Session 5.05 Advancements and needs for developing and validating 3R alternatives for ocular irritancy testing

Use of *In Vitro* Data and (Q)SARs to Classify Eye Irritating Chemicals in the EU – Experience at the BfR

Matthias Herzler, Horst Spielmann, Ingrid Gerner, Manfred Liebsch and Thomas Hoefer
Federal Institute for Risk Assessment (BfR), Berlin, Germany

Summary

Under the EU's proposed new chemicals legislation REACH, the use of experimental animals will be almost completely banned from eye irritation testing of industrial chemicals. Alternative methods are at hand to identify strong irritants. However, there is still a need for reliable tools able to identify non-irritants and to discriminate between non- and moderately irritating chemicals. In addition, testing strategies are required that are capable of integrating all of these approaches with a more efficient exploitation of existing information. In this paper, these issues are addressed, based on regulatory experience gained under the EU New Chemicals notification programme.

Keywords: eye irritation, Draize test, in vitro testing, QSAR, testing strategies

In vivo testing for eye irritation – the Draize test

Today, the Draize rabbit eye test, which was introduced over sixty years ago (Draize et al., 1944), still forms the basis of internationally agreed protocols for eye irritation/corrosion testing (European Commission, 2004; OECD, 2002). The success of this test is based on its obvious biological relevance and the fact that multiple aspects of ocular irritation/corrosion, i.e. different target sites within the eye, as well as the level of severity and the reversibility of effects, are covered in a single test.

However, the Draize test has also been subject to criticism, not only because of the pain and suffering caused to the animals, but also on scientific grounds, among others because of the great variability and low reproducibility of results, an allegedly unrealistic application procedure, and because of differences in the anatomy, physiology, and biochemistry of the rabbit vs. the human eye (Spielmann, 1997; York and Steiling, 1998).

Strategies for classification and labelling

Different systems are in use for the translation of Draize test results data into a classification for eye irritation/corrosion, such as the EU system (European Commission, 2001) or the Globally Harmonised System (GHS, United Nations, 2003). Despite some minor differences (slightly lower thresholds for corneal opacity and conjunctival redness for moderate irritants in the GHS, an additional second sub-category, 2B, for 'mild irritants', is provided for effects that are reversible within 7 days), classification results of these two systems have been proven to be almost identical, while differing significantly from those obtained using other concepts, such as the MMAS approach (Prinsen, 1999).

Today, classification for eye irritation is no longer based on the results of the Draize test alone. Instead, stepwise testing strategies have been developed, such as that summarised in figure 1, which has been annexed to major internationally accepted test guidelines (European Commission, 2004; OECD, 2002; United Nations, 2003).

In this strategy, the need for a Draize test is waived if any other data or information sources, including physico-chemical considerations, (Q)SARs, and *in vitro* testing, can be used for classification with respect to this hazard. However, in the case of negative or inconclusive results *in silico* or *in vitro*, up to three animals must still be tested. An obvious paradoxon lies in the fact that, according to this approach, most of the Draize tests 'needed' will be carried out with non-irritants which, moreover, are much more common than irritants anyway, at least in the EU New Chemicals database (Hoffmann and Hartung, 2005). Thus, for a large majority of the Draize tests to be performed, a negative test result can be expected.

Demands resulting from the REACH legislation

In 2003, the EU commission published its proposal for a new chemicals legislation, also known as the 'REACH' proposal (European Commission, 2003), which includes among others a fundamental change in the testing strategy for eye irritation that will *de facto* result in an almost total ban of *in vivo* testing of industrial chemicals with respect to this endpoint, if the alternative tools needed to support this new approach can be provided:

- All existing data (*in vitro, in vivo*, historical, validated QSARs, data on chemical analogues) must be considered prior to performing any new test.
- Performing *in vivo* tests will no longer be allowed for obtaining information on corrosive properties (EU risk phrases R34/35), or risk of serious damage to eyes (R41).
- For substances produced or imported in the EU in quantities of 1-10 t/a, an *in vivo* test must not be performed, even for testing potential moderate irritants (EU risk phrase R36). Instead, an animal-free testing strategy will be used, including the evaluation of existing human and animal data, the identification of extreme pH, and a suitable *in vitro* test. Even this test may be waived for known corrosives, if pH is below 2.0 or above 11.5, if the chemical is flammable upon contact with air at room temperature, or if the producer has voluntarily assigned a classification for eye irritation based on skin effects.
- For substances produced or imported in the EU in quantities above 10 t/a, an additional *in vivo* test may only be performed if an inconclusive result was returned by the testing strategy described in the preceding paragraph.

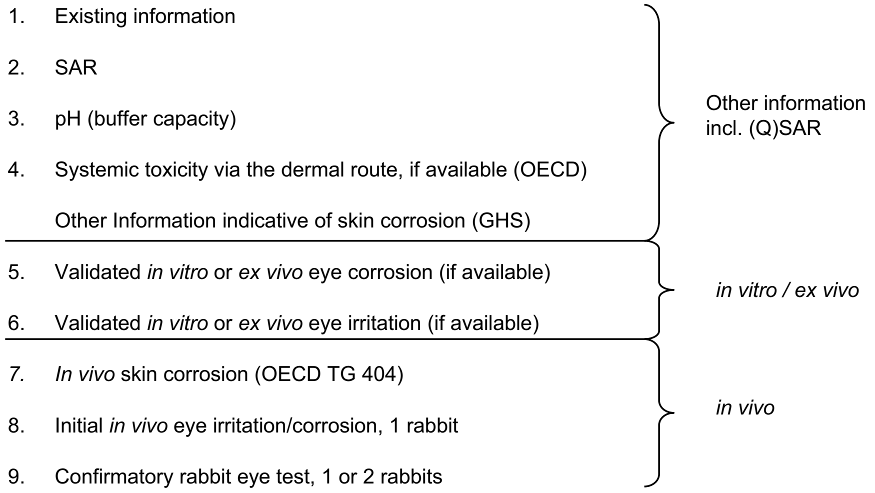
Alternative tools that need to be (further) developed to meet the demands of the future

In order to fulfil the demands of the proposed new policy, the following tools need to be established:

- Simple, animal-free test methods (such as *in vitro* tests or (Q)SARs) for confirming the ABSENCE of irritation/corrosion potential,
- Simple, animal-free test methods (such as *in vitro* tests or (Q)SARs) for discriminating between classification as R41/GHS 1 or R36/GHS 2,
- information management/decision support systems integrating results from different information sources/test methods (*in vivo, in vitro, in silico*) and of different levels of quality to aid industry and regulators in decision-making,
- a comprehensive use of existing information; this would mainly refer to the systematic collection, digitalisation, and evaluation/peer review of already existing toxicological (animal test) data and the development and implementation of efficient data mining techniques.

