EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection
European Centre for the Validation of Alternative Methods (ECVAM)

ECVAM
SCIENTIFIC
ADVISORY
COMMITTEE
(**ESAC**)

ESAC Working Group Peer Review Consensus Report

on

an ECVAM-coordinated study concerning the transferability and reliability of the Direct Peptide Reactivity Assay (DPRA) for skin sensitisation testing

Title page information			
File name	**TEMPLATE_ESAC-WG_REPORT-v4.doc**		
Abbreviated title of ESAC request	DPRA		
Relating to ESAC REQUEST Nr.	2011-03		
Request discussed through	ESAC 34 (March 2011) and through written procedure (mandate). Request adopted by ESAC on 27.1.2012		
Report to be handed over to ESAC Chair and Secretariat by	**14. May 2012**		
Version tracking			
Date	**Version**	**Author(s)**	**Description**
16.05.2012	1	Erwin L Roggen	
22.06.2012	2	Erwin L Roggen	
22.10.2012	3	Erwin L Roggen	Consensus report

TABLE OF CONTENTS

ESAC Working Group

This report was prepared by the "ESAC Working Group (ESAC WG) for Sensitization/DPRA, charged with conducting a detailed scientific peer review of three ECVAM-coordinated prevalidation study on three *in vitro* test methods for skin sensitisation testing: (1) the Direct Peptide Reacitvity Assay, DPRA; (2) the human Cell Line Activation assay, h-CLAT; (3) the MYELOID U937 Skin Sensitisation Test, MUSST. The primary goal of this study was to evaluate the transferability and reliability of these assays in view of assessing their potential usability within a testing strategy for skin sensitisation. This WG report focuses on the part of the ECVAM study addressing the DPRA assay.

The ESAC WG had been set up by the ESAC during its meeting on March 2011 (ESAC 34). Basis for the scientific review was the ECVAM request to ESAC concerning the scientific peer review of the DPRA and the formulation of scientific advice based on three principal questions outlined by ECVAM (ESAC request ER2011-03 ESAC Request DPRA-v4, see Annex 5).

Following availability of the Validation Study Report (VSR) early in 2012, the ESAC WG conducted the peer review from February 2012 to October 2012. This report was endorsed by the ESAC WG on September 2012 and represents the consensus view of the ESAC WG.

This ESAC WG peer review consensus report was endorsed by the ESAC on XXX.

<u>The ESAC WG had the following members:</u>

- Dr. Erwin Roggen (ESAC member, Chair of ESAC WG and rapporteur)
- Prof. Walter Pfaller (ESAC member, ESAC Vice Chair)
- Prof. A. Wallace Hayes (ESAC member)
- Dr. Maja Alecsic (external expert)
- Dr. Emanuela Corsini (external expert)
- Dr. David Lovell (external expert)
- Dr. Michael Woolhiser (external expert)
- Prof. Yong Heo (external expert)

<u>ESAC Coordination / Scientific Secretariat:</u>

- Dr. Claudius Griesinger
- Dr. Alexandre Angers (specific support)

ABBREVIATIONS USED IN THE DOCUMENT

- **BLR** Between-laboratory reproducibility
- **ECVAM** European Centre for the Validation of Alternative Methods
- **ESAC** ECVAM Scientific Advisory Committee
- **ESAC WG** ESAC Working Group
- **GCCP** Good Cell Culture Practice
- **GLP** Good Laboratory Practice
- **GPMT** Guinea Pig Maximisation Test (OECD Test Guideline 406)
- **LLNA** Local Lymph Node Assay (OECD Test Guideline 429)
- **PC** Positive Control
- **PM** Prediction Model
- **SOP** Standard Operating Procedure (used here as equivalent to 'protocol')
- **VC** Vehicle Control
- **VMT** Validation Management Team
- **WLR** Within-laboratory reproducibility

Executive summary

Following a request from ECVAM to ESAC for peer review of and scientific advice on an ECVAM-coordinated study focusing on the assessment of the transferability and reproducibility of the Direct Peptide Reactivity Assay (DPRA), ESAC, supported by the ESAC Secretariat, set-up an ESAC Working Group (ESAC WG) to prepare a detailed scientific peer review report and to draft an ESAC opinion for consideration by the ESAC.

The WG was specifically charged to assess the transferability and reproducibility (within- and between-laboratories) of the DPRA (primary objectives of the study) in view of its possible future use as part of a non-animal testing strategy for skin sensitization hazard assessment. As the study had also been designed to provide *preliminary* information on a) the predictive capacity of the test method and b) its potential use for assessing the sensitization potency of substances, the ESAC WG was requested to review the information on these two aspects. Finally, in view of the potential use of the DPRA within an Integrated Testing Strategy (ITS), the ESAC WG was requested to comment, if possible, on the particular strengths and weaknesses of the assay and its potential placing within such an ITS.

The ESAC WG met twice in person at ECVAM. The first meeting was in February (1-3) 2012. However, the DPRA study report and appendices were available for distribution to the ESAC only a few days in advance of the meeting leaving insufficient time for the WG to prepare for the meeting. Hence, a 2nd two-day meeting was organized in May (10-11) 2012. The DPRA study report was reviewed and the first version of this ESAC report (16.05.2012) was circulated amongst the WG. Following input from the ESAC WG and the ESAC Secretariat, a second version was circulated on 14.08.2012. The final report was adopted by the ESAC WG on 22/10/2012.

The ESAC WG considered the scientific work presented of good quality, despite (1) the fact that the WG had some concerns about the statistical calculations underlying the determination of an adequate sample size to analyse reproducibility as a primary study goal; (2) the WG's concern that possible limitations of the assay were not described in sufficient detail in the report.

Strengths of the study:

- The DPRA is a mechanistically based *in chemico* test method that intends to address one of the key events of the induction of skin sensitisation: haptenation. The DPRA thus has the potential to identify chemicals that are directly reactive with most, but not all, common protein nucleophiles believed to be the causative event of a Type 4 allergic response.
- The primary criterion for the chemical selection was the availability of robust *in vivo* data (LLNA and GPMT primarily, but also human data, allowing for adequate comparisons. The chemical selection also included substances which were previously reported as tested in the method and those that were not reported as tested using this method.
- The study design was considered appropriate for the purpose of addressing the first objective of the study: Assessing the transferability, WLR and BLR of the DPRA. The preliminary results on predictive capacity (secondary goal of the study) were considered very promising. However, in agreement with the Validation Management Group, the WG considered the number of chemicals as insufficient for allowing more than purely preliminary indications on the predictive capacity (in terms of S/NS as well as contributing to potency classification).
- The statistical approach chosen to analyse the data in view of performance was considered appropriate.

Weaknesses:

- It was unclear why a 75% power to detect a change from an assumed concordance had been selected rather than the more conventional 80% or 90%. It should be noted that a power of 75% means that a change of 25% in both directions can be detected. This has the consequence that the lower border of the confidence interval is 65% (90% - 25%) based on the assumption of 90% concordance (historical P&G values). This may be considered a rather low value.
- The limitations imposed on the test by the selection of reactive peptides (lysine, cysteine) were not sufficiently addressed.

Conclusions:

Overall, the conclusions made by the WG correspond well with the conclusions formulated in the report by the VMG.

