ESAC PEER REVIEW PANEL FOR THE SKIN IRRITATION
VALIDATION STUDY – CONSENSUS REPORT

1. PANEL MEMBERS

<table>
<thead>
<tr>
<th>External experts</th>
<th>Internal ESAC members</th>
</tr>
</thead>
</table>
| Dr Johannes van de Sandt  
TNO Nutrition and Food Research  
Department of Explanatory Toxicology  
Utrechtseweg 48  
3700 AJ Zeist, The Netherlands  
vandesandt@voeding.tno.nl  |
| Dr Rob Roggeband  
Product Safety & Regulatory Affairs  
Procter & Gamble Eurocor NV/SA  
Temselaan 100  
1853 Strombeek-Bever, Belgium  
roggeband.r@pg.com  |
| Dr Enzo Berardesca  
Istituto Dermatologico San Gallicano  
Via Elio Chianesi 53  
00144 Roma, Italia  
berarde@unipv.it  |
| Dr Bart De Wever  
Business Development Director  
Phenion GmbH & Co.KG  
Marie Curie Strasse 9  
60439 Frankfurt, Germany  
b.deweever@phenion.uni-frankfurt.de  |
| Dr Els Adriaens  
Laboratory of Pharmaceutical Technology  
Ghent University  
Harelbekestraat, 72  
9000 Gent, Belgium  
els.adriaens@UGent.be  |
| Dr Robert Combes (Chair)  
FRAME  
Russell & Burch House  
96-98 North Sherwood Street  
Nottingham NG1 4EE, UK  
bob@frame.org.uk  |
| Dr Nathalie Alépée  
PFIZER PDRD  
MCT Laboratory  
BP 159  
F-37401 Amboise, France  
Nathalie.alepee@pfizer.com  |
| Dagmar Jírová  
National Institute of Public Health  
National Reference Centre for Cosmetics  
Šrobárova 48  
CZ-100 42 Prague 10, Czech Republic  
djirova@iol.cz  |

Page 1 of 13
2. PEER REVIEW EVALUATION CRITERIA

Criteria for evaluation

1. Goal of the study
   Clearly understandable?
   Scientific rationale given?
   Regulatory rationale?

2. Quality of overall study design
   Test methods protocols: complete?
   SOPs available, quality of the SOPs?
   Prediction model defined?

3. Study management
   Study design?
   Clearly documented planning?
   Secondary modifications of study protocols justified?

4. Test materials
   Is the selection representative?
   Is the number of chemicals sufficient?
   Were the coding and distribution independent?

5. Quality of in vivo reference data
   Source of data?
   Quality of data?

6. Within-laboratory variability
   Was the reproducibility of the data in the same laboratory properly assessed?

7. Transferability
   How easy is it to transfer the tests to a second laboratory?
   Time and cost considerations

8. Between-laboratory variability
   Was the reproducibility of experimental data in different laboratories properly assessed?
   Consider different protocols and cell lines

9. Predictive capacity
   Has the predictive capacity of the methods been properly assessed in the study?
   Are the tests relevant for their stated purpose?

10. Definition of applicability domain

11. Definition of minimum performance standards
12. Readiness for regulatory purposes

3. **BACKGROUND ON THE VALIDATION STUDY PROVIDED**
(as part of the criteria)

During 1999-2001, some of the most promising *in vitro* methods were evaluated in an ECVAM prevalidation study (1). The five tests included were: EpiDerm™, EPISKIN™, Prediskin™, the non-perfused pig ear model, and the *in vitro* mouse skin integrity function test (SIFT). The outcome of these studies was that none of the tests was ready for progression to formal validation. Various follow-up activities took place where appropriate modifications were made to enable some test protocols to meet the criteria for inclusion in a formal validation study (2-5). On the basis of additional work, the EPISKIN™, EpiDerm™ and SIFT test protocols and/or prediction models were evaluated in a formal validation study funded by ECVAM during 2003-2006 (6).

**References**


4. **OVERVIEW OF THE SIVS**

4.1 **Objective**

The overall objective of the SIVS was to develop a replacement test for the Draize Skin Irritation Test (OECD TG 404 & Method B.4 of Annex V to Directive 67/548/ECC).

