The ECVAM International Validation Study on *In Vitro* Tests for Acute Skin Irritation: Report on the Validity of the EPISKIN and EpiDerm Assays and on the Skin Integrity Function Test[a]

Horst Spielmann,[1] Sebastian Hoffmann,[2] Manfred Liebsch,[1] Phil Botham,[3] Julia H. Fentem,[4] Chantra Eskes,[2] Roland Roguet,[5] José Cotovio,[5] Thomas Cole,[6] Andrew Worth,[6] Jon Heylings,[3] Penny Jones,[4] Catherine Robles,[7] Helena Kandárová,[1] Armin Gamer,[8] Marina Remmele,[8] Rodger Curren,[9] Hans Raabe,[9] Amanda Cockshott,[10] Ingrid Gerner[11] and Valérie Zuang[2]

[1]National Centre for Alternative Methods (ZEBET), Berlin, Germany; [2]ECVAM, Institute for Health & Consumer Protection, European Commission Joint Research Centre, Ispra, Italy; [3]Syngenta, Macclesfield, UK; [4]Unilever, Sharnbrook, UK; [5]L'Oréal, Clichy, France; [6]ECB, Institute for Health & Consumer Protection, European Commission Joint Research Centre, Ispra, Italy; [7]Sanofi Aventis, Montpellier, France; [8]BASF, Ludwigshafen, German; [9]Institute for In Vitro Sciences, Gaithersburg, MD, USA; [10]Health and Safety Executive (HSE), Bootle, UK; [11]Federal Institute for Risk Assessment (BfR), Berlin, Germany

Summary — ECVAM sponsored a formal validation study on three *in vitro* tests for skin irritation, of which two employ reconstituted human epidermis models (EPISKIN™, EpiDerm™), and one, the skin integrity function test (SIFT), employs *ex vivo* mouse skin. The goal of the study was to assess whether the *in vitro* tests would correctly predict *in vivo* classifications according to the EU classification scheme, "R38" and "no label" (i.e. non-irritant). 58 chemicals (25 irritants and 33 non-irritants) were tested, having been selected to give broad coverage of physico–chemical properties, and an adequate distribution of irritancy scores derived from *in vivo* rabbit skin irritation tests. In Phase 1, 20 of these chemicals (9 irritants and 11 non-irritants) were tested with coded identities by a single lead laboratory for each of the methods, to confirm the suitability of the protocol improvements introduced after a prevalidation phase. When cell viability (evaluated by the MTT reduction test) was used as the endpoint, the predictive ability of both EpiDerm and EPISKIN was considered sufficient to justify their progression to Phase 2, while the predictive ability of the SIFT was judged to be inadequate. Since both the reconstituted skin models provided false predictions around the *in vivo* classification border (a rabbit Draize test score of 2), the release of a cytokine, interleukin-1α (IL-1α), was also determined. In Phase 2, each human skin model was tested in three laboratories, with 58 chemicals. The main endpoint measured for both EpiDerm and EPISKIN was cell viability. In samples from chemicals which gave MTT assay results above the threshold of 50% viability, IL-1α release was also measured, to determine whether the additional endpoint would improve the predictive ability of the tests. For EPISKIN, the sensitivity was 75% and the specificity was 81% (MTT assay only); with the combination of the MTT and IL-1α assays, the sensitivity increased to 91%, with a specificity of 79%. For EpiDerm, the sensitivity was 57% and the specificity was 85% (MTT assay only), while the predictive capacity of EpiDerm was not improved by the measurement of IL-1α release. Following independent peer review, in April 2007 the ECVAM Scientific Advisory Committee endorsed the scientific validity of the EPISKIN test as a replacement for the rabbit skin irritation method, and of the EpiDerm method for identifying skin irritants as part of a tiered testing strategy. This new alternative approach will probably be the first use of *in vitro* toxicity testing to replace the Draize rabbit skin irritation test in Europe and internationally, since, in the very near future, new EU and OECD Test Guidelines will be proposed for regulatory acceptance.

Key words: *Draize eye test, ECVAM, EpiDerm, EPISKIN, human epidermis model, mouse skin, replacement alternative, SIFT, skin irritation, test guidelines. validation.*

Address for correspondence: *Horst Spielmann, c/o Federal Institute for Risk Assessment (BfR), National Centre for Alternative Methods (ZEBET), 12277 Berlin, Germany.*
E-mail: *horst.spielmann@bfr.bund.de*

[a]*This report reflects the experience of the authors as individual scientists engaged in the study, and not as representatives of the institutions to which they are affiliated.*

1. Introduction

1.1. The background to the ECVAM Skin Irritation Validation Study (SIVS)

The chronology and conduct of the ECVAM SIVS study are summarised in Table 1 and in Figure 1.

In 1998, the ECVAM Skin Irritation Task Force (TF) published a report on the status of *in vitro* skin irritation testing, and proposed 10 "challenge chemicals", for which promising, concordant *in vivo* data from the rabbit test, *in vivo* data from the 4-hour human patch test, and *in vitro* data from the human skin model, EpiDerm™, were available (1). The proponents of new *in vitro* test systems were encouraged to submit data obtained with optimised *in vitro* skin irritation test protocols for the same set of chemicals, to permit an assessment of whether these tests could be considered to be ready to enter an ECVAM prevalidation study. At the same time, the suitability of various endpoints for predicting human skin irritation was evaluated in an EU 4th Framework Programme collaborative project on several human reconstructed skin models, which revealed that the reduction of cell viability (evaluated by the MTT reduction assay) and interleukin-1α (IL-1α) release, were the most promising endpoints. Since the results for cell viability and IL-1α release showed high *in vitro*/*in vivo* correlations, and since IL-1α release was the more variable endpoint, cell viability (MTT reduction) was proposed as the most promising endpoint for use with the human skin models (2).

Among the test systems for which data were submitted to the ECVAM TF, five tests were judged to be sufficiently promising to enter the ECVAM prevalidation study, namely, a perfused pig-ear test, the skin integrity function test (SIFT), a test using human skin (Prediskin™), and two reconstructed human skin models (EPISKIN™ and EpiDerm). During the prevalidation study, two tests (the pig ear test and Prediskin) failed, due to insufficient reproducibility, whereas the three other tests (SIFT, EPISKIN and EpiDerm) showed sufficient intra-laboratory and inter-laboratory reproducibility, but their predictive abilities were insufficient for correctly predicting the skin irritation potentials of the 20 chemicals that were tested in the prevalidation study (3). The Management Team (MT) of the study therefore proposed that refinement and optimisation of the three tests was necessary, before they entered a formal ECVAM validation study.

In 2001, the ECVAM TF and the laboratories responsible for the refinement of the three tests, discussed ways of approaching a formal validation study. Since a *post hoc* analysis of the prevalidation data from the MTT reduction assay for EPISKIN and EpiDerm revealed similar sensitivities, it was recommended that a common test protocol for the two skin models should be developed, before a formal validation study was undertaken (4).

In 2002, the TF analysed the improvements which had been made to the SIFT and the skin model test protocols, and recommended to ECVAM that a formal validation study should now be undertaken. However, since all the refinements were made by using the 20 chemicals that had already been used in the prevalidation study, the TF recommended that the ECVAM Skin Irritation Validation Study (SIVS) should be performed in two phases: in Phase 1, the protocol refinements would be confirmed by the lead laboratories for the three methods, i.e. Syngenta (SIFT), L'Oréal (EPISKIN), and ZEBET (EpiDerm), by the ring testing of another set of coded chemicals. If the outcome of Phase 1 was sufficiently promising, the SIVS should proceed to Phase 2, which would involve extending the ring trial to include two additional laboratories for each skin model.

The EPISKIN test was further refined by L'Oréal, by extending the incubation period for the tissues (after a 15-minute exposure to the test chemical) to 42 hours, which permitted recovery from weak effects and also allowed significant effects to increase in severity. L'Oréal and ZEBET collaborated in the development of a common test protocol for the EPISKIN and EpiDerm models, for use in the SIVS. The data obtained with the two skin models, when the optimised common protocol was applied, were very promising, and were published back-to-back in 2005 (5, 6).

In 2003, an ECVAM stakeholder workshop confirmed the recommendation that a formal validation study should be conducted, and further recommended that it should focus on the prediction of skin irritancy according the EU classification system (i.e. R38 *versus* no label; 7), since the tests to be validated had been developed and optimised in relation to this system.

1.2. The goals and objectives of the SIVS

The SIVS MT defined the goals of the study, as follows:

1. The overall aim of the study was to validate *in vitro* skin irritation tests in a formal inter-laboratory study, in order to replace the Draize skin irritation test performed on rabbits according to Method B.4 of Annex V to the EU Dangerous Substances Directive, *Directive 67/548/EEC* (7) or OECD Test Guideline (TG) 404 (8).

2. The primary goal of the validation study was a scientific evaluation of the ability of the *in vitro* tests to reliably discriminate skin irritants (I) from non-irritants (NI), as defined by the EU

Table 1: Chronology and management of the ECVAM Skin Irritation Validation Study

2001	**ECVAM Skin Irritation TF Meeting**: decision on the way forward after finalisation of ECVAM Skin Irritation Prevalidation Study (see report of Zuang *et al*. [4])
2002	**Test optimisation**: EPISKIN and SIFT (L'Oréal, Syngenta)
	ECVAM TF Meeting: recommendation of a two-phased validation study
2003	**ECVAM Stakeholder Workshop**: positive decision for start of skin irritation validation study (SIVS)
	EC/JRC call for tenders on conduct and management of SIVS
	Tender of main contractor, BfR (Berlin, Germany), accepted, including 9 partner institutes
	Development of optimised, common test protocol for EpiDerm and EPISKIN models (L'Oréal and ZEBET)
	1st SIVS MT Meeting: report on common skin model protocol; decision to test further chemicals at L'Oréal and ZEBET; agreement on study goal: prediction of EU classification system; agreement that chemical selection should equally represent the three categories of the GHS classification system
	Contract signed between JRC and BfR
	2nd SIVS MT Meeting: agreement on project plan provided by the contractor BfR and subcontracts with Syngenta and L'Oréal
2004	**1st MT Teleconference**: final agreement on BfR project plan; report on status of CSSC chemicals selection
	2nd MT Teleconference: for Phase 1: approval of 20 test chemicals, SOPs (for SIFT, EpiDerm & EPISKIN) and data spreadsheets. Decision of GLP/GMP audits by BfR at the skin model production facilities
	Distribution of 20 test chemicals each, to L'Oréal, Syngenta and ZEBET
	GLP/GMP Audits: at MatTek Corporation (Ashland, MA, USA) and EPISKIN SNC (Lyon, France)
	Phase 1 testing (20 chemicals, three times) at the lead laboratories, L'Oréal, Syngenta and ZEBET, and **IL-1α-determination** by L'Oréal for both skin model systems
	3rd MT Meeting: discussion of results of Phase 1 and audits of skin model production facilities. **Conclusion**: overall performance of EpiDerm and EPISKIN sufficiently promising for progress to Phase 2. SIFT needs further investigation of test limitations. Chemicals selection: re-use of 19–20 chemicals of Phase 1, plus 40 new chemicals
	Training of laboratories for Phase 2 (EPISKIN and EpiDerm)
	3rd MT Teleconference: approval of: training reports, coded biostatistical analysis of Phase 1 (incl. IL-1α data), and chemicals selection
	4th MT Teleconference: agreement on IL-1α endpoint inclusion (tiered approach: all laboratories to keep frozen media; lead laboratories to conduct testing, and, if results are promising, all laboratories will determine IL-1α)
	1st distribution of 30 chemicals for Phase 2, to six laboratories
	Start of Phase 2 testing: L'Oréal, Sanofi, Unilever, ZEBET, IIVS, BASF
2005	**2nd distribution of 30 chemicals** for Phase 2, to six laboratories
	IL-1α-determination training for ZEBET by L'Oréal
	IL-1α-determination by L'Oréal (EPISKIN) and ZEBET (EpiDerm)
	Submission of Phase 2 data: lead laboratories: MTT + IL-1α; other laboratories: MTT only
	4th MT Meeting: discussion of preliminary Phase 2 analysis: EpiDerm: MTT not sensitive enough (special study at ZEBET with extended exposure) and IL-1α not promising. EPISKIN: MTT more balanced prediction and IL-1α is promising. Thus, testing of frozen samples by Sanofi and Unilever is required
	IL-1α training and testing by EPISKIN laboratories 2 and 3 (Sanofi and Unilever)
	Submission of IL-1α data by EPISKIN laboratories 2 and 3 (Sanofi and Unilever)
2006	**5th MT Meeting**: discussion of 1st drafts of ECVAM Biostatistical Report and CSSC Report on misclassifications. ANOVA deemed not adequate. Data retrieval at ECVAM needs independent audit by BfR, data used need approval by testing laboratories
	Data Quality Control by laboratories, both for MTT and IL-1α
	5th MT Teleconference on actions needed for finalisation of the study (e.g. communication with EpiDerm laboratories 2 and 3 about quality of individual determinations)

MT = Management Team; TF = Task Force; SIFT = Skin Integrity Function Test; CSSC = Chemicals Selection Sub-Committee.

Table 1: continued

2006	**6th MT Meeting**: Conclusion of the MT, based on 3rd version of Biostatistical Report: both tests sufficiently reproducible. Because of balanced predictivity, EPISKIN validated as a stand alone replacement test based on the MTT assay; IL-1α assay recommended as a complementary endpoint, to increase the sensitivity of the method. Because of high specificity and low sensitivity, EpiDerm recognised as a validated assay usable in tiered strategies, if the sensitivity cannot be improved (see ESAC Statement [12])
	6th MT Teleconference: agreement on study communication and actions needed for submission of documents that permit ECVAM Peer Review

MT = Management Team; TF = Task Force; SIFT = Skin Integrity Function Test; CSSC = Chemicals Selection Sub-Committee.

risk phrases, "R38" and "no label", according to *Directive 67/548/EEC* (7).

3. A secondary goal of the study was a retrospective analysis of the data obtained in the study, in order to evaluate the ability of the *in vitro* tests to reliably discriminate skin irritants from mild-irritants and from non-irritants, as defined by the Globally Harmonised System (GHS) for classification and labelling, adopted by the United Nations (9).

In addition, the MT defined the objectives of the study, as follows. Two different kinds of *in vitro* test system, one employing reconstituted human epidermis models (EPISKIN, EpiDerm), and the other employing *ex vivo* mouse skin (SIFT), having progressed satisfactorily through prevalidation and test optimisation, were considered to be ready to enter a formal validation study. Therefore, the objective was to assess the relevance (predictive ability) and the reliability (reproducibility within and between laboratories) of these test systems with a larger set of coded test chemicals, for which high quality *in vivo* data were available. The validation study was undertaken in accordance with the principles and criteria documented in the draft OECD *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment* (10).

2. The Management and Conduct of the SIVS

2.1. Management

After a call for tenders for the SIVS was issued by the European Commission in June 2003, an offer was submitted by the BfR in July 2003, and the BfR received the management contract in November 2003. The study started formally with the 1st Meeting of the SIVS MT (Table 2) on 17–18 November 2003, and with discussion and approval of a project plan drafted by the BfR. During the course of the SIVS (11/2003 to 6/2006), six face-to-face MT meetings and six MT teleconferences were held (Table 1).

The test chemicals were selected by an independent Chemical Selection Sub-Committee (CSSC). To ensure the high quality of the commercially-produced human skin models, the facilities of the producers of EPISKIN (L'Oréal, France) and EpiDerm (MatTek, USA) were evaluated by independent auditors at the beginning of the SIVS. The biostatistical analysis of the *in vitro* experimental data was the responsibility of an independent biostatistician, at ECVAM. The coding and shipment of the selected chemicals was sub-contracted to an unaffiliated intermediary, RCC–CCR (Research and Consulting Company–Cytotest Cell Research GmbH, Rossdorf, Germany).

The following decisions of the MT had a significant impact on the management and conduct of the SIVS:

1. to audit the skin model production facilities, in order to address requirements of the OECD Guidance Document (10);

2. not to proceed with the SIFT after completion of Phase 1, due to its insufficient predictive power; and

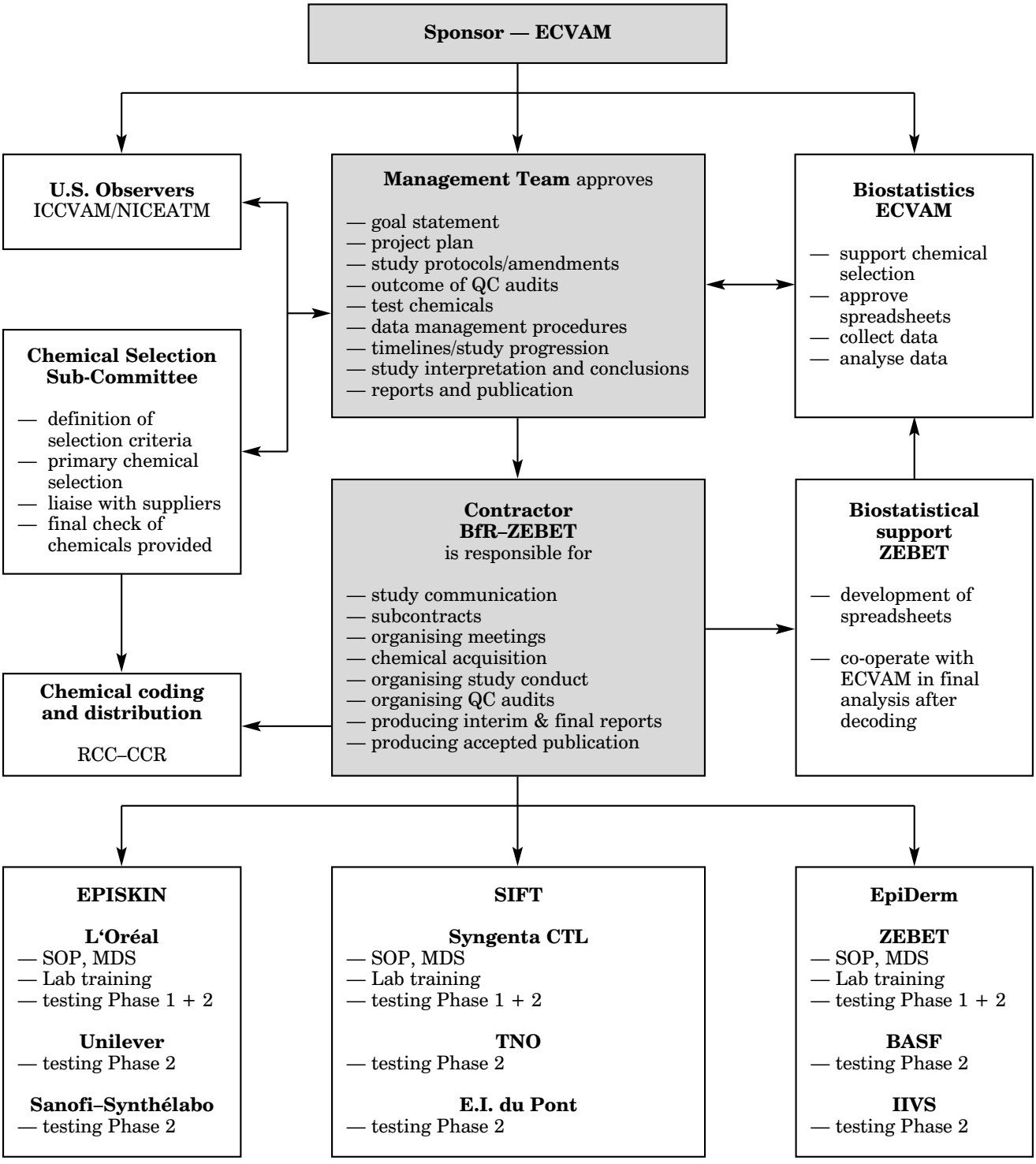
3. to allow for the inclusion of a second endpoint (IL-1α release) in the skin model tests in a tiered manner; hence, all the laboratories were required to collected and freeze the culture medium of each treated tissue, so that, if it was decided that this endpoint could provide an improvement of the predictive power, all the laboratories could be asked to determine the IL-1α contents of their frozen samples.

