EUROPEAN COMMISSION
DIRECTORATE GENERAL JRC
JOINT RESEARCH CENTRE
Institute for Health and Consumer Protection
European Centre for the Validation of Alternative Methods (ECVAM)

PEER REVIEW PANEL REPORT ON THE RETROSPECTIVE VALIDATION OF CYTOTOXICITY/CELL-FUNCTION BASED *IN VITRO* ASSAYS (EYE IRRITATION)

Background

To satisfy in-house decision making and statutory requirements, a range of substances that might either by intention or accidentally come into contact with the eye (for example pharmaceuticals, cosmetics, agrochemicals) have to be evaluated to determine and manage any potential human ocular hazard. Where the testing of substances is undertaken to inform the regulatory assessment of ocular hazard studies involving the *in vivo* rabbit eye test (Draize eye test) data is generally required.

As yet, no full *in vitro* replacement for the rabbit Draize eye test is available. Two *ex vivo* organotypic test methods, the Bovine Corneal Opacity and Permeability (BCOP) and the Isolated Chicken Eye (ICE) test methods[1] have been scientifically validated at international level and can be accepted for regulatory purposes to identify severe ocular irritants and corrosive materials.

This validation study report evaluates the existing evidence and analysis with respect to four cell-based assays to either screen-out severe irritants in a Top-Down approach, and/or non-irritants in a Bottom-Up approach as part of a weight of evidence approach in a tiered testing strategy.

The test methods, and protocols, evaluated are:

- Neutral red release test (INVITTOX 54 and Predisafe),
- Fluorescence leakage (INVITOX 71, 82, 86 and120),
- Cytosensor microphysiometer (INVITOX 97 and 102), and
- Red blood cell haemolysis test (INVITOX 37 and 99).

A retrospective validation study has been undertaken, based mainly on data obtained from peer reviewed literature. As a consequence, as concurrent animal studies were not generally undertaken or reported, few direct contemporary *in vivo/ in vitro* result comparisons can be made. Only in a few instances is the original *in vivo* data available, but information on reproducibility of the *in vivo* test method with the relevant range of test materials is scarce.

These alternative cell-based test methods were devised, developed and evaluated with specific types of test material in mind: as a result, the test database contains results predominantly about test performance with mainly liquid, water soluble, surfactant materials and multi-component test materials weighted towards low irritancy potential. The materials referenced in the database do not represent a balanced cross-section of all types of material which are evaluated for eye irritancy potential, or the full range of mechanisms and responses that might be seen.

The predictive capacities of each test method and prediction model were assessed according to their ability to categorise test materials for eye irritancy potential based on the EU, GHS and EPA classifications systems, with test performance being evaluated against classification using ocular irritation data generated using the *in vivo* rabbit test.

The Peer Review Panel

Before the ECVAM Peer Review Panel (PRP) was established, relevant information was gathered and evaluated by an ECVAM-appointed Validation Management Group (VMG) which produced a report and recommendations for further consideration by the PRP.

ECVAM produced a background review document for each of the four tests, and established the PRP to advise ESAC on whether and on what terms these tests might be considered to have been scientifically validated.

The PRP comprised two members from the ECVAM Scientific Advisory Committee (ESAC), one European expert nominated by the ESAC, and two US experts nominated by ICCVAM. JacVam was invited to nominate experts but declined.

The PRP held one teleconference, communicated by e-mail to produce independent individual reports, and then held a two-day face-to-face meeting to review the available evidence and analysis, with a view to preparing a consensus report with recommendations on the validation status of the test methods for submission to ESAC for further consideration.

There are no published Performance Standards for the evaluation of this class of *in vitro* replacement tests, and no guidance has been provided to the Peer Review Panel (PRP) on the performance outcomes required to consider the *in vitro* tests to have been scientifically validated. The PRP considers that to ensure international acceptance of the resulting ESAC statement there should have been international agreement on these between ECVAM, ICCVAM and JacVAM before the PRP was asked to offer its advice.

In evaluating the available information in the absence of an internationally agreed Performance Standard the PRP applied the following general principles:

- even accepting that the rabbit test is imperfect, in the absence of authoritative human data significant discordance from the available *in vivo* data cannot be considered a benefit:
- any test used in a Bottom-Up approach should give no false negatives, and no false negative should be produced by high-moderate or strong irritants;
- any test used in a Top-Down approach must balance specificity and sensitivity to correctly identify a substantial proportion of severe irritants, with a false negative rate that would not lead to the over-classification of an unreasonable number of materials of lower ocular irritancy potential – an over-classification rate of <10% was considered acceptable.

General Comments

The PRP evaluated the test methods with a view to their being incorporated into weight of evidence and tiered-testing strategies: the individual tests were evaluated both with a view to their relevance for use in a Bottom-Up approach (to reliably identify ocular non-irritants) and/or a Top-Down approach (to reliably identify strong ocular irritants).

The PRP gave due weight to both the ECVAM Background Review Documents and the VMG report.

The interpretation of the Background Review Document was hampered by the fact that, although these followed the modular validation approach, several protocols were considered for some of the test methods and the BRD text (and in some cases the conclusions) did not always clearly or sufficiently distinguish between the different protocols.

The VMG made recommendations on the readiness for regulatory acceptance of the models largely on the basis of the frequency of false positives in the Top-Down approach and false negatives in the Bottom-Up approach, and the range of test materials for which data was available. The VMG report did not discuss in detail other relevant elements such variability, transferability, applicability domain and it is not clear how these were taken into account in the final VMG recommendations.

Due to the number of protocols to be assessed, and the short time stipulated by ECVAM for their evaluation, the PRP, consistent with the VMG findings, did not fully re-evaluate the Red Blood Cell haemolysis test due to established discrepancies in specificity and sensitivity confirmed in the studies reviewed; the INVITOX 82 and 86 (Fluorescence Leakage) due to proven inadequacies of the datasets; the NRR IIVS due to the very limited dataset; and INVITOX 97 since it is technically obsolete.

GENERAL RECOMMENDATIONS

The PRP wishes to draw two general points relating to ECVAM validation studies and PRP review to the attention of ESAC and ECVAM.

1. Several of the test methods on which the PRP was asked to advise were known in advance not to be supported by data relating to all of the essential components of modular validation. In the view of the PRP under these circumstances, as the outcome of any test method PRP review must be “not scientifically validated to the standards required for regulatory consideration”, there no PRP opinion is required.

