Unclassified **ENV/JM/MONO(2015)27**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **23-Jul-2015**

English - Or. English

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND
THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

PERFORMANCE STANDARDS FOR THE ASSESSMENT OF PROPOSED SIMILAR OR MODIFIED IN VITRO RECONSTRUCTED HUMAN EPIDERMIS (RhE) TEST METHODS FOR SKIN IRRITATION TESTING AS DESCRIBED IN TG 439

(Intended for the developers of new or modified similar test methods)
Series on Testing and Assessment
No. 220

JT03380411

Complete document available on OLIS in its original format

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

ENV/JM/MONO(2015)27
Unclassified

English - Or. English

OECD Environment, Health and Safety Publications

Series on Testing and Assessment

No. 220

PERFORMANCE STANDARDS FOR THE ASSESSMENT OF PROPOSED SIMILAR OR MODIFIED

IN VITRO RECONSTRUCTED HUMAN *EPIDERMIS* (RhE) TEST METHODS FOR SKIN IRRITATION TESTING

AS DESCRIBED IN TG 439[1]

(Intended for the developers of new or modified similar test methods)

IOMC
INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS
A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT
Paris 2015

[1] Proposed new similar or modified test method following the PS of this Test Guideline should be submitted to the OECD for adoption and inclusion into the Test Guideline before being used for regulatory purposes.

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 34 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in eleven different series: **Testing and Assessment; Good Laboratory Practice and Compliance Monitoring; Pesticides; Biocides; Risk Management; Harmonisation of Regulatory Oversight in Biotechnology; Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; and Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (http://www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organisations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

Also published in the Series on Testing and Assessment link

**For this and many other Environment,
Health and Safety publications, consult the OECD's
World Wide Web site (www.oecd.org/chemicalsafety/)**

or contact:

**OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France**

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

This document contains the Performance Standards (PS) for the validation of similar or modified RhE methods for skin irritation testing as described in TG 439. In the past, PS were usually annexed to TGs. However, in view of separating information on the *use* of a test method as contained in the TG from information needed to *validate* test methods as contained in the PS, TGs and PS will now both be stand-alone documents. This approach had been agreed by the Working Group of the National Coordinators of the Test Guidelines Programme (WNT). In case of the current PS for skin *in vitro* irritation methods according to TG 439, the text was reviewed in regard to harmonising with other relevant documents addressing skin irritation and skin corrosion. The PS were reviewed by the OECD Expert Group on Skin Irritation/Corrosion in November 2014. The PS are intended for the developers of new or modified similar test methods to the validated reference method. The present document was approved by the WNT in April 2015, declassified and published under the responsibility of the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides, and Biotechnology on 10 July 2015.

INTRODUCTION

1. This document contains Performance Standards which allow, in accordance with the principles of Guidance Document No. 34 (1), determining the validation status (reliability and relevance) of similar and modified skin irritation test methods that are structurally and mechanistically similar to the RhE test method in OECD Test Guideline 439 (2).

2. These PS include the following sets of information: (i) Essential Test Method Components that serve to evaluate the structural, mechanistic and procedural similarity of a new similar or modified proposed test method, (ii) a list of 20 Reference Chemicals to be used for validating new or modified test methods and (iii) defined target values of reproducibility and predictive capacity that need to be met by proposed test methods in order to be considered similar to the validated reference methods.

3. The purpose of Performance Standards (PS) is to provide the basis by which new similar or modified test methods, both proprietary (*i.e.* copyrighted, trademarked, registered) and non-proprietary, can be deemed to be structurally and mechanistically similar to a Validated Reference Method (VRM) and demonstrate to have sufficient reliability and relevance for specific testing purposes (i.e., scientifically valid), in accordance with the principles of Guidance Document No. 34 (1). The PS, based on scientifically valid and accepted test method(s), can be used to evaluate the reliability and relevance of test methods that are based on similar scientific principles and measure or predict the same biological or toxic effect (1). Such methods are referred to as *similar* or *"me-too"* test methods. Moreover, the PS may be used to evaluate *modified* test methods, which may propose potential improvements in comparison to approved earlier versions of a method. In such cases the PS can be used to determine the effect of the proposed changes on the test method's performance and the extent to which such changes may affect the information available for other components of the validation process (e.g. relating to Essential Test Method Components). However, depending on the number and nature of the proposed changes as well as the data and documentation available in relation to these changes, modified test methods may : i) either be found unsuitable for a PS-based validation (e.g. if the changes are so substantial that the method is not any longer deemed sufficiently similar with regard to the PS), in which cases they should be subjected to the same validation process as described for a new test method (1), or ii) suitable for a limited assessment of reliability and relevance using the established PS (1). Similar or modified new test methods (i.e., "me-too" tests) successfully validated according to Performance Standards can be added to TG 439. However, Mutual Acceptance of Data (MAD) will only be guaranteed for those test methods reviewed and adopted by the OECD. Proposed similar or modified test methods validated according to these PS should therefore be submitted to the OECD for adoption and inclusion into TG 439 before being used for regulatory purposes.

4. These PS are based on the EC-ECVAM PS, updated according to the UN GHS systems on classification and labelling (3) (4) (5) (6). The original PS (7) had been defined upon completion of the validation study (8) and were based on the EU classification system as described in the 28th amendment to the Dangerous Substances Directive (9). Due to the adoption of the UN GHS system for classification and labelling in EU, which took place between the finalisation of the validation study and the completion of this Test Guideline, the PS were updated (5) (6). This update concerned: *i*) the composition of the PS Reference Chemicals and *ii*) the defined reliability and accuracy values (5) (6) (10). The PS consists of: (i) Essential Test Method Components; (ii) Recommended Reference Chemicals, and; (iii) Defined Reliability and Predictive Capacity Values that the proposed similar or modified test method should meet or exceed. The VRM used to develop the present PS is the EpiSkin^{TM} test method as described in TG 439 (2); whereas the EpiDerm^{TM} SIT (EPI-200) test method was used as a VRM together with EpiSkin^{TM} test method only to define the Essential Test Method Components. Definitions are provided in Annex I.