- The WLR of the test method with respect to concordance of classification (S/NS) met the target of 85% (as defined by the VMG) and was considered by the ESAC WG sufficient for the purpose of this study.
- The data were considered strong enough to support transferability of the test to properly equipped, trained and staffed laboratories with the appropriate analytical capabilities.
- In spite of a BLR (75%) below the target of 80% (as defined by the VMG), the BLR of the test method with respect to concordance of classification was considered sufficient by the ESAC WG. On the other hand, the BLR assessment alone argued against the possibility to use the DPRA for assigning chemicals to one of the 4 reactivity classes (62.5%).
- As indicated by the VMG, the number of chemicals (N=24) was considered insufficient to draw firm conclusions on the predicative capacity of the test method. The preliminary data were, however, considered promising.
- The number of chemicals tested was insufficient to draw a conclusion about the applicability domain of the test. Empirically the applicability domain seems to exclude pre-/pro-haptens and metal salts.

With respect to the last item, the WG suggested that chemicals that preferably react with amino acids other then cysteine and lysine may fall outside the applicability domain of DPRA. If such preferential reactivity is known, testing of these chemicals can be avoided. However, it should be kept in mind that in cases such preferential reactivity with amino acids other than lysine/cysteine is not known, the test methods may generate false negative results. The same holds true for pre- and pro-haptens (i.e. possibility of false negative results). As there is some existing information (not from the study) that some pre-haptens have been correctly identified by the DPRA, the extent to which the assay systematically does or does not detect pre-haptens would warrant further investigation.

The limitations of the assay (either intrinsic ones: reactivity limited to lysine/cysteine or potential limitations: pre-/pro-haptens) should be carefully taken into account when interpreting, in particular, negative results from the DPRA assay.

Recommendations:

The DPRA addresses a key mechanism (haptenation) in the development of skin sensitization/allergic contact dermatitis. Overall the provided data support transferability and reproducibility of the test to

qualified laboratories. The predictive capacity of the test is not defined yet, but the preliminary data profile the test as a useful tool for early decision making during product development (screening) and a component in a weight-of-evidence approach or integrated testing strategy for safety/hazard assessment.

The WG recommends that the limitations of the DPRA that may lead to uncertainties concerning negative results be further investigated, either by additional prospective testing or through analysis of existing information.

1. Study objective and design

1.1 Analysis of the clarity of the definition of the study objective

(a) ESAC WG summary of the study objective as outlined in the VSR

General observations:

The DPRA study has to be seen as a brick in a larger study including the the human Cell Line Activation assay (h-CLAT) and the MYELOID U937 Skin Sensitisation Test (MUSST).

The primary overall objective of the report was the evaluation of the transferability and reliability (reproducibility within and between laboratories) of the method with a view to its future use in an integrated non-animal approach for replacing the currently used regulatory animal tests.

The secondary objectives were (a) a preliminary evaluation of the capacity of the method to distinguish Sensitizers (S) from Non-Sensitizers (NS) for classification and labelling, and (b) if possible a preliminary consideration of the ability of the assay to contribute to sub-categorisation of sensitizers (1A, 1B) according to the provisions of the UN Globally Harmonised System (GHS) for Classification and Labelling.

UN (2011) UN Globally Harmonised System of Classification and Labelling of Chemicals, 4th revised edition. Available at: http://www.unece.org/trans/danger/publi/ghs/ghs_rev04/04files_e.html [accessed August 2012])

(b) Appraisal of clarity of study objective as outlined in the VSR

General observations:

The report refers to a large extent to a larger study involving DPRA, MUSST and h-CLAT. The objectives as formulated seemed to reflect an expectation that the DPRA results were to be evaluated in the context of (a yet non-existing) integrated approach (VSR, section 1, §a and §b). This created confusion within the WG and turned out to be counterproductive.

The DPRA report was reviewed by the WG as an individual report on the transferability and reliability of the DPRA, and not ''in view of its future use in an integrated non-animal approach for replacing the currently used regulatory animal tests'' as stated in the primary overall objective (p.8). Although (even preliminary) conclusions on how to place the assay in an ITS were not evaluated by the study, ECVAM had requested the WG to make, if possible, recommendations on the potential use of the DPRA within an ITS. The WG however felt that this was not possible without qualified information about the predictivity of the DPRA, as well as the MUSST and h-CLAT assays.

Once this was clarified, the WG judged the primary and secondary objectives as clearly articulated. The study design was generally well thought through and planned.

Specific observations:

In agreement with the VSR, the number of chemicals selected for the study was judged bythe WG as insufficient to allow drawing robust conclusions on the predictive capacity of the DPRA (''Part (a)'' of the secondary objective (sub-categorisation) and its potential ability to contribute to subcategorization (''Part (b)'' of the secondary objective (sub-categorisation). The WG hence agreed with the view expressed in the VSR that these aspects could only be addressed in a preliminary manner by this particular study.

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

(a) Analysis of the scientific rationale provided in the VSR

General observations:

The scientific rationale of the study was adequately addressed in ''Background'' and ''4. Scientific basis – biological and/or mechanistic relevance of the DPRA''.

DPRA is a mechanistically based *in chemico* test method that addresses one of the key events of the induction of skin sensitisation: haptenation. The current understanding of chemical allergy indicates the need for binding of small-molecular weight chemicals to proteins through nucleophilic reactions (haptenation) to form an immunologically recognizable allergen. This protein modification is a characteristic of most chemical allergens and thus, an assessment of the reactivity of a chemical to a peptide *in chemico* is a logical surrogate for this feature of allergenicity.

The DPRA employs two peptides specifically designed to contain either lysine or cysteine as the main reactive amino acid, thereby covering the majority of such reactive chemicals. The designed peptide also contains arginine, but not histidine, which could be a target for certain chemicals.

Thus, the DPRA has the potential to identify chemicals that are directly reactive with most, but not all, common protein nucleophiles which is thought to be the causative event of a Type 4 allergic response. Further limitations of the DPRA include the detection of pre- and pro-haptens (possible false-negative test outcomes) and the observation that in case of cysteine-containing peptides it is difficult to distinguish between chemical induced oxidation (depletion of the peptide due to dimer formation) and haptenation.

Specific observation:

The ''Background'' section provides the general rationale for development of such alternatives in general. There are a few minor things to note here:

- Given that the present assay is based on chemical reactivity with lysine and cysteine residues in amino acids, it has to be highlighted that a) reactivity with amino acids other than lysine and cysteine and b) non-reactive inductive mechanisms of sensitisation belong to the limitations of the assay based on the known mechanisms of action.

- The key biological mechanisms are listed, however there is no mentioning of antigen presentation by the dendritic cells and the interaction with the T cell receptor.
- Point 6 of the biological mechanisms ends with the term 'haptenised chemical'. This term is most likely a mistake. A hapten is the chemical itself therefore it haptenises the target amino acid on a protein rather than itself.

(b) Analysis of the regulatory rationale provided in the VSR

General observations:

The relevant regulatory documents were referenced in the text.

The DPRA, in combination with other non-animal methods potentially addressing key events in sensitization, aims to reduce and eventually replace the currently used regulatory animal tests for hazard identification. It is not intended as a stand-alone but for use in a testing strategy to be developed in the future.