2. No internationally agreed and harmonised performance standards for the performance of these tests in bottom-up and top-down approached were developed before the PRP review. The PRP has therefore determined and displayed elements of its own standard. As the intention is to evaluate alternative methods with a view to their being international agreement on their status, such performance standards should be internationally agreed and harmonised before peer review.

[1] *http://ecvam.jrc.it/ft_doc/ESAC26_statement_Organotypic_20070510_C.pdf*

FLUORESCEIN LEAKAGE ASSAY: INVITTOX120 BOTTOM UP APPROACH

1. Data Collection

Data were collected by retrospective compilation of publicly available information and information submitted by ECVAM and COLIPA for *in vivo* animal ocular irritation tests and *in vitro* data generated for previous validation studies.

For INVITTOX 120 there were data for 11 surfactants and 23 surfactant-based formulations, the tests having been conducted in two laboratories.

2. Goal

The goal of the validation study was to evaluate whether and on what terms Protocol INVITTOX 120 can be considered for use as part of a weight of evidence or tiered-testing strategy in a Bottom-Up approach to correctly identify and classify ocular non-irritants (i.e., EPA Category IV, GHS Not Classified, EU Not Labelled) with little to no likelihood for false negatives caused by materials proving to be more than very mild irritants. No Performance Standard was developed before the PRP review.

3. Test definition

The Fluorescein Leakage (FL) assay measures leakage of a fluorescein dye across an epithelial monolayer using cells that form tight and desmosomal junctions.

The assay as described is performed using a monolayer of non-proliferating Madin-Darby Canine renal tubular epithelial cells that form tight and desmosomal junctions similar to those on the apical side of conjunctival and corneal epithelium. After a short exposure period, the test material is removed and a fluorescent dye is added. Test chemicals that disrupt tight junctions allow the dye to leak through the monolayer into the well below where it can be measured spectrophotometrically. FL_{20} is measured, the concentration that causes 20% fluorescein leakage relative to untreated controls.

Although this is the only one of the four test systems evaluated in the validation study and previously considered by the VMG that may be capable of assessing recovery, its potential to do this was not fully explored in the information provided to the PRP. The PRP therefore offers no opinion as to its usefulness for this purpose.

The FL release method is technically straightforward and reasonably and practicably available: relying largely on commonly available skills and equipment. There are some performance issues relating to potential challenges in culturing this specific cell type; and some unanswered questions related to exposure times, potential reversibility, and dose-response considerations.

A range of factors have been identified that may compromise the performance of the test: these include the relationship of medium calcium concentration to tight junction integrity; cell density requirements; potential read-out interference from highly-coloured test materials; chemical binding to the monolayer or the insert membrane; the effects of multiple rinsing steps; and serum protein binding. Chemical binding to the monolayer was reported to be more likely for cationic surfactants.

The nature of the inert inserts used for the assay should be more clearly defined.

There is a VMG recommendation to test tight junction integrity before performing FL assay; but no explanation as to how this would be done, or what the pass-fail criteria might be.

It is the opinion of the PRP that the test method should also allow users to establish before and after testing that there is a confluent monolayer and an intact membrane. The information supplied to the PRP does not explain how this might be done.

4. Data quality

Original data were available from the ECVAM prevalidation study, but only summarised data were available from the COLIPA study.

The BRD evaluated four protocols: in places the text did not clearly distinguish between them.

No reliable human data were available.

The majority of animal data were obtained in non-GLP studies, reported in the peer-reviewed literature.

The PRP could not confirm for INVITTOX 120 that the *in vivo* and *in vitro* data were obtained simultaneously, and with the same batch and lot of test materials.

5. Test materials

INVITTOX 120 evolved from INVITTOX 71 (previously evaluated in the HO/EC study).

Using INVITTOX 120, a COLIPA study was previously performed with 11 surfactants and 18 surfactant formulations in two laboratories. This protocol was also used for two phases of a previous ECVAM study (Southee, 1998); Phase II (4 chemicals and a shampoo formulation tested in 4 laboratories) and Phase III (10 surfactants, 3 laboratories).

The prediction model was only for surfactants, and was devised, empirically based on the findings, during or after the studies.

The VMG restricted its evaluation of test performance to the COLIPA data. The reasons for this were not clear in the VMG document.

6. Within-laboratory variability

The within-laboratory variability was high for INVITTOX 120 with a CV range of 28-36%, and very high variability for the positive control ranging from 17-603%.

7. Transferability

The protocol did not transfer well.

The CV values varied considerably. For the COLIPA study, the CV mean was 40.1% and the range was 0.9-133.9%.

8. Between-lab variability

The COLIPA study in two laboratories evaluating test performance with surfactants and formulations had a CV of 47% for surfactants, and was slightly better for formulations (33.6%). 89% of the chemicals would have been identically classified on the basis of the information obtained by both laboratories.

The two ECVAM studies (Phase II and III) had CV range of 32-57%; possibly due to poor protocol adherence or SOP ambiguities (Southee, 1998). Original data were not included in the BRD; only mean values were provided.

9. Predictive capacity

The prediction model was restricted to the domain of water-soluble surfactants or surfactant-based formulations.

This is insufficient to evaluate or determine applicability to a broader domain of test materials.

The prediction model was developed post-hoc.

The predictive capacity was evaluated on the basis of test material classification using *in vivo* data from the peer-reviewed literature; no original *in vivo* data produced on the same batch and lots of test materials used for the *in vitro* studies were available.

10. Applicability domain

In view of the findings of the ECVAM study, consideration of the applicability domain must be limited to water-soluble surfactants or surfactant-based formulations on the basis of the COLIPA study.

There are other potential limitations related to test material colour, viscosity, solids, and ability to bind to the cell layers and inserts.

11. Performance standards

No Performance Standards are available.

INVITTOX 120 uses standard tissue culture equipment and competencies.

12. Readiness for regulatory purposes

In the view of the PRP INVITTOX 120 is not considered ready for regulatory acceptance in a Bottom-Up approach for the reasons set out above.

Nor is it recommended as a full replacement method for the *in vivo* Draize test.