2.2. Phase 1 and Phase 2 of the SIVS

In Phase 1 of the SIVS, 20 coded chemicals (9 irritant, 11 non-irritant), selected on the basis of information

in the New Chemicals Database (NCD) of the European Chemicals Bureau (ECB), backed by quality-assured *in vivo* rabbit skin irritation data, were tested by the lead laboratories (EPISKIN: L'Oréal; EpiDerm: ZEBET; SIFT: Syngenta). The methods applied (with Standard Operating Procedures, SOPs) were the refined, optimised protocols developed following the ECVAM prevalidation study. When cell

Figure 1: Management structure of the ECVAM Skin Irritation Validation Study



SOP = Standard Operating Procedure; MDS = Method Documentation Sheet.

Table 2: Management Team (MT) and Chemicals Selection Sub-Committee (CSSC)

Management Team (MT)	
Chair	Dr Phil Botham
Co-chair	Dr Julia Fentem
Sponsor representatives	Dr Valérie Zuang and Dr Chantra Eskes
Independent biostatistician	Sebastian Hoffmann
Representatives of the test systems	
EpiDerm	Dr Manfred Liebsch
EPISKIN	Dr Roland Roguet
SIFT	Dr Jon Heylings
Chair of the Chemicals Selection Sub-committee (CSSC)	Dr Andrew Worth
Representative of the main contractor (BfR)	Dr Horst Spielmann
ECB customer	Dr Tom Cole
Two observers from ICCVAM (USA)	
ICCVAM	(Dr Karen Hamernik; *alternate*: Dr Abby Jacobs)
NICEATM	(Dr William Stokes; *alternate*: Dr Ray Tice)
Chemicals Selection Sub-Committee (CSSC)	
Chair	Dr Andrew Worth
Representatives of ECVAM	Dr Chantra Eskes and Dr Valérie Zuang
Representative of ECB	(Dr Tom Cole)
Representatives from Competent Authorities	
BfR, Germany	Dr Ingrid Gerner
HSE, UK	Dr Amanda Cockshott

viability (MTT reduction) was used as the endpoint, the two skin models met the acceptance criteria set by the MT of the study for within-laboratory reproducibility, since identical predictions were obtained in each independent test run with the same chemical with both models. When the prediction model (PM) was applied, whereby chemicals were classified as irritants when reducing cell viability to 50% or below, and otherwise as non-irritants, the predictive performances of the tests met the criteria set by the MT for accuracy (EpiDerm 75%, EPISKIN 80%), sensitivity (EpiDerm 56%, EPISKIN 67%) and specificity (EpiDerm 91%, EPISKIN 91%). For both skin models, false predictions were only obtained around the *in vivo* classification threshold, i.e. a dominant median score of 2. In contrast, the SIFT test did not meet the acceptance criteria set by the MT.

The results of Phase 1 of the SIVS indicated that false negative results were the major problem, when the MTT reduction protocol was applied. Therefore, the measurement of IL-1α release (11), which had been investigated by L'Oréal for the EPISKIN model (5), was added as an additional endpoint for both skin models. Briefly, in samples providing negative MTT results (chemicals classified as non-irritant), the release of IL-1α was determined, in order to investigate whether the use of the additional endpoint would increase the sensitivities of the two assays. Due to the encouraging results obtained with this second endpoint, the MT decided to apply this improved testing strategy in Phase 2 of the SIVS.

In Phase 2, 58 coded test chemicals (18 from Phase 1, and 40 chemicals selected by the CSSC, including both new chemicals from the NCD and existing chemicals) were tested with the two human skin models. Each chemical was tested with the MTT test in each laboratory, on three parallel tissue replicates per test in three independent runs, and test medium samples were frozen to allow for the subsequent determination of IL-1α.

After the necessary training and the successful transfer of the methods to two additional laboratories for each test, the EpiDerm and EPISKIN tests were each performed in three laboratories (Table 3). The EpiDerm test was conducted by ZEBET (lead laboratory), the IIVS and BASF, and the EPISKIN test was conducted by L'Oréal (lead laboratory), Unilever and Sanofi–Synthélabo. The prediction model (PM) applied in the formal validation study, employed data from the following endpoints: MTT reduction (threshold of 50% reduction of cell viability); IL-1α release (threshold of 60pg/ml). The PM for 1L-1α release, developed by L'Oréal (5), was improved by taking into account the results of the formal validation study.

Since the IL-1α release protocol of the EpiDerm test was introduced rather late in the study, there

was not sufficient time to allow for optimisation of the protocol. When IL-1α release was determined in the lead laboratory for the EpiDerm test, it did not contribute to an improvement in the predictive capacity of the EpiDerm test. Therefore, IL-1α release was not analysed in test samples from the other two laboratories conducting the EpiDerm test.

The SIVS, as outlined in Table 1, was conducted between 2003 and 2006. In 2006, the MT submitted a summary report to ECVAM, the sponsor of the study. After an evaluation of the results of the study by an independent peer review panel, in April 2007, the ECVAM Scientific Advisory Committee (ESAC) endorsed the scientific validity of the EPISKIN test as a replacement for the rabbit skin irritation method, and of the EpiDerm method for identifying skin irritants as part of a tiered testing strategy (12).

3. Materials and Methods

3.1. The skin model tests

The skin models used in the SIVS, i.e. EpiDerm and EPISKIN, are commercially-available reconstituted human epidermis models. To assess skin irritation, cell viability was determined with the MTT reduction assay (13), according to the common protocol developed by the lead laboratories for the two methods (5, 6). Essentially, the incubation period of the tissues after a 15-minute exposure to a test chemical was extended to 42 hours, which allowed recovery from weak effects, whilst also permitting significant effects to increase in severity. Thus, identical data analysis procedures could be applied to data obtained with the two skin models.

Since validation of the SIFT model did not proceed to Phase 2, further details of this assay are not presented in this report.

3.2. The endpoints and PMs

In view of the results from Phase 2, the MT decided to add 1L-1α determination to the MTT protocol of the two human skin models. This decision was based on the fact that L'Oréal had applied a promising protocol for the second endpoint (11), in which the release of the cytokine IL-1α into the EPISKIN assay medium, was determined in test samples giving a negative MTT result (5). For epidermal tissues showing a cell viability of greater than 50%, the amount of IL-1α released into the tissue culture medium by the end of the incubation period (i.e. after 42 hours of incubation) was measured in the medium (immediately, or after being stored frozen) by using ELISA kit DLA 50 from R&D Systems (11). Since strongly-irritating chemicals may affect the measurement of IL1-α release, the MT decided that this endpoint should only be measured in substances giving a non-irritant result in the MTT test, which is characterised by a cell viability of greater than 50%. Thus, according to the PM, a test substance is considered to be an irritant, if the viability after 15 minutes of exposure and 42 hours of further incubation is: a) lower than 50%; or b) higher than 50%, and the amount of IL-1α release is increased by more than three-fold (EpiDerm) or more than five-fold (EPISKIN), when compared to the negative control. Conversely, a test substance is considered to be non-irritant to skin, if the viability after 15 minutes of exposure and 42 hours of further incubation is higher than 50%, and the amount of IL-1α release is increased by less than three-fold (EpiDerm) or less than five-fold (EPISKIN), in comparison to the negative control.

Taking into account the promising results from L'Oréal (5), a tiered testing strategy was supported by the MT, in which MTT reduction was determined in tier 1 and IL-1α in tier 2, in samples from chemicals for which the MTT results indicated a viability above the 50% threshold. The results obtained with IL-1α in the EPISKIN model were used by ZEBET to develop a protocol for the EpiDerm test, in which IL-1α release was also determined in samples from chemicals which provided MTT test results above the 50% threshold.

3.3. Test design

The PM for the primary MTT endpoint for both tests was designed to correctly predict the current EU classifications for skin irritation (R38 *versus* no label): a test substance was predicted to be a skin

Table 3: Laboratories participating in Phase 2 of the ECVAM SIVS

	EpiDerm	**EPISKIN**
Lead laboratory (LL)	ZEBET (Germany)	L'Oréal (France)
Additional laboratory 1 (AL1)	IIVS (USA)	Unilever (UK)
Additional laboratory 2 (AL2)	BASF (Germany)	Sanofi (France)

irritant (R38), if it reduced the mean cell viability to below 50%, compared to the mean cell viability of the negative control; if the mean cell viability was above 50%, it was considered not to be a skin irritant (no label).

Each laboratory tested the same set of 58 chemicals, each in three runs. The chemicals were coded and distributed by RCC–CCR, and the codes were provided to RCC–CCR by ECVAM. Contact between the laboratories during the testing period was not permitted.

3.4. Test chemicals

The CSSC comprised European Commission officials from ECVAM and the ECB, in consultation with two representatives of two EU competent authorities (Table 2). The criteria for the chemicals selection for Phase 1 were maintained in Phase 2. A detailed report on the chemicals selection procedures has been provided in a separate publication (14).

The test chemicals selected on the basis of information obtained from four data bases: 33 chemicals from the NCD of the ECB, 19 from a database compiled by ECETOC (15), five from the TSCA database (16), and one from the CIR (see 14). In addition to the IUPAC name of each substance, the CAS numbers and the database sources are listed in Table 4, along with the assigned number, which was used throughout the study to identify the chemical. Permission for the publication of proprietary chemical identities was not obtained for two of the 60 chemicals originally selected, so they were omitted from the Phase 2 inventory.

Classification and *in vivo* Draize score dominant medians for the 58 test chemicals (Table 4) were supplemented by individual observations on the 25 chemicals from the ECETOC, TSCA and CIR databases only (Table 5). Permission from the respective notifiers to publish individual *in vivo* Draize scores registered in notification files for the 33 NCD chemicals was not obtained, since these are proprietary confidential data (14).

3.5. The *in vivo* reference test data

Since the Draize rabbit test for skin irritation (8) is the standard test for regulatory purposes, the MT decided to use this test as the *in vivo* reference test. The rabbit dominant median values of the 58 test chemicals are shown in Table 4, e.g. their classification and dominant endpoints. The dominant median is a concept developed by Hoffmann *et al.* (17) to model the rabbit data. It is determined by calculating the median of the individual rabbit mean scores for each dermal effect and then choosing the larger one, i.e. the dominant one. This median value permits the classification of a chemical according to both the European classification scheme (ECS; 7) and the Globally Harmonised System (GHS; 9) by comparison with the classification cut-off points, e.g. two animals out of three with a score of 2 in the case of the ECS, or 1.5 and 2.3 in the case of the GHS.

For chemicals selected from the ECETOC, TSCA and CIR databases (Table 5), irritant classifications corresponding to the Draize score observations were assigned. For the NCD chemicals, classifications according to the EU system are registered in notification files, and were used as an initial screening criterion for candidate chemicals. Subsequently, the correlation of the classifications with the Draize test observations was confirmed by the CSSC (14). In addition, all the test chemicals were allocated an irritant classification, according to GHS definitions (9).

3.6. Data submission

A data submission template in Excel® was developed for each test, in a collaborative effort by the lead laboratories and ECVAM. The final version was password-protected by ECVAM, then provided to the lead laboratories, which passed them on to their partner laboratories. The spreadsheet containing the test data had to be returned solely to the biostatistician of the MT. The data were accepted only if the password-protection was still in place.

3.7. Quality criteria

Although the quality criteria for consistency and interpretation of the MTT reduction measurements are defined in the SOPs of the two tests, they are presented here, since the level of compliance might reveal useful information. The quality criteria addressed the responses and the variabilities of the negative control (NC, phosphate-buffered saline [PBS]) and the positive control (PC, 0.5% sodium dodecyl sulphate [SDS]), as well as the variability of the test samples (Table 6). The variability criterion was established by an analysis of the Phase 1 data, and adjusted following a coded interim analysis by the MT. A threshold value for the standard deviation (SD) of 18% was chosen, to ensure that replicate measurements covered less than a third of the response scale of cell viability, i.e. from 0 to about 100%.

Data analysis was performed after all the experimental data had been compiled and submitted, and by excluding data which did not conform to the sample variability criterion of three “valid” runs per set.

Table 4: 58 Phase 2 chemicals: EU and GHS classifications, dominant median scores and dominant endpoints

No.	Chemical name	CAS number	Source	Classifications		Dominant median	Dominant endpoint
				EU	GHS		
1	2-chloromethyl-3,5-dimethyl-4-methoxypyridine hydrochloride	86604-75-3	NCD	R38	I	2.7	B
2	1-bromo-4-chlorobutane	6940-78-9	ECETOC	no label	NI	0	B
3	1-bromohexane	111-25-1	ECETOC	R38	I	2.7	E
4	1-decanol	112-30-1	ECETOC	R38	I	2.3	E
5	3-chloro-4-fluoronitrobenzene	350-30-1	ECETOC	no label	NI	1.0	E
6	3-diethylaminopropionitrile	5351-04-2	ECETOC	no label	NI	0	B
7	3-mercaptohexanol	51755-83-0	NCD	no label	NI	0	B
8	4-methylthio-benzaldehyde	3446-89-7	ECETOC	no label	NI	1.0	E
9	2,6-dimethyl-4-nitrobenzeneamine	16947-63-0	NCD	no label	NI	0.3	E
10	allyl heptanoate	142-19-8	ECETOC	no label	MI	1.7	E
11	allyl phenoxyacetate	7493-74-5	ECETOC	no label	NI	0.3	E
12	2-ethylhexyl 4-aminobenzoate	26218-04-2	NCD	no label	NI	0.7	E
13	1-[4-(2-dimethylaminoethoxy)phenyl]-2-phenylbutan-1-one	68047-07-4	NCD	R38	MI	2.0	E
14*							
15	α-terpineol	98-55-5	ECETOC	R38	I	2.7	O
16	capryl-isostearate	209802-43-7	NCD	no label	NI	1.0	E
17	2-methyl-3-[(1,7,7-trimethylbicyclo[2.2.1]hept-2-yl)oxy]-1-propanol, bornyl isomer	128119-70-0	NCD	no label	MI	1.7	E
18	butyl methacrylate	97-88-1	TSCA	R38	I	3.0	E
19	2,5-dimethyl-4-oxo-4,5-dihydrofuran-3-yl acetate	4166-20-5	NCD	no label	NI	0	B
20	cyclamen aldehyde	103-95-7	ECETOC	R38	I	2.3	O
21	Mixture of: 5-exo-decylbicyclo[2.2.1]hept-2-ene; 5-endo-decylbicyclo[2.2.1]hept-2-ene	22094-85-5	NCD	no label	MI	1.7	E
22	diethyl phthalate	84-66-2	ECETOC	no label	NI	0	E
23	di-n-propyl disulphide	629-19-6	ECETOC	R38	I	3.0	E
24	di-propylene glycol	25265-71-8	ECETOC	no label	NI	0	E
25	dipropylene glycol monobutyl ether	29911-28-2	CIR	no label	NI	0	E

** = confidential chemical. Classifications: NI = non-irritant; MI = mild irritant; I = irritant. Dominant endpoint: E = erythema; O = oedema; B = both endpoint reactions were identical.*

Table 4: continued

No.	Chemical name	CAS number	Source	Classifications		Dominant median	Dominant endpoint
				EU	GHS		
26	3,4-dimethyl-1H-pyrazole	2820-37-3	NCD	no label	NI	0	B
27	2-isopropyl-2-isobutyl-1,3-dimethoxypropane	129228-21-3	NCD	R38	I	4.0	E
28	ethyl *cis*-4-[4-[[2-(2,4-dichlorophenyl)-2-(1H-imidazol-1-ylmethyl)-1,3-dioxolan-4-yl] methoxy]phenyl]piperazine-1-carboxylate	67914-69-6	NCD	no label	NI	0	B
29	Mixture of: 2-methyl-4-(2´,2´,3´-trimethyl-3´-cyclopenten-1´-yl)-4-penten-1-ol 56% (1´R,2R) & 40%(1´R,2S) isomer	014864-90-6	NCD	R38	MI	2.0	B
30	Mixture of: diethyl *cis*-1,4-cyclohexanedicarboxylate; diethyl *trans*-1,4-cyclohexanedicarboxylate	0072903-27-6	NCD	no label	NI	1.3	E
31	Mixture of isomers: ethyl exo-tricyclo[5.2.1.0(2,6)]decane-endo-2-carboxylate; ethyl endo-tricyclo[5.2.1.0(2,6)]decane-exo-2-carboxylate	80657-64-3 (mixture)	NCD	R38	MI	2.0	O
32	2S-(2-furyl)-5R-hydroxy-4R-(1R,2-dihydroxy)ethyl-6S-hydroxymethyl-1,3-dioxane	7089-59-0	NCD	no label	NI	0	B
33	heptyl butyrate	5870-93-9	ECETOC	no label	MI	1.7	E
34	hexyl salicylate	6259-76-3	ECETOC	R38	MI	2.0	B
35	cyclohexadecanone	2550-52-9	NCD	no label	NI	0	B
36	isopropanol	67-63-0	ECETOC	no label	NI	0.3	E
37	[2-(cyclopentyloxy)ethyl]benzene(cyclopentyl 2-phenylethyl ether)	not allocated	NCD	R38	I	3.0	E
38*							
39	methyl stearate	112-61-8	ECETOC	no label	NI	1.0	E
40	1-methyl-3-phenyl-1-piperazine	5271-27-2	NCD	R38	I	3.3	E
41	naphthalene acetic acid	86-87-3	TSCA	no label	NI	0	B
42	disodium 2,2´-(1,4-phenylene)bis-1H-benzimidazole-4,6-disulfonic acid or monosulfonic acid, monosulfonate or disulfonate	180898-37-7	NCD	no label	NI	0	B
43	Mixture of isomers: 1-(1,1-dimethylpropyl)-4-ethoxy-*cis*-cyclohexane; 1-(1,1-dimethylpropyl)-4-ethoxy-*trans*-cyclohexane	181258-87-7 (*cis*) 181258-89-9 (*trans*)	NCD	R38	MI	2.0	B
44	phenylethylalcohol	60-12-8	ECETOC	no label	NI	1.0	E
45	(+/–) *trans*-3,3-dimethyl-5-(2,2,3-trimethyl-cyclopent-3-en-1-yl)-pent-4-en-2-ol	107898-54-4	NCD	R38	I	2.7	E

* = *confidential chemical. Classifications: NI = non-irritant; MI = mild irritant; I = irritant. Dominant endpoint: E = erythema; O = oedema; B = both endpoint reactions were identical.*

Table 4: continued

No.	Chemical name	CAS number	Source	Classifications		Dominant median	Dominant endpoint
				EU	GHS		
46	4-methyl-8-methylenetricyclo[3.3.1.1(3,7)]decan-2-ol	122760-84-3	NCD	R38	MI	2.0	B
47	4-methyl-8-methylenetricyclo[3.3.1.1(3,7)]dec-2-yl acetate	122760-85-4	NCD	R38	MI	2.0	B
48	2-(formylamino)-3-thiophenecarboxylic acid	43028-69-9	NCD	no label	NI	0	B
49	isostearic acid monoisopropanolamide	152848-22-1	NCD	R38	MI	2.0	E
50	2-phenylhexanenitrile	3508-98-3	NCD	no label	MI	1.7	E
51	Mixture of isomers: 1-(2-isopropylphenyl)-1-phenylethane (CAS# 191044-60-7) 1-(3-isopropylphenyl)-1-phenylethane (CAS# 191044-59-4) 1-(4-isopropylphenyl)-1-phenylethane (CAS# 2320-06-1)	52783-21-8 (mixture)	NCD	R38	MI	2.0	E
52	propyl (2S)-2-(1,1-dimethylpropoxy)-propanoate	0319002-92-1	NCD	no label	NI	0.7	E
53	silane A-1430	2530-87-2	TSCA	no label	NI	0.0	B
54	Mixture of isomers: 1-(spiro[4.5]dec-7-en-7-yl)pent-4-en-1-one (CAS# 224031-70-3) 1-(spiro[4.5]dec-6-en-7-yl)pent-4-en-1-one (CAS# 224031-71-4)	224031-70-3	NCD	no label	NI	1.3	E
55	terpinyl acetate	80-26-2	ECETOC	R38	MI	2.0	B
56	benzenethiol, 5-(1,1-dimethylethyl)-2-methyl (NB: CAS name from company)	7340-90-1	NCD	R38	I	3.3	O
57	triethylene glycol	112-27-6	TSCA	no label	NI	0	B
58	tri-isobutyl phosphate	126-71-6	TSCA	R38	MI	2.0	E
59	(E,E)-3,7,11-trimethyldodeca-1,4,6,10-tetraen-3-ol	125474-34-2	NCD	R38	I	4.0	E
60	bis[(1-methylimidazol)-(2-ethyl-hexanoate)], zinc complex	not allocated	NCD	R38	MI	2.0	E