4.2 **Organisation of the SIVS**

The SIVS was sponsored by ECVAM and had an independent Management Team (MT). The study was initiated in 2003 and completed toward the end of 2006. It was undertaken in two phases. In Phase 1, three test methods (EpiDerm, EPISKIN & SIFT) were evaluated, with MTT reduction as the endpoint measured. In Phase 2, the SIFT method was withdrawn leaving only 2 tests (EpiDerm & EPISKIN), but the addition of the IL-1α release endpoint. Three independent laboratories were involved in the study and all worked to GLP guidelines. A range of irritant (R38) and non-irritant test chemicals was used, selected on the basis of information mainly derived from the New Chemicals Database and the ECETOC database.

4.3 **Overall conclusion of the MT**

The Management Team of the SIVS had in 2006 agreed on three conclusions:

1. In this study, both the sensitivity and specificity of the EPISKIN assay were acceptable. The method can therefore be recommended as a replacement for the Draize skin irritation test (EU Annex V B.4; OECD TG 404).

2. For EPISKIN, the additional endpoint IL-1α has provided an improved prediction in a strategic combination with MTT and is therefore regarded as an appropriate adjunct endpoint.

3. In this study, only the specificity of the EPIDERM assay was acceptable. Therefore, the assay cannot be recommended as a replacement for the Draize skin irritation test but could be considered for use within a tiered testing strategy (e.g. as proposed in EU Annex V B.4; OECD TG 404).

5. **DOCUMENTATION PROVIDED TO THE PRP**

The following documents were provided for review:

i. A Summary Report of the SIVS.

ii. Project plan for validating the 3 methods.

iii. SOPs for the two methods and the two endpoints.

iv. A Biostatistical Report of all the data for Phases 1 and 2*

v. Several Annexes comprising: a) a report on misclassifications of irritants/non-irritants; b) proposed performance standards for human skin models; c) individual laboratory results of Phase 1 for the three tests conducted; and d) background published literature (including some publications of phase 1 results).
The following documents were received following a meeting held between the PRP and the MT for the SIVS on 13.02.07 in Berlin:

i. Performance Standards for Applying Human Skin Models to In Vitro Skin Irritation Testing [ECVAM SIS Final Version; 2007-04-13]

ii. ECVAM Skin Irritation Validation Study: Predictive Capacity for Chemical Subsets [Sebastian Hoffmann, 16/04/2007]

iii. Interleukin 1 alpha (IL-1 α); Rationale and use Role in inflammation/irritation processes; Measurement conditions specificities [Roland Roguet and José Cotovio]

iv. SOP - Episkin Skin Irritation Test- 42 hours; Determination of IL 1-1 α Concentration in the Culture medium.

v. Draft manuscript on “ECVAM International Validation Study on In Vitro Tests for Acute Skin Irritation: Selection of Test Chemicals” [Eskes et al.; sent to MT and PRP on 7/02/2007 ]

6. PRP ASSESSMENT ACCORDING TO ECVAM CRITERIA

6.1 Goal of the Study

6.1.2 Test definition

This was clearly defined for human cell based tests, although less so for SIFT.

6.1.3 Scientific rationale

This was clear, although the relationship between the MTT endpoint (a measure of cytotoxicity) and irritancy needs more explanation.

6.1.4 Regulatory rationale

This was clearly explained as being to replace an in vivo regulatory assay to discriminate between skin irritants and non-irritants.

6.2 Quality of overall study design

6.2.1 Test methods and protocols

These were generally well defined, except for IL-1α release. More guidance is needed as to what to use as a vehicle control, other than PBS.

6.2.2 SOP availability/quality

These were generally satisfactory, although some were labelled as being in draft form. The final versions of the SOPs should be submitted. Also, SOPs should be checked for clarity before being provided for general usage.

At a meeting held between the PRP and the MT for the SIVS on 13.02.07 in Berlin, it was agreed that the SOPs now need to be revised according to experience made during the SIVS. This was particularly the case for the IL-1 α method, and this has now been addressed.
6.2.3 Prediction models

These were well-defined, except more explanation is required concerning the importance of the IL-1α release endpoint to the prediction of irritancy using either method, when used in conjunction with MTT. There is, in fact, inadequate documentation for IL-1α release in the report. Also, the MT should take into account the results of recent work on SkinEthic which appear to raise doubts about the relationship between IL-1α release and morphological changes in skin models.

At a meeting held between the PRP and the MT for the SIVS on 13.02.07 in Berlin, it was agreed that the weak correlation between IL-1α-release and morphological changes upon exposure with irritants in the SkinEthic model should not be discussed and considered in connection with the SIVS. This is because histopathological scoring of morphological changes is not easy to standardise.

It was noted that the prediction models do not rely on any statistical parameters, and the PRP felt that there is a need for more justification for the cut-off values identified.