Regulatory acceptance of such tools will depend on proper validation as well as their ability to translate into internationally accepted classification and labelling systems. However, a number of associated problems arise:

- scoring of eye lesions according to the Draize method is a subjective process resulting in high variability, especially for mild to moderate irritants,
- for validation purposes, these variable Draize scores have to be used, as human reference data are scarce,



1. Existing information
2. SAR
3. pH (buffer capacity)
4. Systemic toxicity via the dermal route, if available (OECD)

   Other Information indicative of skin corrosion (GHS)

Other information incl. (Q)SAR

5. Validated *in vitro* or *ex vivo* eye corrosion (if available)
6. Validated *in vitro* or *ex vivo* eye irritation (if available)

in vitro / ex vivo

7. *In vivo* skin corrosion (OECD TG 404)
8. Initial *in vivo* eye irritation/corrosion, 1 rabbit
9. Confirmatory rabbit eye test, 1 or 2 rabbits

in vivo

Fig. 1: Current EU/OECD/GHS testing scheme for eye irritation/corrosion

- a principal difference between the Draize rabbit test and alternative techniques lies in the fact that two substances may receive the same classification and labelling based on different effects in different tissues, caused by different mechanisms. In contrast, a single *in vitro* or *in silico* method generally has a more specific focus and will therefore not be able to cover all effects that may lead to a given classification result.

Current acceptance of *in vitro* tests for eye irritation/corrosion in the EU

In the EU, several *in vitro/ex vivo* tests are accepted for the prediction of serious/irreversible eye damage (R41), e.g. (Spielmann 1997):

- the Bovine Corneal Opacity and Permeability (BCOP) test, the Hen Egg Test on the Chorioallantoic Membrane (HET-CAM), the Isolated Rabbit Eye test (IRE), or the Chicken Enucleated Eye Test (CEET).

A survey of data collected at the BfR under the EU New Chemicals notification programme between 1982 and 2004 demonstrates that out of 285 chemicals labelled for risk of serious eye damage (risk phrase R41), 24 had not been tested in the Draize test:

- 10 were classified due to extreme pH (with one also tested in a HET-CAM test),
- 9 were classified according to *in vitro/ex vivo test* results (5 BCOP, 2 IRE, 1 EYETEX, 1 unspecified),
- 3 were classified based on their severe skin irritation potential,
- two more substances were classified based on SAR considerations.

In contrast to serious eye damage, moderate irritation (corresponding to EU risk phrase R36 or GHS Cat. 2) is much more difficult to predict *in vitro*. No validated and generally acknowledged method is available at present. In 1998, at a workshop organised by the European Centre for the Validation of Alternative Methods (ECVAM), experience from previous validation exercises with *in vitro* testing methods for eye irritation/corrosion was evaluated (Balls et al., 1999). As a result, it was recommended to use a battery of tests rather than a single method, to integrate them into step-wise testing strategies, to validate the results against the EU/GHS classification scheme, and – finally – to learn more about the mechanistic basis of eye irritation/corrosion. These conclusions were more or less confirmed in a more recent review (Huggins, 2003).

Use of existing information – the BfR New Chemicals database for acute effects

As described previously (Gerner et al., 2000a; Gerner et al., 2005), quality-reviewed acute systemic and topical toxicity data of about 1,700 substances from the EU New Chemicals notification programme were collected between 1982 and 2002 by Ingrid Gerner at the BfR, primarily as a means of more effectively storing and using existing information for toxicological hazard assessment. All data were assessed for validity and quality by using identical assessment criteria, e.g. only substances of > 95% purity were included, and materials containing impurities suspected to be reactive were excluded. The database was accompanied by a software tool ('Estoff'), for data administration as well as for finding decision rules and structural alerts, and a Decision Support System (DSS) allowing for the application of these rules and alerts to new substances of unknown skin/eye irritation/corrosion potential without providing the actual confidential training set data (Gerner et al., 2000b; Zinke et al., 1999). In detail, the database contains information (if available/applicable) on:

- chemical structure and molecular weight, physico-chemical properties (differentiating between measured vs. estimated or calculated values): pH, aqueous/lipid solubility, melting/boiling points, log P_{OW}, surface tension, vapour pressure,
- potential for hydrolysis and/or thermal decomposition,
- acute systemic toxicity: oral and dermal LD_{50}, inhalative LC_{50}, and EU risk phrases, if assigned,
- skin irritation/corrosion: intensity and reversibility separately for erythemata and oedemata, time to reversibility, and EU risk phrases, if assigned,

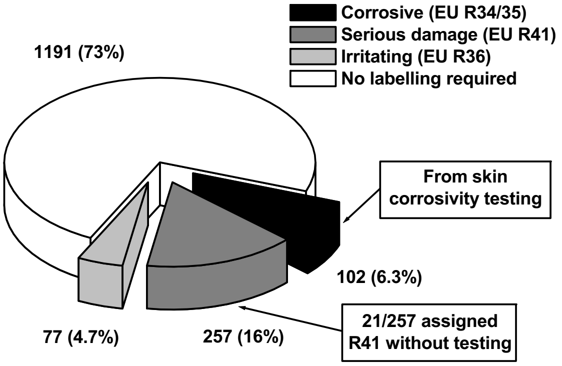


Fig. 2: Distribution of the substances (n = 1627) in the BfR New Chemicals database over different EU classification categories for eye irritation

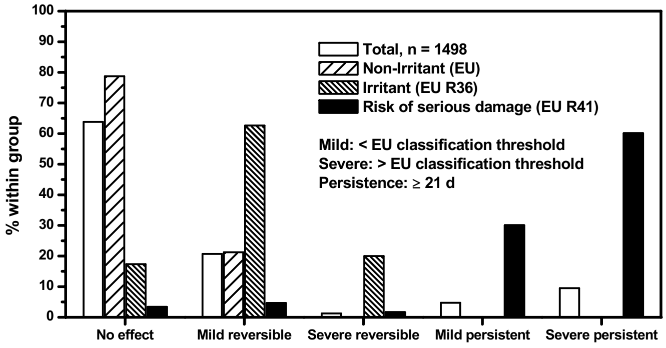


Fig. 3: Discrimination power of the sub-endpoint corneal opacity

- eye irritation/serious damage: intensity and reversibility separately for corneal, iridal, and conjunctival effects, time to reversibility, and EU risk phrases, if assigned,
- sensitisation potential: yes/no, *in vivo* test method used (including percentage of sensitised animals), route (dermal/respiratory).