* = *confidential chemical. Classifications: NI = non-irritant; MI = mild irritant; I = irritant. Dominant endpoint: E = erythema; O = oedema; B = both endpoint reactions were identical.*

Table 5a: *In vivo* data for 25 test chemicals from the ECETOC, CIR and TSCA (US EPA) data bases* — erythema scores

No.	Source	Classifications EU	Classifications GHS	Dominant median	Dominant endpoint	No. of experiments	No. of rabbits	Rabbit erythema scores (average of scores after 24h, 48, 72h) 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Erythema median/ mean	Remarks
2	ECETOC	no label	NI	0.0	B	1	3	0	0	0													0	
3	ECETOC	R38	I	2.7	E	1	3	2.7	2.0	2.7													2.7	
4	ECETOC	R38	I	2.3	E	1	4	2.3	2.3	2.3	1.7												2.3	Single scores of 0.5, 1.5, 2.5
5	ECETOC	no label	NI	1.0	E	1	6	1.0	1.0	1.0	1.7	1.7	1.0										1.0	
6	ECETOC	no label	NI	0.0	B	1	3	0	0	0													0	
8	ECETOC	no label	NI	1.0	E	1	3	1.0	1.3	0.3													1.0	
10	ECETOC	no label	MI	1.7	E	1	4	1.3	2.0	1.7	2.0												1.7	Single scores of 0.5, 1.5
11	ECETOC	no label	NI	0.3	E	1	4	0.3	0.7	0.3	0.3												0.3	Single scores of 0.5
15	ECETOC	R38	I	2.7	O	3	11	1.7	2.0	2.3	2.0	2.7	2.0	1.7	2.0	2.0	1.7	2.0					2.0	
18	TSCA	R38	I	3.0	E	1	6	3.0	3.0	2.7	3.0	3.0	3.0										3.0	
20	ECETOC	R38	I	2.3	O	4	15	2.7	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.7	2.0	2.0	2.0	2.0	2.0	
22	ECETOC	no label	NI	0.0	E	2	7	0.7	0	0	0	0	0	0									0	
23	ECETOC	R38	I	3.0	E	1	3	1.7	3.0	3.0													3.0	
24	ECETOC	no label	NI	0.0	E	2	7	1.0	0	0	0	0	0	0									0	
25	CIR	no label	NI	0.0	E	1	6	0.3	0	0	0	0	0										0	
33	ECETOC	no label	MI	1.7	E	1	4	1.7	2.0	0.7	2.0												1.7	Single scores of 0.5, 1.5
34	ECETOC	R38	MI	2.0	B	4	15	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	
36	ECETOC	no label	NI	0.3	E	1	3	1.7	0.3	0.3													0.3	
39	ECETOC	no label	NI	1.0	E	1	3	1.0	2.3	1.0													1.0	
41	TSCA	no label	NI	0.0	B	1	6	0	0	0	0	0	0										0	No scores after 48h; two application sites per rabbit (scores were averaged)
44	ECETOC	no label	NI	1.0	E	2	7	1.3	1.3	2.0	1.0	1.0	0.7	0.3									1	
53	TSCA	no label	NI	0.0	B	1	6	0	0	0	0	0	0										0	
55	ECETOC	R38	MI	2.0	B	3	11	1.7	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1.7	2.0	1.3					2.0	
57	TSCA	no label	NI	0.0	B	1	6	0	0	0	0	0	0										0	
58	TSCA	R38	MI	2.0	E	1	6	2.0	2.3	1.7	2.0	2.3	2.0										2.0	Two application sites per rabbit (scores were averaged)

*In vivo *data for the 33 test chemicals from the NCD data base are not available in the open literature. Classifications: NI = non-irritant; MI = mild irritant; I = irritant. Dominant endpoint: E = erythema; O = oedema; B = both endpoint reactions were identical.*

Table 5b: *In vivo* data for 25 test chemicals from the ECETOC, CIR and TSCA (US EPA) data bases* — oedema scores

No.	Source	Classifications		Dominant median	Dominant endpoint	No. of experiments	No. of rabbits	Rabbit oedema scores (average of scores after 24h, 48, 72h)															Oedema median/ mean	Remarks
		EU	GHS					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
2	ECETOC	no label	NI	0.0	B	1	3	0	0	0													0	
3	ECETOC	R38	I	2.7	E	1	3	0	2.7	2.0													2.0	
4	ECETOC	R38	I	2.3	E	1	4	2.0	1.0	1.0	1.0												1.0	Single scores of 0.5, 1.5, 2.5
5	ECETOC	no label	NI	1.0	E	1	6	0.7	0	0	0.7	0.7	0										0.3	
6	ECETOC	no label	NI	0.0	B	1	3	0	0	0													0	
8	ECETOC	no label	NI	1.0	E	1	3	0	0	0													0	
10	ECETOC	no label	MI	1.7	E	1	4	0.3	0.7	0.7	0.7												0.7	Single scores of 0.5, 1.5
11	ECETOC	no label	NI	0.3	E	1	4	0.3	0	0	0.3												0	Single scores of 0.5
15	ECETOC	R38	I	2.7	O	3	11	2.0	2.3	3.0	3.0	3.0	2.7	1.7	2.7	2.0	0.7	3.0					2.7	
18	TSCA	R38	I	3.0	E	1	6	2.7	3.7	2.3	2.7	2.7	4.0										2.7	
20	ECETOC	R38	I	2.3	O	4	15	3.0	3.0	2.7	2.3	2.7	2.0	1.7	2.7	2.7	2.3	3.0	1.3	1.0	2.0	1.3	2.3	
22	ECETOC	no label	NI	0.0	E	2	7	0	0	0	0	0	0	0									0	
23	ECETOC	R38	I	3.0	E	1	3	0	0	0													0	
24	ECETOC	no label	NI	0.0	E	2	7	0	0	0	0	0	0	0									0	
25	CIR	no label	NI	0.0	E	1	6	0	0	0	0	0	0										0	
33	ECETOC	no label	MI	1.7	E	1	4	0	0.3	0	0.7												0.3	Single scores of 0.5, 1.5
34	ECETOC	R38	MI	2.0	B	4	15	1.0	1.3	2.0	0.7	2.0	2.0	2.0	2.7	1.7	2.0	2.7	1.3	2.0	1.0	1.0	2.0	
36	ECETOC	no label	NI	0.3	E	1	3	0	0	0													0	
39	ECETOC	no label	NI	1.0	E	1	3	0	2.0	0													0	
41	TSCA	no label	NI	0.0	B	1	6	0	0	0	0	0	0										0	No scores after 48h; two application sites per rabbit (scores were averaged)
44	ECETOC	no label	NI	1.0	E	2	7	0	1.0	1.0	0	0.7	0	0									0	
53	TSCA	no label	NI	0.0	B	1	6	0	0	0	0	0	0										0	
55	ECETOC	R38	MI	2.0	B	3	11	1.0	2.0	2.0	3.0	2.0	2.3	2.0	2.0	1.0	0.7	0.3					2.0	
57	TSCA	no label	NI	0.0	B	1	6	0	0	0	0	0	0										0	
58	TSCA	R38	MI	2.0	E	1	6	1.0	1.3	0.7	1.3	1.0	0										1.0	Two application sites per rabbit (scores were averaged)

*In vivo *data for the 33 test chemicals from the NCD data base are not available in the open literature. Classifications: NI = non-irritant; MI = mild irritant; I = irritant. Dominant endpoint: E = erythema; O = oedema; B = both endpoint reactions were identical.*

Table 6: Quality criteria of the EpiDerm and EPISKIN skin models, according to their SOPs: example of the range of model responses to a positive control (5% SDS)

	Viability	Range (95% confidence interval)	SD
EPISKIN	< 40%	1.5–32.2 (1.3–41.6)	≤ 18%
EpiDerm	< 20 %	3.7–13.8 (4.7–13.6)	≤ 18%

3.8. Data analysis and statistics

3.8.1. Special considerations for the analysis of the results of a formal validation study

In 2006, the MT agreed to include only the results calculated according to two options, i.e. either considering the data from all runs, or considering the data for which three valid runs (according to the variability criterion) were available. Thus, any *post hoc* rationalisation on how the data should be analysed, was avoided. For completeness, it should be pointed out that, in the few cases when more than three valid runs were available, the first option was used.

At the final meeting in 2006, the MT agreed that, for evaluating the overall outcome for each skin model, the main emphasis should be placed on analysing the data for chemicals with three valid runs in one laboratory.

3.8.2. Within-laboratory variability

In the analysis of within-laboratory variability, the test run was considered to be the experimental unit which better reflected current use and potential applications in the future. Two measures of within-laboratory variability are reported here. The within-laboratory SD, a measure of repeatability according to ISO standards (18), was determined for each chemical. It should be noted that this measure is not completely transferable from the ISO guidance, since variable aspects of the testing, e.g. the operators and materials used, were not systematically included in the study design. In addition, the predicted classifications resulting from the PMs were compared between the runs by a simple measure of similarity, i.e. the proportion of identical predictions.

In general, these parameters were calculated by considering all the available experimental runs — allowing direct comparisons of all the laboratories — and by considering the runs for those chemicals for which three runs met the variability criterion. The latter approach might result in unbalanced data sets for individual laboratories, since one laboratory might not have obtained three valid runs for a given chemical, while the other two laboratories had.

An additional analysis was conducted on the 18 chemicals which were tested in the lead laboratories in both phases of the study. Depending on the specific test method, chemical and run, there was a time difference of five to twelve months between the two sets of runs. For comparing the results of Phase 1 and Phase 2, a *t*-test (with a significance level of 1%) was applied for each test chemical. Furthermore, a paired *t*-test was calculated with the mean run results of the phases for all the chemicals.

3.8.3. Between-laboratory variability

The variability between the three laboratories of the primary endpoint was assessed by applying three statistical techniques. The SD of the three means of runs per laboratory, was determined. As the second measure, the proportion of identical run classifications and identical median run classifications over the three laboratories, were evaluated. In addition, the SDs and coefficients of variation (CVs) of the mean IL1-α release data of the three laboratories were calculated.

3.8.4. Predictive capacity

The abilities of the test systems to predict the EU risk phrases, i.e. R38 for skin irritants, and no label for non-irritants, were analysed in 2 × 2-contingency tables, and the following measures of predictive capacity were calculated: sensitivity, specificity, accuracy, positive predictive value (PPV) and negative predictive value (NPV). These parameters were determined for the cell viability (MTT reduction) endpoint, in comparison to the negative control. For the IL1-α release endpoint, the fold-increase in comparison to the negative control, as well as the total amounts of IL1-α, were analysed.

To analyse the impact of shifting the PM thresholds — to discriminate irritant from non-irritant test chemicals — on the predictive capacity of a test, an analysis of the receiver operating characteristic (ROC) curve was performed. This approach is well

established in clinical chemistry, to assess the quality of the determination of specific agents (19, 20). The parameters, sensitivity and (1–specificity), are calculated for each measured value and are plotted against each other. A ROC curve close to the line of identity would be useless, while a test characterised by a curve approaching the upper left corner of the plot would be particularly useful. The sum of sensitivity and specificity, which gives equal weight to the two parameters, was chosen to assess the ROCs.

In addition, the *in vivo* rabbit data used to classify the test chemicals, were correlated with data from the endpoints of the new test systems. For this purpose, the concept of the dominant median (17) was applied to reduce the *in vivo* data to a one-dimensional measure, while the loss of information was minimised. Since the PM had been developed only in a single laboratory and with a limited set of test chemicals, the PM of the IL1-α release endpoint was improved by taking into account all the 58 chemicals tested in Phase 2 of the study, in the three EPISKIN laboratories.

The second aim of the SIVS, to assess performance of the test systems according to GHS classification, was performed as a *post hoc* analysis. Since the results from Phase 1 did not permit the definition of a PM for the GHS classification system, two thresholds, maximising the accuracy, were chosen for each of the test methods. For chemicals with three valid runs, the median classifications were chosen for this analysis.

3.8.5. Statistics

All the calculations were performed with Microsoft Excel 2002, Graphpad Prism 4.02 or S-Plus 6.2. In the context of the acceptance criteria and reproducibility aspects, 1-way ANOVA techniques (with a confidence level of 1%) and descriptive measures, e.g. the SD or the CV, and similarity measures were applied. To assess the predictive capacity, contingency tables were applied, and confidence intervals are reported for the estimated parameters, when appropriate. ROC analysis was used to provide a detailed description of the predictive capacity of the test systems.

4. Results

4.1. Phase 1

The validation study was conducted in two phases. In Phase 1, 20 coded chemicals were tested in the lead laboratories for the three test systems, i.e. EPISKIN (L'Oréal), EpiDerm (ZEBET) and SIFT (Syngenta), in three independent runs. The Phase 1 data permitted a preliminary assessment of the within-laboratory reproducibility and the predictive capacities of the tests. EPISKIN and the EpiDerm are commercially-available reconstituted human epidermis models, and the endpoint measured in these assays is cell viability. The SIFT measures two endpoints after the application of the chemicals, namely, trans-epithelial water loss (TEWL) and electrical resistance (ER).

The balanced set of 20 coded chemicals tested in Phase 1 (Table 4) comprised 11 non-irritant (EU classification: no label) and 9 irritant (EU classification: R38) chemicals, selected by the CSSC.

4.1.1. SIFT

A good reproducibility was indicated by 1-way ANOVA and a *post hoc* Bonferroni post-test (with a significance level of 1%) for the two endpoints, TEWL and ER. The application of the PM, a threshold of 10 for TEWL and a threshold of 4 for ER, and comparison of the classifications between the runs for each chemical, resulted, for TEWL, in ambiguous classifications between runs for three chemicals, and for five chemicals, in the case of ER.

The predictive capacity of the SIFT in the lead-laboratory and the 2×2-contingency table are given in Table 7a. The accuracy of 45% indicates a discouraging overall performance of the SIFT method. Moving the thresholds did not substantially improve the assay's performance.

4.1.2. EpiDerm

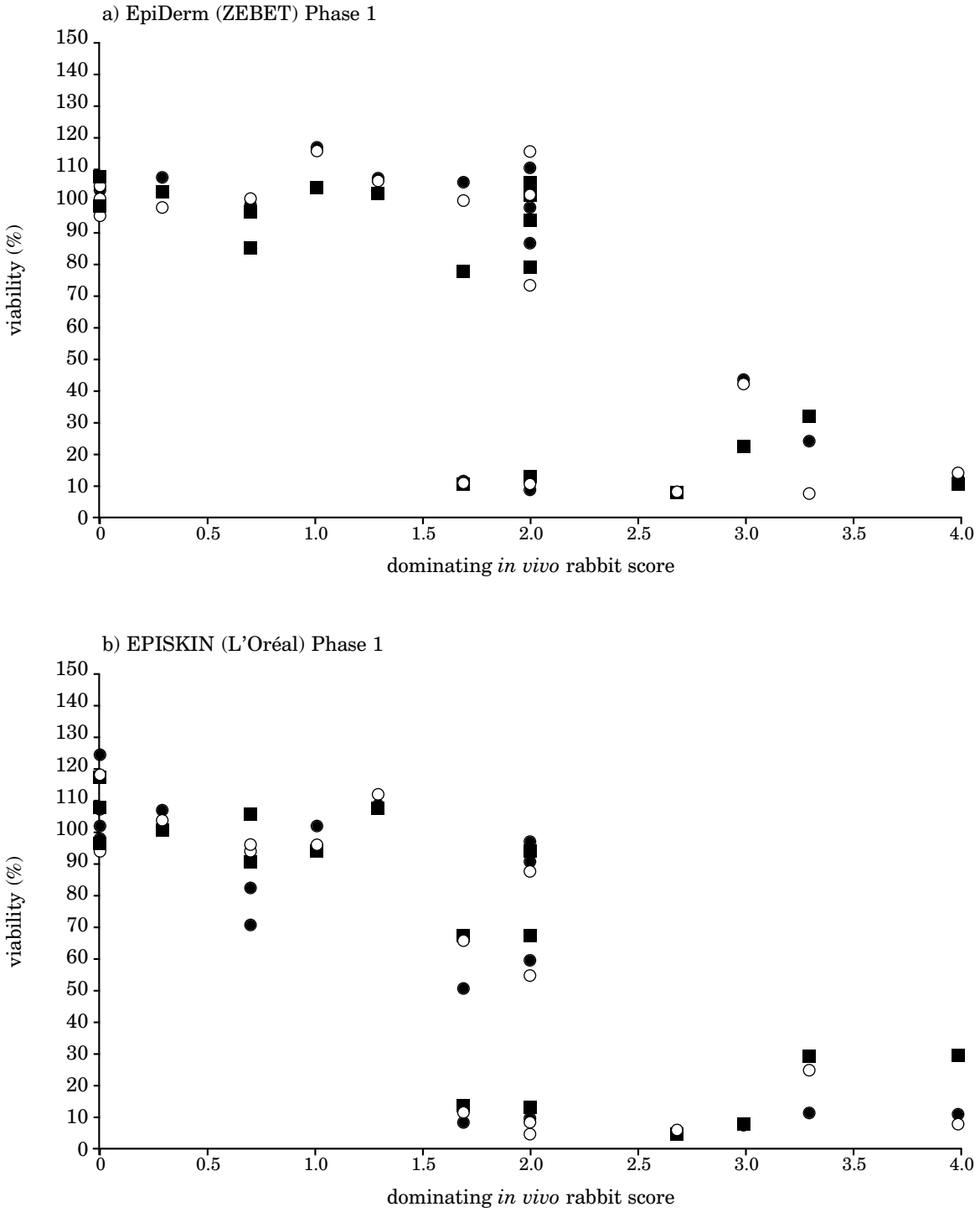
The within-laboratory variability was low, as indicated by identical classifications in individual runs. The measures of predictive capacity, e.g. sensitivity, specificity, accuracy, PPV and NPV, for the application of the EpiDerm model in the lead-laboratory are shown in Table 7b, along with the 2×2-contingency table. The accuracy of 75% indicates a promising overall performance of the EpiDerm test.

Figure 2 reveals a tendency for misclassifications to cluster near the dominant median Draize score of 2, which is, by definition, the threshold which separates irritants from non-irritants, according to the European classification system. Otherwise, and more typically, chemicals characterised by a dominant median *in vivo* score below 1.5 (< 1.5) or above 2.5 (> 2.5), correlated with the classification by the *in vitro* assay, as either non-irritant or irritant, respectively.

A preliminary ROC analysis revealed that thresholds of between 43% and 74% viability, would result in the maximum sum of sensitivity and specificity. Therefore, the chosen PM threshold of 50% was able to provide a reproducible and optimised test performance.

Despite an accuracy of 75%, the analysis also indicates that the EpiDerm model is not capable of dis-

Figure 2: Phase 1 — correlation of mean MTT viability data with *in vivo* scores in the rabbit, for the EpiDerm and EPISKIN skin models



MTT data are given for the 20 phase 1 chemicals (3 runs for each) and compared to the dominating in vivo *score on the rabbit skin.*

■ = *run 1;* ○ = *run 2;* ● =*run 3.*

Table 7: Phase 1 — contingency tables for SIFT, EpiDerm and EPISKIN

a) 2 × 2-contigency table and predictive capacity for SIFT in Phase 1

	European classification		
SIFT	**no label**	**R38**	**Σ**
PM non-irritant	7	7	14
PM irritant	4	2	6
Σ	11	9	20

Sensitivity:	2/9	= 22%
Specificity:	7/11	= 64%
Accuracy:	9/20	= 45%
PPV:	2/6	= 33%
NPV:	7/13	= 50%

b) 2 × 2-contigency table and predictive capacity for EpiDerm in Phase 1

	European classification		
EpiDerm	**no label**	**R38**	**Σ**
PM non-irritant	10	4	14
PM irritant	1	5	6
Σ	11	9	20

Sensitivity:	5/9	= 56%
Specificity:	10/11	= 91%
Accuracy:	15/20	= 75%
PPV:	5/6	= 83%
NPV:	10/14	= 71%

c) 2 × 2-contigency table and predictive capacity for EPISKIN in Phase 1

	European classification		
EPISKIN	**no label**	**R38**	**Σ**
PM non-irritant	10	3	13
PM irritant	1	6	7
Σ	11	9	20

Sensitivity:	6/9	= 67%
Specificity:	10/11	= 91%
Accuracy:	16/20	= 80%
PPV:	6/7	= 86%
NPV:	10/13	= 77%

PM = prediction model; PPV = positive predictive value; NPV = negative predictive value.

tinguishing between the three GHS classes, since the mild-irritants were assigned to all of the GHS classes.