1.3 Appraisal of the appropriateness of the study design

General observations:

The project was described and designed in clearly recognizable and well described phases including Test Definition (Module 1), Transferability (Module 3), Within Laboratory Reproducibility (WLR) (Module 2), Between Laboratory Reproducibility (BLR) (Module 4) and Predictive Capacity (Module 5).

Transferability activities were divided into Training, Transferability and Quality Control. The WLR was formulated for each partner to include 1) concordance in prediction, 2) depletion values for cysteine and lysine, as well as 3) control values. The BLR was assessed in terms of 1) concordance in prediction and 2) depletion values for cysteine and lysine.

The number of test chemicals was sufficient for the primary objective to assess transferability and within-laboratory reproducibility (n=15 chemicals) and between-laboratory reproducibility (n=24 chemicals), and allowed for preliminary evaluation of the ability of the tests to discriminate between sensitizers and non-sensitizers as stated in the VSR (Secondary objective ''Part a''). Moreover, it should be noted that the chemicals sets used to assess transferability and reproducibility were different apart from two overlapping chemicals (p-benzoquinone and formaldehyde).

The project management was adequate.

Specific observations:

It was noticed that the work was performed under GLP, with the exception of the P&G lead laboratory.

1.4 Appropriateness of the statistical evaluation

a. Statistical calculations underlying sample size

General observations:

Overall, the chosen statistical approach was considered appropriate, but a number of areas were unclear making it difficult to fully assess its validity.

The sample size for the evaluation of WLR and BLR was calculated based on the expected proportion of concordant classifications obtained in different experiments performed within the same laboratory or in different laboratories, respectively. The methodology applied is the one for comparing two proportions, when one is known, as detailed in Equation (3.10), *Machin D, Campbell M, Fayers P, Pinol A. Sample size tables for clinical studies. 2nd ed. Oxford: Blackwell Science; 1997.*

Within laboratory reproducibility

Similar calculations were performed for the WLR. However, in absence of experimental historical data on WLR, the VMG assumed, based on previous experiences in validation studies that the WLR concordance would be higher than that for BLR and was set at 95%. The statistical power chosen was 80%.

Between laboratory reproducibility

The 'expected proportion' of concordant classifications (between laboratories) was calculated to be 90% on the basis of available data on between-laboratory reproducibility as submitted to ECVAM (see Appendix 2 of VSR, page 5). However, it was not clear for the WG why a power of 75% rather than the more conventional 80% or 90% power had been applied. This power allows for detecting 25% changes in each direction and, as a consequence, leads to a lower limit of the confidence interval of 65 % (90%-25%).

Specific observations:

The information provided in the VSR contained a justification for the choice of a sample size of at least 21 for the evaluation of BLR and at least 13 for the evaluation of WLR (Appendix 2 of VSR, page 3). This sample size is described as being calculated based on the information submitted to ECVAM in the Test Submission Template which include data from ring trials organised by Cosmetics Europe.
Sample sizes for the study were based on a power calculation carried out by an ECVAM statistician at the study design stage. There is a description of the calculations in Appendix 2 of VSR (pp. 2-3). The ESAC WG however found that the explanations were not detailed enough to fully comprehend how the minimal number of chemicals had been derived and, for this reason, requested ECVAM to provide additional information clarifying the calculations used to determine the sample size for WLR and BLR. This clarification (outlined by the statistician) was provided by ECVAM to the WG on 4 June 2012 and is reproduced in the box below.

Box 1: Clarification of the sample size calculation for within- and between-laboratory reproducibility performed during the study design phase. This summary by the ECVAM statistician in charge of the statistical aspects of study design had been forwarded on 4 June 2012 to the ESAC WG which had requested clarification on the methodology used to calculate the sample size required to conclude on sufficient BLR and WLR.

"The sample size for the evaluation of Between-Laboratory Reproducibility (BLR) and Within-Laboratory Reproducibility (WLR) was calculated based on the expected proportion of concordant classifications obtained in different experiments performed in different laboratories or within the same laboratory, respectively.
The methodology applied is the one for comparing two proportions, when one is known, as detailed in Equation (3.10), *Machin D, Campbell M, Fayers P, Pinol A. Sample size tables for clinical studies. 2nd ed. Oxford: Blackwell Science; 1997.*

Assuming that:
(1) π is the expected proportion of concordant classifications among laboratories
(2) $\pi - \delta$ is the lower border of the Confidence Interval for the expected proportion π (i.e. the reference proportion to which the expected proportion should be compared)

And considering Equation (3.10):

$$N = \frac{\{z_{1-\alpha/2}\sqrt{[\pi_1(1-\pi_1)]} + z_{1-\beta}\sqrt{[\pi_2(1-\pi_2)]}\}^2}{\delta^2}$$

It can be assumed that $\pi - \delta$ corresponds to π_1 of Equation (3.10), while π corresponds to π_2 of Equation (3.10). Therefore δ of Equation (3.10) corresponds to $\pi_2 - \pi_1$.

Equation (3.10) can then be written in terms of π and $\pi - \delta$ as follows:

$$n = \frac{[z_{1-\beta}\sqrt{\pi(1-\pi)} + z_{1-\alpha/2}\sqrt{(\pi-\delta)(1-\pi+\delta)}]^2}{\delta^2}$$

Considering the case of BLR and the assumptions reported in the Experimental Design:

$\pi = 0.9$
$\pi - \delta = 0.65$
$\alpha = 0.05$
$1 - \beta = 0.75$

The previous formula gives the following:

$N = [0.67 * \sqrt{(0.9*0.1)} + 1.96 * \sqrt{(0.65*0.35)}]^2 / [0.25]^2 =$
$= [0.67 * 0.3 + 1.96 * 0.477]^2 / [0.0625] =$
$= [1.13592]^2 / [0.0625] = 1.2903 / 0.0625 = 20.64 \approx 21$ chemicals
The same procedure can be applied for the WLR"

The 'historic' value of the concordance was set to be 90% on the basis of available information on BLR contained in the test submission to ECVAM. The sample size is based upon having 75% power to detect a change in either direction of 25%, specifically a reduction in the concordance from 90% to 65% (a difference of 25%).

Specifically the following points are not sufficiently clear from the description provided in the VSR:

- It is not completely clear why a power of 75% rather than the more conventional 80 or 90% power was used. It should be noted that a power of 75% will result in smaller sample sizes than, for instance, using 90% power for the formulation presented in the memo (i.e. 21 v. 28 chemicals). It should be noted that the VMG decided, on the basis of this calculation, to use

24 chemicals instead of the 21 calculated for a power of 75%. Thus, the power of the sample size used for empirical testing was 83%.

- There is a complication in that the statistical test is two-sided and the CI for the concordance is asymmetric. It is not possible to have a concordance greater than 100%. This, however, is a widely recognised problem in diagnostic tests with values close to 100%.
- It is not completely clear why the comparison was made with a concordance of 65%. The implication is that this was chosen to be "the lower border of the CI for the expected proportion" i.e. the lower limit on the 95% confidence interval. If this is the case, the CI seems rather wide and results consequently in a small sample size. The implication is that the CI of the estimate of the proportion would extend from 65% to 100% as the CI is asymmetrical.
- The implications of the sample size chosen is that even if the 'sample' had a concordance of about 65% it would be difficult to rule out the possibility that the 'true' underlying 'population' concordance could be 90%.