6.3 Study management

6.3.1 Study design

This was felt to be generally satisfactory with independence of the participating laboratories from each other and from the MT. However, it was noted that the data analysis was undertaken by the sponsor, ECVAM.

6.3.2 Clearly documented planning

The planning was clearly documented and there were clear protocol modifications, with the omission of a test method at the end of Phase 1, and the subsequent use of IL-1α release as an additional toxicity endpoint all clearly explained.

6.3.3 Justification for protocol changes

The protocol changes were all clearly and convincingly justified.

6.4 Test materials

6.4.1 Representative selection

The rationale for selection of test materials was detailed and clearly explained. It was noted that: a) the numbers of R38 vs. non-irritant chemicals were clearly disproportionate for IL-1α release; and b) there was a lack of information for the panel to readily assess the inclusion of different types of test materials (e.g. acids, and bases, and physically different substances, such as emulsions and volatiles).

Additional information was given on the different types of test materials in a draft manuscript on the Chemicals Selection and at the joint PRP–MT meeting held in February 2007 in Berlin.

6.4.2 Sufficient numbers
The following points were noted by the PRP: a) the numbers of non-irritants was sufficient; b) the numbers of clearly classified irritants was insufficient; and c) the number of substances with borderline classification was too high.

In the joint MT and PRP panel meeting it was acknowledged that the low proportion of clear irritants and the high proportion of borderline irritants was due to the prevalence of substances in the set of 5000 substances screened, and that the CSSC made all efforts to include as many clear irritants as possible, but only 13 were found out of 5000 substances.

6.4.3 **Independent coding and distribution**
As far as can be ascertained, the coding and distribution were independent.

6.5 **Quality of in vivo reference data**

6.5.1 **Source of data**

The reference data were mainly from recognised databases of rabbit Draize skin irritation test results. The submitted documentation contained much information on how test material selection was undertaken. However, it was noted by the PRP that: a) individual animal data for about 50% of the test materials used were lacking from the documentation supplied; and b) there appeared to have been no consideration taken of human data.

With regard to the latter point, it was explained at the joint MT and PRP meeting in Berlin that, all efforts had been made to make use of available human data.

6.5.2 **Quality of data**

The criteria for information acceptance and rejection, and whether the data were based on quantitative information, were all unclear. Also, as noted above, information on the rabbit data was lacking. However, the PRP acknowledges that these issues might be considered unimportant if prior criteria for classification used in public databases are deemed acceptable.

At a meeting held between the PRP and the MT for the SIVS on 13.02.07 in Berlin, it was explained that a total of approximately 3500 New Chemicals from the New Chemicals Database (NCD) of the European Chemicals Bureau (ECB/IHCP/JRC/Ispra) and about 1500 existing chemicals from alternative databases (TSCA, CIR and the ECETOC database) were screened by the Chemical Selection Sub-Committee (CSSC) for the SIVS. Only few chemicals were found to fulfil the CSSC selection criteria. Particular difficulties were observed for identifying GHS category 2 chemicals. Finally, a total set of 60 chemicals were selected and proposed to be tested in the SIVS. However, five of the existing chemicals (source: ECETOC) had been tested already in studies preceding the SIVS, and two of the New Chemicals finally had to be excluded from the SIVS assessment, because one company did not agree on disclosure of chemical information on the chemicals. This left 58 chemicals to be used in the SIVS divided into 25 Existing Chemicals and 33 New Chemicals. As the in vivo information available for the latter group is restricted to dominant median scores, irritancy classifications, and whether this was based on erythema or oedema. As such, it was noted that the absence of raw data did not fully allow for an in depth comparison of *in vitro* and *in vivo* data. In particular, an assessment of reversibility,
the main parameter that determines classification of a test material as irritant, was not possible.

To overcome this problem, it was suggested that analyse, publish and discuss in depth the data of the 25 Existing Chemicals (ECETOC and TSCA) as the “primary data set” separately from the data of the 33 New Chemicals (NCD - to be considered as the “supporting data set”).

The PRP agrees with this suggestion, but requires that the validation dossier should contain a re-analysis of the data on the two separate data sets. This has now been done by Sebastian Hoffmann, on behalf of the MT, and the results are presented below.

