date/time	Action	Deliverable
1	Mid April	Teleconference to discuss/decide 1. the list of proposed external (non-ESAC) experts for the ESAC WG 3T3 NRU 2. the more detailed questions to put forward to the ESAC WG	1. List with 3 preferred options (3 external experts + 3 ESAC members = 6 experts in total) 2. Consolidated list of questions
2	Mid April	Both deliverables of item 1 to go to the ESAC for approval / amendment	Amended deliverables as listed under item 1 (if appropriate)
3	Kick-off teleconference in May or June	Discuss the organisation of review and drafting of report, distribution of work. Discuss the studies. Agree on the meeting date and further timelines.	Minutes and agreed meeting date/timelines, work organisation.
	WG meeting in September 2011	Finalisation of draft WG report. Preparation of presentation to ESAC.	1. Preliminary draft report. 2. Presentation of key elements (ESAC)

7.6 QUESTIONS WHICH SHOULD BE ADDRESSED BY THE ESAC WG

The ESAC WG is requested to address the three questions posed to the ESAC which have been broken down further in more specific questions by the ESAC chair, the chair of the ESAC WG and the Secretariat (see section 4.2).

When preparing the final ESAC WG report to address these questions, the ESAC WG is requested to use a pre-defined reporting template. This template (see appendix 1) follows ECVAM's modular approach and addresses to which extent the standard information requirements have been addressed by the study. The template allows moreover for addressing the issues specific studies outlined in section 4.2. The Secretariat will provide guidance if necessary.

APPENDIX 1 REPORTING STRUCTURE FOR THE ESAC WG REPORT

The following suggested structure follows the ECVAM information requirements ("modules") for scientific review following validation and allows at the same time for the description of the analysis and conclusions concerning more specific questions. A template has been created on the basis of the structure below and this template will be made available to the ESAC.

The template can be used for various types of validation studies (e.g. prospective full studies, retrospective studies, performance-based studies and prevalidation studies). Depending on the study type and the objective of the study, not all sections may be applicable. However, for reasons of consistency and to clearly identify which information requirements have not been sufficiently addressed by a specific study, this template is uniformly used for the evaluation of validation studies.

1. Data collection
    - 1.1 Information / data sources used
    - 1.2 Search strategy
    - 1.3 Selection criteria applied to the available information

2. Study objective and design
    - 2.1 Clarity of the definition of the study objective
    - 2.2 Analysis of the scientific rationale provided
    - 2.3 Analysis of the regulatory rationale provided
    - 2.4 Appropriateness of the study design
    - 2.5 Appropriateness of the statistical evaluation

3. Test definition (Module 1)
    - 3.1 Quality and completeness of the overall test definition
    - 3.2 Quality of the background provided concerning the purpose of the test method
    - 3.3 Quality of the documentation and completeness of (a) standardised protocols (SOPs) and (b) prediction models

4. Data quality
    - 4.1 Overall quality of the evaluated data
    - 4.2 Sufficiency of the evaluated data in view of the study objective
    - 4.3 Quality of the reference data for evaluating reliability and relevance[3]

5. Test materials
    - 5.1 Sufficiency of the number of evaluated test items in view of the study objective
    - 5.2 Representativeness of the test items with respect to applicability

6. Within-laboratory reproducibility (Module 2)

[3] OECD guidance document Nr. 34 on validation defines relevance as follows: "Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It sis the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of accuracy (concordance) of a test method."

6.1 Assessment of repeatability and reproducibility in the same laboratory
6.2 Conclusion on within-laboratory reproducibility as assessed by the study

7. Transferability (Module 3)
7.1 Quality of design and analysis of the transfer phase
7.2 Conclusion on transferability to a second laboratory/other laboratories as assessed by the study

8. Between-laboratory reproducibility (Module 4)
8.1 Assessment of reproducibility in different laboratories
8.2 Conclusion on reproducibility as assessed by the study

9. Predictive capacity (Module 5)
9.1 Adequacy of the assessment of the predictive capacity in view of the purpose
9.2 Overall relevance (biological relevance and accuracy) of the test method in view of the purpose

10. Applicability domain (Module 6)
10.1 Appropriateness of study design to conclude on applicability domain, limitations and exclusions
10.2 Quality of the description of applicability domain, limitations, exclusions

11. Performance standards (Module 7)
11.1 Adequacy of the proposed Essential Test Method Components
11.2 Adequacy of the Reference Chemicals
11.3. Adequacy of the defined Accuracy Values

12. Readiness for standardised use
12.1 Assessment of the readiness for regulatory purposes
12.2. Assessment of the readiness for other uses (in house screening etc.)
12.3 Critical aspects impacting on standardized use
12.4 Gap analysis

13. Other considerations

14. Conclusions on the study
14.1 Summary of the results and conclusions of the study
14.2 Extent to which conclusions are justified by the study results alone
14.3 Extent to which conclusions are plausible in the context of existing information

15. Recommendations

15.1 General recommendations concerning the study
15.2 Recommendations concerning the test method (test system, protocol, prediction model)

16. References

17. Annexes