An alternative to the hypothesis testing approach would have been the estimation approach of sample size estimation for a specific precision. This is basically the approach used in sampling designs, such as opinion polls, to provide appropriate margins of error.

The approximate sample sizes needed to provide margins of errors of +/- 5%, 7.5%, 10% and 12.5% on an estimate of concordance of 90% are 139, 62, 35 and 23, respectively. The margin of error is half the width of the confidence interval. Therefore, the margin of error returned by the sample size used in the study (n=24) appears >12.5%.

(*b) Statistical analysis of the experimental data for assessing reproducibility*

General observations:

Prediction model not stated in VSR, thus underlying algorithm required for WLR & BLR assessment via concordance was not readily available.

Specific observations:

Appendix 2 (Experimental Design) specified some statistical analyses that were not performed according to Appendix 15 (Statistical Analysis). The possibility for amendments (if scientifically justified) was mentioned in Appendix 2, but the actual implementation of such amendments was not highlighted in the report. It has been assumed that the amendments were accepted by the VMT.

2. Collection of existing data

2.1 Existing data used as reference data

General observations:

Two recognised databases were used as a convenient source of authoritative peer-reviewed data for chemical selection: i) the ICCVAM database containing information about 103 chemicals, and ii) the LLNA database of 341 chemicals (Gerberick et al, Kern et al) (p15) . These databases include all the relevant reference data.

The primary criterion for the chemical selection was the availability of robust *in vivo* data (primarily LLNA and GPMT, but also human data) to allow for adequate comparisons to be made. The chemical selection also included substances (about one third) which were previously reported tested in the method and those that were not reported as tested using this method.

2.2 Existing data used as testing data

General observations:

The number of chemicals, and the ratio of sensitizers (S) to non sensitizers (NS), was selected based on statistical considerations. According to these, the VSR states that at least 21 chemicals are required for BLR assessment, and at least 13 for WRL assessment. Therefore, 24 chemicals (16 sensitizers and 8 non sensitisers) were selected for assessing BLR, and 15 chemicals (10 sensitizers and 5 non sensitizers) were selected for assessing WLR. The same 24 chemicals are going to be evaluated in the other two methods currently under validation at ECVAM.

2.3 Search strategy for retrieving existing data

General observations:

The rationale, the strategy and the procedure followed for the chemicals selection and associated reference data are exhaustively described in Appendixes 2 and 3.

The two recognised databases (i.e., ICCVAM performance standards and P&G publications) provided a convenient source of authoritative data for selection of substances meeting criteria of being associated with reference testing data of sufficient quality. These datasets were considered reliable and sufficient for the purposes of the study design and objectives. A sufficient diversity of test substances to satisfy selection criteria was identified from these lists. No further search strategy was deemed necessary.

The historical dataset (used for the comparison and evaluation of the transfer experiments in each naïve laboratory) was acquired by P&G during a ring trial performed prior to this study.

2.4 Selection criteria applied to existing data

General observations:

The rationale for the selection of reference substances was adequately described under the criteria defining the data reliability (see: Selection of Test Chemicals).

3. Quality aspects relating to data generated during the study

3.1 Quality assurance systems used when generating the data

General observations:

The quality systems in the participating laboratories were clearly described and are thought sufficient.

Specific observations:

Ricerca Biosciences was fully Good Laboratory Practice (GLP) accredited and performed the work according to GLP.

IVMU performed the work described in the study in compliance with the OECD Principles of Good Laboratory Practice.

The P&G site followed a set of quality assurance requirements defined by the VMG and considered essential for the acceptance of information and data produced in the validation process (see pg 13 of VMG report for detailed requirements).

3.2 Quality check of the generated data prior to analysis

General observations:

Statistical analysis of data was performed by an independent group (Adriaens Consulting BVBA, Aalter, Belgium). The statistical analysis plan was decided before the start of testing phase.

Only data from valid experimental runs were considered. The frequency of invalid runs was reported. Quality check procedure was developed in the form of a checklist (described in appendix 7). It focused on the acceptance criteria for the run and for each of the chemicals to ensure the results were valid. Internal checks were used during the compilation of the summary template for the statistician, in order to ensure that no transcription errors were made in the transfer of data. As an additional check, the final conclusions for each chemical were also compared to the conclusion of the reports sent by the individual laboratory.

4. Quality of data used for the purpose of the study (existing and newly generated)

4.1 Overall quality of the evaluated testing data (newly generated or existing)

General observations:

The measures put in place by the VMG and ECVAM (training of the participating laboratories and a 2-step assessment of the protocol performance) resulted in a data set with an overall high quality supporting reproducibility with respect to S/NS decisions.

The evaluated data included a number of results (IVMU laboratory) which were derived from runs which did not strictly meet test acceptance criteria. In short, the the values for Reference Control C (cysteine peptide) were consistently below but very close to the lower acceptance limit (the lower acceptance limit was 0.45 mM and the non-qualified results of IVMU were in average 0.44 mM). Considering that this acceptance limits were based on the criteria developed in one single laboratory (lead lab) and considering the proximity of these reported values to the lower acceptance limit, the VMG decided to accept these values. The rationale for this decision is reproduced on pages 51-52 of the VSR.

The VMG made this decision prior to the ultimate analyses and reproducibility comparisons for the coded test substances. Thus, the WG agreed that the final analysis of the S/NS concordance between experimental runs and laboratories was unaffected by this decision.

It was noted that for a test system which is largely non-biological, requiring only direct chemical and peptide reactivity following complete solubility/miscibility, the absolute depletion values were unexpectedly variable. In the context of the primary study objective (i.e., reproducibility of S/NS), however, this did not present an issue.

Specific observations:

The acceptance criteria for the method were clearly described. The mean cysteine and lysine concentration for control A, control C in water and control C in acetonitrile should be between 0.45 mM and 0.55 mM. The mean cysteine depletion for the positive control should be between 60.8 % and 100 %, and for the lysine depletion between 40.2 % and 69.4 %.

The acceptance criteria for all reference and positive controls were always met by P&G and Ricerca Biosciences, but not by IVMU. At IVMU the data for the cysteine reference control C and lysine depletion for the positive control were systematically very close to the lower limit of the acceptance range. On request of the WG concerning further clarification of this issue, ECVAM provided the information that VMT had decided during its meeting held at Ispra on October 6th, 2011, to accept all the results of the experiments performed at IVMU, and that the acceptance criteria would need to be revised at the end of the study.

This is likely to occur in the real life, but the WG has the concern that this decision has been influenced by the fact that the laboratory was the one in ECVAM. Consequently, it is not clear how any other naïve laboratory would have performed (Ricerca Biosciences was not a naïve laboratory). In the previous ring trial sponsored by Cosmetics Europe such problems were not encountered.

4.2 Quality of the reference data for evaluating reliability and relevance[1]

General observations:

Having come from expert-reviewed sources, the reference data used to classification S/NS were considered to be highly reliable. The reference data were considered to be (largely) accurate in the characterisation of skin sensitisation classification, again based on previous expert-reviews. There were exceptions in which a few test substances which have given inconsistent or ambiguous results *in vivo* were strategically included in the study design so that the DPRA might be assessed for its potential improvement in the prediction of sensitisation potential.

Furthermore historical DPRA data were used during the transfer phase. These historical data originated from P&G (lead laboratory). For the purposes of the training and transfer phases, these data were also considered to be reliable and appropriate.