Although the false negative rate for the three classification systems were low, ranging from 0-5% in the Background Review Documents, there are other compelling reasons for not considering this test to have been scientifically validated and shown to be ready for regulatory consideration on the basis of the information available.

The intra- and inter-laboratory variability and poor transferability are causes for concern.

For the purposes of GSH classification, one significant irritant was misclassified. Although the false negative rate was low, one of the ocular irritants produced irritation that cleared only by seven days and would have been misclassified under GHS. This would present an unacceptable risk to humans if such results could occur with other products intended for widespread human exposure.

The prediction model and protocol only applied to surfactants and surfactant-based formulations.

In addition, in the view of the PRP there is a need for a secondary means to perform a qualitative assessment of any negative prediction. Consideration was given as to whether a microscopic visual inspection of the cells would suffice. However, given that a Bottom-Up approach would be the approach where no or mild irritancy would be expected and strong irritancy would not expected, many compounds selected for evaluation in a Bottom-Up approach would be expected to have at most only very subtle effects, thus there may be little likelihood of this being an adequate safeguard in practice.

Expansion of the range of compounds and formulations considered in the applicability domain is needed. INVITTOX 120 should be considered for further validation by testing of a balanced data set of clearly classified chemicals in at least three laboratories in order to further evaluate the variability and limitations of the assay.

Further evaluation would benefit from additional data, preferably acquired by further testing, including recovery up to 72 hr. Performance standards and any positive and negative controls should be clearly defined and agreed at international level.

FLUORESCEIN LEAKAGE ASSAY, TOP-DOWN APPROACH INVITTOX PROTOCOL 71

1. Data collection

Data were collected by a retrospective compilation of data from the peer-reviewed literature. An extensive literature search was undertaken, and the relevant publications were taken into account.

For the INVITTOX 71 protocol, the key reference is Balls et al., 1995 paper (EC/HO study).

2. Goal of the study

The goal of the validation study was to evaluate whether and on what terms Protocol INVITTOX 71 can be considered for use in a weight of evidence or tiered-testing strategy in a Bottom-up approach to correctly identify and classify ocular non-irritants, with little or no likelihood for false negatives caused by materials proving to be more than mild irritants- and/or in Top-Down approach to identify severe irritants.

On the basis of the available evidence the VMG focused on the validity of the test method for a Top-Down approach, discriminating severe irritants from all other classes, or discriminating non-irritants from all other classes in a bottom-up approach, based on EU, GSH and EPA classifications.

3. Test definition

The Fluorescein Leakage (FL) assay measures leakage of a fluorescein dye across an epithelial monolayer using cells that form tight and desmosomal junctions. The BRD confirms that disruption of such inter-cellular junctions, increasing permeability of the cell membrane, is relevant to the mode of action of some ocular irritants.

Although this is the only one of the four test systems evaluated in the validation study and previously considered by the VMG that may be capable of assessing recovery, its potential to do this was not fully explored in the information provided to the PRP. The PRP therefore offers no opinion as to its usefulness for this purpose.

The FL assay was designed and developed initially to identify possible irritating surfactants and surfactant-based formulations.

The assay as described is performed using a monolayer of non-proliferating Madin-Darby Canine renal tubular epithelial cells that form tight and desmosomal junctions similar to those on the apical side of conjunctival and corneal epithelium. After a short exposure period, the test material is removed and a fluorescent fluorescein dye is added. Test chemicals that disrupt tight junctions allow the dye to leak through the monolayer into the well below where it can be measured spectrophotometrically. FL_{20} is measured, the concentration that causes 20% fluorescein leakage relative to untreated controls.

A prediction model (pm) for protocol 71 was devised based on the HO/EC study. Based on *in vitro* and *in vivo* data a prediction model (pm) was then constructed to facilitate and evaluate EU, GHS and EPA classification for the 60 chemicals for which data was available. These represent a number of different chemical classes, including surfactants, and in varying physical states (solids, and liquids of unknown viscosity).

The FL release method is technically straightforward and reasonably and practicably available: relying largely on commonly available skills and equipment. There are some performance issues relating to potential challenges in culturing this specific cell type; and some unanswered questions related to exposure times, potential reversibility, and dose-response considerations.

A range of factors have been identified that may compromise the performance of the test: these include the relationship of medium calcium concentration to tight junction integrity; cell density requirements; potential read-out interference from highly-coloured test materials; chemical binding to monolayer or the insert membrane; the effects of multiple rinsing steps; and serum protein binding. Chemical binding to the monolayer was reported to be more likely for cationic surfactants.

The nature of the inert inserts used for the assay should be more clearly defined.

There is a VMG recommendation to test tight junction integrity before performing FL assay; but no explanation as to how this would be done, or what the pass-fail criteria might be.

It is the opinion of the PRP that the test method should also allow users to establish before and after testing that there is a confluent monolayer and an intact membrane. The information supplied to the PRP does not explain how this might be done.

4. Data quality

Data for 60 test materials were obtained from the peer-reviewed literature.

For various reasons the quality of the data is regarded by the PRP as not sufficiently robust, and there is concern about protocol breaches.

5. Test materials

Data from 60 chemicals is available (mainly Ball et al, 1995).

These represent a number of different chemical classes, including surfactants, and in varying physical states (solids, and liquids of unknown viscosity).

There is no data on test performance with multi-component formulations.

6. Within-laboratory variability

There was a large variation in results for this protocol (CV 56.5 – 63.2%).

The variation between operators ranged from 0% to 50% and did not seem to improve over time or with experience.

7. Transferability)

The PRP was provided with the results of only one study relevant assessing transferability.

In that publication the protocol did not transfer well (consequently, refinements were made to the protocol which was later accepted as INVITTOX 120).

The VMG recommended training for technicians to improve operator reproducibility if this test method is further developed: but as it not clear what caused the problems in practice, the PRP cannot comment on whether or to what extent this might improve test performance.

8. Between-laboratory variability

The PRP was provided with data from only one study where between-laboratory variability of protocol 71 was assessed and all laboratories fully adhered to the test protocol.

An agreement/concordance score of only 52% was obtained.

9. Predictive capacity

The predictive capacity was evaluated mainly against rabbit *in vivo* data from the literature and obtained with the LVET. Although it is accepted that the LVET does not provide results that are in complete agreement with the standard Draize eye test, the significance of this is not fully evaluated in the BRD or VMG report.