5. Similar (me-too) or modified test methods proposed for use under Test Guideline 439 (2) should be evaluated to determine their reliability and predictive capacity using Reference Chemicals represent the full range of the TG 404 *in vivo* irritation scores (Table 3), prior to their use for testing other chemicals, in order to ensure that these methods are able to correctly discriminate UN GHS No Category chemicals (for member countries that do not adopt optional UN GHS Category 3) from Irritant chemicals (UN GHS Category 2) (3). The proposed similar or modified test methods should have reproducibility, sensitivity, specificity and accuracy values which are equal or better than those derived from the VRM and as described in paragraphs 26 to 28 of these PS (Table 4) (4) (8).

ESSENTIAL TEST METHOD COMPONENTS

6. The Essential Test Method Components consist of essential structural, functional, and procedural elements of the scientific valid test methods (the VRMs) that should be included in the protocol of a proposed, mechanistically and functionally similar or modified test method. These components include unique characteristics of the test method, critical procedural details, and quality control measures. Adherence to essential test method components will help to assure that a similar or modified proposed test method is based on the same concepts as the corresponding VRMs (1) (2). The essential test method components to be considered for similar or modified test methods related to TG 439 are described in detail in the following paragraphs.

7. For specific parameters (e.g. Tables 1 and 2), or modified procedures, adequate values or procedures should be provided for the proposed similar or modified test method, these specific values or procedures may vary depending on the specific test method and/or its modification.

General conditions

8. Non -transformed human keratinocytes should be used to reconstruct the epithelium. The RhE model is prepared in inserts with a porous synthetic membrane through which nutrients can pass to the cells. Multiple layers of viable epithelial cells (basal layer, *stratum spinosum, stratum granulosum*) should be present under a functional *stratum corneum*. The test chemical is applied topically to the three-dimensional RhE model, which should have a surface in direct contact with air so as to allow for an exposure similar to the *in vivo* situation. The *stratum corneum* should be multi-layered containing the essential lipid profile to produce a functional barrier with robustness to resist rapid penetration of cytotoxic benchmark chemicals, *e.g.* sodium dodecyl sulphate (SDS) or Triton X-100. The barrier function should be demonstrated and may be assessed either by determination of the concentration at which a benchmark chemical reduces the viability of the tissues by 50% (IC_{50}) after a fixed exposure time, or by determination of the exposure time required to reduce cell viability by 50% (ET_{50}) upon application of the benchmark chemical at a specified, fixed concentration (see paragraph 8). The containment properties of the RhE model should prevent the passage of test chemical around the *stratum corneum* to the viable tissue, which would lead to poor modelling of skin exposure. The RhE model should be free of contamination by bacteria, viruses, mycoplasma and fungi.

Functional conditions

Viability

9. The assay used for quantifying tissue viability is the MTT-assay (11). The viable cells of the RhE tissue construct can reduce the vital dye MTT into a blue MTT formazan precipitate which is then extracted from the tissue using isopropanol (or a similar solvent). The optical density (OD) of the

extraction solvent alone should be sufficiently small, *i.e.* OD < 0.1. The extracted MTT formazan may be quantified using either a standard absorbance (OD) measurement or an HPLC/UPLC-spectrophotometry procedure (12). The RhE model users should ensure that each batch of the RhE model used meets defined criteria for the negative control. An acceptability range (upper and lower limit) for the negative control OD values should be established by the RhE model developer/supplier. Acceptability ranges for the negative control OD values for the RhE VRMs are given in Table 1. An HPLC/UPLC-Spectrophotometry user should use the negative control OD ranges provided in Table 1 as the acceptance criterion for the negative control. It should be documented that the tissues treated with the negative control are stable in culture (provide similar viability measurements) for the duration of the test exposure period.

Table 1: Acceptability ranges for negative control OD values of the VRMs

	Lower acceptance limit	**Upper acceptance limit**
EpiSkin™ (SM)	≥ 0.6	≤ 1.5
EpiDerm™ SIT (EPI-200)	≥ 0.8	≤ 2.8

Barrier function

10. The *stratum corneum* and its lipid composition should be sufficient to resist the rapid penetration of certain cytotoxic benchmark chemicals (*e.g.* SDS or Triton X-100), as estimated by IC_{50} or ET_{50} (Table 2).

Morphology

11. Histological examination of the RhE model should be performed demonstrating human multi-layered human *epidermis*-like structure containing *stratum basale*, *stratum spinosum*, *stratum granulosum* and *stratum corneum* and exhibits lipid profile similar to lipid profile of human epidermis.

Reproducibility

12. The results of the positive and negative controls of the test method should demonstrate reproducibility of the test method over time.

Quality control (QC)

13. The RhE model should only be used if the developer/supplier demonstrates that each batch of the RhE model used meets defined production release criteria, amongst which those for *viability* (paragraph 7), *barrier function* (paragraph 8) and *morphology* (paragraph 10) are the most relevant. An acceptability range (upper and lower limit) for the IC_{50} or the ET_{50} (see paragraphs 6 and 8) should be established by the RhE model developer/supplier. The acceptability range of the VRMs are given in Table 2. Adequate ranges should be provided for any new similar or modified test method. These may vary depending on the specific test method. Data demonstrating compliance with all production release criteria should be provided by the RhE model developer/supplier. Only results produced with tissues fulfilling all of these production quality criteria can be accepted for reliable prediction of irritation classification.