<table>
<thead>
<tr>
<th></th>
<th>EpiDerm</th>
<th>EPISKIN (MTT)</th>
<th>EPISKIN (MTT+IL1-α)</th>
</tr>
</thead>
<tbody>
<tr>
<td>specificity</td>
<td>87.5%</td>
<td>79.2%</td>
<td>83.8%</td>
</tr>
<tr>
<td>sensitivity</td>
<td>33.3%</td>
<td>74.1%</td>
<td>88.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECETOC + TSCA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCD</td>
<td>80.4%</td>
<td>82.4%</td>
<td>74.5%</td>
</tr>
<tr>
<td>overall</td>
<td>83.8%</td>
<td>80.8%</td>
<td>78.8%</td>
</tr>
</tbody>
</table>

The overall predictivity of the assay differs little whether the data are split or combined, except for the slightly lower specificity obtained for the new chemicals. However, it should be remembered that, should the validation status of the assay be judged on only existing data set (due to these chemicals having raw data available), then the overall number of chemicals used for the assessment is more than halved, and this could have implications for the applicability domain.

### 6.6 Within-laboratory variability

The PRP made the following observations concerning intra-laboratory variability: a) this was carefully assessed; b) there are some concerns about the statistical methods used, especially the adoption of a 1% significance level (with concomitant reduction in statistical power); c) the variability of EpiDerm was rather high; d) the variability for EPISKIN was generally satisfactory, except for one laboratory (Sanofi) with the positive control; and e) several issues in the biostatistics report need addressing.

Point (b) above was satisfactorily explained at the joint MT and PRP panel meeting in Berlin.

### 6.7 Transferability

Transferability was assessed satisfactorily, and the documentation contained sufficient information for effective training of personnel to use the MTT test method to be transferred readily. However, the PRP raised some issues with: a) transferability of the IL-1α release method suggesting that it might be more difficult to transfer (e.g. the variability of the negative control was significant between laboratories); and b) performance standards that need to be addressed to facilitate a naïve laboratory to assess its results for the first time.

### 6.8 Between-laboratory variability
The data were analysed in a variety of ways, and the PRP had some reservations about using 1% level (as for the analysis of intra-laboratory variability), despite explanations provided at the joint MT and PRP panel meeting in Berlin. The results for EPISKIN were somewhat better than for EpiDerm (overall 86.2% vs. 74.1%). However, one reviewer noted important interlaboratory differences for the predictions of the same chemicals in different laboratories, and felt that these need to be discussed with the MT. It was also noted by another reviewer that some of the median classifications for rabbit data relating to the reference chemicals used could have led to a distortion in the assessment of inter-laboratory variability.
6.9  **Assessment of predictive capacity**

The sensitivity, specificity and overall accuracy of the methods were all assessed adequately. The sensitivity of EpiDerm was low (60%) whereas its specificity was high (89%) using MTT. The overall predictivity of the test method was not enhanced when IL-1α release was used. The sensitivity of EPISKIN was 77.6% while its specificity was 80.7% using MTT. When IL-1α release was used; only sensitivity was enhanced. The PRP noted: a) some concerns about the interpretation of IL-1α release data using the recommended threshold level for this; and b) both toxicity endpoints would be needed to give acceptable predictivity when using EPISKIN.

At a meeting held between the PRP and the MT for the SIVS on 13.02.07 in Berlin, it was acknowledged that there are insufficient data to conclude that use of IL-1α is valid. It was, therefore, proposed that the IL-1α endpoint should be recommended as a useful adjunct to the MTT assay endpoint, as it has the potential to increase sensitivity of the test, without reducing its specificity. This is supported now with the addition of a substantial document on the background to the rationale for using this endpoint, as well as for how it is measured.

The PRP accepts this suggestion.

6.10  **Assays fit for purpose**

The PRP concluded that EpiDerm is not fit for its intended purpose since its predictivity was too low. However, the EPISKIN test method is probably fit for purpose, but: a) the interlaboratory discrepancies in false positives and negatives need to be addressed; and b) the purpose of the assay needs to be carefully defined with respect to concerns raised about the test materials used.

As a result of discussions with the MT, and receipt of new and revised documentation, the PRP is now satisfied that the EPISKIN test method is now fit for purpose.

6.11  **Applicability domain**

The PRP felt that insufficient information was given regarding the applicability domain of the test methods. In particular, any constraints due to the selection of test materials need to be discussed, including the fact that the AD for volatiles, emulsions, mixtures, hydrolysing and polymerising chemicals, acids and bases needs to be discussed. The PRP noted that the report on an analysis of the mis-classified chemicals provided in no simple explanations. Lastly, there should be some guidance on any chemicals that proved difficult to test.