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

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

General observations:

The Reference Control C data variability did not meet the intended assay criteria for some runs in one particular laboratory. Some experimentation was undertaken to investigate this but in the end the reason(s) for this phenomenon remains unknown.

Specific observations:

For Reference Control A the reproducibility and variability look good. Evaluation of the Reference Control C data (n=4) reveal a decline in the HPLC signal for the unreacted cysteine peptide over the time course of the assay (i.e., 24hr). It was pointed out that there are Reference Control B samples in the analysis sequence that are measured at the beginning of an HPLC run and again at the end. These data were not presented although they might give important information about the degree of specificity of the phenomenon for Reference Control C.

It has been observed that oxidation/dimerization of the cysteine peptide is possible, thus appearing as depletion via HPLC-UV/VIS. This was not proposed or discussed in the report. The ESAC WG expressed the concern that one reason for Reference C not meeting the assay criteria may be unstability of the cysteine stability in the presence of certain test substance resulting in either promotion or inhibition of Cys-Cys dimerization.

5. Test definition (Module 1)

5.1 Quality and completeness of the overall test definition

General observations:

The overall test definition was considered of good quality, containing well described test system, protocol (SOP, version 2), prediction models and biological/mechanistic relevance of the test method.

Known limitations and drawbacks of the method (including solubility and co-elution) were listed. In order to cope with co-elution of the test material with the lysine peptide a modified prediction model allowed for classification of the chemical into reactivity classes based on the depletion value for the cysteine peptide alone.

The VMG did however not consider that cysteine peptide depletion of the parent peptide may also occur as a function of the oxidation of the peptide in the interaction with the chemical resulting in a peptide dimerisation. As it is not possible to distinguish between haptenation and oxidation of cysteine in this assay, there remains some uncertainty regarding positive results, which may result

from such an overestimation of the positive reaction, thus the reactivity class of the compound and which may lead to false positive predictions.

The acceptance criteria are divided in criteria for assay run, test chemical and data acceptance. The acceptance criteria are formulated clearly. The system suitability is stringent. The SD and CV values mentioned in the run acceptance criteria were provided upfront by P&G and were based on historical data.

Specific observations:

The WG expressed concern whether the solvent buffer (ammonium acetate) can react and scavenge the test chemical potentially leading to false negative test results.

Table 6 lists misclassifications reported by P&G in their submission. The first part of the table contains chemicals which are mostly weakly reactive and their reactivity could only be confirmed with a method that observes the modification of the peptide formed as part of the reaction. The second part of the table lists mainly chemicals which are reactive (and would be positive in most reactivity assays), but fail to sensitise for different reasons – either they are volatile compounds and would fail to penetrate the skin sufficiently or they are effectively detoxified before they can generate any antigens via modification of the proteins.

Applicability domain of the DPRA assay is stated to exclude metals and pro-haptens, while definitive conclusions on whether or not the DPRA can reliably detect pre-haptens cannot be yet be drawn (pre-haptens are described as "not systematically detected"; page 30 of VSR). It should be noted though that, when using the assay in practice, it will be easy to exclude metal, while both pre- and pro-haptens will be difficult to identify. They are thus a potential source of uncertainty regarding negative test results (false negatives).

5.2 Quality and completeness of the documentation concerning SOPs and prediction models

General observations:

The protocol (version 1) submitted by the lead-laboratory was improved on the basis of discussions with ECVAM and, subsequently, with the participating laboratories during training and technology transfer. The modifications that were introduced were described in Table 5 of the VSR.

The resulting protocol (version 2) (Appendix 8) was used for the WLR and BLR assessment. The Working Group concluded that basic analytical experience is essential for performance of this assay. The protocol (SOP, version 2) is described adequately, but the laboratory needs to have expertise in HPLC-based analyses.

Protocol version 3 (Appendix 9) evolved from the presented study for future use is clear and detailed, but the Prediction Model for binary classifications was not described in the VSR.

The ‘expected proportion’ was based upon the application of a set of algorithms/decision rules (e.g. Fig. 3). A clarity issue arises from this figure since it tells one how to arrive at "P&G reactivity classes" but not how to get to the binary prediction (S/NS) which is used for reproducibility assessment. The information required (i.e. the actual prediction model of the test method for dichotomous predictions) is available only from papers published by Gerberick et al, but was not described in the report. At least, the publications describing the prediction model should have been referred to in the report.

6. Test materials

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

General observations:

Sample sizes for the study were based on power calculation carried out by ECVAM statisticians at the study design stage. There is a description of the calculations in Appendix 2 (pp. 2-3), but it was not completely clear to the WG (See also Section 1.4, pp 11-13).

The WG considered the number of evaluated test items sufficient to assess the transferability and reproducibility of the DPRA.

6.2 Representativeness of the test items with respect to applicability

General observations:

The main objective was to assess WLR and BLR, whereas defining the applicability domain of the test was not the purpose of this study.

On the basis of the preliminary data presented in the report, the WG concluded that the DPRA has the potential to identify chemicals that are directly reactive with the most common protein nucleophiles (lysine, cysteine) which is currently accepted to be the initiation event of an allergic response.

Since the DPRA test method does not contain a metabolic/bioactivation system, pre-haptens (i.e. chemicals requiring biochemical activation) and pro-haptens (i.e. requiring metabolic activation to become reactive) cannot be excluded from testing when applying the assay in a routine context. In addition, metals which sensitize via non-covalent coordination bonds involving also other amino acids than cysteine and lysine (e.g. nickel) fall outside the applicability domain.
Both groups of chemicals constitute a potential source of false negative results.

The WG wants to highlight that there are substances that, although non-reactive, lead to sensitisation, potentially leading to false negative predictions in the DPRA.

Specific observations:

Five compounds (nickel chloride, beryllium sulphate, 4-Phenylendiamine, R(+)-Limonene and dihydroeugenol) anticipated to be outside the applicability domain of the test were included in the list of chemicals in order to provide a consistent chemical list allowing for the comparison with other (cell-based test) that currently are under prevalidation. 4-Phenylendiamine and R(+)-Limonene were identified as reactive in the DPRA despite being outside the applicability domain as they are spontaneously oxidized.

In page 19, the report states that dihydroeugenol is a recognised pro-hapten and that it was included in the chemical selection regardless of the suggested applicability domain of DPRA to allow adequate comparison with the other two tests undergoing prevalidation.

7. Within-laboratory reproducibility (Module 2)

7.1 Assessment of repeatability and reproducibility in the same laboratory

General observations:

The main determinant of reliability assessment of the test method was the concordance of classification (S/NS), which was determined from the peptide depletion values. Furthermore the concordance of classification with regard to the four reactivity classes was considered.

The WLR was assessed with data generated with a subset of 15 chemicals tested in three independent experiments in three laboratories. The data are presented in the VSR laboratory by laboratory and the three required independent assessments of each chemical are referred to as experiment 1, experiment 2, and experiment 3.

The WLR of the test method with respect to concordance of classification was considered sufficient for the purpose of this study. The WLR of the concordance of classification with regard to the four reactivity classes could not be assessed properly due to the limited number of compounds.

Additionally, descriptive and inferential statistical analyses (ANOVA) were performed on the raw peptide depletion data. However, because of the limited number of replicates (n=3), the results of the inferential tests applied cannot be consistently interpreted and were only considered as additional descriptive information.