4.1.3. EPISKIN

The within-laboratory variability was low, as indicated by identical classifications between the runs, i.e. a similarity of 100%. The measures of predictive capacity of the EPISKIN model in the lead-laboratory are shown in Table 7c, as is the 2 × 2-contingency table. An accuracy of 80% indicates a promising overall performance of the test method. Figure 2 shows a result for EPISKIN which is similar to that obtained with EpiDerm, since there is again a tendency for misclassifications to cluster near the dominant median Draize score of 2, which is the threshold separating irritants from non-irritants in the EU classification system. Again, chemicals characterised by a dominant median *in vivo* score below 1.5 (< 1.5) or above 2.5 (> 2.5), correlated with the classification by the EPISKIN assay, as either non-irritant or irritant, respectively.

The ROC analysis revealed that all the thresholds between 30% and 50% viability would result in comparably high sensitivities and specificities. It can be concluded that the PM threshold of 50% was appropriate.

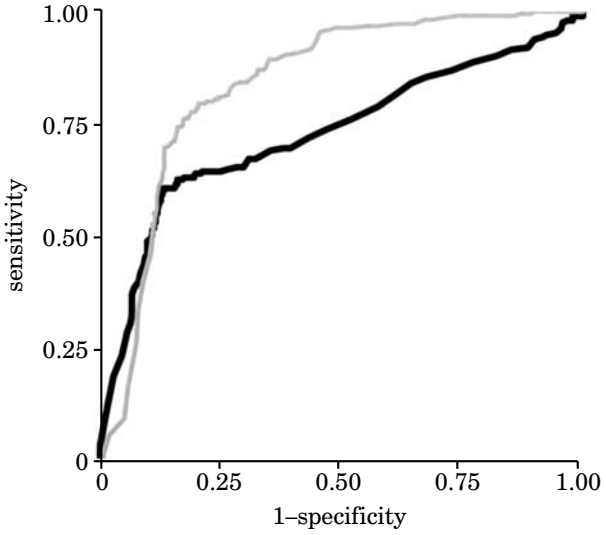
Despite an accuracy of 80%, the data analysis also indicated that the EPISKIN model is not capable of distinguishing between the three GHS classes, since the mild-irritants were assigned to all of the GHS classes.

4.1.4. The additional endpoint, IL1-α release

In a test development study, the application of the two-tiered testing strategy resulted in an increase

in sensitivity to 95%, and the rate of false positive results was reduced to 4.3% (5). At the L'Oréal laboratory, the 20 Phase 1 chemicals were also tested with the EPISKIN model in the two-tiered testing strategy (MTT reduction + IL-1α release). This refined approach improved the overall predictive ability of the EPISKIN test, when compared with the results summarised above, enabling 8 out of 9 irritants to be identified correctly as rabbit skin irritants. On the basis of this evidence, it was agreed by the MT that there was a need to investigate the potential value of incorporating IL-1α release measurements, in improving the sensitivities (i.e. the prediction of irritants) of both human skin model assays in Phase 2 of the SIVS. L'Oréal assisted ZEBET in establishing a protocol for the additional endpoint with the EpiDerm test, after which the 20 Phase 1 chemicals were tested at ZEBET with the new protocol for IL-1α release. Due to the encouraging results obtained, the MT decided that the two-tiered testing strategy should be used in Phase 2 of the SIVS. However, since strongly-irritant chemicals might affect the generation of IL1-α, the MT decided that this endpoint should only be measured in substances classified as non-irritating in the MTT test.

Figure 3: Receiver operating characteristic (ROC) curves of all valid runs in EpiDerm and EPISKIN laboratories, for the endpoint MTT



——— = *EpiDerm;* ——— = *EPISKIN.*

In the ROC curve, the sensitivity and (1–specificity) parameters are calculated for each measured value, and plotted against each other. Thus, a ROC curve close to the line of identity would indicate that a test is useless, while a ROC curve approaching the upper left corner of the plot indicates that a test is particularly useful. The results of all valid runs are summarised for the two skin models.

4.1.5. Conclusions from Phase 1

Taking into account the good within-laboratory reproducibility and the acceptable predictive capacities of the EpiDerm and EPISKIN tests in Phase 1 (Figure 3 and Table 7), notwithstanding the problem represented by chemicals with Draize scores near the mid-range threshold of 2, which by definition separates irritants (R38) from non-irritants (no label), the MT recommended that these two test systems be assessed in Phase 2 of the SIVS. In addition, taking into account the encouraging results with the determination of the additional endpoint IL-1α, the MT decided that the described two-tiered testing strategy (MTT reduction and IL-1α release) should be tested in Phase 2 of the ECVAM SIVS study. In contrast, due to the insufficient predictive capacity of the SIFT, the MT decided that this assay needed further development, and that it should not proceed to Phase 2.

The *post hoc* analysis of the EpiDerm and the EPISKIN data showed that these two test systems were sufficient to meet the needs of the European classification system for skin irritation. However, chemicals that are classified as "mild irritants" according to the GHS classification system cannot be discriminated from the other two GHS-classes by the two skin models. Nevertheless, the MT recommended that a similar *post hoc* analysis should be conducted on the larger data set generated in Phase 2.

A detailed outline of the statistical analysis of the SIVS data is provided on the ECVAM website (21).

4.2. Phase 2

4.2.1. EpiDerm cell viability data

4.2.1.1. Analysis of data quality

In total, five data-related quality criteria were included in the EpiDerm SOP — the first four addressed the controls, and the fifth concerned the test samples. The first criterion required a mean response (in OD) of the negative control larger than 0.6 OD. Differences in sample sizes per laboratory were the consequence of tests which did not meet the quality criteria, and thus triggered additional testing. Secondly, a run was only considered valid according to the SOP, when the mean viability of the positive control was below 40% of the viability of the negative control. The data analysis proved that this criterion was always met. The third crite-

rion required that the variability of the SD of the negative control replicates was smaller than 18%. The fourth criterion was identical to this, but referred to the variability of the positive control. The last criterion focused on the variability of a test sample — it required that a sample had to be retested, if three replicates did not meet the variability criterion. However, retesting could only be performed once.

In addition, a run was only considered valid according to the SOP, when the mean relative viability of the positive control was below 20% of the viability of the negative control. The data obtained from the three laboratories, and for Phases 1 and 2, showed that this criterion was always met. However, a more variable response to the positive control occurred at ZEBET in Phase 2, so the variation in the data was significantly larger than that for the respective responses at IIVS and BASF.

The number of test substances that were retested at ZEBET was lower than the numbers at IIVS and BASF. At ZEBET, 193 tests in total were performed with the 58 chemicals, including 18 tests which did not meet the variability criterion. At IIVS, 196 tests were conducted, of which 32 did not meet the SD criterion, and at BASF, 200 tests were performed, 36 of which did not meet the quality criteria. At BASF, all the tests of the fourth runs were triggered by failure to meet the SD criterion.

4.2.1.2. Within-laboratory variability

The results of the within-laboratory variability of all the applied measures are summarised for the three laboratories in Table 8. As far as the number of determinations (sample size) is concerned, it is obvious that, in the more experienced lead laboratory (ZEBET), three valid runs were achieved for more chemicals than in the other two laboratories. When considering a SD of 18% as the threshold for acceptable variability, the results for two to four chemicals per laboratory were not reproducible. The numbers of chemicals not consistently classified in the test runs in a given laboratory ranged, for the analysis of three runs, between one and three.

The comparison of the results of the 18 chemicals tested in Phase 1 and Phase 2 of the SIVS was performed on the ZEBET data with the three valid runs. This information adds to the assessment of the within-laboratory variability, since the time difference between determinations was up to twelve months. All the tests were conducted by the same operator, but for Phase 1 and Phase 2, separate sets of test chemicals were provided by the distributor.

When a *t*-test with a significance level of 1% was applied to the data for each run with each of the chemicals, only one of the substances provided significantly different results in the Phase 1 and Phase 2 tests. Testing the mean viabilities of the phases for the 18 chemicals by a paired *t*-test resulted in a non-significant p value of 0.298, which indicated good within-laboratory reproducibility.

4.2.1.3. Between-laboratory variability

The first measure of between-laboratory variability was the SD of the means of the runs per laboratory (Table 9). In the analysis of all the runs, applying the *below 18* variability criterion showed that, for 13 chemicals, the SD was above 18%. Of the 52 chemicals with three valid runs, seven showed an SD of 18% or above.

Applying the second measure — the proportion of identical chemical classifications, taking the median classification per laboratory into account — to the 52 chemicals characterised by three valid runs in at least two laboratories, 47 chemicals were classified identically. For six of these, there were three valid runs in only two laboratories. The test chemicals which provided irreproducible results according to this measure, were identical to the ones identified in the between-laboratory SD analysis. Considering all the runs, 15 out of 58 chemicals were not classified consistently.

In a pair-wise comparison of the classification results in the EpiDerm laboratories, the concordance of classifications was 47/58 = 81.0% for ZEBET–IIVS, 51/58 = 87.9% for ZEBET–BASF,

Table 8: Summary of the evaluation of within-laboratory variability for EpiDerm, in three laboratories

	ZEBET		IIVS		BASF	
Variability measure	**All runs**	**Three valid runs**	**All runs**	**Three valid runs**	**All runs**	**Three valid runs**
Sample size	58	54	58	48	58	48
Number of chemicals with SD >18	4	2	7	2	12	4
Proportion of identically-classified chemicals	52/58	53/54	48/58	46/48	48/58	45/48

Table 9: EpiDerm: between-laboratory variability — the standard deviation of mean MTT values

	All runs				Three valid runs			
No.	**ZEBET**	**IIVS**	**BASF**	**SD**	**ZEBET**	**IIVS**	**BASF**	**SD**
1	6.39	5.61	6.43	0.47	6.39	5.61	6.43	0.47
2	88.85	48.22	23.20	**33.14**	87.82		23.20	**45.70**
3	96.39	100.41	95.92	2.47	96.39	100.41	95.92	2.47
4	12.66	7.11	58.65	**28.29**	12.66	7.11		3.93
5	101.67	53.19	84.82	**24.61**	101.67		84.82	11.91
6	76.80	23.41	30.33	**29.03**	88.24		29.15	**41.78**
7	51.44	43.02	63.39	10.24				
8	96.12	83.85	52.76	**22.35**	96.12	94.99	46.21	**28.49**
9	97.21	107.38	101.30	5.12	97.21	107.38	107.49	5.91
10	101.07	102.09	97.01	2.68	101.07	102.09	97.01	2.68
11	95.46	99.66	100.79	2.81	95.46	99.66	105.57	5.08
12	104.55	91.94	107.34	8.20	104.55	91.94	107.34	8.20
13	68.26	43.27	46.16	13.67	66.12			
14*								
15	65.34	17.86	70.87	**29.14**				
16	101.06	108.02	99.46	4.55	101.06	108.02	99.46	4.55
17	11.43	7.90	10.57	1.84	11.43	7.90	10.57	1.84
18	98.11	92.64	96.88	2.87	98.11	92.64		3.87
19	89.99	80.30	91.20	5.97	89.99	80.30	91.20	5.97
20	21.60	8.67	26.39	9.16	23.41	8.67	24.17	8.73
21	101.24	109.22	104.61	4.01	101.24	109.22	104.61	4.01
22	99.22	98.90	102.15	1.79	99.22	98.90	102.15	1.79
23	95.68	100.78	92.09	4.36	95.68	100.78	92.09	4.36
24	100.35	91.79	99.67	4.76	100.35	91.79	99.67	4.76
25	98.19	102.54	99.34	2.26	98.19	102.54	102.93	2.63
26	8.70	14.07	7.58	3.47	8.70	14.07	7.58	3.47
27	101.56	65.02	103.58	**21.70**	101.56	52.38	103.58	**28.99**
28	95.08	88.58	82.95	6.07	95.08	89.34	82.95	6.07
29	11.41	8.53	14.47	2.97	11.41	8.53	8.44	1.69
30	97.75	100.08	106.67	4.63	97.75	100.08	106.67	4.63
31	18.27	58.78	20.39	**22.80**	14.84		8.16	4.72
32	100.19	97.66	106.13	4.35	100.19	96.96	106.13	4.65
33	100.48	93.76	98.67	3.48	100.48	92.28	98.67	4.31
34	102.17	96.66	89.72	6.24	102.17	96.66	80.70	11.15
35	94.82	113.61	7.85	**56.42**	94.82	113.61	7.85	**56.42**
36	97.88	85.07	93.94	6.56	97.88	85.07	93.94	6.56
37	42.60	47.28	31.00	8.38	42.60	47.28	33.08	7.24
38*								
39	97.42	101.57	104.00	3.33	97.42	101.57	104.00	3.33
40	21.96	55.07	39.88	16.57	21.96	76.67	36.67	**28.31**
41	98.32	98.65	103.69	3.01	98.32	98.65	103.69	3.01
42	101.49	106.21	92.87	6.77	101.49	106.21	92.87	6.77
43	103.72	101.71	100.12	1.80	103.72	101.71	102.28	1.03
44	71.36	45.68	92.91	**23.64**			98.16	
45	11.02	7.80	14.07	3.14	11.02	7.80	14.07	3.14

*Bold: SD > 18; * = confidential chemical.*

Table 9: continued

	All runs				Three valid runs			
No.	**ZEBET**	**IIVS**	**BASF**	**SD**	**ZEBET**	**IIVS**	**BASF**	**SD**
46	11.33	17.12	8.25	4.50	11.33	17.12	8.25	4.50
47	17.60	19.52	12.73	3.50	20.90		6.11	10.46
48	100.44	97.80	105.70	4.02	100.44	97.80	105.70	4.02
49	96.59	94.24	99.43	2.60	96.59	95.28	99.43	2.12
50	73.87	82.14	80.92	4.46	74.80	82.25	80.92	3.98
51	107.62	116.25	116.45	5.04	107.62	116.25	116.45	5.04
52	96.78	99.34	86.88	6.58	96.78	97.99		0.85
53	98.62	98.55	61.72	**21.29**	97.41	98.55		0.81
54	79.36	51.21	87.65	**19.10**	81.25	46.02	88.20	**22.62**
55	81.53	10.16	80.45	**40.90**		10.16		
56	14.58	12.18	9.89	2.34	14.58	12.18	9.89	2.34
57	94.95	94.99	95.29	0.19	94.95	94.99	95.29	0.19
58	29.97	44.63	24.33	10.48	10.81			
59	17.46	7.76	40.27	16.69	17.46	7.76		6.86
60	8.82	6.55	13.70	3.65	8.82	6.55	11.44	2.44

*Bold: SD > 18; * = confidential chemical.*

and 46/58 = 79.3% for IIVS–BASF.

A more-detailed outline of the within-laboratory and between-laboratory variability of the EpiDerm test results in the SIVS, is provided on the ECVAM website (21).

4.2.1.4. Predictive capacity

For the predictions obtained with the EpiDerm model, two approaches were applied, taking into account the median of all the available test runs, or the median of three valid runs (in this case, if more than three valid runs had been conducted, the first three were considered). As an example, the EpiDerm test outcomes for the two parameters, specificity and sensitivity, at ZEBET are presented in Table 10, for each single run and for the summarising approach. In addition to sample sizes, the exact lower 5% confidence limits are also presented.

Table 10: Example of the predictive capacity of EpiDerm at the lead laboratory (ZEBET) — specificity and sensitivity, taking into account sample sizes and 5% lower confidence limits for each valid run, and summaries for all runs

		Specificity			Sensitivity		
		n	**%**	**LB–5%**	**n**	**%**	**LB–5%**
Run 1	All	33	87.9	74.4	25	52.0	34.1
	Valid	32	90.6	77.5	21	57.1	37.2
Run 2	All	33	90.9	78.1	25	56.0	37.9
	Valid	32	90.6	77.5	23	56.5	37.5
Run 3	All	33	87.9	74.4	25	60.0	41.7
	Valid	30	91.5	80.5	23	65.2	46.0
All runs (median)		33	90.9	78.1	25	56.0	37.9
Three valid runs (median)		31	93.6	81.1	23	60.9	41.9

LB–5% = lower boundary of 95% confidence interval.

Both the approaches provided a specificity of around 90%. However, Table 10 also demonstrates that the specificity of the valid runs meeting the quality criteria was slightly higher. A similar effect was observed for sensitivity, which ranged between 52.0 and 65.2%, depending on the analysis performed.

A comparison of the data for chemicals misclassified at least once in one of the three EpiDerm laboratories, is presented in Table 11. Eleven of the 33 non-irritant chemicals (33%), which are not classified as irritant *in vivo*, were misclassified as "irritant (R38)" at least once, while 16 of the 25 irritant (R38) chemicals (64%) were misclassified at least once as "non-irritant". Table 11 also demonstrates that two non-irritant chemicals were consistently classified as irritant in all the runs and by all the laboratories, while seven irritants were consistently classified as non-irritants in all the runs and by all the laboratories.

The predictive capacities of the EpiDerm test for classifying test chemicals according to their irritation potentials in the three laboratories are presented in Table 12, which summarises specificity and sensitivity over all the runs in each laboratory, according to the two approaches outlined above.

In addition, sample sizes, as well as specificity and sensitivity, from all the runs in all the laboratories, were calculated for the two approaches, considering either all the individual classifications or the median classification for a given chemical (Table 12). Taking all the individual classifications into account, a low predictivity was obtained, when compared to the results obtained when three valid runs were considered. This indicates that the chances of misclassification are increased by high variability, e.g. when the exclusion criteria for high variability are not applied. When the chemicals for which there were three valid runs were evaluated, a specificity of 89% and a sensitivity of 60% were obtained. The results were similar when the median classifications were used. Due to the strong influence of the reproducibility of the determinations on the results, no estimation of confidence limits was performed.

In Figure 3, a ROC curve summarises the performance of the PM for the MTT reduction endpoint in the three EpiDerm laboratories. The angular shape of the curve indicates the successful test optimisation to specifically distinguish the classification of the two classes of skin irritation.

The same effect is evident in Figure 4, which shows the sensitivity and specificity curves, as well as their summary curve. In the threshold range of between 45% and 73%, the sensitivity and specificity curves are almost in a plateau, and the sum of the two values remains approximately constant. The maximum summary value of 1.454 was reached at a threshold of 54%, which is close to the 50% threshold defined in the PM.

Finally, the performance of the EpiDerm test in predicting the three classes of the GHS for the set of 58 test chemicals was evaluated. To keep this analysis simple, and disregarding reproducibility, only the median run classifications of chemicals with three valid runs in all the laboratories were considered.

The resulting set of 150 eligible cases comprised 33 GHS irritants, 43 GHS mild-irritants, and 74 GHS non-irritants. As described for Phase 1, no satisfactory PM existed, so a *post hoc* approach was used to derive a new PM. The new PM was based on a 60% viability threshold, below which chemicals would be classified as GHS irritants, and an 81% viability threshold, above which chemicals would be classified as GHS non-irritants. Chemicals characterised by viabilities between the two thresholds would be classified as GHS mild-irritants. Applying the PM provided correct classifications for 66.7% of the GHS irritants, 9.3% of the GHS mild-irritants, and 87.8% of the GHS non-irritants; viabilities between 60% and 81% were present for only six chemicals. This confirmed the indications from Phase 1, that the EpiDerm test is not able to correctly predict the classification of chemicals into the three GHS classes. That either high or low cell viabilities were found for the GHS mild-irritants, reflects the fact that the EpiDerm protocol had been optimised for the European classification system, which is based on two, rather than on three, classes of irritation.