Table 2: QC batch release criteria of the VRMs

	Lower acceptance limit	Upper acceptance limit
EpiSkin^{TM} (SM) (18 hours treatment with SDS) (13)	IC_{50} = 1.0 mg/ml	IC_{50} = 3.0 mg/ml
EpiDerm™ SIT (EPI-200) (1% Triton X-100) (14)	ET_{50} = 4.0 hr	ET_{50} = 8.7 hr

Procedural Conditions

Application of the Test Chemical and Control Substances

14. At least three tissue replicates should be used for each test chemical and each control substance in each run. For liquid as well as solid chemicals, sufficient amount of test chemical should be applied to uniformly cover the *epidermis* surface while avoiding an infinite dose (*e.g.* a minimum of 26 $\mu L/cm^2$ or mg/cm^2 for the VRMs). Whenever possible, solids should be tested as a fine powder. The exposure periods and temperatures are optimized for each individual RhE model and are related to the different intrinsic properties of the RhE model (*e.g.* barrier function). Furthermore, the viability measurement should not be performed immediately after exposure to the test chemical, but after a sufficiently long post-treatment incubation period of the rinsed tissue in fresh medium. This period allows both for recovery from weak cytotoxic effects and for appearance of clear cytotoxic effects. A 42 hours post-treatment incubation period was found optimal for the VRMs (13) (14).

15. Concurrent negative control (NC) and positive control (PC) should be used in each run to demonstrate that viability (using the NC), and sensitivity (using the PC) of the tissues are within a defined historical acceptance range. The concurrent negative control also provides the baseline (100% tissue viability) to calculate the relative percent viability of the tissues treated with the test chemical. The PC suggested for the VRMs is 5% aqueous SDS. The suggested VRMs NCs is phosphate buffered saline (PBS).

Cell Viability Measurements

16. The MTT assay, which is a quantitative assay, should be used to measure tissue viability. It is compatible with use in a three-dimensional tissue construct. The tissue sample is placed in MTT solution of an appropriate concentration (*e.g.* 0.3 - 1 mg/mL in the VRMs) for 3 hours. The vital dye MTT is reduced into a blue formazan precipitate by the viable cells of the RhE model. The precipitated blue formazan product is then extracted from the tissue using a solvent (*e.g.* isopropanol, acidic isopropanol), and the concentration of formazan is quantified by determining the OD at 570 nm using a filter band pass of maximum ± 30 nm or, by using an HPLC/UPLC-spectrophotometry procedure (12). The same procedure should be employed for the concurrently tested negative and positive controls.

17. Optical properties of the test chemical or its chemical action on MTT may interfere with the measurement of MTT formazan leading to a false estimate of tissue viability. Test chemicals may interfere with the MTT assay, either by direct reduction of the MTT into blue formazan, and/or by colour interference if the test chemical absorbs, naturally or due to treatment procedures, in the same OD range of formazan (i.e. 570 ± 30 nm, mainly blue and purple chemicals). Pre-checks should be performed before testing to allow identification of potential direct MTT reducers and/or colour interfering chemicals. The corresponding procedures should be standardised and part of the SOP. Additional controls should be used to correct for a potential interference from these test chemicals such as the non-specific MTT reduction (NSMTT) control and the non-specific colour (NSC) control (see paragraphs 16 to 19). This is especially important when a specific test chemical is not completely removed from the tissue by rinsing or when it

penetrates the epidermis, and is therefore present in the tissues when the MTT viability test is performed. For coloured test chemicals or test chemicals that become coloured in contact with water or isopropanol, which are not compatible with the standard absorbance (OD) measurement due to too strong interference with the MTT assay (i.e., strong absorption at 570 ± 30 nm), an HPLC/UPLC-spectrophotometry procedure to measure MTT formazan may be employed (30). A detailed description of how to correct direct MTT reduction and colour interferences by the test chemical should be available in the test method's SOP. A description of the control measures used in the VRMs are summarised in paragraphs 16 to 19 below.

18. To identify direct MTT reducers, each test chemical should be added to freshly prepared MTT solution. If the MTT mixture containing the test chemical turns blue/purple, the test chemical is presumed to directly reduce MTT and a further functional check on non-viable RhE tissues should be performed, independently of using the standard absorbance (OD) measurement or an HPLC/UPLC-spectrophotometry procedure. This additional functional check employs killed tissues (by e.g., exposure to low temperature ("freeze-killed" tissues) or by other means), that possess only residual metabolic activity but absorb and retain the test chemical in a similar way as viable tissues. Each MTT reducing test chemical is applied on at least two killed tissue replicates which undergo the entire testing procedure. The true tissue viability is calculated as the percent tissue viability obtained with living tissues exposed to the MTT reducer **minus** the percent non-specific MTT reduction obtained with the killed tissues exposed to the same MTT reducer, calculated relative to the negative control run concurrently to the test being corrected (%NSMTT).

19. To identify potential interference by coloured test chemicals or test chemicals that become coloured when in contact with water or isopropanol and decide on the need for additional controls, spectral analysis of the test chemical in water (environment during exposure) and/or isopropanol (extracting solution) should be performed. If the test chemical in water and/or isopropanol absorbs light in the range of 570 ± 30 nm, further colorant controls should be performed or, alternatively, an HPLC/UPLC-spectrophotometry procedure should be used in which case these controls are not required (see paragraph 19). When performing the standard absorbance (OD) measurement, each interfering coloured test chemical should be applied on at least two viable tissue replicates, which undergo the entire testing procedure but are incubated with medium instead of MTT solution during the MTT incubation step to generate a non-specific colour (NSC_{living}) control. The NSC_{living} control needs to be performed concurrently to the testing of the coloured test chemical (in each run) due to the inherent biological variability of living tissues. The true tissue viability is calculated as the percent tissue viability obtained with living tissues exposed to the interfering test chemical and incubated with MTT solution **minus** the percent non-specific colour obtained with living tissues exposed to the interfering test chemical and incubated with medium without MTT, run concurrently to the test being corrected ($\%NSC_{living}$).