Figure 2 gives an overview of the distribution of the substances (n=1627) in the database over different EU classification categories for eye irritation.

About 3 out of 4 substances in the database do not require hazard classification with respect to eye irritation/corrosion. This is in good agreement with figures reported by ECVAM from the EU commission's New Chemicals database (Hoffmann and Hartung, 2005). Only about 1 in 5 performed Draize tests actually led to a classification of moderate or severe eye irritation (313 chemicals, i.e. 77 substances classified as R36 and 236 substances classified as R41 based on the Draize test results), while 4 out of 5 tests (1191 substances not classified after a Draize test had been performed) could have been spared, if animal-free test methods/strategies for the exclusion of eye irritation potential had been available.

As already mentioned, the intensity and reversibility of effects are stored separately in the BfR database for the different sub-endpoints covered by the Draize test. These 'existing data' can be used to gain some insight into the potential and limitations of these sub-endpoints for discriminating between different categories of irritants (non-irritant/moderate/severe according to current EU criteria):

Figure 3 demonstrates that corneal opacity is well-suited for discriminating between non- or moderate irritants on the one hand, and severe irritants on the other. A large majority of severe irritants (all R41 substances causing severe reversible or persistent corneal damage, a total of approx. 92%) would have been classified correctly based on corneal opacity scores alone. In contrast, between non-irritants and moderate irritants, there is some degree of overlap and a lower discrimination power can be assumed.

This situation becomes more difficult with conjunctival effects: already almost 2/3 of all non-irritants in the database show some degree of mild, reversible irritation, and the same holds true for substances that were – overall – classified as moderate irritants (fig. 4).

About 80% of all substances and more than 90% of the non-irritants caused at least mild, reversible conjunctival redness (fig. 5).

Finally, for iritis (fig. 6), overlap between all groups is fairly high and, conversely, predictivity for classification becomes rather low, with 20% of the R41 substances displaying either no effect at all or evoking only a sub-threshold reaction. However, in this context it should be noted that iris effects are often camouflaged by the substantial corneal opacity typically caused by this group.

When looking at the data presented in figures 3-6 and considering again that scoring, especially at the border between mild and severe but reversible effects, is subjective and its results are highly variable, it is not surprising that problems with the use of the Draize test (and, consequently, with all alternative methods that are validated against it) must arise when non-irritants have to be discriminated from moderate irritants.

In summary, two conclusions might be drawn from the preceding considerations:

- Difficulties with discriminating between moderate and non-irritants based on Draize test data are an inherent problem of the scoring method and classification thresholds used and not of the alternative methods proposed for replacing the *in vivo* experiment.
- Rather than validating new alternative methods against the overall outcome of a Draize test, it appears more rewarding to perform such validations for each sub-endpoint separately. This approach would also be in line with the previous statement, i.e. that a single *in vitro* test or QSAR model for eye irritation/corrosion will not be able to cover all aspects of a Draize test with the same adequacy. Separate validations could also serve to gain a deeper understanding of the mechanistic aspects of eye irritation/corrosion and to facilitate the identification of the specific potentials and limitations of particular *in vitro* tests.

Using the BfR database to build (Q)SARs

The BfR database was used to establish structural alerts for the prediction of eye irritation potential as well as for building

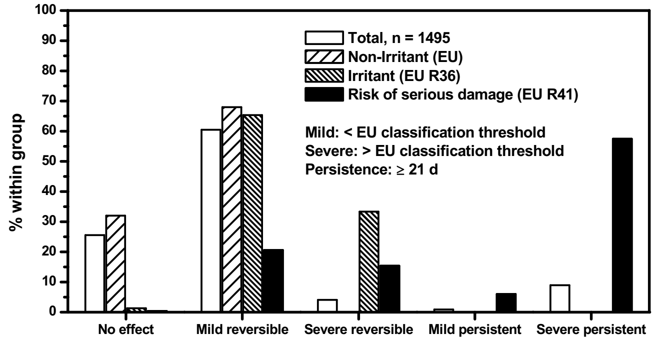


Fig. 4: Discrimination power of the sub-endpoint conjunctival oedema (chemosis)

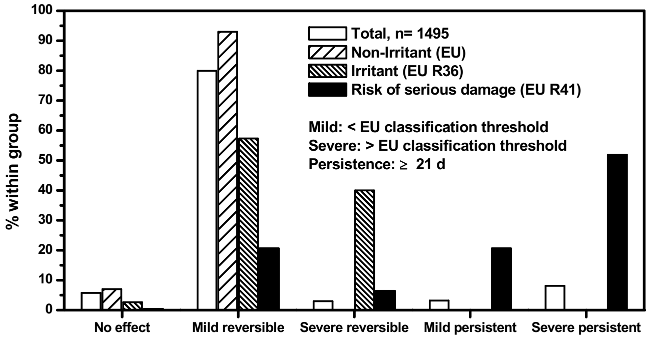


Fig. 5: Discrimination power of the sub-endpoint conjunctival erythema (redness)

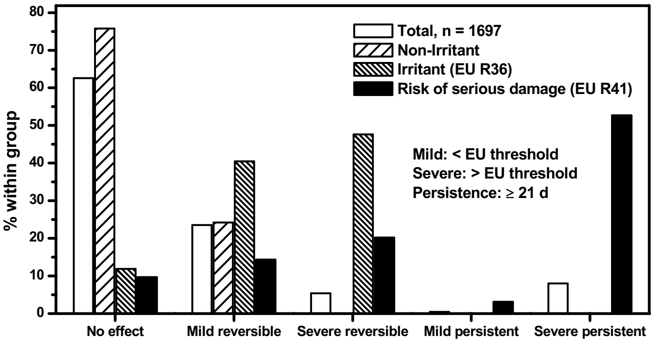


Fig. 6: Discrimination power of the sub-endpoint iritis

physico-chemical rules for the EXCLUSION of such potential. Both alerts and rules have been published before (Gerner et al., 2000b) and were updated recently (Gerner et al., 2005), based on the inclusion of additional data into the database. Examples of some physico-chemical exclusion rules are given in figure 7.