A detailed outline of the predictive capacity of the EpiDerm model in the ECVAM SIVS is provided on the ECVAM website (21).

4.2.2. EPISKIN cell viability data

4.2.2.1. Analysis of data quality

As for the EpiDerm SOP, five data-related quality criteria were included in the EPISKIN SOP — the first four addressed the controls, and the fifth concerned the tested samples. The first criterion required a mean response (in OD) of the negative control larger than 0.6 OD. Differences in sample sizes per laboratory were the consequence of tests which did not meet the quality criteria, and thus triggered additional testing. Secondly, a run was only considered valid according to the SOP, when the mean viability of the positive control was below 40% of the viability of the negative control. The data analysis proved that this criterion was always met. The third criterion required that the variability of the SD of the negative control replicates was smaller than 18%. The fourth criterion was identical to this, but referred to the variability of the positive control. The last criterion focused on the variability of a test sample — it required that a sample had to be retested, if three replicates did not meet the variability criterion. However, retesting could only be performed once.

Table 11: EpiDerm: summary of chemicals which were misclassified at least once

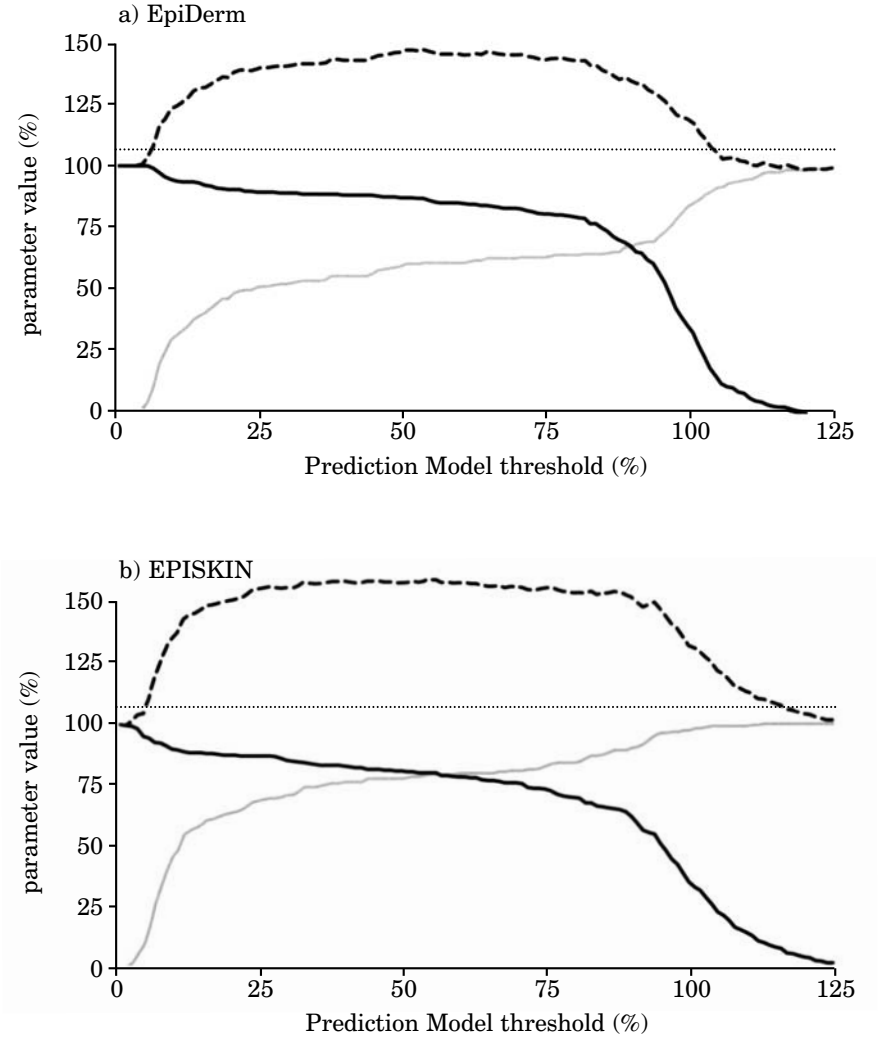
No.	EU class	Dominant median	ZEBET run 1	2	3	4	5	6	IIVS run 1	2	3	4	5	BASF run 1	2	3	4	Total number of runs	Misclassifying runs (%)
17	no label	1.7	**1**	**1**	**1**				**1**	**1**	**1**			**1**	**1**	**1**		9	100.00
26	no label	0	**1**	**1**	**1**				**1**	**1**	**1**			**1**	**1**	**1**		9	100.00
6	no label	0	0	0	**1**	0			**1**	**1**	**1**	**1**		**1**	1	0	**1**	12	66.67
7	no label	0	**1**	**1**	**1**	0			**1**	0	**1**	**1**		0	0	0	**1**	12	58.33
2	no label	0	0	0	0	0			**1**	0	0	**1**		**1**	**1**	**1**		11	45.45
35	no label	0	0	0	0				0	0	0			**1**	**1**	**1**		9	33.33
5	no label	1.0	0	0	0				**1**	0	**1**	0		0	0	0		10	20.00
8	no label	1.0	0	0	0				0	0	0	0		**1**	**1**	0	0	11	18.18
54	no label	1.3	0	0	0	0			**1**	0	**1**	0		0	0	0	0	12	16.67
44	no label	1.0	**1**	0	0	0	0	0	0	**1**	0	0		0	0	0	0	14	14.29
53	no label	0	0	0	0	0			0	0	0			**1**	0	0	0	11	9.09
34	R38	2.0	**0**	**0**	**0**				**0**	**0**	**0**			**0**	**0**	**0**	**0**	10	100.00
49	R38	2.0	**0**	**0**	**0**				**0**	**0**	**0**	**0**		**0**	**0**	**0**		10	100.00
51	R38	2.0	**0**	**0**	**0**				**0**	**0**	**0**			**0**	**0**	**0**		9	100.00
23	R38	3.0	**0**	**0**	**0**				**0**	**0**	**0**			**0**	**0**	**0**		9	100.00
43	R38	2.0	**0**	**0**	**0**				**0**	**0**	**0**			**0**	**0**	**0**	**0**	10	100.00
18	R38	3.0	**0**	**0**	**0**				**0**	**0**	**0**			**0**	**0**	**0**	**0**	10	100.00
3	R38	2.7	**0**	**0**	**0**				**0**	**0**	**0**			**0**	**0**	**0**		9	100.00
27	R38	4.0	**0**	**0**	**0**				1	1	**0**	**0**		**0**	**0**	**0**		10	80.00
55	R38	2.0	**0**	**0**	**0**	**0**			1	1	1			**0**	**0**	**0**	**0**	11	72.73
15	R38	2.7	**0**	1	**0**	**0**	**0**		1	1	1	1		1	**0**	**0**	**0**	13	53.85
13	R38	2.0	**0**	**0**	1	**0**	1		**0**	1	1	**0**		**0**	1	1	1	13	46.15
58	R38	2.0	**0**	**0**	1	1	1		1	**0**	**0**	1	**0**	1	1	1	1	14	35.71
31	R38	2.0	1	1	1	1			**0**	1	**0**	**0**		1	**0**	1	1	12	33.33
40	R38	3.3	1	1	1				**0**	**0**	**0**	1	1	**0**	1	1	1	12	33.33
4	R38	2.3	1	1	1				1	1	1			1	**0**	**0**	**0**	10	30.00
59	R38	4.0	1	1	1				1	1	1			1	**0**	**0**	1	10	20.00

Bold = misclassified; underlined = SD > 18; 0 = non-irritant in vitro*; 1 = irritant* in vitro.

Table 12: EpiDerm: summary of specificity and sensitivity, considering results either of all determinations or only of three valid determinations

Laboratory	Specificity (%)		Sensitivity (%)	
	All runs	Three valid runs	All runs	Three valid runs
ZEBET	90.90	93.60	56.00	60.90
IIVS	78.80	89.30	60.00	60.00
BASF	87.80	80.00	56.00	61.10
All runs (individual classification)	84.76 (328 runs)	88.76 (267 runs)	56.32 (261 runs)	60.11 (183 runs)
All runs (median classification)	83.83 (99 runs)	87.66 (89 runs)	57.33 (75 runs)	60.66 (61 runs)

Figure 4: MTT: Curves for sensitivity, specificity and for their sum, depending on the *in vitro* Prediction Model (PM) threshold [%], when considering all valid EpiDerm and EPISKIN determinations



——— = *specificity;* ——— = *sensitivity;* - - - - - - = *sum of sensitivity and specificity.*

Table 13: Summary of the evaluation of within-laboratory variability for EPISKIN, in three laboratories

Variability measure	L'Oréal		Unilever		Sanofi	
	All runs	Three valid runs	All runs	Three valid runs	All runs	Three valid runs
Sample size	58	55	58	56	58	54
Number of chemicals with SD >18	4	2	7	5	5	2
Proportion of identically-classified chemicals	52/58	52/55	52/58	53/56	50/58	50/54

At L'Oréal, a total of 178 tests were conducted on the 58 chemicals; four chemicals were tested four times, and ten tests did not meet the variability criterion. Unilever performed 187 tests, 13 of which did not meet the variability quality criterion, so retesting was performed, once only. At Sanofi, 182 tests were performed, eight of which did not meet the variability criterion, so retesting was conducted, once only.

4.2.2.2. Within-laboratory variability

The results of the within-laboratory variability of all applied measures are summarised for the three laboratories in Table 13. The numbers of test runs, and the total test runs providing three valid runs, were similar.

Only minor differences between the laboratories became evident, when the numbers of chemicals failing to meet the quality criterion of an SD of below 18% were assessed.

The numbers of chemicals not consistently classified in the test runs in a given laboratory were three or four.

A comparison of the results for the 18 chemicals tested by L'Oréal in Phase 1 and Phase 2 was performed for chemicals with three valid runs. This information adds to the assessment of the within-laboratory variability, since the time difference between the determinations was up to twelve months. All the tests were conducted by the same operator, but for Phase 1 and Phase 2, separate sets of test chemicals were provided by the distributor. When a *t*-test with a significance level of 1% was applied to the run data for each of the chemicals, it was found that there were no significant differences between the results obtained in the two phases. Testing the mean viabilities of the two phases for the 18 chemicals by using a paired *t*-test, resulted in a non-significant *p* value of 0.458, which indicated good within-laboratory reproducibility.

4.2.2.3. Between-laboratory variability

The first measure of between-laboratory variability was the SD of the means of the runs per laboratory (Table 14). In relation to the variability criterion, Table 14 shows that the results for nine chemicals did not meet this acceptance criterion.

Applying the second measure — the proportion of identical chemical classifications, taking the median classification per laboratory into account — 51 of 57 chemicals (chemical 53 was excluded, as no laboratory produced three valid runs) were classified identically. Seven of them were characterised by three valid runs in two laboratories only. The test chemicals for which reproducible data were not produced according to this measure, also did not have reproducible results in the between-laboratory SD. Considering all runs, 8 out of 58 chemicals were not consistently classified. In a pair-wise comparison of the classifications obtained in the individual laboratories, the concordance of classifications was 50/58 = 86.2% for L'Oréal–Unilever, 56/58 = 96.6% for L'Oréal–Sanofi and 52/58 = 89.7% for Unilever– Sanofi.

A detailed outline of the within-laboratory and between-laboratory variability of the EPISKIN model in the SIVS, is provided on the ECVAM website (21).

4.2.2.4. Predictive capacity

The predictions obtained with the EPISKIN model were analysed in two ways, taking into account the median of all the available test runs, or the median of three valid runs. However, differences between these approaches were minor. As an example, the EPISKIN test outcomes for two parameters, specificity and sensitivity, at L'Oréal are presented in Table 15, for each single run and for the summarising approach. In addition to sample sizes, the exact lower 5% confidence limits are given.

Irrespective of analytical method, the specificity was always well above 80%. Nevertheless, the analysis of valid runs shows that test runs meeting the variability criterion are characterised by a slightly higher specificity value. A similar effect was observed for sensitivity, which ranged from 72.0 to 76.0%, depending on the mode of analysis.

To compare the misclassified chemicals in the three EPISKIN laboratories, the chemicals which

Table 14: EPISKIN: between-laboratory variability — the standard deviation of mean MTT values

	All runs				Three valid runs			
No.	**L'Oréal**	**Unilever**	**Sanofi**	**SD**	**L'Oréal**	**Unilever**	**Sanofi**	**SD**
1	5.67	4.66	3.92	0.88	5.67	4.66	3.92	0.88
2	5.95	4.02	4.40	1.02	5.95	4.02	4.40	1.02
3	26.22	18.88	46.79	14.47	26.22	11.90	46.79	17.54
4	7.31	6.45	6.94	0.43	7.31	6.45	6.94	0.43
5	54.90	10.96	21.62	**22.92**		10.96	10.17	0.56
6	27.66	51.20	63.07	**18.03**	24.95	51.20		**18.56**
7	62.63	15.52	44.36	**23.75**	62.63	15.52	44.36	**23.75**
8	51.51	13.91	34.69	**18.83**	51.51	11.66	34.69	**20.01**
9	107.13	100.53	104.36	3.31	107.13	98.58	104.36	4.36
10	101.13	99.53	101.33	0.98	101.13	99.53	101.33	0.98
11	96.82	96.39	102.14	3.20	96.82	96.39	102.14	3.20
12	111.44	90.87	105.37	10.57	111.44	90.87	105.37	10.57
13	4.97	4.84	13.44	4.93	4.97	4.84	13.44	4.93
14*								
15	15.16	2.75	6.40	6.38	14.72	2.75	6.40	6.14
16	99.00	95.95	102.33	3.19	99.00	95.95	102.33	3.19
17	10.81	4.44	9.73	3.41	10.81	4.44	9.73	3.41
18	11.25	21.13	27.93	8.39	11.25	10.65		0.43
19	114.34	109.48	98.18	8.29	114.34	109.48	101.48	6.50
20	24.37	8.53	42.48	16.99	24.37	8.53		11.21
21	103.47	107.00	104.22	1.86	103.47	107.00	104.22	1.86
22	95.31	75.04	92.54	10.99	95.31	75.04	92.54	10.99
23	52.04	11.17	76.12	**32.83**		7.48	77.01	**49.16**
24	106.24	93.56	90.59	8.31	106.24	93.56	97.44	6.50
25	105.32	93.24	103.08	6.43	105.32	93.24	103.08	6.43
26	7.02	31.30	2.60	15.45	7.02	31.30	2.60	15.45
27	81.03	21.48	89.54	**37.09**	81.03	11.21	89.54	**42.98**
28	118.51	116.10	113.74	2.39	118.51	116.10	113.74	2.39
29	11.09	6.20	9.01	2.46	11.09	6.20	9.01	2.46
30	81.90	67.94	99.92	16.03	81.90	67.70	99.92	16.15
31	12.05	11.43	7.78	2.31	12.05	11.43	7.78	2.31
32	99.72	111.79	102.27	6.36	99.72	111.79	102.27	6.36
33	103.99	102.32	111.54	4.91	103.99	102.32	111.54	4.91
34	99.85	101.87	94.56	3.78	99.85	100.49	94.56	3.26
35	121.85	112.38	114.87	4.91	121.85	112.38	114.87	4.91
36	100.41	81.58	82.39	10.65	100.41	80.46	86.96	10.17
37	9.15	9.62	11.09	1.01	9.15	9.62	11.09	1.01
38*								
39	103.99	90.33	101.08	7.19	103.99	90.33	101.08	7.19
40	9.47	17.40	11.04	4.20	9.47	17.40	4.30	6.59
41	96.39	88.60	91.96	3.91	96.39	88.60	91.96	3.91
42	100.30	93.40	95.25	3.57	100.30	101.39	95.25	3.28
43	53.55	31.32	48.18	11.60	53.55	31.32	46.94	11.42
44	92.55	83.03	95.99	6.72	92.55	77.80	95.99	9.67
45	11.71	9.79	8.83	1.46	11.71	9.79	8.83	1.46

* = *confidential chemical; bold = SD > 18.*

Table 14: continued

No.	All runs				Three valid runs			
	L'Oréal	Unilever	Sanofi	SD	L'Oréal	Unilever	Sanofi	SD
46	12.36	7.55	4.54	3.94	12.36	7.55	4.54	3.94
47	14.82	31.50	13.33	10.08	14.82	31.50	13.33	10.08
48	95.82	89.02	93.43	3.45	95.82	89.02	93.43	3.45
49	95.69	84.85	96.59	6.53	95.69	84.85	96.59	6.53
50	116.18	86.67	112.93	16.18	116.18	93.78	112.93	12.10
51	85.67	84.20	66.65	10.59	85.67	84.20	66.65	10.59
52	69.37	58.06	85.07	13.56	77.95	65.37	85.07	9.97
53	68.90	57.50	76.04	9.35			76.04	
54	66.43	40.40	55.86	13.09	66.43	40.40	55.86	13.09
55	52.96	6.26	53.77	**27.20**	52.96	6.26		**33.02**
56	12.74	87.75	14.19	**42.89**	12.74		14.19	1.02
57	97.97	100.56	101.49	1.83	97.97	100.56	101.49	1.83
58	7.12	5.89	6.90	0.65	7.12	5.89	6.90	0.65
59	21.91	14.76	17.12	3.64	21.91	14.76	17.12	3.64
60	88.36	6.34	78.81	**44.85**	88.36	6.34	78.81	**44.85**

** = confidential chemical; bold = SD > 18.*

were misclassified at least once in one of the laboratories, are shown in Table 16. Ten of the 33 non-irritant chemicals (30%), which are not classified as irritant *in vivo*, were misclassified as "irritant (R38)" at least once, while 12 of the 25 irritant (R38) chemicals (48%) were misclassified at least once as "non-irritant". Three of the non-irritant chemicals were consistently misclassified as irritants in all the runs in all the laboratories, and three of the irritant chemicals were classified as non-irritants in all the runs, in all the laboratories.

The predictive capacity of the EPISKIN test for classifying test chemicals according to their irritation potential in the three laboratories of the SIVS is given in Table 17, which summarises specificity and sensitivity over all the runs in each laboratory, according to the two approaches outlined above.

In another analysis, sample sizes, as well as specificity and sensitivity, from all the runs in all the laboratories, were calculated for the two approaches,

Table 15: Example of predictive capacity of EPISKIN at the lead laboratory, L'Oréal — specificity and sensitivity, taking into account sample sizes and 5% lower confidence limits for each valid run, and summaries for all runs

		Specificity			Sensitivity		
		n	%	LB–5%	n	%	LB–5%
Run 1	All	33	81.8	67.2	25	72.0	53.8
	Valid	30	83.3	68.1	24	75.0	56.5
Run 2	All	33	81.8	67.2	25	72.0	53.8
	Valid	31	83.9	69.0	24	75.0	56.5
Run 3	All	33	84.8	70.8	25	72.0	53.8
	Valid	30	90.0	76.1	23	73.9	54.9
All runs (median)		33	81.8	67.2	25	72.0	53.8
Three valid runs (median)		31	83.9	69.0	24	75.0	56.5

LB–5% = lower boundary of 95% confidence interval.