20. Test chemicals that are identified as producing both direct MTT reduction (see paragraph 16) *and* colour interference (see paragraph 17) will also require a third set of controls when performing the standard absorbance (OD) measurement, apart from the NSMTT and NSC_{living} controls described in the previous paragraphs. This is usually the case with darkly coloured test chemicals interfering with the MTT assay (e.g., blue, purple, black) because their intrinsic colour impedes the assessment of their capacity to directly reduce MTT as described in paragraph 16. These test chemicals may be retained in both living and killed tissues and therefore the NSMTT control may not only correct for potential direct MTT reduction by the test chemical, but also for colour interference arising from the retention of the test chemical by killed tissues. This could lead to a double correction for colour interference since the NSC_{living} control already corrects for colour interference arising from the retention of the test chemical by living tissues. To avoid a possible double correction for colour interference, a third control for non-specific colour in killed tissues (NSC_{killed}) needs to be performed. In this additional control, the test chemical is applied on at least two killed tissue replicates, which undergo the entire testing procedure but are incubated with medium instead of MTT solution during the MTT incubation step. A single NSC_{killed} control is sufficient per test chemical

regardless of the number of independent tests/runs performed, but should be performed concurrently to the NSMTT control and, where possible, with the same tissue batch. The true tissue viability is calculated as the percent tissue viability obtained with living tissues exposed to the test chemical **minus** %NSMTT **minus** %NSC_{living} **plus** the percent non-specific colour obtained with killed tissues exposed to the interfering test chemical and incubated with medium without MTT, calculated relative to the negative control run concurrently to the test being corrected (%NSC_{killed}).

21. NSC_{living} or NSC_{killed} controls are never required when using HPLC/UPLC-spectrophotometry, independently of the chemical being tested. NSMTT controls should nevertheless be used if the test chemical is suspected to directly reduce MTT or has a colour (intrinsic or when mixed with water) that impedes the assessment of the capacity to directly reduce MTT (as described in paragraph 16). When using HPLC/UPLC-spectrophotometry to measure MTT formazan, the percent tissue viability is calculated as percent MTT formazan peak area obtained with living tissues exposed to the test chemical relative to the MTT formazan peak obtained with the concurrent negative control. For test chemicals able to directly reduce MTT, true tissue viability is calculated as the percent tissue viability obtained with living tissues exposed to the test chemical minus %NSMTT. Finally, it should be noted that in very rare cases, direct MTT-reducers or MTT-reducers that are also colour interfering and are retained in the tissues after treatment, may not be assessable by the VRMs if they lead to ODs (using standard OD measurement) or peak areas (using UPLC/HPLC-spectrophotometry) of the tested tissue extracts that fall outside of the linearity range of the spectrophotometer.

Acceptability Criteria

22. For each run, tissues treated with the negative control should exhibit OD reflecting the quality of the tissues that followed shipment, receipt steps and all protocol processes and should not be outside of the historically established boundaries (see paragraph 7 and table 1). Similarly, tissues treated with the PC should show mean tissue viability (relative to the negative control) within an historically established range, thus reflecting the ability of the tissues to respond to an irritant chemical under the conditions of the test method. The variability between tissue replicates of test chemicals and/or control substances should fall within the accepted limits also established from historical values (*e.g.* SD ≤ 18 for the VRMs (13) (14). If either NC or PC included in a run falls outside of the accepted ranges, the run is considered non-qualified and should be repeated. If the variability between tissue replicates of test chemicals falls outside of the accepted range, the test chemical should be re-tested. Paragraph 29 provides more details on re-testing in case of non-qualified runs during validation studies. Importantly, an increased frequency of non-qualified runs may indicate problems with either the test system (e.g. the intrinsic RhE tissue quality) or with the handling (e.g. shipment, SOP execution). Therefore, occurrence of non-qualified runs in validation studies should be carefully monitored and all non-qualified runs need to be reported.

Interpretation of Results and Prediction Model

23. The OD values obtained with each test chemical should be used to calculate the percentage of viability relative to the negative control, which is set to 100%. In case HPLC/UPLC-spectrophotometry is used, the percent tissue viability is calculated as percent MTT formazan peak area obtained with living tissues exposed to the test chemical relative to the MTT formazan peak obtained with the concurrent negative control. The cut-off value of percentage cell viability distinguishing irritant from non-classified test chemicals and the statistical procedure(s) used to evaluate the results and identify irritant chemicals should be clearly defined, documented, and proven to be appropriate (see SOPs of adopted test methods for information). The cut-off values of the VRMs for the prediction of irritation are given below (13) (14):

- The test chemical is identified as requiring classification and labelling according to UN GHS (Category 2 or Category 1) if the mean percent tissue viability after exposure and post-treatment incubation is less than or equal (≤) to 50%. Since the VRMs cannot discriminate between UN GHS Categories 1 and 2, further information on skin corrosion is required to decide on the test chemical's final classification [see also (2)]. In case the test chemical is found to be non-corrosive (e.g., based on TG 430, 431 or 435), and tissue viability after exposure and post-treatment incubation is less than or equal (≤) to 50%, the test chemical is considered to be irritant to skin in accordance with UN GHS Category 2.

- Depending on the regulatory framework in member countries, the test chemical may be considered as not requiring classification and labelling (UN GHS No Category) if the tissue viability in the VRMS after exposure and post-treatment incubation is more than (>) 50%.

MINIMUM LIST OF REFERENCE CHEMICALS

24. Reference Chemicals are used to determine whether the reliability and predictive capacity of a proposed similar or modified test method, proven to be structurally and functionally sufficiently similar to the VRM, or representing a minor modification of the VRM, are equal or better than those derived from the VRM (4) (5) (6) (8) (10). The 20 recommended Reference Chemicals listed in Table 3 include chemicals representing different chemical classes (i.e. chemical categories based on functional groups), and are representative of the full range of Draize skin irritancy scores (from non-irritant to strong irritant). The chemicals included in this list comprise 10 UN GHS Category 2 chemicals and 10 non-categorised chemicals, of which 3 are optional UN GHS Category 3 chemicals. Under this Test Guideline, the optional Category 3 is considered as No Category. The Reference Chemicals were selected using the selection criteria as described in Table 3 (foot-note 1) on the basis of data from the VRMs and relate to chemicals used for the prospective validation study (8) as well as chemicals used in the optimisation phases following pre-validation. Due regard has been given to e.g. chemical functionality and physical state when composing this list (15) (16).