The rules were derived empirically by analysing the distribution of substances over the respective descriptors and setting the limit values such as to exclude 100% of the active (i.e. irritant) substances. Their use is easy and straightforward (in fact, they could be built into a simple Excel spreadsheet): descriptors are not estimated from structural features (e.g. by quantum-mechanical approximations), instead, 'real-life' physico-chemical properties are used for which measured data are readily available, as is at least the case for New Chemicals, pesticides and biocides in the EU.

In 2004, structural alerts and physico-chemical rules for both skin and eye effects have been submitted for validation to the (Q)SAR group at the European Chemicals Bureau (ECB). This validation project might be seen as a model for future validation exercises with models built on confidential data: The collection of data and model development were performed in one EU member state (BfR, Germany), the process was organised by a supranational European body (EU Joint Research Centre/ECB), and the validation itself was contracted to independent experts from a second EU member state (RIVM, the Netherlands).

In the meantime, the first part of the validation, i.e. the validation of skin irritation/corrosion physico-chemical rules has been completed (Rorije and Hulzebos, 2005), and some preliminary conclusions can be drawn that will most likely also apply to the rules for eye irritation:

- In general, the rules are in good agreement with the OECD principles for (Q)SAR validation.
- High predictivity was obtained in an external validation with a second dataset of EU New Chemicals (also compiled at the BfR, but not used to build the rules).
- Skin irritation/corrosion testing for up to approx. 40% of the test set substances could have been waived based on the validated rules. However, this percentage might be lower for test datasets from other domains of the chemical space (e. g. pharmaceuticals, pesticides, etc.).

After successful validation, the exclusion rules are proposed to act as early-stage 'exclusion filters' that are combined with both structural alerts and *in vitro* testing under the frame of an integrated testing strategy, such as that depicted in figure 8 (Gerner et al., 2005; Gerner and Schlede, 2002).

Recommendations for future work

Improved *in vitro* tools need to be developed for classification of moderate eye irritation (EU R36/GHS Category 2) as well as for predicting the absence of irritation potential. Internationally recognised test guidelines and standard operating procedures are required for the harmonised use of *in vitro* tests to predict eye irritation/corrosion. Validation of both the structural alerts and physico-chemical exclusion rules developed at the BfR should be extended to substances outside of the 'New Chemicals Space', e.g. pesticides, biocides or cosmetic ingredients. Further exploitation of the BfR database with different

Rules appropriate for all groups of chemicals

Basis
Evaluation of data for 1627 chemicals with purity ≥ 95%

If	melting point > 200 ºC	Then not	(skin corrosion R34 or R35) (true for 245/252 chemicals tested)[a]
If	log P_{ow} > 9	Then not	(lesions R34, R35, R36 or R41) (true for 32/32 chemicals tested)
If	log P_{ow} < -3.1	Then not	(skin corrosion R34 or R35) (true for 53/53 chemicals tested)
If	lipid solubility < 0.01 g/kg	Then not	(skin corrosion R34 or R35) (true for 58/58 chemicals tested)
If	aqueous solubility < 0.00002 g/l	Then not	(eye irritation R41) (true for 109/109 chemicals tested)
If	aqueous solubility < 0.000005 g/l	Then not	(eye irritation R36) (true for 38/38 chemicals tested)
If	molecular mass > 650 g/Mol	Then not	(eye irritation R36) (true for 139/139 chemicals tested)[b]

[a] *The seven skin corrosive substances are organic salts which release strong inorganic acids or bases when in contact with aqueous substrates/organic media*

[b] *Chemicals with molecular mass > 650 g/Mol may elicit severe tissue damage resulting in local corrosion (labelled R41)*

Fig. 7: Example of physico-chemical exclusion rules derived from the BfR database (from Gerner et al., 2005, where a more detailed description can be found)

(multivariate) statistical approaches could give deeper insight into the interdependency of both the physico-chemical descriptor variables and the different sub-endpoints of the Draize test (cf. e.g. Lovell, 1996). More sophisticated concepts to exploit existing *in vivo* test data more efficiently should be developed. A large collection of historical animal test data is stored in the archives of both industry and governmental authorities – a treasure of information still waiting to be systematically evaluated for its potential to be used as an aid to further reduce animal testing for eye irritation.

Summary and conclusions

In order to avoid the use of experimental animals in eye irritation/corrosion testing, integrated testing strategies are needed