No original contemporary *in vivo* data for the same test material dataset was available.

For protocol 71 no prediction model (pm) was available and therefore a new pm using the data from the EC/HO study was constructed for chemicals only.

Since no data were available for formulations its suitability for these could not be assessed.

The concordance for non-irritants versus the rest was 62%-70% and 78%-83% for the irritants versus the rest. Specificity was in both situations higher than sensitivity indicating a greater predictive capacity for the non-irritants.

Concordance did not improve when only surfactants and surfactant based formulations were taken into account.

In general, the predictive capacity was better for surfactants than for alcohols. Also, the predictive capacity was better for the non-irritants than for the irritants.

10. Applicability domain

Due to short exposure times, relatively high chemical concentrations are used.

As the test materials must be removed from the test system after exposure, the retention of, or damage caused by the remove of, viscous substances might therefore affect the test performance and predictive capacity. However, as there was no information on viscosity of chemicals used in protocol 71 these effects could not be evaluated by the PRP.

The tested chemicals generally were of low irritancy potential. Strong acids and bases, fixatives and highly volatile chemicals should be excluded from the applicability domain as they are incompatible with the cell mono-layers used.

Although INVITOX 71 was tested with chemicals from different chemical classes, in the view of the PRP too few chemicals of each class were tested and therefore no predictive capacity for each class could be ascertained.

If the test is further developed the PRP recommends that more chemicals in different classes should be tested to improve information on the applicability domain.

It is hypothesized that some highly- coloured substances may affect the results of protocol 71.

Furthermore, the fact that the some test materials may bind to the insert and affect the predictive capacity is a point of concern.

In the view of the PRP, on the basis of the available evidence, other factors potentially affecting test performance include the relationship of medium calcium concentration to tight junction integrity; cell density requirements; coloured test materials; and serum protein binding. A further potential confounding factor was chemical binding to the insert membranes despite multiple rinsing steps, as well as damage to the insert membranes because of multiple washing steps. Chemical binding to the membranes was found to be more likely for cationic surfactants.

The PRP could assess the potential use of the test to evaluate recovery on the basis the information provided.

11. Performance standards

Performance standards are not provided in the material presented to the PRP for consideration.

12. Readiness for regulatory purposes

The available evidence is inclusive.

The evidence available has not satisfied the PRP that the test can be considered to have been scientifically validated and to be ready for regulatory consideration.

However, the available information has, in the view of the PRP, not demonstrated conclusively that the test method does not have the potential to be further developed for consideration for use in a Top-Down approach.

Protocol 71 may have potential for a Top-Down approach (the BRD data suggest a false positive rate of only 7-9%), as part of a tiered testing strategy but needs additional testing and further refinement, in particular with respect to variability and definition of the applicability domain by expanding the dataset of tested chemicals and direct comparison with *in vivo* data before it can be regarded ready for consideration for regulatory purposes.

Recovery data might improve the predictive capacity of the test.

NEUTRAL RED RELEASE ASSAY PROTOCOLS (PROTOCOL 54)

1. Data Collection

Data from a number of sources have been identified, obtained and evaluated.

The best dataset is from the study carried out by CTFA in their Phase I, II, and III studies, and includes quality *in vitro* and *in vivo* data.

Even with this dataset, it is the view of the PRP that additional data are needed to fully evaluate Protocol 54. Currently the dataset is limited to 10 alcohol formulations, 18 oil/water emulsions and 25 surfactant-based formulations.

2. Goal of the study

The goal of the study was to consider the scientific validity in the context of a testing strategy for a Bottom-Up approach. The goal was clearly understandable and an acceptable scientific rationale was given in the BRD. No agreed performance standard was provided to the PRP.

3. Test definition

The test and its purpose are well-defined.

The test has mechanistically relevant features for modelling eye irritation, namely cell membrane damage. The test was not designed to assess reversibility or recovery.

Originally there was no prediction model for using the NRR assay (Protocol 54) to predict the ocular irritancy of chemicals according to the EU, GHS or EPA classifications. Various prediction models have been empirically devised and evaluated in order to use NRR assay data to predict ocular irritation potentials of the test materials

4. Data Quality

In general, the data comprised results from peer-reviewed articles, unpublished data (with hand-written corrections of results in the case of the FRAME submission (1995)) and in-house data.

Only one set of original high quality data (CTFA) is provided in the BRD. The other data are of varying quality.

All of the *in vitro* data and the majority of the *in vivo* data were obtained in non-GLP compliant studies. The *in vivo* data in the CTFA study was GLP compliant. Both the *in vivo* and *in vitro* data in the CTFA study were obtained concurrently.

5. Test materials

The assay was designed and developed for the in-house evaluation of the ocular irritancy potential of cosmetic ingredients and formulations with expected mild or no irritancy.

Consequently test materials represented in the testing database are predominantly surfactants and surfactant based cosmetic formulations, and two pesticides.

Materials tested included 10 hydro-alcoholic formulations, 25 surfactant based formulations, and 18 oil/water emulsions.

A broader range of materials needs to be tested using Protocol 54 to better establish its applicability domain.

6. Within-laboratory variability

Seven reported studies based on INVITTOX Protocol 54 had appropriate data-sets for within-laboratory reproducibility analyses. The test materials were restricted to surfactants or surfactant-based formulations, with one study also containing data obtained with 2 pesticides.

The number of test runs varied for each test material, with the tendency of higher repeats for chemicals giving variable results: thus in interpreting the test performance it is necessary to keep in mind that the data-sets are biased in favour of such materials.

Other data limitations include the absence of “greater-than” values; and in some cases only a single value was used for analyses, which is not sufficient.

The within-laboratory variability could be calculated only for eight data-sets and all but one were generated by testing surfactants or surfactant-based formulations without detained information on their chemical composition (and one by testing two pesticides).

Overall, the within–laboratory variability for INVITTOX Protocol 54 was high, with mean CVs ranging from 22% - 85%.

Only the data for INVITTOX Protocol 54, provided by FRAME Alternatives Laboratory, enabled analyses of time and operator variability. The operator variability was found to be 7% - 32%, independent of the number of experiments performed. The data indicated that the INVITTOX Protocol 54 has different levels of variability according to the type of test material; specifically variability was particularly high with surfactants (see Table 3.2.21, page 78-81 of the BRD).