25. The 20 Reference Chemicals listed in Table 3 represent the minimum number of chemicals that should be used to evaluate the reliability and predictive capacity of a proposed similar or modified test method. The exclusive use of these Reference Chemicals for the development/optimization of new similar test methods should be avoided to the extent possible. In situations where a listed reference chemical is unavailable or cannot be used for other justified reasons, another chemical could be used provided it fulfils the selection criteria as described in Table 3 (foot-note 1) and adequate *in vivo* reference data need to be available, e.g. preferentially from the test chemicals used during optimisation following pre-validation or from the validation study of the VRMs (8) (15) (16). To gain further information on the predictive capacity of the proposed test method, additional chemicals representing other chemical classes and for which adequate *in vivo* reference data are available may be tested in addition to the minimum list of Reference Chemicals.

Table 3: Minimum List of Reference Chemicals for Determination of Reproducibility and Predictive Capacity of similar or modified RhE Skin Irritation Test Methods

Chemical[1]	**CAS Number**	**Physical state**	***In vivo* score**	**VRM* Cat. based on *in vitro***	**UN GHS Cat. based on *in vivo* results**
NON-CLASSIFIED CHEMICALS					
1-Bromo-4-chlorobutane	6940-78-9	Liquid	0	Cat. 2	No Cat.
Diethyl phthalate	84-66-2	Liquid	0	No Cat.	No Cat.
Naphthalene acetic acid	86-87-3	Solid	0	No Cat.	No Cat.
Allyl phenoxy-acetate	7493-74-5	Liquid	0.3	No Cat.	No Cat.
Isopropanol	67-63-0	Liquid	0.3	No Cat.	No Cat.
4-Methyl-thio-benzaldehyde	3446-89-7	Liquid	1	Cat. 2	No Cat.
Methyl stearate	112-61-8	Solid	1	No Cat.	No Cat.
Heptyl butyrate	5870-93-9	Liquid	1.7	No Cat.	No Cat. (*Optional Cat. 3*)
Hexyl salicylate	6259-76-3	Liquid	2	No Cat.	No Cat. (*Optional Cat. 3*)
Cinnamaldehyde	104-55-2	Liquid	2	Cat. 2	No Cat. (*Optional Cat. 3*)
CLASSIFIED CHEMICALS					
1-Decanol[2]	*112-30-1*	*Liquid*	*2.3*	*Cat. 2*	*Cat. 2*
Cyclamen aldehyde	103-95-7	Liquid	2.3	Cat. 2	Cat. 2
1-Bromohexane	111-25-1	Liquid	2.7	Cat. 2	Cat. 2
2-Chloromethyl-3,5-dimethyl-4-methoxypyridine HCl	86604-75-3	Solid	2.7	Cat. 2	Cat. 2
Di-n-propyl disulphide[2]	*629-19-6*	*Liquid*	*3*	*No Cat.*	*Cat. 2*
Potassium hydroxide (5% aq.)	1310-58-3	Liquid	3	Cat. 2	Cat. 2
Benzenethiol, 5-(1,1-dimethylethyl)-2-methyl	7340-90-1	Liquid	3.3	Cat. 2	Cat. 2
1-Methyl-3-phenyl-1-piperazine	5271-27-2	Solid	3.3	Cat. 2	Cat. 2
Heptanal	111-71-7	Liquid	3.4	Cat. 2	Cat. 2
Tetrachloroethylene	127-18-4	Liquid	4	Cat. 2	Cat. 2

*) VRM = validated reference methods (EpiSkin^{TM}, see paragraph 2 for explanations).

[1] The Reference Chemicals selection was based on the following criteria; (i), the chemicals are commercially available; (ii), they are representative of the full range of Draize irritancy scores (from non-irritant to strong irritant); (iii), they have a well-defined chemical structure; (iv), they are representative of the chemical functionality used in the validation process; and (v), they are not

associated with an extremely toxic profile (*e.g.* carcinogenic or toxic to the reproductive system) and they are not associated with prohibitive disposal costs.

[2] Reference Chemicals that are irritant in the rabbit but for which there is reliable evidence that they are non-irritant in humans (17) (18) (19).

DEFINED RELIABILITY AND PREDICTIVE CAPACITY VALUES

26. For purposes of establishing the reliability and predictive capacity (i.e., sensitivity, specificity and accuracy) of proposed similar or modified RhE test methods to be used by several laboratories, all 20 Reference Chemicals listed in Table 3 should be tested in at least three laboratories. However, if the proposed test method is to be used in a single laboratory only, multi-laboratory testing will not be required for validation. In each laboratory, all 20 Reference Chemicals should be tested in three independent runs performed with different tissue batches and at sufficiently spaced time points. Each run should consist of at least three concurrently tested tissue replicates for each test chemical, negative control, positive control and adapted controls for direct MTT reduction and/or colour interference.

27. The calculation of the within-laboratory reproducibility, between-laboratory reproducibility, accuracy, sensitivity and specificity values of the proposed test method should be done according to the rules described below to ensure that a predefined and consistent approach is used:

1. Only the data of runs from complete run sequences qualify for the calculation of the test method within, and between-laboratory variability and predictive capacity (accuracy).
2. The final classification for each Reference Chemicals in each participating laboratory should be obtained by using the mean value of viability over the different runs of a complete run sequence.
3. Only the data obtained for chemicals that have complete run sequences in all participating laboratories qualify for the calculation of the test method between-laboratory variability.
4. The calculation of the predictive capacity (sensitivity, specificity and accuracy values) should be done on the basis of the individual laboratory predictions obtained for the 20 Reference Chemicals by the different participating laboratories.