Table 16: EPISKIN: summary of chemicals which were misclassified at least once

No.	EU class	Dominant median	L'Oréal run 1	L'Oréal run 2	L'Oréal run 3	L'Oréal run 4	Unilever run 1	Unilever run 2	Unilever run 3	Unilever run 4	Sanofi run 1	Sanofi run 2	Sanofi run 3	Sanofi run 4	Total number of runs	Misclassifying runs (%)
2	no label	0	**1**	**1**	**1**		**1**	**1**	**1**		**1**	**1**	**1**		9	100.00
17	no label	1.7	**1**	**1**	**1**		**1**	**1**	**1**		**1**	**1**	**1**		9	100.00
26	no label	0	**1**	**1**	**1**		**1**	**1**	**1**		**1**	**1**	**1**		9	100.00
8	no label	1.0	**1**	**1**	0		**1**	**1**	**1**	**1**	**1**	**1**	**1**		10	90.00
5	no label	1.0	0	**1**	**1**		**1**	**1**	**1**		0	**1**	**1**	**1**	10	80.00
7	no label	0	0	0	0		**1**	**1**	**1**		**1**	**1**	**1**		9	66.67
6	no label	0	**1**	**1**	**1**	**1**	**1**	0	0		0	0	**1**		10	60.00
52	no label	0.7	**1**	0	0	0	0	0	**1**	1	0	0	0		11	27.27
53	no label	0	0	0	0		0	**1**	0		**1**	0	0		9	22.22
54	no label	1.3	0	0	0		**1**	0	0		**1**	0	0		9	22.22
34	R38	2.0	**0**	**0**	**0**		**0**	**0**	**0**	**0**	**0**	**0**	**0**		10	100.00
49	R38	2.0	**0**	**0**	**0**		**0**	**0**	**0**		**0**	**0**	**0**		9	100.00
51	R38	2.0	**0**	**0**	**0**		**0**	**0**	**0**		**0**	**0**	**0**		9	100.00
27	R38	4.0	**0**	**0**	**0**		**0**	1	1	1	**0**	**0**	**0**		10	70.00
60	R38	2.0	**0**	**0**	**0**		1	1	1		**0**	**0**	**0**		9	66.67
23	R38	3.0	**0**	**0**	**0**	1	1	1	1	1	**0**	**0**	**0**	**0**	12	58.33
55	R38	2.0	1	**0**	**0**		1	1	1		1	**0**	**0**		9	44.44
56	R38	3.3	1	1	1		**0**	**0**	**0**		1	1	1		9	33.33
43	R38	2.0	**0**	1	1		1	1	1		**0**	1	**0**	1	10	30.00
20	R38	2.3	1	1	1		1	1	1		1	**0**	**0**	1	10	20.00
3	R38	2.7	1	1	1		1	1	1	1	**0**	1	1		10	10.00
18	R38	3.0	1	1	1		**0**	1	1	1	1	1	1		10	10.00

Bold = misclassified; underlined = SD > 18; 0 = non-irritant in vitro; *1 = irritant* in vitro.

Table 17: EPISKIN: summary of specificity and sensitivity, considering results either of all determinations or only of three valid determinations

	Specificity (%)		Sensitivity (%)	
Laboratory	**All runs**	**Three valid runs**	**All runs**	**Three valid runs**
L'Oréal	81.80	83.90	72.00	75.00
Unilever	78.80	81.30	84.00	87.50
Sanofi	81.80	81.30	68.00	72.70
All runs (individual classification)	79.74 (311 runs)	80.70 (285 runs)	73.73 (236 runs)	77.62 (210 runs)
All runs (median classification)	80.80 (99 runs)	82.15 (95 runs)	74.67 (75 runs)	78.56 (70 runs)

considering either all the individual classifications or the median classification for a given chemical. Taking all the individual classifications into account, a lower predictivity was obtained, when compared to the results obtained when three valid runs were considered. This indicates that the chances of misclassification are increased by high variability, e.g. when high variability exclusion criteria are not considered. Slightly increased parameter estimations and a similar pattern were obtained when summarising the median classifications.

Taking all the valid runs into account, similar performance parameters were obtained. Due to the strong dependencies between the data in terms of reproducibility, confidence boundaries were not calculated.

In Figure 3, the ROC curve summarises the performance of the PM for the MTT reduction endpoint in the three EPISKIN laboratories. As outlined in the Materials and Methods section, the angular shape of the ROC curve indicates the successful test optimisation to specifically distinguish the classification of the two classes of skin irritation. Moreover, a comparison of the ROC curves for the two skin model tests, shows that the PM of the EPISKIN model seems to be better suited to distinguishing between the two classes of skin irritation than is the PM developed for the EpiDerm model.

Figure 4 shows the sensitivity and specificity curves, as well as their summary curve. In the threshold range of between 24% and 77%, the sensitivity and specificity curves are almost parallel to the x-axis, and the sum of the two values remains approximately constant, i.e. larger than 1.55. The summary curve reaches a maximum of 1.593 at 55%, close to the threshold of 50% defined in the PM.

Finally, the performance of the EPISKIN test in predicting the three classes of the GHS for the set of 58 test chemicals was evaluated. To keep this analysis simple and disregarding reproducibility, only the median run classifications of chemicals with three valid runs for all laboratories were considered. Thus, the data set was confined to 165 test results obtained with 35 GHS irritants, 50 GHS mild-irritants and 80 GHS non-irritants. As described for Phase 1, no robust PM had been established, so a *post hoc* approach was used to derive a new PM. The new PM was based on a viability threshold of 30%, below which chemicals would be classified as GHS-irritants, and a 50% viability threshold, above which chemicals would be classified as GHS non-irritants. Chemicals characterised by viabilities between the two thresholds would be classified as GHS mild-irritants.

Application of this PM provided correct classifications for 88.6% of the GHS irritants, 6.0% of the GHS mild-irritants, and 88.6% of the GHS non-irritants; viabilities between 30% and 50% were indicated for only eight chemicals. This confirmed the indications from Phase 1 that the EPISKIN test is not able to correctly predict the classification of chemicals into the three GHS classes. That either high or low cell viabilities were found for the GHS mild-irritants, reflects the fact that the EPISKIN protocol was optimised for the European classification system. Interestingly, the threshold of 50%, which is the cut-off value in the PM for the European classification system, was also almost optimal for GHS classification, when discriminating GHS non-irritants from the other two classes.

A detailed outline of the predictive capacity of the EPISKIN model in the ECVAM SIVS, is provided on the ECVAM website (21).

4.2.3. EpiDerm IL-1α release data

The MT decided that IL1-α release should only be measured in substances with a non-irritant indication in the MTT reduction test, i.e. resulting in a viability of more than 50%. ZEBET submitted data for 43 chemicals, 42 of which triggered IL1-α release according to the agreed criteria. One of these chemicals was reproducibly indicated as irritant in the MTT test (with a mean viability of 43%), so it was therefore not included in the analysis of

the total set of 42 chemicals. The tests were performed on the stored supernatants from at least three runs, conducted by the operator who had performed all the MTT reduction tests. Chemicals producing high IL1-α concentrations (numbers 6, 7, 13, 15, 25, 44, 49 and 54) had to be diluted before they could be tested.

To assess intra-assay variability, SDs and CVs were calculated. Irrespective of the measure, substantial intra-assay variability was observed. The SD increased with increasing concentrations, while the CV decreased with increasing response levels.

The negative control provided an average IL1-α release of 40pg/ml (a minimum of 24 and a maximum of 50pg/ml; CV = 20.6%), while the positive control induced a mean IL1-α release of 919pg/ml (a minimum of 773, and a maximum of 1172pg/ml; CV = 30.19% and SD = 15.46). The CV was the preferred measure, since it did not reveal substantial concentration dependence.

Applying the threshold value of a three-fold increase for the classification of chemicals as irritant, reproducible data were not obtained for eight of the 42 chemicals. A comparison of the data for chemicals tested in both Phase 1 and Phase 2 provided additional information on within-laboratory reproducibility. It showed that the data from the two phases were very similar.

Of the 42 chemicals, four were correctly classified as positives, seven were "false negatives", nine were "false positives", and 22 were correctly classified as negatives. Taking also into account the 16 chemicals which had only been tested in the MTT test, the tiered testing strategy provided a specificity of 66.6% (22/33) and a sensitivity of 72% (15/18). Compared to the predictive capacity of the MTT data (specificity: 90.9%; sensitivity: 56.0%), the increase in sensitivity was accompanied by a severe loss in specificity. The MT concluded from this result that determination of the second endpoint IL1-α did not offer any advantage in addition to the primary endpoint, MTT reduction, so IL1-α release should not be determined in test samples produced in the two additional EpiDerm laboratories.

A detailed outline of the analysis of the predictive capacity of the second endpoint, IL1-α release, with the EpiDerm model in the SIVS, is provided on the ECVAM website (21).

4.2.4. EPISKIN IL-1α release data

Since the determination of IL1-α release was the second step in a tiered testing strategy, the culture media of 21 chemicals, which had provided a positive result in the MTT reduction test, were not tested for this second endpoint. Of the 24 "positive" chemicals, 17 were correctly classified as irritant, while four were wrongly classified as irritant.

Of the remaining 37 chemicals analysed in all three EPISKIN laboratories, eight were classified as R38 according to rabbit *in vivo* data, and 29 as non-irritant. The latter group included three chemicals which did not provide three acceptable MTT test results (numbers 5, 23 and 53); one of them was an irritant, and the two others were classified as non-irritant in the *in vivo* test.

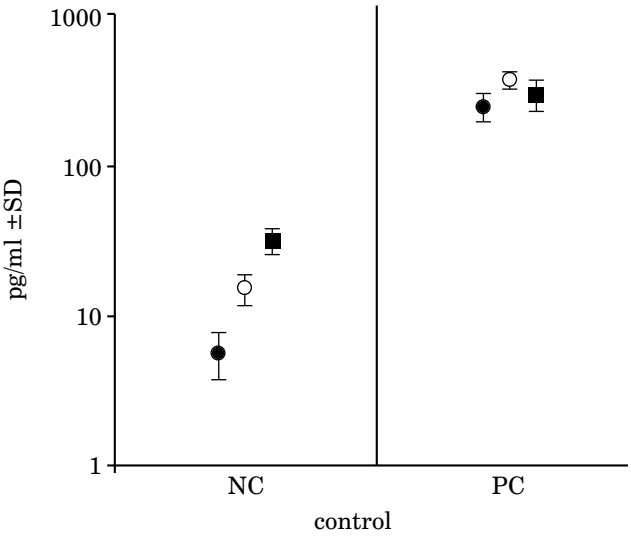
Upon analysing the Phase 2 data from the three EPISKIN laboratories, it became obvious that the PM originally suggested (5) and applied in Phase 1, which was based on the five-fold increase above the NC value, could not be applied to the Phase 2 data. This was due to the significant differences between the negative control values in the three laboratories (see Figure 5), which may indicate problems with the standardisation of the determination of this endpoint.

As a result of these considerations, a modification of the PM was developed, as described in section 4.2.4.3.

4.2.4.1. Results obtained in the L'Oréal laboratory

L'Oréal determined IL1-α release into the EPISKIN culture media of 39 test chemicals, for which negative results had been obtained in the MTT reduction test. Since two confidential chemi-

Figure 5: EPISKIN: summary of IL1-α data for the negative and positive controls in the three laboratories, on a logarithmic scale



● = *L'Oreál;* ○ = *Unilever;* ■ = *Sanofi.*

NC = negative control (PBS);
PC = positive control (5% SDS).

cals subsequently had to be excluded, the data obtained with 37 chemicals were analysed (Table 18). Although positive results had been obtained in two of three MTT determinations for two of the test chemicals (numbers 8 and 43), the mean MTT values of all the runs were higher than 50%. Moreover, for two chemicals (numbers 23 and 52), the MTT test had been repeated due to high variability of the IL1-α release data.

To assess the intra-assay variability, SDs and CVs were calculated. Regardless of the measurement made, substantial intra-assay variability was observed. The SD increased with increasing concentrations, while the CV decreased with increasing response levels. The negative control provided an average IL1-α release of 5.8pg/ml (a minimum of 3.0 and a maximum of 8.6pg/ml; CV = 34.9%), while the positive control induced an average IL1-α release of 254pg/ml (a minimum of 183 and a maximum of 335pg/ml; CV = 21.9%). The within-laboratory variability was determined via the SD and the CV. Furthermore, a 1-way ANOVA was calculated for each chemical on logarithmically-transformed data. The CV was the preferred measure, since it did not reveal substantial concentration dependence. Applying the threshold value of five-fold increase for the classification of irritating properties, reproducible data were not obtained for three of the 37 chemicals (numbers 5, 49 and 51). A comparison of the IL1-α release data obtained with the chemicals tested in both Phase 1 and Phase 2 showed that the data from the two phases were very similar.

Since the determination of IL1-α release was the second step in a tiered testing strategy, the culture media of 21 chemicals, which had provided a positive result in the MTT reduction test, were not tested for this second endpoint. In this group, 17 chemicals were correctly classified as R38, while four were wrongly classified as positive. Of the remaining 37 chemicals, eight had been classified as R38 according to rabbit *in vivo* data, and 29 had the no label classification. The latter group included three chemicals, for which three acceptable MTT test results were not available (numbers 5, 23 and 53); one of them was an irritant and the two others were classified as non-irritant.

Table 18 summarises the data obtained at L'Oréal in the EPISKIN test with two PMs, one based on the five-fold increase compared to the negative control criterion, and the second on 1-way ANOVA and a Dunnett's post test for comparing the individual results for response to test chemicals and to the negative control. The test results for each chemical were combined: a chemical was classified as irritant, when the mean increase was larger than five-fold, or when two runs for a chemical provided a significantly ($p < 0.05$) higher response than the negative control (log-transformed data). Since the overall classifications were identical for the two PMs, the predictive capacity obtained with EPISKIN, presented in Table 18, is only for the PM based on the five-fold increase. Of the 37 chemicals, six were correctly classified as positive, two were "false negatives", five were "false positives", and 24 were correctly classified as negative. The two chemicals for which positive results were obtained in two runs in the MTT reduction test, also provided a positive IL1-α release response. Also taking into account the 21 chemicals which had only been tested in the MTT reduction test, the tiered testing strategy provided a specificity of 72.7% (24/33) and a sensitivity of 92% (23/25). Compared to the predictivity of the MTT reduction endpoint, the increase in sensitivity (20.0%) was accompanied by a loss of specificity (11.7%). Since the determination of the second endpoint considerably improved the predictivity of the EPISKIN test, the MT recommended that the two additional EPISKIN laboratories should also measure IL1-α release in their test samples.

4.2.4.2. Between-laboratory variability in the three EPISKIN laboratories

A comparison of the positive control (PC) and negative control (NC) values for IL1-α release was performed on ten NC and PC values from each laboratory, and the results are summarised in Figure 6. The NC values were distributed over a fairly broad dose range, from 2.95 to 46pg/ml (maximum/minimum ratio = 15), while the PC values showed a greater reproducibility in the dose range from 199 to 421pg/ml (maximum/minimum ratio = 2). The NC values were significantly different ($p < 0.001$) in the three laboratories, while the PC values were more reproducible, except for the results measured at L'Oréal and Unilever, which showed a significant difference ($p < 0.001$). A mean NC value of 32.1pg/ml was found at Sanofi, and 15.4pg/ml at Unilever, while the lowest value of 5.8pg/ml was obtained at L'Oréal.

With regard to reproducibility, the results obtained with chemicals tested in at least two laboratories are presented in Table 19. Reproducibility was assessed by a comparison of the mean IL1-α values in the three laboratories, which revealed a mean SD of 14.37 and a mean CV of 50%. The large between-laboratory variability reflected the insufficient standardisation of the conditions for determining the second endpoint.

4.2.4.3. Improving the PM for the IL-1α release endpoint

The significant differences in the NC values in the three laboratories did not permit the application in Phase 2 of the original PM proposed by Cotovio *et al*.

Table 18: EPISKIN IL1-α determinations at L'Oréal — within-laboratory variability of IL1-α values

No.	Chemical class	Run 1	Run 2	Run 3	Run 4	Mean	SD	CV [%]
5	no label	9.32	31.97	17.73		19.67	11.45	58.20
7	no label	68.86	56.55	50.27		58.56	9.46	16.15
8	no label	38.48	59.71	23.06		40.42	18.40	45.53
9	no label	8.47	4.40	5.45		6.11	2.11	34.60
10	no label	5.34	8.94	12.91		9.06	3.79	41.78
11	no label	6.04	11.19	6.95		8.06	2.75	34.10
12	no label	10.21	12.90	10.37		11.16	1.51	13.52
16	no label	11.53	12.06	17.18		13.59	3.12	22.96
19	no label	11.62	7.58	9.18		9.46	2.03	21.51
21	no label	6.42	11.33	9.80		9.18	2.51	27.36
22	no label	13.62	10.73	17.03		13.79	3.15	22.86
23	R38	53.95	86.89	93.69	104.50	84.76	21.78	25.70
24	no label	13.30	11.13	11.52		11.98	1.16	9.65
25	no label	17.21	16.89	11.19		15.10	3.39	22.44
27	R38	90.05	83.32	117.19		96.85	17.93	18.51
28	no label	8.82	9.97	8.60		9.13	0.74	8.06
30	no label	16.67	10.43	18.10		15.07	4.08	27.07
32	no label	3.31	5.27	6.95		5.18	1.82	35.19
33	no label	13.69	14.92	12.50		13.70	1.21	8.83
34	R38	3.87	14.01	6.71		8.20	5.23	63.82
35	no label	3.99	4.24	4.10		4.11	0.13	3.05
36	no label	7.84	10.40	18.56		12.27	5.60	45.64
39	no label	10.07	13.61	9.46		11.05	2.24	20.28
41	no label	6.89	6.94	10.35		8.06	1.98	24.61
42	no label	12.57	7.18	8.81		9.52	2.76	29.04
43	R38	39.46	107.30	122.97		89.91	44.39	49.37
44	no label	7.93	13.29	8.81		10.01	2.87	28.72
48	no label	12.44	9.79	9.96		10.73	1.48	13.82
49	R38	45.17	11.15	64.94		40.42	27.21	67.31
50	no label	17.12	8.98	11.38		12.49	4.18	33.48
51	R38	50.40	21.99	28.36		33.58	14.91	44.39
52	no label	24.44	29.72	24.62	99.43	44.55	36.67	82.30
53	no label	60.45	54.25	40.96		51.89	9.96	19.19
54	no label	86.93	26.53	116.36		76.61	45.80	59.78
55	R38	162.64	59.91	142.60		121.72	54.46	44.74
57	no label	4.69	15.11	8.52		9.44	5.27	55.83
60	R38	6.30	5.40	3.37		5.02	1.50	29.88

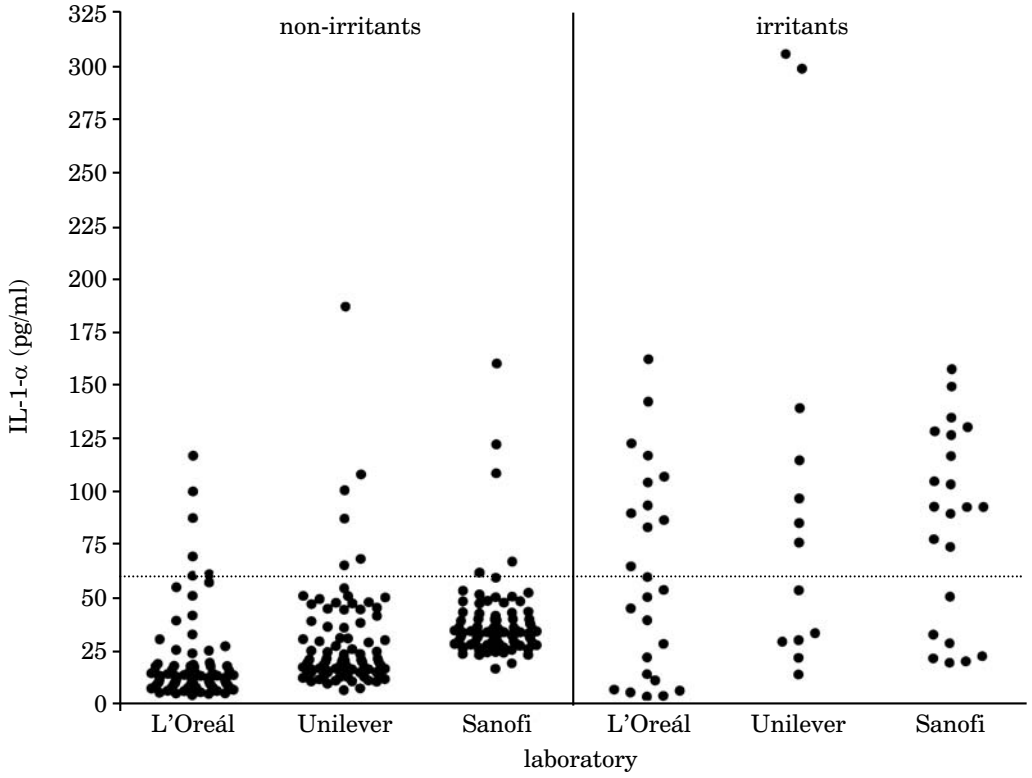
Chemicals characterised by MTT-variability of SD >18 are underlined.