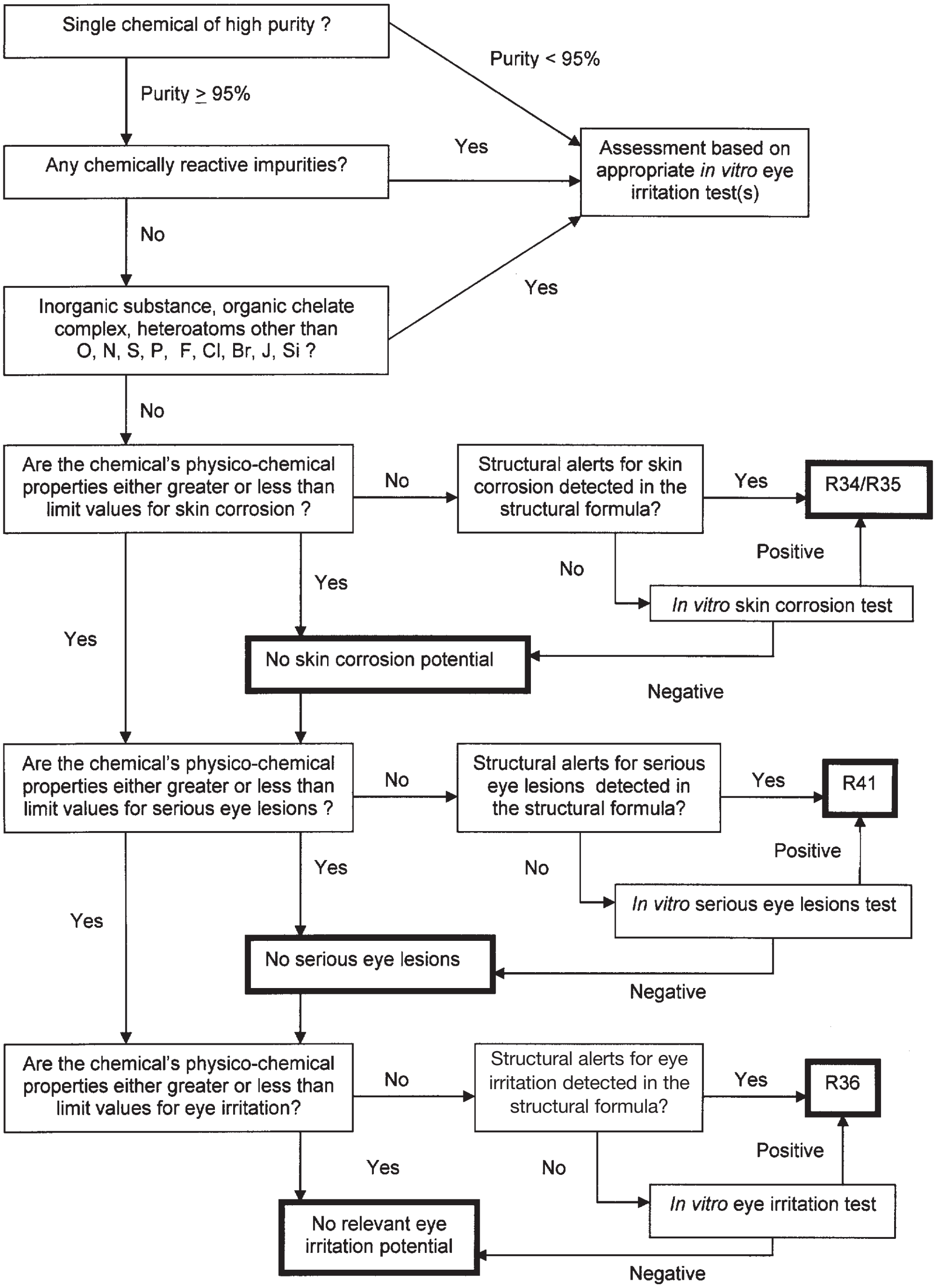

Single chemical of high purity ?
Purity < 95%
Purity ≥ 95%
Any chemically reactive impurities?
Yes
Assessment based on appropriate *in vitro* eye irritation test(s)
No
Inorganic substance, organic chelate complex, heteroatoms other than O, N, S, P, F, Cl, Br, J, Si ?
Yes
No
Are the chemical's physico-chemical properties either greater or less than limit values for skin corrosion ?
No
Structural alerts for skin corrosion detected in the structural formula?
Yes
R34/R35
Positive
No
In vitro skin corrosion test
Yes
Yes
No skin corrosion potential
Negative
Are the chemical's physico-chemical properties either greater or less than limit values for serious eye lesions ?
No
Structural alerts for serious eye lesions detected in the structural formula?
Yes
R41
Positive
Yes
No
In vitro serious eye lesions test
Yes
No serious eye lesions
Negative
Are the chemical's physico-chemical properties either greater or less than limit values for eye irritation?
No
Structural alerts for eye irritation detected in the structural formula?
Yes
R36
Positive
Yes
No
In vitro eye irritation test
No relevant eye irritation potential
Negative

Fig. 8: Proposed integration of physico-chemical exclusion rules, structural alerts and *in vitro* testing in an animal-free testing strategy for skin/eye irritation/corrosion

that combine the use of existing information, chemical knowledge, batteries of *in vitro* tests and (Q)SARs. For the prediction of serious/irreversible eye damage, validated *in vitro* methods are at hand that may most efficiently be used in a battery approach, however, valid methods are missing for reliable identification of moderate and non-irritants. During 20 years of New Chemicals notification in the EU, a database on the acute systemic and topical toxicity of some 1,700 chemicals was compiled at the BfR. From this database, structural alerts for the prediction of eye irritation/corrosion potential were derived as well as physico-chemical exclusion rules to predict the ABSENCE of such effects. Both the alerts and exclusion rules were submitted to the ECB for validation and first experience obtained with the validation of exclusion rules for skin irritation/corrosion produced promising results.

References

Balls, M., Berg, N., Bruner, L. H. et al. (1999). Eye irritation testing: the way forward. *Alternatives to Laboratory Animals 27*, 53-77.

Draize, J. H., Woodard, G. and Calvery, H. O. (1944). Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membranes. *Journal of Pharmacology and Experimental Therapeutics 82*, 377-390.

European Commission (2001). Commission Directive 2001/59/EC of 6 August 2001 adapting to technical progress for the 28th time Council Directive 67/548/EEC on the approximation of the laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances, OJ L225/1.

European Commission (2003). Proposal for a regulation of the regulation of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency and amending Directive 1999/45/EC and Regulation (EC) {on Persistent Organic Pollutants}, COM (2003) 644 final.

European Commission (2004). Commission Directive 2004/73/EC of 29 April 2004 adapting to technical progress for the 29th time Council Directive 67/548/EEC on the approximation of the laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances, OJ L216/3.

Gerner, I., Graetschel, G., Kahl, J. and Schlede, E. (2000a). Development of a decision support system for the introduction of alternative methods into local irritancy/corrosivity testing strategies. Development of a relational database. *Alternatives to Laboratory Animals* 28, 11-28.

Gerner, I., Liebsch, M. and Spielmann, H. (2005). Assessment of the eye irritating properties of chemicals by applying alternatives to the Draize rabbit eye test: the use of QSARs and in vitro tests for the classification of eye irritation. *Alternatives to Laboratory Animals 33*, 215-237.

Gerner, I. and Schlede, E. (2002). Introduction of in vitro data into local irritation/corrosion testing strategies by means of SAR considerations: assessment of chemicals. *Toxicology Letters 127*, 169-175.