In this context, a **run sequence** consists of three independent runs from one laboratory for one test chemical. A **complete run sequence** is a run sequence from one laboratory for one test chemical where all three runs are valid. This means that any single invalid run invalidates an entire run sequence of three runs.

Within-laboratory reproducibility

28. An assessment of within-laboratory reproducibility should show in one single laboratory, a concordance of predictions (UN GHS Category 2 and No Category) obtained in different, independent test runs of the 20 Reference Chemicals equal or higher (≥) than 90%.

Between-laboratory reproducibility

29. An assessment of between-laboratory reproducibility is not essential if the proposed test method is to be used in a single laboratory only. For methods to be transferred between laboratories, the concordance of predictions (UN GHS Category 2 and No Category) obtained in different, independent test runs of the 20 Reference Chemicals between a minimum of three laboratories should be equal or higher (≥) than 80%.

Predictive capacity

30. The predictive capacity (sensitivity, specificity and accuracy) of the proposed similar or modified test method should be equal or better than the target values derived from the VRM, taking into consideration additional information relating to effects in the species of interest (Table 4). The sensitivity should be equal or higher (≥) than 80% (4) (5) (6) (10). However, a further specific restriction applies to the sensitivity of the proposed *in vitro* test method in as much as only two *in vivo* Category 2 Reference Chemicals, *1-decanol* and *di-n-propyl disulphide,* may be misclassified as No Category by more than one participating laboratory. The specificity should be equal or higher (≥) than 70% (4) (5) (6) (10). There is no further restriction with regard to the specificity of the proposed *in vitro* test method, *i.e.* any participating laboratory may misclassify any *in vivo* No Category chemical as long as the final specificity of the test method is within the acceptable range. The accuracy should be equal or higher (≥) than 75% (4) (5) (6) (10). Although the sensitivity of the VRM calculated for the 20 Reference Chemicals listed in Table 3 is equal to 90%, the defined minimum sensitivity value required for a similar or modified test method to be considered valid is set at 80% since both *1-decanol* (a borderline Reference Chemical) and *di-n-propyl disulphide* (a false negative of the VRM) are known to be non-irritant in humans (17) (18) (19), although being identified as irritants in the rabbit test. Since RhE models are based on cells of human origin, they may predict these Reference Chemicals as non-irritant (UN GHS No Category).

<u>Table 4:</u> Required sensitivity, specificity and accuracy values for similar or modified RhE test method to be considered valid to discriminate skin irritants (UN GHS Category 2) from non-classified (UN GHS No Category)

Sensitivity	Specificity	Accuracy
≥ 80%	≥ 70%	≥ 75%

Study Acceptance Criteria

31. It is possible that one or several tests pertaining to one or more Reference Chemicals does/do not meet the test acceptance criteria (non-qualified tests) or is/are not acceptable for other reasons such as technical reasons or because they were obtained in a non-qualified run due to failure of the concurrent positive and/or negative control. To complement missing data, a maximum of two additional runs are admissible ("re-testing"). More precisely, since in case of re-testing also the positive and negative control substances have to be concurrently tested, a maximum number of two additional runs may be conducted for each Reference Chemical in each laboratory. Non-qualified tests should be documented and reported. Importantly, each laboratory should not produce more than three qualified tests per Reference Chemical. Excess production of data and subsequent data selection are regarded as inappropriate. All tested tissues should be reported. The extent of unacceptable tests/runs should be documented and the basis for the likely cause of each should be provided.

32. It is conceivable that even after re-testing, the minimum number of three valid runs is not obtained for every Reference Chemical in every participating laboratory, leading to an incomplete data matrix. In such cases the following three criteria should all be met in order to consider the datasets acceptable for purposes of PS-based validation studies:

1. All 20 Reference Chemicals should have at least one complete run sequence in one laboratory;

2. Each at least three participating laboratories, should have a minimum of 85% of complete run sequences (for 20 Reference Chemicals: 3 invalid run sequences are allowed per laboratory);

3. At least 90% of all run sequences from at least three laboratories need to be complete (for 20 Reference Chemicals tested in 3 laboratories: a total of 6 invalid run sequences are allowed).

LITERATURE

1. OECD (2005), OECD Series on Testing and Assessment No. 34. Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. Available at: [http://www.oecd.org/document/30/0,3343,en_2649_34377_1916638_1_1_1_1,00.html].

2. OECD (2015), OECD Guideline for the Testing of Chemicals No. 439. *In vitro* skin irritation: reconstructed human epidermis test method. OECD, Paris.

3. United Nations (UN) (2013), Globally Harmonized System of Classification and Labelling of Chemicals (GHS), Second revised edition, UN New York and Geneva, 2013. Available at: [http://www.unece.org/trans/danger/publi/ghs/ghs_rev05/05files_e.html].

4. EC-ECVAM (2009), Statement on the "Performance under UN GHS of three in vitro assays for skin irritation testing and the adaptation of the Reference Chemicals and Defined Accuracy Values of the ECVAM skin irritation Performance Standards", issued by the ECVAM Scientific Advisory Committee (ESAC30), 9 April 2009. Available at: [http://ihcp.jrc.ec.europa.eu/our_activities/alt-animal-testing]

5. EURL-ECVAM (2009), Performance Standards for *in vitro* skin irritation test methods based on Reconstructed human Epidermis (RhE). Available at: [http://ihcp.jrc.ec.europa.eu/our_activities/alt-animal-testing] *N.B. This is the current version of the ECVAM PS, updated in 2009 in view of the implementation of UN GHS. These PS should not be used any longer as replaced by the present updated version.*