Specific observations:

Regarding WLR Phase B2, chemicals resulting in statistically significant variations between experiments are different among laboratories (for examples, formaldehyde at P&G (Table 10), and formaldehyde, benzyl cinnamate, and isopropanol at IVMU (Table 22)).

Acceptance criteria for the Reference control C values were not met in all instances by IVMU (see paragraph 4.1 of the WG report and Fig. 7 in VSR). Acceptance criteria were not revised during the study, the VMG however recommended to revise the acceptance range (0.45 - 0.55 mM) at a later stage (Report, p. 71).

On pages 51-52 of the report it was explained why the VMG took the decision to accept these results from IVMU despite the fact that the acceptance criterion for Reference control C was not met in two out of four runs. The VMG therefore decided to instruct the IVMU laboratory not to undertake supplementary testing in order to avoid generating additional invalid runs and instead to use the results generated for the statistical evaluations.

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

General observations:

The WLR was assessed at the level of concordance in prediction (S/NS). An average reproducibility of 87% (P&G (73,3%), RICERCA (100%), IVMU (86,7%)) met the 85% reproducibility target set by the VMG.

The definition of the reproducibility target (85%) was based upon i) the background and specific objectives of the validation study; ii) the standards of performance that can realistically be expected from an *in vitro* test and standards of performance which have been considered acceptable in previous validation studies; iii) the proposed used of the *in vitro* tests (i.e. as a partial replacement method to become part of a toolbox of tests to be used in combination); and iv) the power of the design of the validation study.

Reactivity class assignment concordance was 66,7% (P&G), 100% (RICERCA) and 73,3% (IVMU). No target was set by the VMG.

Specific observations:

The WG was intrigued by the fact that RICERCA produced a 100% WLR while the other laboratories scored systematically lower. RICERCA was the only laboratory in the study working under GLP (certified), while the other laboratories were working according to a minimum set of quality principles defined by the VMG (Appendix 1 of VSR, page 11) , which were monitored during the study.

8. Transferability (Module 3)

8.1 Quality of design and analysis of the transfer phase

General observations:

There was a training phase (6 chemicals, selected by P&G) followed by a technology transfer phase (15 chemicals, also selected by P&G) followed by discussions. Issues were identified and corrected (related to reagents and instrument configurations).

The criteria for a good transfer phase were set by the lead laboratory (P&G) and considered successful if 14/15 chemicals (Ricerca: 15/15 - IVMU: 15/15) were correctly labelled (S/NS) with 13/15 (Ricerca: 15/15 – IVMU: 15/15) being assigned the same, or no more than one above or below, reactivity class (minimal, low, moderate, high).

The SOP was slightly modified in September 2010 (version 2). Concern was expressed about the cysteine peptide decrease encountered during the transfer phase. The problem was identified to be associated with the source of acetonitrile and resolved around November or December 2010 by including in the SOP a provision to test each new batch of the solvent. . SOP Version 2 was implemented by all laboratories for the purpose of BLR assessment.

A second problem encountered by IVMU concerned the high variability of the measurements. This was found to be due to the particular instrumental setup used (injector loop) and was solved prior to the start of the blind experimental phase.

The stability issue with the cysteine depletion observed during the implementation of the DPRA to the naive laboratories was explained (Report p. 35) as a progressive decrease of the integration peak area over time of the Cys peptide in absence of any test item induced depletion reaction.

8.2 Conclusion on transferability to a naïve laboratory / naïve laboratories as assessed by the study

General observations:

The data are strong enough to support transferability to properly equipped, trained and staffed laboratories with the appropriate analytical capabilities.

IVMU did not meet the first preset criterion stipulating that all the runs had to meet the acceptance criteria defined in the SOP for assessment of success primarily because of the Reference Control C

being outside the acceptance criteria. A possible cause for failure was too stringent test acceptance criteria, but these criteria were not relaxed by the VMG at this point of the study on the basis of lack of sufficient evidence. On the advice of the lead laboratory (P&G) the transferability to IVMU was considered successful by the VMG. The causes of the difficulties to meet the criteria are still not understood. The VMG recommended in the VSR that these acceptance criteria should be relaxed in the future.

9. Between-laboratory reproducibility (Module 4)

9.1 Assessment of reproducibility in different laboratories

General observations:

The primary objective was to assess concordance in prediction (S/NS) with a target of 80%, set by the VMG based upon the same factors considered important for the WLR.

For the purpose of BLR assessment, the 15 WLR test substances were tested in triplicate experiments by each laboratory, while an additional nine materials were only tested once per laboratory. This process assessed test performance from solvent selection to prediction.

The run and data acceptance criteria were fulfilled for all runs except for Reference Control C (see also Section 7.1, pp. 19-20).

Eighteen of the 24 chemicals were consistently classified (S/NS) (75%), while 15/24 were assigned to the same reactivity class (62.5%).

It was noted that patterns of co-elution observed in this study were not consistent between the laboratories, but were reproducible within a laboratory. The WG agreed that this should not have any influence on either quality or accuracy of the data.

9.2 Conclusion on reproducibility as assessed by the study

General observations:

Eighteen of the 24 chemicals were consistently classified (S/NS) by the three laboratories resulting in a BLR reproducibility of 75%, which is below the target (80%). Nevertheless, the BLR of the test method with respect to concordance of classification was considered sufficient. This decision was based upon the observation that the reproducibility assessment included 3 chemicals (beryllium sulphate, nickel chloride and dihydroeugenol) that were considered by the VMG as outside the applicability domain of the test.

For 15 out of the 24 chemicals the laboratories assigned the same reactivity class resulting in a BLR of 62.5%. It was appreciated that reactivity classes do not necessarily correspond to potency classes, but the general feeling of the WG was that, as described in the section regarding the WLR, also the BLR results argue against the possibility to use the DPRA for potency classification as a stand-alone method.

As for the WLR the concern was expressed whether the number of chemicals included in the test was sufficient to allow a valid judgement on the BLR quality (Section 1.4, pp. 11-13).

Data variability was observed to result for chemicals with low or no reactivity. On this background, the WG considered the DPRA reproducible for testing moderately to highly reactive chemicals, but not for chemicals with limited or no reactivity.

10. Predictive capacity and overall relevance (Module 5)

10.1 Adequacy of the assessment of the predictive capacity in view of the purpose

General observations:

The primary goal of this pre-validation study was the evaluation of the transferability and reliability (within and between laboratories reproducibility) of the DPRA.

Secondary goals included a preliminary evaluation of the ability of the test to discriminate skin sensitizers from non-sensitizers, and a preliminary consideration of the ability to contribute to sub-categorization of skin sensitising chemicals (GSH sub-category 1A and 1B).

The VSR report did not present a summary of the predictive capacity based on all 24 chemicals tested, since the VMG judged three of them to fall outside the applicability domain. These were the two metals (beryllium sulphate, nickel chloride) and the pro-hapten dihydroeugenol. The two pre-haptens (4-phenylendeiamine and R(+)-Limonene) were included in the analysis as the VMG felt that there was insufficient evidence to exclude them from the AD and hence from the evaluation of the predictive performance.

In order to assess the effect of all chemicals that potentially do not fall in the applicability domain on the predictive capacity of the assay, the WG prepared two tables summarising the predictive capacity for all 24 substances (Table A) and for the 19 substances left (Table B), once the metals and pre- as well as pro-haptens have been excluded.