A thorough study on a wide range of chemicals regarding their type and irritancy potential is recommended by the PRP if additional studies are undertaken for a further validation study.

7. Transferability

No information is available regarding transferability. This is an essential component of the modular validation requirements.

8. Between laboratory variability

No information is available regarding between laboratory variability. This is an essential component of the modular validation requirements.

9. Predictive capacity

The capacity of the NRR assay to predict ocular irritation *in vivo* was assessed against *in vivo* data from rabbits, as data for humans are virtually nonexistent.

Only the CTFA included original *in vivo* data. In the other studies historical *in vivo* data were used. All available *in vivo* data were compiled and entered into ECVAM template v6, which converted them into EU, GHS and EPA classifications.

In the case of the CTFA Phase I-III studies, there was no prediction model before the study. Threshold values were empirically assigned in the BRD for the CTFA study (cut-off values 250 and 600 mg/ml).

With the prediction model used the assay is highly sensitive (100% sensitivity) with poor specificity as it has a tendency to over predict the irritancy of materials (20-44% false positives). It showed better predictive capacity for severe irritants rather than non-irritants or mild irritants. The predictive capacity for oil in water emulsions could not be assessed as actual NRR values could not be obtained for 83% of materials. The dataset for hydro-alcoholic formulations (10 materials) was too small to draw any conclusions; however, the predictive capacity for hydro-alcoholic formulations seemed to be better than for surfactant-based cosmetic formulations.

A compilation of data on chemicals (5 surfactants, 24 non-surfactants) was performed from the literature and the *in vivo* and *in vitro* results were paired purely on chemical name and concentration. The sensitivity of the assay was confirmed as very high (92.3 – 100%) with a lower specificity (56.3 – 80%) with slightly better specificity for evaluation of severe irritants versus the rest (71.4-72.7% for all classification systems). For the limited number of surfactants (5), 100% sensitivity and 50% specificity was observed. However, in the case of 29 non-surfactants, the sensitivity was 80-100% and specificity 70-76.9% when analyzed for severe irritants versus the rest.

However, due to the lower quality of the data, these results should be treated with caution.

10. Applicability domain

The INVITTOX Protocol 54 was the only NRR assay protocol for which original *in vitro* and *in vivo* data were available to assist in the determination of the applicability domain. The available data were on hydro-alcoholic, oil-in-water and surfactant-based formulations.

The PRP concluded that the assay was not sensitive enough to measure the toxicity of oil-in-water emulsions. Only formulations were tested in the CTFA study and it is not possible to assume that the predictivity for chemicals would be similar.

On the basis of the evidence available to the PRP the applicability domain of this assay, based on high quality data, is restricted to hydro-alcoholic and surfactant-based formulations.

The compilation of lower quality data on chemicals from the literature suggested that the protocol had 100% sensitivity and better predictive capacity for severe irritants than non-irritants and mild irritants. The protocol had varying predictive capacities for different types of formulations with overall tendency to over predict non-irritants as irritants.

Data were lacking on the potential effects of pH, colour or viscosity. Colour did not appear to affect the result, while viscosity may play a role.

The EC Collaborative Study (1991) identified problems using this assay to evaluate certain chemicals (silver nitrate, acetaldehyde and mercury chloride were misclassified).

Too few solids were tested to determine the predictive capacity with this type of material.

11. Performance standards

No performance standards have been defined.

12. Readiness for regulatory purposes

In the opinion of the PRP the INVITTOX Protocol 54 is not ready for consideration of acceptance for regulatory purposes either for a Bottom-Up or Top-Down approach within a tiered-testing strategy without further testing because of the limited number and classes of chemicals/formulations tested, lack of information on transferability and interlaboratory variability.

Conclusions and recommendation

It is important to note that the opinions of the PRP are premised on issues which cannot be resolved on the basis of the available information, rather than the available evidence having established conclusively that the test method should not be further developed and is incapable of being developed for regulatory use.

The INVITTOX Protocol 54 utilizes a commercially available and easily cultured cell line routinely used for other methods worldwide.

Refinement of protocols including modifications and different prediction models designed for different types of materials may be necessary, e.g. reduction of exposure time in case of surfactants may increase specificity, mindful that longer exposure may decrease variability.

Positive and negative controls and the acceptance criteria should be defined in order to ensure that the cells are responding in a predictable and reproducible manner and the results are valid.

Without additional studies using a better defined and balanced set of chemicals and formulations in at least three laboratories, the reproducibility of results, applicability domain and limitations of the assay cannot be determined.

Some false negatives may be the result of the oil/water matrix trapping the neutral red dye and this should be evaluated. The protocol should be optimized to ensure that this trapping does not occur by evaluating all negative samples under the microscope to determine the location of the neutral red dye.

The SOP needs refinement to include details and specific procedures for rinsing and for verifying that the neutral red is not bound by the oil/water matrix but actually inside of the cells.

Dilutions of the sample including the neat material should be tested.

In the opinion of the PRP this method may have potential to be further developed for use in a Bottom-Up approach as part of a tiered-testing strategy or as a screen prior to *in vivo* Draize testing, but not as a full replacement for the Draize eye-test.

However, in the view of the PRP the current BRD provides insufficient evidence to justify its acceptance for consideration for regulatory use.

PREDISAFE KIT (NEUTRAL RED RELEASE ASSAY)

1. Data collection

In the view of the PRP the data collection procedure was described adequately in the Background Review Document and VMG report.

Although the test method is approved in France as an official regulatory method (Methode Officiel Francaise), surprisingly little data were provided on the use of the PREDISAFE kit by its manufacturer or users.

2. Goal of the study

The BRD and VMG papers were prepared with a view to considering the scientific validity and readiness for regulatory consideration of the test method protocol as part of a weight of evidence and tiered-testing strategy to minimise animal use to establish the eye irritancy potential of test materials.

3. Test definition

The test definition is adequately defined in the view of the PRP.

It relies on mechanistically relevant features for modelling eye irritation, namely membrane damage, and is not designed to assess reversibility or recovery.

In the view of the PRP the SOP for Predisafe needs to be improved, in part because of variation in use of the kit, possibly impacting on test performance, has been attributed to ambiguity in the instructions.

The original prediction model was only qualitative and not applicable for predicting ocular irritancy of chemicals according to the EU, GHS and EPA classifications.