Gerner, I., Zinke, S., Graetschel, G., and Schlede, E. (2000b). Development of a decision support system for the introduction of alternative methods into local irritancy/corrosivity testing strategies. Creation of fundamental rules for a decision support system. *Alternatives to Laboratory Animals 28*, 665-698.

Hoffmann, S. and Hartung, T. (2005). Prevalence and test interdependence: pivotal parameters in the design and validation of testing strategies. *ALTEX 22*, 281.

Huggins, J. (2003). Alternatives to animal testing: research, trends, validation, regulatory acceptance. *ALTEX Suppl. 1/03*, 3-61.

Lovell, D. P. (1996). Principal component analysis of Draize eye irritation tissue scores from 72 samples of 55 chemicals in the ECETOC data bank. *Toxicology in Vitro 10*, 609-618.

OECD (2002). Test guideline 405 “Acute Eye Irritation/ Corrosion”.

Prinsen, M. K. (1999). An Evaluation of the OECD Proposal for the Harmonised Classification of Eye Irritants and Corrosives. *Alternatives to Laboratory Animals 27*, 72-77.

Rorije, E. and Hulzebos, E. M. (2005). Evaluation of (Q)SARs for the Prediction of Skin Irritation/Corrosion Potential. Unpublished report. National Institute of Public Health and Environment, Expert Centre for Subsatnces (RIVM-SEC). Bilthoven, The Netherlands.

Spielmann, H. (1997). Ocular Irritation. In J. V. Castell and M. J. Gómez-Lechón (eds.), *In vitro methods in pharmaceutical research* (265-287). San Diego/CA, USA: Academic Press.

United Nations Economic Commission for Europe/UNECE (2003). Globally Harmonized System of Classification and Labelling of Chemicals (GHS), 137-149.

York, M. and Steiling, W. (1998). A critical review of the assessment of eye irritation potential using the Draize rabbit eye test. *Journal of Applied Toxicology 18*, 233-240.

Zinke, S., Gerner, I., Graetschel, G., and Schlede, E. (1999). Local irritation/corrosion testing strategies: development of a decision support system for the introduction of alternative methods. *Alternatives to Laboratory Animals 28*, 29-40.

Correspondence to

Matthias Herzler
Federal Institute for Risk Assessment (BfR)
Thielallee 88-92
14195 Berlin
Germany
e-mail: m.herzler@bfr.bund.de

ICCVAM Progress in Evaluating *In Vitro* Test Methods for Identifying Severe Ocular Irritants/Corrosives

Jill Merrill[1], Karen Hamernik[2], Leonard Schechtman[1], William Stokes[3] and Marilyn Wind[4]

[1]US Food and Drug Administration, Center for Drug Evaluation and Research, Rockville, MD, USA; [2]U. S. Environmental Protection Agency, Washington, DC, USA; [3]National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, National Institute of Environmental Health Sciences, NIH, DHHS, Research Triangle Park, NC, USA; [4]Consumer Product Safety Commission, Bethesda, MD, USA

The views expressed in this paper do not necessarily represent the official positions of any US Federal government agencies.

Summary

In response to a nomination by the US Environmental Protection Agency, the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) initiated a review of the validation status of four in vitro test methods for use in screening chemicals for severe eye irritation or corrosion. The four test methods are the isolated rabbit eye test (IRE), isolated chicken eye test (ICE), hen's egg test on chorioallantoic membrane (HET-CAM) and bovine corneal opacity and permeability assay (BCOP). Background review documents (BRDs) were prepared based on the available data and independently reviewed by an Expert Panel. This paper describes the status of the review process as of August, 2005.

Keywords: bovine corneal opacity and permeability assay (BCOP), hen's egg test on chorioallantoic membrane (HET-CAM), ICCVAM, in vitro ocular toxicity test methods, isolated chicken eye test (ICE), isolated rabbit eye test (IRE)

Introduction

An ongoing ICCVAM review is being conducted with the overall goal of evaluating the validation status of four *in vitro* test methods (IRE, ICE, HET-CAM and BCOP) for their possible regulatory use in a tiered-testing strategy. The use of these tests, once appropriately validated, could reduce and refine animal use. Using a tiered-testing strategy, substances that tested positive in a validated *in vitro* test could be classified and labelled as severe ocular irritants/corrosives with no *in vivo* testing necessary. Substances that tested negative would undergo additional testing, either in the *in vivo* rabbit eye test or another validated *in vitro* test capable of detecting false negatives in the first *in vitro* test (This testing strategy may not apply to some pharmaceuticals). However, the key to this tiered-testing strategy is the availability of appropriately validated *in vitro* tests.

Validation of alternative methods in the United States

In the United States, ICCVAM, which is composed of 15 federal agencies, coordinates the technical review of new or revised alternative test methods as well as issues related to their validation. An important aim of the ICCVAM process is to help facilitate regulatory acceptance of such methods, as appropriate, by relevant member agencies. Priority for ICCVAM activities is generally given to methods that may improve the prediction of adverse human, animal or ecological effects and those that might reduce, refine or replace animal use. Validation has been defined by ICCVAM as the process by which the reliability and relevance of an assay for a specific purpose is established (ICCVAM, 1997). Reliability is defined as the reproducibility of the test method within and among different laboratories. It should be based on performance with different substances, representing the types of chemicals and product classes that are expected to be tested, and the range of responses that needs to be identified. Relevance is defined as the extent to which an assay will correctly predict or measure the biological effect of interest (i.e., via performance characteristic measures compared to a standard). (In the draft version of the BRDs the term "accuracy" rather than the term "relevance" was used to describe the overall performance characteristics of a test method as the comparison was made with respect to rabbit *in vivo* data). Ideally this analysis would have evaluated the ability of the *in vitro* test to correctly predict human ocular toxicity testing. However, severe ocular irritants are not tested in human eyes and in the absence of such data, how well the *in vitro* test predicts the *in vivo* rabbit test was measured.