6. EURL-ECVAM (2009), ESAC Statement on the Performance Standards (PS) for in vitro skin irritation testing using Reconstructed human Epidermis, issued by the ECVAM Scientific Advisory Committee (ESAC31), 8 July 2009. Available at: [http://ihcp.jrc.ec.europa.eu/our_activities/alt-animal-testing]

7. EURL-ECVAM (2007), Performance Standards for applying human skin models to *in vitro* skin irritation testing. Available at: [http://ihcp.jrc.ec.europa.eu/our_activities/alt-animal-testing] *N.B. These are the original PS used for the validation of two test methods. These PS should not be used any longer as replaced by the present updated version.*

8. Spielmann, H., Hoffmann, S., Liebsch, M., Botham, P., Fentem, J., Eskes, C., Roguet, R., Cotovio, J., Cole, T., Worth, A., Heylings, J., Jones, P., Robles, C., Kandárová, H., Gamer, A., Remmele, M., Curren, R., Raabe, H., Cockshott, A., Gerner, I. and Zuang, V. (2007), The ECVAM international validation study on *in vitro* tests for acute skin irritation: Report on the validity of the EPISKIN and EpiDerm assays and on the skin integrity function test, *ATLA* 35, 559-601.

9. EC (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 laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances, Official Journal of the European Union L225, 1-333.

10. OECD (2010), Explanatory background document to the OECD draft Test Guideline on *in vitro* skin irritation testing. Published in OECD Series on Testing and Assessment, No. 137, OECD, Paris.
Available at:

http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2010)36&doclanguage=en

11. Mosmann, T. (1983), Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays, *J. Immunol. Methods* 65, 55-63.

12. Alépée, N., Barroso, J., De Smedt, A., De Wever, B., Hibatallah, J., Klaric, M., Mewes, K.R., Millet, M., Pfannenbecker, U., Tailhardat, M., Templier, M., and McNamee, P. Use of HPLC/UPLC-spectrophotometry for detection of MTT formazan in *in vitro* Reconstructed human Tissue (RhT)-based test methods employing the MTT assay to expand their applicability to strongly coloured test chemicals. Manuscript in preparation.

13. EpiSkin™ SOP, Version 1.8 (February 2009), ECVAM Skin Irritation Validation Study: Validation of the EpiSkin™ test method 15 min - 42 hours for the prediction of acute skin irritation of chemicals. Available at: [http://ihcp.jrc.ec.europa.eu/our_activities/alt-animal-testing]

14. EpiDerm™ SOP, Version 7.0 (Revised March 2009), Protocol for: *In vitro* EpiDerm™ skin irritation test (EPI-200-SIT), For use with MatTek Corporation's reconstructed human epidermal model EpiDerm (EPI-200). Available at: [http://ihcp.jrc.ec.europa.eu/our_activities/alt-animal-testing]

15. Cotovio, J., Grandidier, M.-H., Portes, P., Roguet, R. and Rubinsteen, G. (2005), The *in vitro* acute skin irritation of chemicals: optimisation of the EPISKIN prediction model within the framework of the ECVAM validation process, *ATLA* 33, 329-349.

16. Eskes, C., Cole, T., Hoffmann, S., Worth, A., Cockshott, A., Gerner, I. and Zuang, V. (2007), The ECVAM international validation study on *in vitro* tests for acute skin irritation: selection of test chemicals, *ATLA* 35, 603-619.

17. Basketter, D.A., York, M., McFadden, J.P. and Robinson, M.K. (2004), Determination of skin irritation potential in the human 4-h patch test. *Contact Dermatitis* 51, 1-4.

18. Jirova, D., Liebsch, M., Basketter, D., Spiller, E., Kejlova, K., Bendova, H., Marriott, M. and Kandarova, H. (2007), Comparison of human skin irritation and photo-irritation patch test data with cellular *in vitro* assays and animal *in vivo* data, *AATEX*, 14, 359-365.

19. Jírová, D., Basketter, D., Liebsch, M., Bendová, H., Kejlová, K., Marriott, M. and Kandárová, H. (2010), Comparison of human skin irritation patch test data with *in vitro* skin irritation assays and animal data, *Contact Dermatitis*, 62, 109-116.

20. OECD (2015), OECD Guideline for Testing of Chemicals. No. 404: Acute Dermal Irritation, Corrosion. OECD, Paris.

21. Harvell, J.D., Lamminstausta, K., and Maibach, H.I. (1995), Irritant contact dermatitis, *In*: Practical Contact Dermatitis, pp 7-18, (Ed. Guin J. D.). Mc Graw-Hill, New York.

ANNEX 1

DEFINITIONS

Accuracy: The closeness of agreement between test method results and accepted reference values. It is a measure of test method performance and one aspect of relevance. The term is often used interchangeably with "concordance" to mean the proportion of correct outcomes of a test method (1).

Between-laboratory reproducibility: A measure of the extent to which different qualified laboratories, using the same protocol and testing the same substances, can produce qualitatively and quantitatively similar results. Between-laboratory reproducibility is determined during the prevalidation and validation processes, and indicates the extent to which a test can be successfully transferred between laboratories, also referred to as inter-laboratory reproducibility (1).

Cell viability: Parameter measuring total activity of a cell population *e.g.* as ability of cellular mitochondrial dehydrogenases to reduce the vital dye MTT (3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide, Thiazolyl blue), which depending on the endpoint measured and the test design used, correlates with the total number and/or vitality of living cells.

Chemical: means a substance or a mixture.

Complete test sequence: A test sequence containing three qualified tests. A test sequence containing less than 3 qualified tests is considered as incomplete (see also definition of "test sequence" below).

Concordance: This is a measure of test method performance for test methods that give a categorical result, and is one aspect of relevance. The term is sometimes used interchangeably with accuracy, and is defined as the proportion of all chemicals tested that are correctly classified as positive or negative. Concordance is highly dependent on the prevalence of positives in the types of test chemical being examined (1).