(5) and applied in Phase 1 (based on a 5-fold increase, compared to the NC values). Therefore, a PM was applied, which takes into account only the absolute amount of IL1-α released into the culture medium. With this PM, the significant differences in the NC values did not contribute to the classification.

Since irritating chemicals tended to induce the release of higher amounts of IL1-α, the usefulness of this PM was evaluated in more detail by ROC analysis, in order to identify the optimal threshold value. When IL1-α release is used as single endpoint, the highest summary values of specificity and sensitivity were obtained for threshold values of between 48 and 54pg/ml, with a maximum at a threshold value of 51pg/ml.

The result of applying this PM to the results of all the available runs in each of the laboratories is given in Table 19. In comparison to the classification

Figure 6: EPISKIN: IL1-α release, determined with the MTT-negative substances tested in three laboratories



Dotted line = threshold for IL1-α (60 pg/ml).

obtained with the MTT alone at L'Oréal, four additional chemicals were correctly classified as irritant, while three non-irritants were falsely classified as irritant. At Unilever, in addition to the MTT classification, three chemicals were correctly identified as irritant, and three were "false positives". At Sanofi, in comparison with the MTT results, five additional chemicals were correctly classified as irritant, and three were "false positives".

Table 19 also demonstrates that the variability issues persisted. At L'Oréal, three chemicals (numbers 50, 49 and 51) led to unusually low IL1-α release values. Nevertheless, this PM provided good between-laboratory reproducibility for predicting non-irritant chemicals: 22 of the 25 chemicals tested in three laboratories were classified consistently. In contrast, only one of the three irritant chemicals was classified consistently.

The overall predictivity of the strategic combination of the two endpoints, MTT reduction and IL1-α release, was evaluated by taking into account the mean IL1-α release values shown in Table 19. The results are summarised in Figure 7. When the optimum IL1-α release threshold of the stand-alone approach, 51pg/ml, was used, a specificity of 73.7% (73/99) and a sensitivity of 90.7% (68/75) were obtained. However, the maximum of the sum of the two parameters is correlated to a PM threshold of between 59 and 60pg/ml. When applying this threshold as the second endpoint in the tiered PM, a specificity of 78.8% (78/99) and a sensitivity of 90.7% (68/75) were obtained.

Thus, the use of the second endpoint, IL1-α release, considerably improved the overall predictive capacity of the PM, in comparison to a PM based on application of only the MTT reduction endpoint, and also improved the balance between specificity and sensitivity for the Phase 2 data. However, since the new PM for IL1-α release was only optimised after the performance the study, it should be confirmed by testing an additional set of different test chemicals.

4.2.5. The overall predictivity of the two skin models

To evaluate the predictive capacity of the EPISKIN model, the two endpoints were considered in the

Table 19: EPISKIN: the mean IL1-α values of all determinations per laboratory, and the between-laboratory variability of IL1-α values, in three laboratories

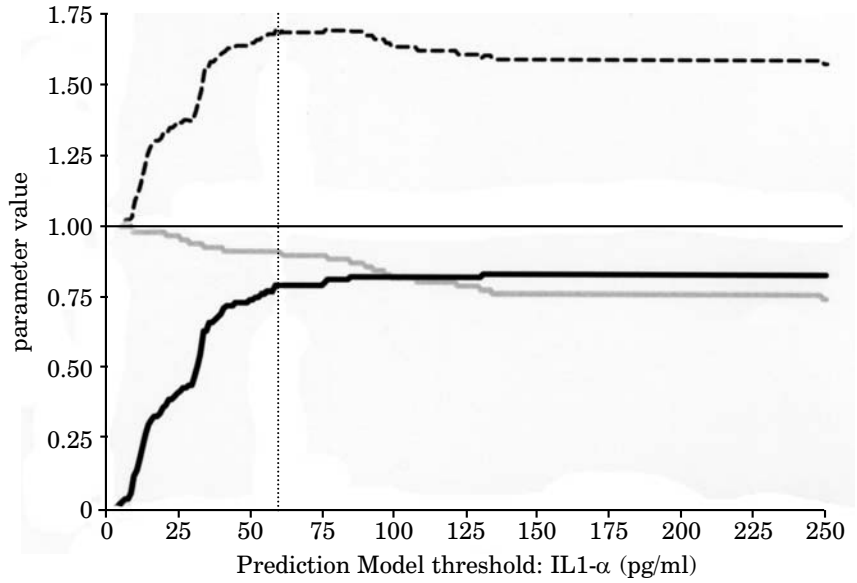
		Mean IL1-α of all runs [pg/ml]					
No.	**Chemical class (EU)**	**L'Oréal**	**Unilever**	**Sanofi**	**Mean**	**SD**	**CV (%)**
5	no label	19.67	n.d.	n.d.	n.a.	n.a.	n.a.
6	no label	n.d.	37.64	49.36	43.50	8.29	19.06
7	no label	**58.56**	n.d.	n.d.	n.a.	n.a.	n.a.
8	no label	40.42	n.d.	n.d.	n.a.	n.a.	n.a.
9	no label	6.11	23.55	25.29	18.31	10.60	57.88
10	no label	9.06	35.10	32.25	25.47	14.28	56.07
11	no label	8.06	14.80	33.73	18.86	13.31	70.56
12	no label	11.16	33.16	41.72	28.68	15.76	54.96
16	no label	13.59	13.95	35.63	21.05	12.62	59.94
19	no label	9.46	19.54	30.64	19.39	10.65	54.93
21	no label	9.18	32.91	39.01	27.03	15.76	58.30
22	no label	13.79	21.00	33.54	22.78	9.99	43.86
24	no label	11.98	16.33	31.18	19.83	10.07	50.78
25	no label	15.10	33.20	33.95	27.42	10.67	38.92
28	no label	9.13	14.30	30.25	17.89	11.01	61.53
30	no label	15.07	32.38	38.14	28.53	12.01	42.09
32	no label	5.18	11.32	33.86	16.79	15.10	89.95
33	no label	13.70	31.47	32.69	25.95	10.63	40.96
35	no label	4.110	18.80	33.62	18.84	14.76	78.33
36	no label	12.27	21.68	31.88	21.94	9.81	44.71
39	no label	11.05	12.63	30.48	18.05	10.79	59.77
41	no label	8.06	12.84	26.05	15.65	9.32	59.54
42	no label	9.52	20.04	24.60	18.05	7.73	42.82
44	no label	10.01	31.32	35.77	25.70	13.77	53.58
48	no label	10.73	14.28	33.20	19.40	12.08	62.26
50	no label	12.49	**84.17**	53.33	50.00	35.96	71.92
52	no label	44.55	**75.61**	55.23	58.46	15.78	26.99
53	no label	**51.89**	**58.69**	40.01	50.20	9.45	18.83
54	no label	**76.61**	n.d.	**130.38**	103.5	38.02	36.74
57	no label	9.44	13.55	28.47	17.15	10.01	58.36
23	R38	**84.76**	n.d.	**98.74**	91.75	9.89	10.78
27	R38	**96.85**	n.d.	**133.12**	114.99	25.65	22.31
34	R38	8.20	25.10	28.63	20.64	10.92	52.91
43	R38	**89.91**	n.d.	n.d.	n.a.	n.a.	n.a.
49	R38	40.42	**92.53**	**107.87**	80.27	35.36	44.05
51	R38	33.58	**60.35**	**76.71**	56.88	21.77	38.28
55	R38	**121.72**	n.d.	**129.33**	n.a.	n.a.	n.a.
56	R38	n.d.	**248.56**	n.d.	125.53	5.38	4.29
60	R38	5.02	n.d.	20.97	13.00	11.29	86.85

n.d. = not done; n.a. = not applicable. Classification "R38/irritant", considering 50pg/ml (bold + underlined) or 60 pg/ml (bold) as the threshold.

strategic manner described above. The resulting predictive capacities for MTT reduction alone, and in combination with IL1-α release, are shown in Table 20. A substantial increase in sensitivity by 16% (from 74.7% to 90.7%) was achieved, which was accompanied by a small loss in specificity of 1.2% (from 80.0% to 78.8%).

The overall results of the SIVS are summarised in Table 20, and clearly demonstrate that the use of the second endpoint improved the predictive capac-

Figure 7: EPISKIN: Receiver operating characteristic (ROC) curves for sensitivity, specificity and for their sum, depending on the *in vitro* Prediction Model (PM) threshold for IL1-α, when applied in combination with MTT



——— = *specificity;* ——— = *sensitivity;* - - - - - - = *sum of sensitivity and specificity; vertical dotted line = threshold for IL1-α (60pg/ml).*

ity of the EPISKIN test. With regard to the EpiDerm test, the use of the MTT reduction endpoint alone provided a specificity of 84.7% and a sensitivity of 56.3%, while the use of the additional endpoint did not improve the predictive capacity.

Finally, the MTT reduction and IL1-α release results were used to establish a PM for the GHS classification system. A dataset of 162 chemicals was used, which were backed by three valid MTT reduction runs. It became evident that the measurement of the effects of chemicals on cell viability in the skin models, did not permit discrimination between non-irritants and mild-irritants according to GHS classification.

Table 20: Overall performance characteristics obtained with the EpiDerm and EPISKIN assays in the ECVAM SIVS, for 58 test chemicals, tested in three laboratories (n = 174)

Test	Reproducibility (based on identical classifications)	Predictive Ability		
		Accuracy	**Sensitivity** (I)	**Specificity** (NI)
EpiDerm (MTT)	**Acceptable** — within-lab: 96% between-lab: 89%	**72.0%**	56.3%	**84.7%**
EPISKIN (MTT)	**Acceptable** — within-lab: 94% between-lab: 90%	**78.0%**	**74.7%**	**80.8%**
MTT + IL-1α	IL-1α endpoint more variable than MTT	**83.0%**	**90.7%**	**78.8%**

All runs (individual classifications) were considered for this summary table (for comparison, see Tables 12 and 17). Bold font indicates values meeting the acceptance criteria set by the management team.

4.2.6. The predictive capacity of the two skin model test for subsets of chemicals

As described above, in order to provide a balanced data set, the test chemicals used in the SIVS study were carefully selected by the CSSC from the NCD, as well as from additional sources, such as the ECETOC, TSCA and CIR databases. The existing chemicals selected from the ECETOC or TSCA databases were considered to be readily commercially available. Nine of these 25 chemicals were classified as skin irritants (Table 5). Of the 33 new chemicals selected from the NCD (Table 4), which is concerned with chemicals which are less readily available from major manufacturing or distribution sources, sixteen were classified as skin irritants.

The analysis conducted, was based on all the available independent experiments, and the median classification of these experiments for each chemical in a given laboratory was calculated. The respective specificities and sensitivities for the two skin models are summarised in Table 21.

For EpiDerm, with MTT reduction as the endpoint, a specificity of 87.5% and a sensitivity of 33.3% were obtained for the existing chemicals, and a specificity of 80.4% and a sensitivity of 70.8% for the NCD chemicals.

For EPISKIN, with MTT reduction as the endpoint, a specificity of 79.2% and a sensitivity of 74.1% were obtained for the existing chemicals, and a specificity of 82.4% and a sensitivity of 75.0% for the NCD chemicals.

Finally, when EPISKIN was used with both the MTT reduction and IL1-α release endpoints, a specificity of 83.8% and a sensitivity of 88.9% were obtained for the existing chemicals, and a specificity of 74.5% and a sensitivity of 91.7% for the NCD chemicals.

Thus, it can be concluded from the results summarised in Table 21, that the overall performance of the EPISKIN test was similar for the two subsets of chemicals, whether one or two endpoints were used. In contrast, in the case of the EpiDerm test, the sensitivity was significantly better with the new chemicals from the NCD, than with the existing chemicals from the ECETOC, TSCA and CIR databases.

5. Discussion

5.1. Observations on the *in vivo* data

Crucial criteria for the selection of chemicals in the SIVS were the availability and quality of the *in vivo* data. It was agreed that, in general, individual Draize scores had to be available for at least three rabbits. However, substances were also considered when, especially in case of the NCD chemicals, mean scores were reported for tests with more than three rabbits. Furthermore, the *in vivo* data had to be in line with the current European classification scheme (R38 *versus* no label). In this way, the *in vivo* classifications could be traced back to the respective *in vivo* data.

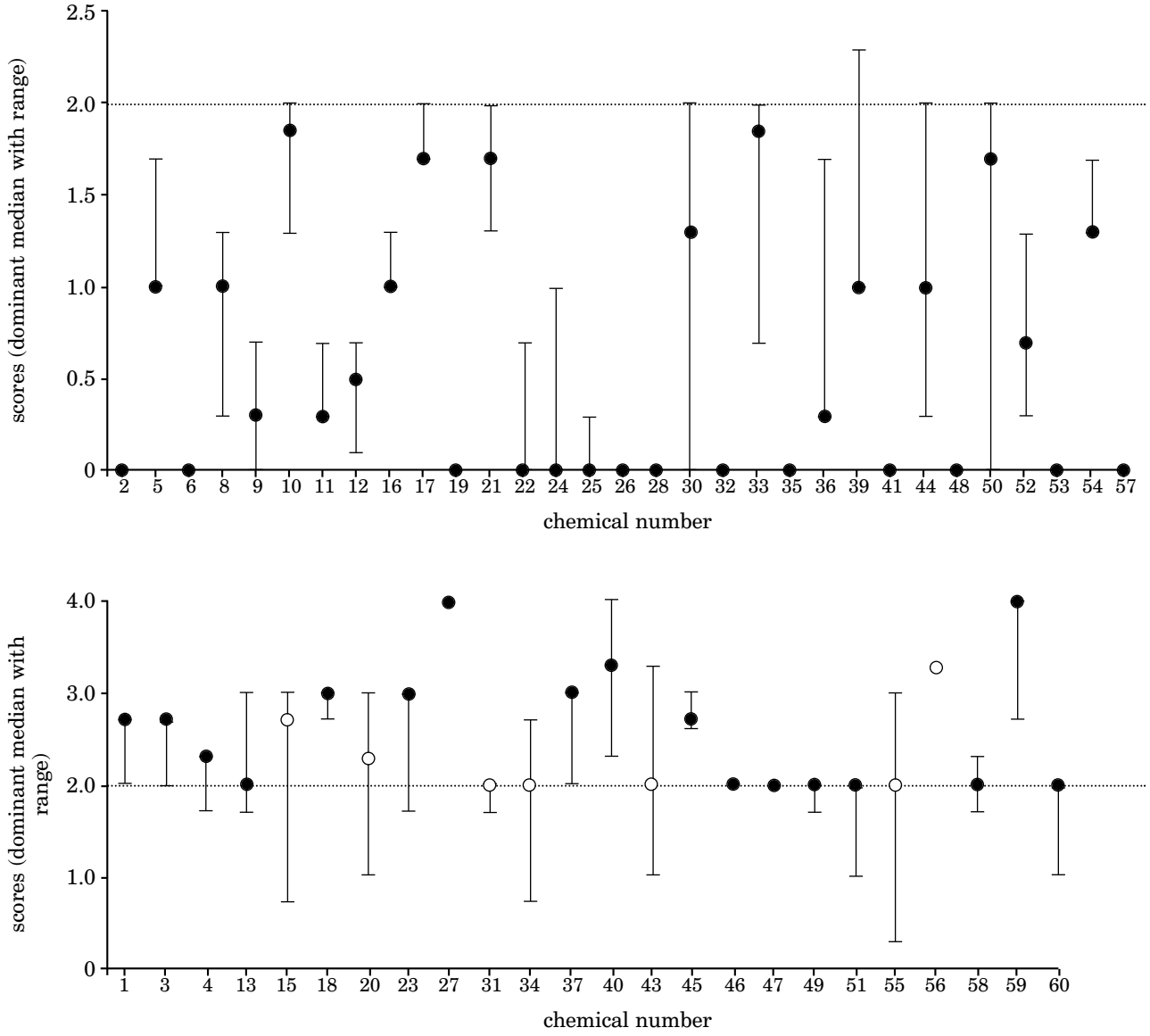
The dominating median values, with the respective range of individual scores, are shown for all the selected chemicals in Figure 8. It is evident that the selected chemicals displayed a range of *in vivo* scores, including values close to the classification threshold, which is 2. Special consideration should be given to chemicals which have *in vivo* scores close to this threshold, and which show a tendency to fall on one side or the other side of the threshold value, e.g. due to variability in responses of the animals. For example, chemicals with Draize scores of 0 for both endpoints (erythema and oedema) in all the treated animals (e.g. chemicals 6 and 19; Figure 8), can be classified as no label with greater certainty than can chemicals with a dominant median of 1.7, which present scores which cross or fall on the threshold of 2 (e.g. chemicals 21 and 50). Therefore, special care is necessary when interpreting the results of the study. This was taken into account when potential reasons for *in vitro* misclassification were considered in section 5.2, below.

Table 21: Summary of predictive capacities for existing and new chemicals

	EpiDerm		EPISKIN (MTT)		EPISKIN (MTT+IL1-α)	
	Specificity	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**
ECETOC, TSCA, CIR	87.5%	33.3%	79.2%	74.1%	83.8%	88.9%
NCD	80.4%	70.8%	82.4%	75.0%	74.5%	91.7%
Overall	84.7%	56.3%	80.8%	74.7%	78.8%	90.7%

Corresponding overall data taken from Table 20; existing chemicals from the ECETOC, TSCA and CIR data bases, see Table 5; new chemicals from the NCD, see Table 4.

Figure 8: *In vivo* Draize skin irritation test data in the rabbit, expressed as the dominant median with the corresponding range



● = *indicates that erythemal effects were dominant;* ○ = *indicates that oedemal effects were dominant.*

5.2. Chemicals misclassified in both the EpiDerm and EPISKIN models

The CSSC investigated the possible reasons for the misclassification of 21 chemicals in at least one of the participating laboratories (non-valid runs were not taken into account, because misclassification could have been due to technical issues in such cases). For that purpose, the following points were carefully investigated:

— The relationship between *in vitro* misclassifications and *in vivo* data.

— The value of QSAR analysis of misclassified chemicals via the use of descriptors, the use of expert systems in a weight-of-evidence approach, and the use of the BfR rulebase for skin irritation.

— The combined physico–chemical properties of the misclassified chemicals.

— The potential relationships between misclassification of chemicals and risk phrases related to human health effects other than skin irritation, including skin sensitisation (EU R43) and eye irritation (EU R41 and R36).

— Observations made by the participating laboratories on anomalies encountered in Phase 2 of the SIVS (not including potential technical and biological interactions with the tissue model and with the MTT reduction assay).

Some chemicals were misclassified in both the EPISKIN and EpiDerm models, while others were misclassified in one of the two models. In the summarising conclusions below, the chemicals were broken down into these three categories. In addition, it was considered whether they were misclassified in only one, or up to all, of the six participating laboratories.

Five chemicals were misclassified with both the EPISKIN and EpiDerm tests in all six laboratories: chemicals 17, 49 and 51 were identified as compounds with a variability in the *in vivo* response, close to the threshold of 2, and chemical 26 showed signs of possible test material deterioration. Chemical 34 (hexyl salicylate) did not show any clear reasons for misclassification, but it is a GHS mild-irritant, situated in the middle range of the *in vivo* scores. Furthermore, being a salt of salicylic acid, it may have an anti-inflammatory effect on the tissue.

Four chemicals were misclassified in four laboratories with both the EPISKIN and EpiDerm tests: chemical 8 showed signs of possible test material deterioration, and chemicals 2, 23 and 27 showed no discernible reasons for misclassification.

Three chemicals were misclassified in two or three laboratories with the EpiDerm test alone (chemicals 3, 18 and 43). No clear reasons were found for these misclassifications, but these chemicals all had lower melting points and higher boiling points than water, as well as positive Log K_{ow} values, tending towards the upper range of values, and low water solubility. Further investigation is recommended into the possibility of tissue-specific incompatibility with a certain sets of physico–chemical properties.