A technical review was initiated to assess the current validation status of each of the four test methods. An ICCVAM Ocular Toxicity Working Group (OTWG) was established to work with the NTP (National Toxicology Program) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) to carry out these evaluations. NICEATM provides support to ICCVAM. ICCVAM also collaborated closely with the European Centre for the Validation of Alternative Methods

(ECVAM). Draft BRDs were prepared for each of the four test methods. The BRDs were designed to be comprehensive, reviewing the available data and information for each of the four methods. As such, the information included in the BRDs describes what is known about the accuracy and reliability of the test method, the scope of substances tested and the availability of a standardised protocol. In addition, based on existing data, the BRDs describe the known usefulness and limitations of each test method. Recommendations for any future optimisation or validation studies and a list of recommended reference substances for those validation studies as well as a standardised test method protocol are included in each BRD.

In order to assess the validation status of the test methods, high quality *in vivo* rabbit eye data and high quality *in vitro* data for each test method was needed. Some of this came from the peer-reviewed literature and some was submitted to NICEATM in response to a Federal Register (FR) notice (March 24, 2004). To be considered in the analysis, the *in vivo* data had to be based on the Draize scoring system and had to meet the following acceptance criteria: At least three rabbits were tested in the study unless a severe effect was noted in a single animal, in which case substance classification could be based on effects observed in less than three animals; 0.1 mL or 0.1 g was tested in each animal, unless a severe effect was noted with lesser amounts; minimally, observations had to have been made at 24, 48, and 72 h, unless a severe effect was observed earlier. If any of the above three criteria were not met, the data for that substance could not be used for the accuracy analyses.

The goal was to be able to classify each substance using the three major hazard classification systems: the US Environmental Protection Agency (EPA), the European Union (EU) and the UN Globally Harmonized System (GHS). However, each of these systems uses different decision criteria to identify corrosives/severe irritants based on the same rabbit data. For example, although all three systems make classifications based on the magnitude of the individual rabbit response, the EPA and the GHS also use the time taken for ocular lesions to clear. Therefore, individual rabbit data, collected at the different observation times, were needed if the *in vivo* data were to be used in the accuracy analyses. This was not always available, resulting in some submitted data not being included in the accuracy analyses. The accuracy of each proposed test method, compared to the *in vivo* rabbit test, was evaluated by calculating various statistics, of which the key ones are described below. The first of these is also termed accuracy, but in this context it refers to accuracy in the sense of concordance and measures the proportion of correct outcomes, either positive or negative, of a test method. Sensitivity measures the proportion of all positive substances that are correctly classified as positive relative to the standard of comparison, whereas specificity measures the proportion of all negative substances that are correctly classified as negative. The false positive rate measures the proportion of all negative substances that are incorrectly identified as positive, whereas the false negative rate measures the proportion of all positive substances that are incorrectly identified as negative. Reliability includes intra-laboratory repeatability and intra- and inter-laboratory reproducibility. Repeatability refers to the closeness of agreement between results on the same substance, using identical conditions within a given time period. Intra-laboratory reproducibility refers to the extent to which qualified personnel, within the same laboratory, can replicate results using a specific protocol at different time periods. Inter-laboratory reproducibility refers to the extent to which qualified personnel can replicate results in different laboratories (i.e., transferability).

Based on each of the three classification systems, an accuracy and reliability analysis was performed for each of the four test methods, except where the available data did not permit a complete analysis. Once these analyses were complete, the draft BRDs were made publicly available on the ICCVAM/NICEATM website (http://iccvam.niehs.gov) in November 2004.

Expert Panel review of the BRDs

The BRDs were then reviewed by an independent international expert panel. The Expert Panel was selected with input from ICCVAM, the OTWG and ECVAM. Members of the Panel are recognised experts in the field and come from different backgrounds with representatives from academia, government, industry and animal welfare. Panel members were specifically chosen for both their scientific expertise and their lack of direct involvement with the test methods under consideration. A public meeting of the Panel was held on January 11-12, 2005 at the National Institutes of Health, Bethesda, Maryland. The Panel was charged with evaluating the extent and adequacy that each of the applicable ICCVAM validation and acceptance criteria had been addressed, based on the available information and data, or will be addressed in proposed studies. They were also charged with developing conclusions and recommendations on the usefulness and limitations of the assays, the protocol that should be used for any future testing and validation studies, the adequacy of the proposed optimisation and/or validation studies and the adequacy of reference substances proposed for future validation studies.

Results of the Expert Panel review

Although the Panel's final report is available on the ICCVAM/NICEATM website (http://iccvam.niehs.nih.gov), the key points for each test method will be discussed here.

IRE

Based on the available data, an evaluation of intra-lab repeatability and reproducibility could not be performed for this method. The Panel concluded that the BRD-proposed test method appears useful in a tiered-testing strategy to identify severe ocular irritants/corrosives. However, the test method accuracy must be corroborated with a larger number of substances and a reliability analysis should be conducted when additional data become available. With respect to optimisation and validation, the Panel recommended additional data be requested from the test method users and that a reanalysis of all the data be conducted subsequently. The Panel also proposed several modifications to the standardised protocol, including the use of positive and negative controls and benchmark substances, the inclusion of methods to detect cellular damage and death, the development of a standardised histopathology scoring system for corneal damage, and the use of reference photographs for all subjective endpoints.

ICE

The Panel concluded the ICE test method can be used in a tiered-testing strategy to identify severe ocular irritants/corrosives. However, the Panel noted that alcohols tend to be overpredicted, while surfactants tend to be underpredicted. The Panel suggested that solids and insoluble substances may not come in adequate contact with the corneal surface, resulting in underprediction. The Panel also proposed several modifications to the proposed protocol, including adding centering lights to the optical pachymeter; the use of histopathology when the standard ICE endpoints produce borderline results; the use of reference photographs for all subjective endpoints; the use of concurrent negative and positive control eyes (at least three eyes per group). The Panel suggested control eyes be spread throughout the superperfusion apparatus such that replicate eyes are placed randomly. This would make order effects in dosing less likely. Given the limited amount of ICE reliability data, additional studies were suggested to better characterise the repeatability and reproducibility of the test method.