ET_{50}: Can be estimated by determination of the exposure time required to reduce cell viability by 50% upon application of the benchmark chemical at a specified, fixed concentration, see also IC_{50}.

GHS (Globally Harmonized System of Classification and Labelling of Chemicals): A system proposing the classification of chemicals (substances and mixtures) according to standardized types and levels of physical, health and environmental hazards, and addressing corresponding communication elements, such as pictograms, signal words, hazard statements, precautionary statements and safety data sheets, so that to convey information on their adverse effects with a view to protect people (including employers, workers, transporters, consumers and emergency responders) and the environment (3).

HPLC: High Performance Liquid Chromatography.

IC_{50}: Can be estimated by determination of the concentration at which a benchmark chemical reduces the viability of the tissues by 50% (IC_{50}) after a fixed exposure time, see also ET_{50}.

Infinite dose: Amount of test chemical applied to the *epidermis* exceeding the amount required to completely and uniformly cover the *epidermis* surface.

Me-too test: A colloquial expression for a test method that is structurally and functionally similar to a validated and accepted reference test method. Such a test method would be a candidate for catch-up validation (1). The term is interchangeably used with similar test method.

Mixture: means a mixture or solution composed of two or more substances in which they do not react (3).

MTT: 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide; Thiazolyl blue tetrazolium bromide.

NSC$_{\text{killed}}$: Non-Specific Colour in killed tissues.

NSC: Non-Specific Colour in living tissues.

NSMTT: Non-Specific MTT reduction.

OD: Optical Density

PC: Positive Control, a replicate containing all components of a test system and treated with a substance known to induce a positive response. To ensure that variability in the positive control response across time can be assessed, the magnitude of the positive response should not be excessive.

Performance standards (PS): Standards, based on a validated test method, that provide a basis for evaluating the comparability of a proposed test method that is mechanistically and functionally similar. Included are; (i) essential test method components; (ii) a minimum list of Reference Chemicals selected from among the chemicals used to demonstrate the acceptable performance of the validated test method; and (iii) the similar levels of reliability and accuracy, based on what was obtained for the validated test method, that the proposed test method should demonstrate when evaluated using the minimum list of Reference Chemicals (1).

Prediction Model: a formula or algorithm (*e.g.,* formula, rule or set of rules) used to convert the results generated by a test method into a prediction of the (toxic) effect of interest. Also referred to as decision criteria. A prediction model contains four elements: (i) a definition of the specific purpose(s) for which the test method is to be used; (ii) specifications of all possible results that may be obtained, (iii) an algorithm that converts each study result into a prediction of the (toxic) effect of interest, and (iv) specifications as to the accuracy of the prediction model (*e.g.,* sensitivity, specificity, and false positive and false negative rates). Prediction models are generally not used in *in vivo* ecotoxicological tests (1).

Predictive Capacity: The predictive capacity reflects the test method performance in terms of correct and incorrect predictions in comparison to reference data. It gives quantitative information (e.g. correct prediction rate) on the relevance of the test method. It comprises, amongst others, the sensitivity and specificity of the test method.

Qualified run: A run that meets the test acceptance criteria for the NC and PC, as defined in the corresponding SOP. Otherwise, the run is considered as non-qualified.

Qualified test: A test that meets the criteria for an acceptable test, as defined in the corresponding SOP, and is within a qualified run. Otherwise, the test is considered as non-qualified.

Reference Chemicals: Chemicals selected for use in the validation process, for which responses in the *in vitro* or *in vivo* reference test system or the species of interest are already known. These chemicals should be representative of the classes of chemicals for which the test method is expected to be used, and should represent the full range of responses that may be expected from the chemicals for which it may be used, from strong, to weak, to negative. Different sets of reference chemicals may be required for the different stages of the validation process, and for different test methods and test uses (1).

Relevance: Description of relationship of the test method to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test method correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of the accuracy (concordance) of a test method (1).

Reliability: Measures of the extent that a test method can be performed reproducibly within and between laboratories over time, when performed using the same protocol. It is assessed by calculating intra- and inter-laboratory reproducibility (1).

Reproducibility: The agreement among results obtained from testing the same substance using the same test protocol (1).

Run: A run consists of one or more test chemicals tested concurrently with a negative control and with a positive control.

Sensitivity: The proportion of all positive/active chemicals that are correctly classified by the test method. It is a measure of accuracy for a test method that produces categorical results, and is an important consideration in assessing the relevance of a test method (1).

Skin irritation *in vivo*: The production of reversible damage to the skin following the application of a test chemical for up to 4 hours (20). Skin irritation is a locally arising reaction of the affected skin tissue and appears shortly after stimulation (21). It is caused by a local inflammatory reaction involving the innate (non-specific) immune system of the skin tissue. Its main characteristic is its reversible process involving inflammatory reactions and most of the clinical characteristic signs of irritation (erythema, oedema, itching and pain) related to an inflammatory process.

Specificity: The proportion of all negative/inactive chemicals that are correctly classified by the test method. It is a measure of accuracy for a test method that produces categorical results and is an important consideration in assessing the relevance of a test method (1).

Substance: means chemical elements and their compounds in the natural state or obtained by any production process, including any additive necessary to preserve the stability of the product and any impurities deriving from the process used, but excluding any solvent which may be separated without affecting the stability of the substance or changing its composition (3).

Test: A single test substance concurrently tested in a minimum of three tissue replicates as defined in the corresponding SOP.

Test chemical: means what is being tested.

Validated Reference Method(s) (VRM(s)): one (or more) test method(s) officially endorsed as scientific valid that was(were) used to develop the related official Test Guidelines and Performance Standards (PS). The VRM is considered the reference test method to compare new proposed similar or modified test methods in the framework of a PS-based validation study.

Within-laboratory reproducibility: determination of the extent that qualified people within the same laboratory can successfully replicate results using a specific protocol at different times, also referred to as intra-laboratory reproducibility (1).