For the purposes of the COLIPA study an alternative prediction model was developed in order to convert the IC50 values into predictive MMAS DET scores.

4. Data quality

Surprisingly, in the view of the PRP, although the test method has been accepted by the French Authorities for testing cosmetics, the data on use and performance of this protocol were scarce (one COLIPA study with quality data).

5. Test materials

The assay was originally designed and developed for the testing of cosmetic ingredients and formulations with expected mild or no ocular irritancy potential.

Consequently, the data-base of test material contains predominantly surfactants and surfactant-based formulations.

The COLIPA study contained 22 ingredients of which 13 were surfactants, 9 chemicals and 32 formulations. The majority of the materials were water soluble or at least miscible in water.

6. Within-laboratory variability

The PRP could not evaluate within-laboratory variability because of a lack of relevant performance data for the PREDISAFE kit.

7. Transferability

There was no available documentation which could be used by the PRP to evaluate transferability.

Although the PREDISAFE Protocol was conducted in more than one laboratory with the same set of materials, there was no information on whether one of the laboratories was naive and so the transferability could not be determined.

No information on PREDISAFE kit storage, shelf life, or overseas shipment was available to the PRP.

The assay uses commercially available cell culture equipment; however, the Predisafe protocol is based on a kit that contains live-cell cultures and reagents for which stability and performance after long-distance shipping and storage is unknown.

8. Between-laboratory variability

The COLIPA study based on the PREDISAFE Protocol investigated the same set of chemicals in three laboratories; however, no detailed information on potential sources of variability, such as the time-frame, lot and identical use of protocol were available to the PRP.

Comparison of IC50 values of 22 ingredients (13 surfactants) and 32 formulations showed CV of 45% (for ingredients) and 22% (for formulations). The use of different batches of test materials might explain the higher variability.

Moreover, the PREDISAFE SOP did not contain detailed information on storage, rinsing and delay between arrival and usage, which might increase between-laboratory variability.

All laboratories classified the irritants and slight irritants similarly with greater variation for moderate irritants. Good agreement in classification was observed, 85.5% of materials were classified identically in all three laboratories.

Further work should be done with a wider range of chemicals to evaluate its predictive capacity for other types of test materials.

9. Predictive capacity

The capacity of the NRR assays to predict ocular irritation *in vitro* was assessed against *in vivo* data from rabbits, as data for humans are virtually nonexistent.

The level of false negatives in the evidence provided to the PRP was up to 16% (EPA classification) and the assay was not able to reliably distinguish ocular irritants from non-irritants sufficiently well to be endorsed by the PRP as a replacement or partial replacement for the rabbit Draize eye-test. This opinion is reinforced by the severity and number of ocular lesions that would have been caused by some of the test materials that the test would have miscategorised as ocular non-irritants.

The COLIPA study provided the only original data available for this assay (i.e. mean *in vitro* data and raw *in vivo* data on 22 chemicals and 32 formulations). The PM was established before the results were known (an algorithm converting IC50 values into predictive MMAS DET scores). The sensitivity was high (92.9- 95.5%) but the specificity was much lower (60.6-70.4%).

Six different surfactants in multiple concentrations confirmed 100% sensitivity, but very poor specificity (0% when analyzed non-irritants versus the rest, 28.6-33.3% when analyzed severe irritants versus the rest). The PREDISAFE Protocol has a good predictive capacity for severe irritants, but tends to over predict non-irritants and mild irritants (non-surfactant chemicals only (9) showed 100% sensitivity and specificity 42.9-75% for non-irritants versus the rest and decreased sensitivity 50-66% and specificity 42.9% for severe irritants versus the rest). In the case of formulations, the best predictive capacity was found for severe irritants versus the rest (sensitivity 100%, specificity 84%).

In the view of the PRP, none of the prediction models used for the purposes of this validation study adequately distinguished between test materials falling into any of the principal regulatory categories.

In the view of the PRP, on the basis of the information provided for consideration, the method can be used for screening to identify likely potential ocular non-irritants prior to *in vivo* Draize testing but not as a replacement for regulatory purposes.

10. Applicability domain

The potential applicability domain discussed in the BRD is restricted.

The PREDISAFE kit did not correctly classify any of the non irritant surfactants and therefore, the test should not be used to evaluate and classify surfactants in a Bottom-Up approach.

The predictive capacity of PREDISAFE was calculated on the basis its performance in the COLIPA study. Its sensitivity was consistently higher than specificity and the concordance was better for non-surfactants than surfactants. There was insufficient data for coloured, solid or viscous chemicals for the PRP to determine the effects of physical features on the predictive capacity of this protocol. The best predictive capacity was found for formulations (100% sensitivity, 84% specificity) when analyzed as severe irritants versus the rest. The applicability domain comprises both chemicals (based on data on 13 surfactants and 10 alcohols and preservatives) and formulations, with better predictive capacity for non-surfactants.

In general, the NRR test requires the use of high concentrations of chemicals which may react with the plastic of wells and will be difficult to remove, thus increasing the apparent toxicity of the tested samples reacting with the cells uncontrollably for prolonged time. The solubility in water seems to be critical. No results are available on materials tested in mineral

oil or other solvents. The assay is not suitable for materials with fixative properties or for strong acids and bases and highly volatile materials.

11. Performance standards

No performance standards were provided to the PRP.

12. Readiness for regulatory purposes

In the view of the PRP, on the basis of the available evidence, this test is not ready for acceptance for regulatory purposes without further testing, including evaluation of performance using a broader range of chemicals in at least three laboratories to establish acceptable intra-laboratory variability and transferability.

Conclusions and recommendation

It is the view of the PRP that further testing should be undertaken to determine if the Predisafe kit can be recommended for screening as part of a battery of tests in a Bottom-Up approach for non-surfactant cosmetics but not for EPA classification and not for Top-Down due to the high number of false positives.

The test has a false negative rate of 6% with a false positive rate greater than 35% (very poor with non-irritant surfactants but better with multi-component mixtures).

The test seems promising, but it tends from the available dataset to over predict the irritancy of materials.

PREDISAFE is supplied in the form of a standardized kit including reagents and cells; however, no information on performance after prolonged shipment e.g. overseas is available. In the view of the PRP the time between delivery and use of the PREDISAFE kit may play a role in its performance, and this is another factor requiring additional evidence to evaluate.