Table A: Predictive capacity calculated on the basis of all 24 substances tested

PARAMETER	P&G	P&G	RICERCA	IVMU	Cumulative

	historical[a]				
Sensitivity	87.0	68.7	68.7	75.0	70.8
Specificity	83.0	100	100	75.0	91.7
Accuracy	86.0	79.2	79.2	75.0	77.8

[a]Data from Table 33 p68 of the DPRA ECVAM validation study report.

Table B: Predictive capacity calculated on the basis of the 19 substances that fall within the applicability domain (i.e. excluding metals and pre- / pro-haptens)

PARAMETER	P&G historical[a]	P&G	RICERCA	IVMU	Cumulative
Sensitivity	87.0	72.7	72.7	75.0	73.5
Specificity	83.0	100	100	81.8	91.7
Accuracy	86.0	84.2	84.2	78.9	82.4

Taking into consideration the applicability domain, the overall accuracy of the DPRA obtained in this pre-validation study is consistent with the performance of the DPRA as evaluated from the results submitted to ECVAM by P&G (historical data) and with published information.

The prerequisite for obtaining the accuracy shown in Table B, when using the test in practice, would require an accurate method for identification of potential pre-/pro-haptens in order to avoid testing these.

10.2 Overall relevance (biological relevance and accuracy) of the test method in view of the purpose

General observations:

The assay itself is biologically relevant as it addresses one of the key events in the induction of sensitisation. Direct reactivity with nucleophiles is not, however, conclusively shown in all instances, as the depletion of the unchanged peptide is used as a surrogate measure for this reaction. The observed variability is prominent for chemicals with low or no reactivity which in combination with the (S/NS) prediction model cut-off creates discrepancies.

Interpretation of the data also does not take into account cysteine peptide oxidation which for some chemicals can be largely responsible for the peptide depletion observed. These points are not likely

to have an influence on the S/NS prediction, but are likely to influence the prediction of the reactivity class.

Due to the limited number of chemicals tested, the predictive capacity was not the main goal of this study. Nevertheless, by removing the five chemicals falling outside the applicability domain of the assay, the overall accuracy of DPRA (82.4%) was consistent with submitted and published information (86%).

As stated by the VMG, the data obtained did not support the possibility to use DPRA as a stand-alone assay for potency classification, which was part (b) of the secondary goal of the study. However, based on preliminary evaluation of a three-class prediction model (1A, 1B, non-classified – see figure X), the sensitivity, specificity and accuracy of the DPRA for distinguishing theses three categories appears promising. Notably the sample size is too low to draw robust conclusions.

Table C: Three-class prediction model satisfying GHS requirements on the basis of the reactivity classes assigned by the DPRA in the three laboratories (21 chemicals x 3 labs = 63 results)

Reactivity class predicted by DPRA assay (decision tree page 26 of VSR)	**Consider GHS category**	**Sensitivity (per predicted sensitiser class) and specificity (non-classified=NC)**
High	1A	**Sensitivity: 71%** (15 of 21 predictions correct, 3 under-predicted as 1B, 3 as NC)
Moderate AND Low	1B	**Sensitivity: 72%** (13 of 18 predictions correct, 5 under-predicted as NC)
Minimal	Non-classified	**Specificity 92%** (22 of 24 predictions correct, 2 over-predicted as 1B)
Accuracy	**79%**	

11. Applicability domain (Module 6)

11.1 Appropriateness of study design to conclude on applicability domain, limitations and exclusions

General observations:

Assessment/description of the applicability domain was not the objective of this study. Consequently, the small number of chemicals used in the validation study, which was set to satisfy

the primary goal of the study, is not sufficient on its own to draw robust conclusions on predictive capacity nor on applicability domain.

Empirically, the test method was set to have the following limitations with respect to the chemicals that can be tested:

- Lacking metabolic activity, the DPRA was considered not to be able to identify pre-/pro-haptens.
- The method being based on covalent binding of a chemical to a protein, metal salts (forming co-ordination bonds with specific amino acids) were also considered as falling outside the applicability domain of the method.

An additional limitation was introduced by the selection of the peptides. The DPRA employs two peptides specifically designed to contain either lysine or cysteine as the main reactive amino acid, thereby covering the majority, but not all, of reactive chemicals.

11.2 Quality of the description of applicability domain, limitations, exclusions

General observations:

The test submitters described (Report, p30) limitations of the test with respect to chemicals that were not compatible with this test. This was already addressed under 'Test Materials'.

The WG did not consider this a complete description of the applicability domain of the DPRA for the following reasons:

- The DPRA was considered not to be able to identify pre-/pro-haptens. However, some pre-haptens were reported as correctly identified (e.g. 4-phenylendiamine). Therefore it cannot be concluded with sufficient confidence whether or not these substances fall outside the applicability domain of the test method. This uncertainty may explain why the evaluation of the predicitive capacity of the DPRA by the VMG included the pre-haptens (VMG report p19).
- The DPRA targets primarily cysteine and lysine-reactive chemicals. The consequences of this selection were not addressed in the report.

12. Performance standards (Module 7)

12.1 Adequacy of the proposed Essential Test Method Components

N.A. in the present context

12.2 Adequacy of the Reference Chemicals

N.A. in the present context

13. Readiness for standardised use

13.1 Assessment of the readiness for regulatory purposes

General observations:

As outlined in the VSR and the ECVAM request for ESAC advice, the DPRA cannot be used as a standalone in a regulatory context but should be considered for use in an Integrated Testing Strategy (ITS). On the basis of the present report, especially negative outcomes have to be considered with care.

- As pre-haptens are not consistently correctly predicted by the DPRA, there remains uncertainty about whether to consider pre-haptens as part of the applicability domain of the method or not.
- Unless there are sufficiently accurate assays available identifying chemicals as pre-/pro-haptens in view of excluding them from routine testing using the DPRA, such compounds will be tested in the DPRA and may cause false negative results.
- The selection of cysteine and lysine-containing peptides selects for the majority, but not all, reactive chemicals.

Regarding reactivity class, the data obtained did not support the possibility to use DPRA as a standalone assay for potency classification. This is in agreement with the statement of the VMG that the assay should be further evaluated for its capacity to "contribute" to a potency classification (VSR page 8).

Information generated by the DPRA can be used to support regulatory decision making when used in the context of a weight-of-evidence approach or Integrated Testing Strategy (ITS). It is important to use the test in a context that allows confident conclusions about the protein-reactivity of the chemical, especially when the chemical in question is negative in the DPRA. As such the method may be helpful to address testing requirements of the REACH legislation and the 7th Amendment of the Cosmetic Directive.

Its inclusion into future integrated testing strategies can be considered for the purpose of an eventual full replacement of current *in vivo* hazard identification assays.

13.2. Assessment of the readiness for other uses

General observations:

DPRA is useful for a variety of possible **screening purposes** of chemicals expected to fall within the applicability domain of the test described in Section 11. As yet, deciding whether or not a chemical falls within the applicability domain of the test seems to be a challenge with regard to pre- and pro-haptens and for substances causing haptenation by reacting with amino acids other than lysine and cysteine. In contrast, excluding metal (salts) from routine testing (not part of the applicability domain) is not expected to cause problems in practice.