HET-CAM

The Panel concluded that HET-CAM is useful in a tiered-testing strategy to identify severe ocular irritants/corrosives. However, the high false positive rate is a limitation of the test method. They suggested retesting positive HET-CAM results in a modified HET-CAM or in a different *in vitro* test method. The Panel stated that optimisation studies could increase the accuracy of HET-CAM, possibly reducing the false positive rate while maintaining an acceptable false negative rate. Therefore, a retrospective analysis should be conducted to determine whether different decision criteria might enhance the accuracy and/or reliability of the test method. The Panel also proposed modifications to the proposed protocol, including the inclusion of different endpoints (e.g. trypan blue absorption, antibody staining, membrane changes, etc.) which may reduce the false positive rate; the inclusion of procedures for applying and removing solids from the chorioallantoic membrane (CAM) which otherwise may adhere to the CAM and damage it upon removal.

BCOP

The Panel concluded that the BRD-proposed test method is useful in a tiered-testing strategy for identifying severe ocular irritants/corrosives. However, the test should not be used for alcohols, ketones, and solids. Further optimisation and validation studies are necessary before these materials can potentially be assessed with this assay. It needs to be confirmed that the BCOP identifies known human ocular irritants as well as or better than the Draize test. The Panel concluded that histology should be added to the test method protocol, unless the test substance is from a class of materials known to be accurately predicted using only opacity and permeability. The Panel also expressed concern that users be aware of the possibility of zoonoses, including Bovine Spongiform Encephalopathy. They recommended that standard universal precautions, such as gloves and glasses, always be used. The Panel also proposed several modifications to the proposed protocol including the use of the holders suggested by Ubels et al., (2002); re-examining the use of the calculated total score when the endpoint is severe injury only; possible changes to the medium used to bathe the eyes, including determination of whether foetal bovine serum is needed. The Panel also suggested the possibility of using the porcine eye as a model for the human eye. They recognised that this change would require a complete validation, but wanted to be sure that it was considered for future work.

For all four methods, the Panel specified that any further optimisation or validation studies should be conducted using existing animal data. Additional animal studies should only be conducted if important data gaps are identified. Such studies would need to be carefully designed to maximise the amount of information obtained and minimise animal usage. The Panel also commented on the accuracy and reliability of the *in vivo* rabbit test, stating there should be more discussion of the variability of the *in vivo* rabbit data. This is particularly important in the determination of accuracy of an *in vitro* test method. Because of the known variability in the rabbit test, it is not possible from the data presented in the BRD to determine whether the inconsistencies between the two tests are due to ‘failure’ of the *in vitro* test method or a misclassification by the single *in vivo* test result. Some public comments also expressed concern that the variability of the rabbit data be more prominently discussed in the BRDs.

The Panel also reviewed the adequacy and completeness of the proposed list of reference substances. They concluded that the list is comprehensive, substances appear to be commercially available in an acceptably pure form, and an appropriate range of ocular responses appears adequately represented. Although the Panel recognised that the list is limited by the availability of *in vivo* reference data, they concluded that surfactants are over-represented, the list is too long and should be shortened, but at the same time more inorganic substances should be added. They also recommended that substances known to induce severe ocular lesions in humans be included in the list, even in the absence of rabbit data.

Additional data and data reanalysis

Public comments made at the meeting indicated that additional data could be made available. The Expert Panel recommended that this data be requested and that a reanalysis of the accuracy and reliability of each test method be conducted. Consequently a second FR notice was published February 28, 2005, requesting all available *in vitro* data on these test methods and any corresponding *in vivo* rabbit data.

An accuracy reanalysis was also required because, after the

BRDs were released, NICEATM received clarification of the classification rules for severe irritants. This change resulted in a small number of substances previously classified as non-severe now being classified as severe. A reanalysis was also required because a standardised chemical structure classification scheme, based on Medical Subject Headings, was used to ensure consistency in classifying substances tested in the *in vitro* ocular test methods. This resulted in some chemicals being reclassified.

Revised performance characteristics

A brief summary of the revised performance characteristics is provided for each test method in table 1. These characteristics are important in considering a method's usefulness in a tiered-testing strategy. This summary is based on the GHS classification system only because the international community appears to be adopting this classification system. Based on the reanalysis, IRE had a sensitivity of 100%, identifying all known corrosives as corrosive. However, it was overly sensitive, with a high false positive rate (56%). Based on the limited number of substances evaluated, IRE had a false negative rate of 0% and may be useful in identifying a substance as non-corrosive. ICE had a low false positive rate and as such it appears to have utility in reliably identifying a substance as corrosive. But, with a sensitivity of 50%, it missed many corrosives. The high false negative rate means that products sent on for confirmatory *in vivo* testing have a high possibility of being severely irritating to rabbit eyes. HET-CAM had a high false positive rate (60%). However, the assay may still be useful in a tiered-testing strategy where positive substances could be retested *in vivo* or in another appropriately validated alternative method to confirm the result. The BCOP assay was fairly good at predicting both known corrosives and non-corrosives. It was also fairly reliable at correctly identifying a product as corrosive with a false positive rate of 20%. In terms of identifying a product as non-corrosive it was fairly reliable with a false negative rate of 16%.

The results of the reanalyses were made publicly available on the ICCVAM/NICEATM website (http://iccvam.niehs.nih.gov) in July, 2005. The Expert Panel has since been asked whether the results of the reanalyses form the basis for any changes to their conclusions. Once the Panel's conclusions are final, ICCVAM and OTWG will consider their report and the public comments received in response to the review process. ICCVAM will then make recommendations regarding these test methods to US Federal Agencies for their consideration and action.

References

ICCVAM (1997). Validation and regulatory acceptance of toxicological test methods: A Report of the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods. NIH Publication No.: 97-3981. Research Triangle Park, NC: National Institute of Environmental Health Sciences.

Ubels, J. L., Paauw, J. D., Casterton, P. L. and Kool, D. J. (2002). A redesigned corneal holder for the bovine cornea opacity and permeability assay that maintains normal corneal morphology. *Toxicol. In Vitro 16*, 621-628.

Acknowledgements

This review represents a collaborative effort between ICCVAM, ECVAM, OTWG and NICEATM. Many people participated with public comments, providing useful information and constructive criticism. The authors would like to acknowledge the time and expertise of all these people and thank them for their hard work in this ongoing collaborative effort.

Correspondence to

Jill Merrill, PhD, DABT
US Food and Drug Administration
Center for Drug Evaluation and Research
Rockville, MD, USA
e-mail: jill.merrill@fda.hhs.gov