Refinement of the protocol including modifications and different prediction models designed for different types of materials may be necessary, e.g. reduction of exposure time in case of surfactants may increase specificity, on the other hand longer exposure may decrease variability.

The acceptance criteria should be defined in order to ensure that the cells are responding in a predictable and reproducible manner and the results are valid.

In the view of the PRP without additional studies of a defined and balanced set of chemicals and formulations in at least three laboratories, the reproducibility of results, application domain and limitations of the assay cannot be determined.

RED BLOOD CELL ASSAY

1. General Comments:

The VMG and PRP reviews of the data and analysis of INVITTOX protocols 37 and 99 confirmed significant shortcomings and inconsistencies with respect to specificity and sensitivity, protocol deviations, and problems with the prediction model.

These shortcomings, and the failure on the basis of the available information to satisfy other required validation modules, in the view of the PRP currently preclude consideration of these test methods for consideration for regulatory use, even as part of a tiered-testing strategy, to determine the eye irritancy potential of test materials.

2. Data collection:

Reasonable efforts have been made to obtain relevant, robust data, and three studies are reviewed in the BRD.

3. Goal of the study:

The BRD and VMG papers were prepared with a view to considering the scientific validity and readiness for regulatory consideration of the test method protocols as part of a testing strategy to minimise animal use to establish the eye irritancy potential of test materials.

4. Test definition:

Test definition is adequately described. The mechanistic basis and its relevance to eye irritancy are explained.

5. Data quality:

Data quality is variable.

In the view of the PRP there may be a case for discarding the least reliable data if any further or supplementary evaluation is made of these test methods and better quality information becomes available.

6. Test materials:

As described in the BRD.

7. Intra-laboratory variability:

The intra-laboratory variability was difficult for the PRP to asses, with one study showing CVs of up to 67%.

A wide range of potential variables were indentified, but cannot be fully evaluated on the basis of the material contained in the reference documents.

8. Transferability:

There was no evidence for transferability available.

9. Inter-laboratory variability:

The reported CVS are unacceptably high.

10. Predictive capacity:

The predictive capacity was poor when judged against the standard classification systems.

The method is likely to over-classify 20% of test materials if used for Top-Down evaluation: and to misclassify more than 10% of irritants with a Bottom-Up approach.

11. Applicability domain:

The applicability domain was evaluated primarily with water-soluble test materials.

The method is known not to be suitable for assessing ketones, alcohols, fixatives, dyes and viscous materials.

12. Minimum performance standards:

Performance standards were not provided to the PRP.

13. Readiness for regulatory purposes:

In the view of the PRP although it may be suitable for in-house decision making, on the basis of the available evidence these protocols have not been shown to be ready for consideration either as a Top-Down or a Bottom-Up screen for regulatory purposes.

CM ASSAY – INVITTOX PROTOCOL 102 – FOR TOP-DOWN APPROACH

1. General Comments:

INVITTOX Protocol 97 is considered by the PRP to be redundant, and is not considered further.

INVITTOX Protocol 102 has been used as in-house screen by the test-developer.

Essential equipment and consumables are no longer manufactured, and the original software now obsolete.

2 Data collection:

In the view of the PRP reasonable efforts have been made to obtain relevant, robust data.

The BRD is based on historical data, in part from the peer reviewed literature, including previous validation studies.

3 Goal of the study:

The BRD and VMG papers were prepared with a view to considering the scientific validity and readiness for regulatory consideration of the test method protocols as part of a testing strategy to minimise animal use to establish the eye irritancy potential of test materials.

4 Test definition:

In the view of the PRP the test system has been adequately described.

The mechanistic basis and its relevance to eye irritancy are explained. However in the view of the PRP it is not clear precisely what class of interaction (e.g. lysis, saponification, coagulation, or reactive chemistry) is modelled in the test system.

5 Data quality:

There is published and unpublished data, not obviously produced to GLP, although some original data have been audited.

The Draize eye-test reference data were not obtained using the standard *in vivo* protocol: topical anaesthesia was used to obtain the CTFA data set, and the LVET was used for the company data. The significance of this is not discussed in detail in the BRD or VMG report.

In the view of the PRP the HO/EC material is particularly weak in a number of places, and in the view of the PRP there may a case for setting it aside if a further evaluation of validation status is undertaken.

6 Test materials:

INVITTOX 102 data relate to 53 water-soluble test materials – 21 chemicals and 32 multi-component mixtures, 29 of which were non-surfactants. These are clearly described in the material made available to the PRP.

7 Intra-laboratory variability:

The intra-laboratory variability is considered acceptable by the PRP.

The low CVs tend to confirm low degree of temporal and operator variability.

8 Transferability:

The method requires specialist equipment and consumables which are no longer currently manufactured.

The performance described seems adequate to the PRP, although there is no agreed Performance Standard, with respect to laboratories with established cell culture expertise.

9 Inter-laboratory variability:

Although there are no agreed Performance Standards in the opinion of the PRP the inter-laboratory variability seems to have been adequately demonstrated with generally good correlations.

10 Predictive capacity:

The prediction model was developed and refined during the evaluation, however the final recommended protocol must more clearly state how it is applied.

For a Top-Down approach, the material contained in the BRD and reviewed by the VMG shows that for the evaluation of water-soluble surfactants and multi-component water-soluble surfactant containing mixtures the false positive rate was 3-10%, and false negative rate 2-22%. It is not known if the mixtures contained other constituents expected to be biologically active.

With water-soluble non-surfactants the false positive rate was 0-6%, and the false negative rate 43-55%.

11 Applicability domain:

Other than the HO/EC data, evaluation was primarily based on water-soluble surfactants, multi-component water-soluble surfactant containing mixtures, and water-soluble chemicals/materials – generally expected to be of low irritancy potential. It is not known if the mixtures contained other constituents expected to be biologically active.

In the opinion of the PRP this test is not considered suitable, on the basis of the available data, for consideration for regulatory use for the evaluation of non-water soluble or non-

aqueous materials, mixtures containing other ingredients likely to be biologically active, viscous materials, suspensions or solids.

12 Minimum Performance Standards:

No performance standards were provided to the PRP.

The PRP has assumed for a Top-Down approach, with negatives being further evaluated before classification, that fewer than 10% of test materials should be over-classified.