At a meeting held between the PRP and the MT for the SIVS on 13.02.07 in Berlin, it was noted that the definition of an AD is complicated by the fact that many of the SIVS chemicals (the New Chemicals) did not fit into established grouping schemes. To alleviate this situation, the possibility of using relevant physicochemical descriptors (e.g. pH, molecular weight, melting point) was discussed. It was also noted that similar difficulties were encountered with the Skin Corrosion Validation Study. Given the wide diversity of the chemicals included in the SIVS, and taking note of those physicochemical properties not covered (as mentioned above), the PRP
agreed to assume that the AD is potentially wide, but that this aspect of the validation study should be continually revised according to results obtained from using the test. In addition, splitting the reference chemicals into two sets (existing and new) could have implications with regard to the AD (see above).

6.12 Minimum performance standards

The PRP felt that there should be standards identified for: a) barrier function; b) reproducibility; and c) predictivity. In addition, more guidance is needed on the responses expected with the suggested positive & negative controls. Concerns were expressed about suggesting reference controls that: a) had only been tested in one laboratory (e.g. heptanol); b) were of borderline activity; c) are negative for human skin irritancy (e.g. α-terpineol); d) were unstable (e.g. 1–bromohexane). The PRP questions whether such controls are acceptable.

The PRP also made the following observations: a) the performance standards for EpiSkin are very specific, which would complicate a later catch-up validation study; b) the 15 mins exposure time may not be optimal; c) the performance standards seem to be clearer for MTT than for IL-1α; d) a lack of information on the use of vehicle controls other than PBS; e) the paucity of positive controls specified for IL-1α; and f) the ambiguity of whether the IL-1α endpoint can be used alone as a criterion.

At a meeting held between the PRP and the MT for the SIVS on 13.02.07 in Berlin, it was explained that the minimum performance measures here were not pre-defined by the MT, but were developed during the study based on analyses of the in vivo data performed by ICCVAM and ECVAM. According to these, a 70-80% balanced sensitivity and specificity was regarded acceptable. HK commented that the lower boundary of this range would also be reached with EpiDerm, if MTT and IL-1α endpoints were combined.

It was further agreed that the draft performance standards documentation required revision since it is too constraining. Until this has been completed, the PRP is unable to make any further comments as to Performance Standards. A revised version of the PF document has now been received and this is much improved, covering most of the original concerns of the PRP, particularly the need for standards for barrier function and differentiation level of the skin model, as well as for the IL-1α release methodology. However, it still includes some recommended positive control reference chemicals, which are not supported by the PRP due to reasons given above. Also, the PS document refers to both Epiderm and Episkin. The PRP is concerned that this could be confusing and that reference to EpiDerm should only be made once the validation of the test has been deemed satisfactory.

6.13 Readiness for regulatory usage

The PRP felt that the EPISKIN protocol involving MTT reduction is the nearest to being ready to use. However, this conclusion is dependent on addressing several issues concerning predictivity, applicability domain and performance standards. In addition, it is considered that it would be preferable if the method could also be used with IL-1α release, although issues related to reproducibility and relevance need to be addressed before this would be feasible.
As a result of discussions with the MT, and receipt of new and revised documentation, the PRP is now satisfied that the EPISKIN test method is now ready for regulatory usage.

7. MAIN CONCLUSIONS OF THE PRP

* Of the 3 methods originally considered for validation, only EPISKIN is suitable as a replacement method at present.

* The MT should require participating laboratories to produce Phase 2 reports. At a meeting held between the PRP and the MT for the SIVS on 13.02.07 in Berlin, it was explained that in the final phases (pure blind trials) of all previous successful validation studies full written reports have never been produced by participating laboratories, because laboratories are unable to interpret data at that stage. Instead, laboratories had submitted test data in electronic format, plus, in a tabulated manner, observations and comments on each of the test chemicals. This information was provided for EPISKIN and EPIDERM. In addition, since tests were conducted under GLP conditions, all additional information (e.g. incubator and fridge temperatures, pipette calibrations) are documented and can be subjected to audits. The PRP accepts these explanations.

* The documents received by the PRP since the Berlin meeting, show that the MT has addressed nearly all the concerns expressed earlier by the PRP. It is therefore now possible to produce a draft ESAC statement.
8. APPENDIX – SUMMARY OF INDIVIDUAL PRP MEMBERS REPORTS

Table 1 – Summary of PRP reviewers comments on the SIVS (Y = important concern; N = no important concern)

<table>
<thead>
<tr>
<th>Criterion number</th>
<th>Reviewer number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>8</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>10</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>