13.3 Critical aspects impacting on standardised use

General observations:

Analytical skills and appropriate facilities are important. In addition, there seem to be indications for a certified GLP environment having a positive impact on test performance.

There is an issue of co-elution. While co-elution did not affect the prediction in this study, the WG expressed concern about the potential occurrence of wrong predictions for some chemicals in the future.

Sensitizing chemicals falling outside the applicability domain of the test (e.g. pre-haptens) may or may not be identified by the DPRA.

The observed variability appears to result from chemicals with low or no reactivity suggesting that the DPRA is reproducible for testing moderately to highly reactive chemicals, but to a lesser extent for chemicals with limited or no reactivity.

One laboratory did not meet the first preset criterion for assessment of successful transferability. The causes of the difficulties to meet one of the criteria are still not understood in detail.

13.4 Gap analysis

General observations:

With respect to the predictive capacity and potency class identification, the obtained values should not be considered as more then indicative.

14. Other considerations

The the limitations of the DPRA may lead to uncertainties concerning negative results. This should be further investigated, either by additional prospective testing or through analysis of existing information.

15. Conclusions on the study

15.1 ESAC WG summary of the results and conclusions of the study

General observations:

The number of chemicals was considered as too small a sample size for allowing a firm conclusion about the predictive capacity (in terms of S/NS as well as potency classification) of the DPRA (secondary objective of the study). The preliminary results were, however, considered very promising.

The conclusions drawn by the VMG as described in the VSR on the basis of the results shown in the report:

- The WLR of the test method with respect to concordance of classification (S/NS) met the target of 85% and was considered sufficient for the purpose of this study.
- The data were considered strong enough to support transferability of the test to properly equipped, trained and staffed laboratories with the appropriate analytical capabilities.
- In spite of a BLR (75%) below the target of 80%, the BLR of the test method with respect to concordance of classification was considered sufficient. On the other hand, the BLR assessment argued against the possibility to use the DPRA for potency classification (62.5% concordance).
- The number of chemicals (N=24) did not provide support for a firm conclusion about the predicitive capacity of the test method. The preliminary data were, however, considered promising.
- The number of chemicals did not allow drawing a conclusion about the applicability domain of the test. Empirically the applicability domain seems to exclude pre-/pro-haptens and metal salts.

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

General observations:

Overall, the test design and the quality of the selected the chemicals (N=24) were considered appropriate for the purpose of addressing the first objective of the study: Assessing the WLR and BLR of the DPRA.

In agreement with the VMG statement, the number of chemicals was considered as too small a sample size for allowing a firm conclusion about the predictive capacity (in terms of S/NS as well as potency classification) of the DPRA (secondary objective of the study). The preliminary results were, however, considered very promising.

Overall, the conclusions made by the WG correspond well with the conclusions drawn by the VMG as described in the VSR, indicating that these conclusions are supported by the results shown in the report (see Section 15.1).

- The WLR was assessed using 15 chemicals in three independent experiments. The acceptance criteria were well described. The average concordance (87%) met the target (85%) set by the VMG. There were issues with respect to the reproducibility of experiments involving the control with Reference Control C (cysteine peptide). These issues are thought to be related to the stability of the peptide. No significant effect on the concordance (S/NS) was expected. Therefore, the WLR of the test method with respect to concordance of classification (S/NS) was considered sufficient for the purpose of this study.
- The data were considered strong enough to support transferability of the test to properly equipped, trained and staffed laboratories with the appropriate analytical capabilities. A note was made about too stringent test acceptance criteria resulting in difficulties to meet the Reference control C acceptance criterion in one of the laboratories. Although the causes of such difficulties are still not understood, the SOP (version 3) suggests that the acceptance criteria could be relaxed in the future.
- Eighteen of the 24 chemicals were consistently classified (S/NS) by the three laboratories resulting in a BLR reproducibility of 75%, which is below the target (80%). Nevertheless, the BLR of the test method with respect to concordance of classification was considered sufficient. This decision was based upon the observation that the reproducibility assessment included chemicals that were considered by the VMG as outside the applicability domain of the test.
- For 15 out of the 24 chemicals the laboratories assigned the same reactivity class resulting in a BLR of 62.5%. The BLR assessment argues against the possibility to use the DPRA as a stand-alone assay for potency classification.
- The number of chemicals did not allow drawing a conclusion about the applicability domain of the test (which, notably, was not one of the study objectives). Empirically the applicability domain seems to exclude pre-/pro-haptens and metal salts. However, some pre-/pro-haptens were reported as correctly identified.

With respect to the last item, the WG made the remark that chemicals that preferably react with amino acids other then cysteine and lysine may fall outside the applicability domain. In addition, some pre-/pro-haptens were reported as correctly identified. Finally, the data seem to indicate that the test method has problems identifying weak sensitizers. The uncertainty about the applicability domain may result in an unacceptable level of false negative results.

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

The DPRA was developed to represent one of the key event s of sensitization using well characterized reference data. The test results are consistent with what is known about the reactivity and sensitization potential of test chemicals (Gerberick et al., 2004, 2007).

On this background the conclusions of the report are plausible in the context of the existing information.

16. Recommendations

16.1 General recommendations

General observations:

The DPRA addresses a key mechanism (haptenation) in the development of sensitization. Overall the provided data support reproducibility and transferability of the test to qualified laboratories. The predictive capacity of the test is not defined yet, but the preliminary data profile the test as a useful tool for early-decision making during product development (screening) and as a component in a weight-of-evidence approach or integrated testing strategy for safety assessment.

The WG recommends to better define ...

(1) the predictive capacity and

(2) the applicability domain of the DPRA (to eliminate the uncertainty currently associated with a negative result)

... either through further testing (i.e. prospective validation) or through restrospective analysis of existing information (retrospective validation: data grouping / meta-analysis).

16.2 Specific recommendations (e.g. concerning improvement of SOPs)

General observations:

The WG advises to modify the validated SOP to the extent possible to address the co-elution issue before continuing with possible further validation (see 16.1).

17. References

1. Gerberick GF, Vassallo JD, Bailey RE, Chaney JG, Morrall SW, Lepoittevin JP, (2004) Development of a peptide reactivity assay for screening contact allergens. Toxicol Sci. 81; 332-43.
2. Gerberick GF, Vassallo JD, Foertsch LM, Price BB, Chaney JG, Lepoittevin JP, (2007). Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. Toxicol Sci. 97, 417-27.

18. ESAC Request concerning the current review

MANDATE OF THE ESAC WG

The EWG is requested to conduct a scientific review of the relevant studies concerning four skin sensitisation test methods (DPRA, MUSST, h-CLAT, Keratinosens). The review needs to address the questions put forward to ESAC by ECVAM.

The review should focus on the appropriateness of design and conduct of the study in view of the study objective and should provide an appraisal to which extent the conclusions of the Validation Management Team (VMT) / test method submitter are substantiated by the information generated during the study and how the information generated relates to the scientific background available.

DELIVERABLE OF THE ESAC WG

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

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

The WG is further asked to prepare a draft ESAC opinion as basis for the discussions by the entire ESAC, which shall adopt its opinion to the extent possible by consensus and on the basis of the ESAC WG report as well as all documents that were made available to the WG as well as to all ESAC members.

19. Annexes

Annex 1 – Addendum DPRA study report