13 Readiness for regulatory purposes:

On that basis, the CM INVITTOX Protocol 102 is considered by the PRP to be suitable for consideration for regulatory use as a Top-Down screen to detect strong irritants (EPA Category 1, GHS Category 1, and EU R41) as part of a tiered-testing strategy for the applicability domain given.

The PRP wishes to emphasise the following:

1. This opinion is based upon the performance standards developed by the PRP itself, and not yet validated or endorsed by an other party.
2. Any validation statement is relevant only to the hardware and software used to generate the data presented in the BRD.
3. Any significant changes to the hardware or software should be subject to a catch-up validation study conducted at a single centre with small number of well-characterised and representative test materials.
4. All negative results must be further evaluated before test materials are classified in a Top-Down testing strategy.
5. Further work is required to establish if the applicability domain can be broadened to additional classes of test materials.
6. The test may not be suitable to evaluate test materials expected to have a direct effect on cell respiration or glycolysis.
7. The VMG has recommended changes to the exposure conditions, but the detail and rationale are not clear. The VMG should be asked for clarification.
8. When interpreting results with water-soluble surfactants users should take account of whether the test materials were evaluated at concentrations above or below the critical micelle concentration (CMC).

CM ASSAY – INVITTOX PROTOCOL 102 – BOTTOM-UP APPROACH

1. General Comments:

INVITTOX Protocol 97 is considered redundant, and is not considered further.

INVITTOX Protocol 102 was previously used as in-house screen by test-developer.

Equipment and consumables are no longer manufactured, and the original software is now obsolete.

2. Data collection:

Reasonable efforts have been made to obtain relevant, robust data.

The BRD is based on historical data, in part from the peer reviewed literature, including previous validation studies.

3. Goal of the study:

The BRD and VMG papers were prepared with a view to considering the scientific validity and readiness for regulatory consideration of the test method protocols as part of a testing strategy to minimise animal use to establish the eye irritancy potential of test materials.

4. Test definition:

The test definition is, in the view of the PRP, adequately described.

The mechanistic basis and its relevance to eye irritancy are explained. However in the view of the PRP it is not clear precisely what class of interaction (e.g. lysis, saponification, coagulation, or reactive chemistry) the test system models.

5. Data quality:

The BRD included both published and unpublished data, not obviously produced to GLP, although some have been audited.

The *in vivo* reference data were not obtained using the standard rabbit Draize eye-test protocol: topical anaesthesia was used to obtain the CTFA data set, and the LVET was used for the company data. Although the significance of this is not discussed in the BRD or VMG report, in the view of the PRP the data must still be considered for the purposes of validation.

In the view of the PRP the HO/EC material is particularly weak in places, and in the view of the PRP there may be a case for setting it aside of any further validation assessment is conducted.

6. Test materials:

INVITTOX 102 data relates to 53 water-soluble test materials – 21 chemicals and 32 multi-component mixtures, 29 of which were non-surfactants.

7. Intra-laboratory variability:

Intra-laboratory variability is considered acceptable by the PRP. Low CVs tended to confirm low degree of temporal and operator variability.

8. Transferability:

Transferability requires essential specialist equipment and consumables which are no longer manufactured.

The performance described seems adequate to the PRP, although there is no agreed Performance Standard, with respect to laboratories with established call culture expertise.

9. Inter-laboratory variability:

Although there are no agreed Performance Standards, inter-laboratory variability seems to have been adequately demonstrated in the view of the PRP with generally good correlations.

10. Predictive capacity:

The prediction model was developed and refined during the evaluation, the final recommended protocol must in the view of the PRP more clearly state how it is applied.

For a Bottom-Up approach the material contained in the BRD and reviewed by the VMG shows that for the evaluation of water-soluble surfactants and multi-component water-soluble surfactant containing mixtures the false negative rate was 0-2% - with the materials misclassified shown to be more than mild irritants when evaluated in the DET. It is not known if the mixtures contained other constituents expected to be biologically active.

In contrast the false negative rate with non-surfactants was 24-38%, in the view of the PRP this is too high for regulatory use in a bottom-up approach.

11. Applicability domain:

Other than HO/EC data, the evaluation was primarily based on data generated with water-soluble surfactants and multi-component water-soluble surfactant containing mixtures – that were expected to be of low irritancy potential. It is not known if the mixtures contained other constituents expected to be biologically active.

Performance was poor with non-surfactants.

It is not suitable in the view of the PRP, on the basis of the available data, for the evaluation of non-aqueous or non-water soluble materials, mixtures containing other ingredients likely to be biologically active, viscous materials, suspensions or solids.

12. Minimum Performance Standards:

Minimum Performance Standards were not available.

The PRP has assumed for a bottom-up approach, with positives being further evaluated before classification, that there should be a negligible false negative rate, with no false negative proving to be more than a mild irritant.

13. Readiness for regulatory purposes:

On that basis CM INVITTOX Protocol 102 is considered by the PRP to be suitable for consideration for regulatory use as a Bottom-Up screen to detect non-irritants (insert classifications) as part of a tiered testing strategy only for water-soluble surfactants, and multi-component mixtures containing water-soluble surfactants.

In the view of the PRP it is not suitable, on the basis of the available data, for the evaluation of non-aqueous or non-water soluble materials, mixtures containing other ingredients likely to be biologically active, viscous materials, suspensions or solids.

The PRP emphasises that it considers:

1. This opinion is based upon the performance standards developed by the PRP itself, and not yet validated or endorsed by an other party.
2. Any validation statement is relevant only to the hardware and software used to generate the data presented in the BRD.
3. Any significant changes to the hardware of software should be subject to a catch-up validation study conducted at a single centre with small number of well-characterised and representative test materials.
4. All positive results obtained in a Bottom-UP approach must be further evaluated before test materials are classified.
5. It is not recommended, on the basis of the available evidence, that the test be used to evaluate non-surfactants.
6. Further work is required to establish if the applicability domain can be broadened to additional classes of test materials.
7. The test may not be suitable to evaluate test materials expected have a direct effect on cell respiration or glycolysis.
8. The VMG has recommended changes to the exposure conditions, but the detail and rationale are not clear. Clarification should be sought from the VMG.
9. When interpreting results with water-soluble surfactants users should take account of whether the test materials were evaluated at concentrations above or below the critical micelle concentration (CMC).