Finally, three chemicals were misclassified in two laboratories with the EPISKIN test alone (chemicals 5, 7 and 60). No clear reasons for misclassification were identified, so further investigation is recommended. In addition, the following observations were made:

1. No clear patterns for possible relationships between R43, R41 and R36 labelling and the overall *in vitro* misclassification for skin irritation were found (the chances of being misclassified were equal to, or less than, one in two).

2. No clear pattern emerged which related physico–chemical property descriptor (e.g. mp, bp, vp, ws, log P) to the correct or incorrect classification of chemicals.

3. The BfR Decision Support System (DSS) for skin and eye irritation (physico–chemical exclusion rules and structural inclusion rules) gave predictions for only four chemicals (two of which were correctly predicted, and two of which were false negatives), so there was no basis for drawing meaningful conclusions.

4. There could be a molecular structural rationale for the likelihood that an incorrect *in vitro* classification will be made; however such chemical alerts could not be identified with statistical significance, due to the small number of chemicals investigated.

5. The expert systems investigated in a weight-of-evidence approach were found to be good at predicting likely irritants, but less capable of predicting non-irritants. On the basis of the limited dataset which was analysed, it appeared that QSARs from the expert systems, and the use of read-across analogues, might be useful in a weight-of-evidence evaluation before testing commences. Such an evaluation could be complementary to the use of the *in vitro* tests, as relevant for further investigation and for the development of tiered testing strategies for skin irritation.

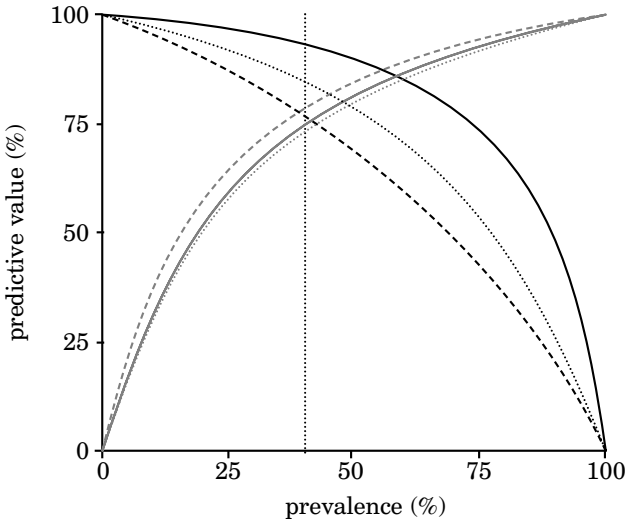
A detailed outline of the reasons for misclassifications is given in the *Report from the Chemicals Selection Sub-Committee (CSSC) to the Management Team on Potential Reasons for the Misclassification of Chemicals in the EPISKIN and EpiDerm Assays* (22).

5.3. The predictive value and regulatory use of the new test systems

To support the evaluation of the predictive capacity of the two tests, the negative predictive values (NPVs) and positive predictive values (PPVs) were calculated for the prevalence of skin irritants, which is the proportion of irritant chemicals in a defined population of chemicals. Prevalence, variability and regulatory classification for the skin irritation endpoint have recently been analysed for the available *in vivo* data for existing chemicals (17). According to this study, the prevalence of skin irritating chemicals among new chemicals is 8%, which must be considered in future evaluations of the tests.

For the EpiDerm test, a specificity of 84.7% and a sensitivity of 56.3% were assumed (see Table 20). For the sum of the two predictive values, a maximum of 1.524 was reached, at a prevalence of 40%; this is at the intersection of the two curves (Figure 9). The prevalence in this study (43.1% = 26/58) is indicated by the vertical line in Figure 9. In the context of the prevalence of new chemicals (17), which is about 8%, the use of EpiDerm as a stand-alone

Figure 9: EpiDerm and EPISKIN: curves for the negative and positive predicted values, over the entire range of prevalence



------ = *NPV EpiDerm (MTT only);*
------ = *PPV EpiDerm (MTT only);*
——— = *NPV EPISKIN both endpoints;*
——— = *PPV EPISKIN both endpoints;*
..... = *NPV EPISKIN (MTT only);*
..... = *PPV EPISKIN (MTT only);*

NPV = negative predictive value; PPV = positive predictive value; endpoints: MMT and IL-1α. The vertical dotted line indicates the prevalence of irritating and non-irritating chemicals in the study (43.1%). Note that this value is only 8%, among the new chemicals of the NCD of the ECB.

test would have a NPV of 96.2% and a PPV of 29.3%, i.e. only 3.9% of the negative results would be false negatives, but over 70% of the positives would be false positives.

For the EPISKIN test, assuming a specificity of 80.7% and a sensitivity of 74.7%, the sum of the two predictive values reached a maximum of 1.584 for a prevalence of 49%, at the intersection point of the two curves. Taking into account the prevalence of 8% for skin irritating chemicals, EPISKIN as a stand-alone test would have a NPV of 97.6% and a PPV of 25.9%. In such a scenario, only 2.4% of the negative results would be false negatives, but over 70% of the positive results would be false positives.

If the same calculations are made for the EPISKIN test with the two endpoints, and assuming a specificity of 78.8% and a sensitivity of 90.7%, a very high NPV of 99% and a low PPV of 27.1% would be obtained. As a consequence, only 1% of the negative results would be false negative, and 22.9% of the positive results would be false positives.

Thus, the information obtained in the SIVS shows that, for the prediction of skin irritation, negative results are significantly more reliable than positive results. However, as a general rule, positive results from validated *in vitro* tests are accepted for regulatory purposes, while negative results have to be confirmed by *in vivo* studies, as outlined in the OECD *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment* (10). In view of the high NPVs of the two *in vitro* skin tests, of more than 95%, and the low prevalence of skin irritating chemicals among new test chemicals, the OECD recommendations should be revised, both from the scientific and the animal welfare perspectives. In particular, negative data should be accepted when the new skin model *in vitro* skin irritation tests are used for regulatory purposes.

5.4. Critical evaluation of the results obtained in the SIVS

All of the SIVS data were submitted to ECVAM in 2006, and have been critically evaluated by an independent peer review panel (PRP) of the ECVAM Scientific Advisory Committee (ESAC). The PRP experts carefully analysed the data and submitted a consensus report to ECVAM. The SIVS MT responded to all of the points raised in the consensus report, and, at its 26th meeting in April 2007, the ESAC unanimously endorsed the following statement (12):

After a review of scientific reports and peer reviewed publications on the following range of in vitro *tests, which had been subjected to a full validation study:*

1. *EpiDerm (with MTT reduction and IL-1α release); and*
2. *EPISKIN (with MTT reduction and IL-1α release);*

of these, the EPISKIN method showed evidence of being a reliable and relevant stand-alone test for predicting rabbit skin irritation, when the endpoint is evaluated by MTT reduction, and for being used as a replacement for the Draize Skin Irritation Test (OECD TG 404 and Method B.4 of Annex V to Directive 67/548/EEC*) for the purposes of distinguishing between R38 skin irritating and no-label (non-skin irritating) test substances. At the present time, the IL-1α endpoint should be regarded as a useful adjunct to the MTT assay, as it has the potential to increase the sensitivity of the test, without reducing its specificity. This endpoint could be used to confirm negatives obtained with the MTT endpoint.*

At this time, due to its high specificity, the EpiDerm model reliably identifies skin irritants, but

negative results may require further testing (e.g. according to the tiered strategy, as described in the OECD TG 404). Improvement of the EpiDerm protocol should be made to increase the level of sensitivity.

This endorsement takes account of the dossiers prepared for peer review; the views of independent experts who evaluated the dossiers against defined validation criteria; supplementary submissions made by the Management Team; and the considered view of the Peer Review Panel appointed to oversee the process.

5.5. Performance standards for applying human skin models to *in vitro* skin irritation testing

After the SIVS was completed and the results had been accepted by the ESAC, performance criteria needed to be defined, which would have to be met by all skin models that are to be used for predicting skin irritation, in the light of the outcome of the SIVS. For this purpose, a document entitled *Performance Standards for Applying Human Skin Models to* In Vitro *Skin Irritation Testing* (23) has been drafted, and has been approved by the ESAC. It is aimed at establishing the basis by which proposed new test methods (sometimes referred to as "me-too" tests), which are based on similar scientific principles and measure or predict the same biological or toxic effects as the validated test methods, could be assessed for their accuracy and reliability for skin irritation testing purposes. It also addresses the extent to which the validation and acceptance criteria have to be met.

The three elements of the proposed performance standards are:

1. Minimum procedural standards that identify the essential structural, functional, and procedural components (e.g. the morphological structure and integrity of the test system, the proper controls, the biological identities of key components, and the expected biological responsiveness) of a validated test method.

2. The minimum procedural standards to be adhered to, to ensure that the proposed test method is based on the same concepts as the validated test method.

3. A list of recommended reference chemicals that should be used to assess the reliability and predictivity of the proposed test method. The list will include 20 commercially-available compounds, tested in the SIVS.

4. Specific test performance requirements: the reliability and predictivity that should be achieved by the proposed test method when the proposed reference chemicals are tested.

6. The SIVS and the Modules of the ECVAM Validation Approach

6.1. Module 1: Definition of test methods

6.1.1. Human skin model skin irritation tests (EPISKIN and EpiDerm)

The two human reconstituted skin model skin irritation tests evaluated in the SIVS are well-defined test methods that have undergone prevalidation and subsequent refinement. Before the start of the SIVS, the lead laboratories, L'Oréal and ZEBET, collaborated in the development of a common test protocol for the use of the MTT reduction assay. The SOPs for the two tests were as identical as was possible (in relation to experimental design, the application and rinsing procedures, the amount of test material applied per area of tissue, and the post-exposure incubation period before the determination of MTT reduction). They differed only in terms of model-specific treatment details, such as the conditioning of the tissues after transport, and the separation of the EPISKIN tissues from the thick collagen layer before performing the MTT test (5, 6). For the secondary endpoint, IL-1α-release, the SOP developed by L'Oréal (5) was used for both the EPISKIN and EpiDerm tests.

Identical acceptance criteria were defined for the two tests, based on the outcome of the concurrently tested PC (5% SDS) and NC (water or PBS), and on a maximum SD obtained from three replicate tissues treated identically. Assays providing data with an inter-tissue SD of 18% or more, were rejected as non-qualified, and were repeated. However, if a fourth run again produced a non-qualified result, no further repetition was performed.

6.1.2. Skin Integrity Function Test (SIFT)

The SIFT is a well-defined test, employing a PC (10% SDS) and NCs as acceptance criteria, which had undergone refinements since the ECVAM prevalidation study. However, because the SIFT did not progress to Phase 2 of the SIVS, literature references describing the refinements made are not given in this report. Nevertheless, further details of the method are given in the previous publications of the ECVAM skin irritation TF (3, 4).

6.1.3. Test chemical selection criteria

Existing chemicals proposed by ECETOC (15) were extensively used in the ECVAM prevalidation study

and in the subsequent test optimisation phases. New sources of test materials were therefore needed for the SIVS. Crucial criteria for the selection of chemicals were their availability and the high quality of *in vivo* data concerning them. The first source of chemicals used was the NCD of the ECB, which comprises data concerning "new" commercial chemicals registered after 1981, and for which skin irritation testing has been performed according to regulatory standards involving the use of official methods and GLP compliance. In addition, to obtain "existing" chemicals (i.e. chemicals in use before 1981) which would be readily available from major manufacturing and/or distribution sources, additional databases were surveyed, including the TSCA database maintained by the US EPA, and the ECETOC database. Chemicals used in the previous optimisation and prevalidation phases were not selected.

A total of approximately 3500 chemicals from the NCD and 1600 from the additional databases were screened by the CSSC. Pre-determined selection criteria were applied, primarily to ensure the quality of the *in vivo* data and the practicability of testing (for details, see 14). In total, 60 chemicals were selected for Phase 2, comprising 18 test materials from Phase 1 (the use of 20 was initially foreseen, but two could not be re-used due to their short shelf-lives), and 42 chemicals selected from the NCD and from the other sources mentioned above. The 60 chemicals were distributed to the laboratories in two deliveries of 30 chemicals each, in September 2004 and in February 2005. The confidentiality of the identities of two chemicals prevented their use in Phase 2, which therefore involved 58 chemicals. The selected chemicals (14):

1. represented statistically-justified sample sizes for distinguishing R38 from no-label chemicals;

2. provided a balanced representation of the three GHS categories, to allow for the *post hoc* evaluation of the performance of the assays for that classification system;

3. acknowledged, to a certain degree, the large prevalence known to exist for chemicals which have oedema and erythema scores of 0; and

4. presented a variety of molecular structures, functional chemical groups, and effect and use categories, as well as a wide range of physico–chemical properties.

6.1.4. Definition of PMs

As outlined in this report, the following common PM was defined for EPISKIN and EpiDerm: a test material is predicted to be an irritant (R38 according to the EU classification scheme), if the mean relative tissue viability of three individual tissues exposed to the test substance is reduced to below 50% of the mean viability of the NCs.

For the IL-1α-release endpoint, before Phase 2, L'Oréal had defined a five-fold increase, as compared to the NC, as the cut-off point at which a chemical would be labelled irritant, but this applied only to chemicals predicted to be non-irritant in the MTT reduction test. However, this cut-off value turned out to be laboratory-specific, but an absolute IL-1α-release value of 60pg/ml medium or above (established in a *post hoc* ROC analysis by ECVAM) was found to be a promising criterion, with acceptable inter-laboratory reproducibility.

6.1.5. Explanation of mechanistic basis

Since reconstituted skin models lack vascularisation, the most important endpoints defined for *in vivo* irritation testing (erythema and oedema) cannot be measured in the *in vitro* systems. However, an analysis of data in the literature (1) and an evaluation of several *in vitro* endpoints (2, 5), revealed cell viability (measured in the MTT reduction assay) as the most promising *in vitro* endpoint, followed by IL-1α-release. The latter showed a higher variability, but was slightly more sensitive in the determination of chemical effect, than was the MTT reduction assay.

6.2. Module 2: Within-laboratory variability

Within-laboratory variability was determined twice — once in Phase 1 in the lead laboratories for the SIFT, EPISKIN and EpiDerm tests with 20 chemicals, and then in Phase 2 in all the skin model laboratories for 58 test chemicals.

In Phase 1, the within-laboratory reproducibility with regard to the consistency of classifications obtained in three independent test runs, was acceptable for all three tests. However, the SIFT did not progress to Phase 2, due to its insufficient predictive capacity.

The within-laboratory variability for the MTT reduction endpoint with EpiDerm and EPISKIN has been carefully analysed in sections 4.2.1.2 and 4.2.2.2, respectively, of this report. It was acceptable for the two models, but there was a significant difference in the number of non-qualified tests in the lead laboratory for the EpiDerm test and in the other two participating laboratories.

6.3. Module 3: Transferability

Both the EPISKIN and the EpiDerm tests were successfully transferred to laboratories that had never

used the test protocols before. The lead laboratories, L'Oréal and ZEBET, held face-to-face meetings in Paris and Berlin during Phase 1 of the study, in order to standardise the test procedures. Chemicals were tested during these meetings, and it was shown that they were classified consistently in both laboratories. In a similar manner, the laboratories which joined the study for Phase 2, attended training meetings for EpiDerm at ZEBET, and for EPISKIN at L'Oréal.

6.4. Module 4: Between-laboratory variability

For the MTT reduction endpoint in the EpiDerm test, the between-laboratory variability, in terms of classifications obtained, is discussed in section 4.2.1.3 of this report. The inter-laboratory concordance was 78.8% for the no-label chemicals, and 74.1% for the R38 chemicals.

For the MTT reduction endpoint in the EPISKIN test, the between-laboratory variability in terms of classifications obtained, is discussed in section 4.2.2.3 of this report. The inter-laboratory concordance was 90.9% for the no-label chemicals, and 80.0% for the R38 chemicals.

The overall reproducibilities (positive and negative predictions) were 74.1% for EpiDerm and 86.2% for EPISKIN.

6.5. Module 5: Predictive capacity

The predictive capacity of the EpiDerm test, for classifying test chemicals according to their irritation potentials in the SIVS and applying the cytotoxicity endpoint in the three laboratories involved, are given in Tables 12 and 20. Based on these classifications, a specificity of 85%, a sensitivity of 56% and an accuracy of 72% were obtained.

For the predictive capacity of the EPISKIN test, two endpoints, MTT reduction and IL1-α release, were considered. The resulting predictive capacities for MTT reduction alone and in combination with IL1-α release, are presented in Table 20. A substantial increase in sensitivity by 16% (from 74.7% to 90.7%), accompanied by a minor loss in specificity of 2% (from 80.0% to 78.8%), was achieved when both endpoints were used, and this was accompanied by an increase in accuracy from 78% to 83%.

6.6. Module 6: Applicability domain

Restrictions in terms of the types of chemistry that could not be tested in the skin model systems had been identified before the SIVS started. For both the EPISKIN and EpiDerm tests, correction techniques were developed for chemicals that could interfere with the endpoint assay determination (5, 6). The testing of volatile chemicals is possible when the trans-well test plates are covered.

It transpired, *post hoc*, that chemicals that react with the plastic material of the test plates may give false negative predictions. The polystyrene used with EpiDerm caused more-significant problems than the polypropylene used with EPISKIN, resulting, for example, in the false negative classification of bromohexane with EpiDerm.

No other applicability restrictions have been identified so far, since the CSSC investigation into possible reasons for misclassification showed no clear reasons associated with the types of chemistry or specific physico–chemical properties of the test materials (see section 5.2).

6.7. Module 7: Performance standards

The lead laboratories, L'Oréal and ZEBET, have cooperated with ECVAM and all the members of the MT, in drafting a document on *Performance Standards for Applying Human Skin Models to* In Vitro *Skin Irritation Testing* (23), which has been approved by the ESAC. A summary of the essential elements of this performance standards document is given in section 5.5 of this report.

7. Acknowledgements

This study was funded by the Joint Research Centre of the European Commission through ECVAM, via Contract Number 21323-2003-10 F1ED ISP DE, entitled *Validation of the EPISKIN™ and EpiDerm™ assays and of the Skin Integrity Function Test for acute skin irritation testing*. We are indebted to the ECB for providing access to the new and existing chemicals database of the ECB, and to ECVAM for providing independent biostatistical support and hosting the meetings of the MT. We are also indebted to the companies which provided test chemicals and supporting toxicity and physico–chemical data.

In addition, we are grateful to our colleagues, Bill Stokes from the US Interagency Coordinating Committee for the Validation of Alternative Methods (ICCVAM) and Karen Hamernik (US EPA and ICCVAM), who served as ICCVAM observers on the Management Team, and who provided data concerning test chemicals from the US TSCA and CIR databases.

Received 12.11.07; accepted for publication 28.11.07.

8. References

1. Botham, P.A., Lesley, K.E., Fentem, J.H., Roguet, R. & van de Sandt, J.J.M. (1998). Alternative methods

for skin irritation testing: The current status. ECVAM Skin Irritation Task Force Report 1. *ATLA* **26**, 195–211.
2. Faller, C., Bracher, M., Dami, N. & Roguet, R. (2002). Predictive ability of reconstructed human epidermis equivalents for assessment of skin irritation of cosmetics. *Toxicology in Vitro* **16**, 557–572.
3. Fentem, J.H., Briggs, D., Chesne, C., Elliott, G.R., Harbell, J.W., Heylings, J.R., Portes, P., van de Sandt, J.J.M. & Botham, P.A. (2001). A prevalidation study on *in vitro* tests for acute skin irritation: results and evaluation by the Management Team. *Toxicology in Vitro* **15**, 57–93.
4. Zuang, V., Balls, M., Botham, P.A., Coquette, A., Corsini, E., Curren, R.D., Elliott, G.R., Fentem, J.H., Heylings, J.R., Liebsch, M., Medina, J., Roguet, R., van de Sandt, J.J.M., Wiemann, C. & Worth, A.P. (2002). Follow-up to the ECVAM prevalidation study on the *in vitro* tests for acute skin irritation. ECVAM Skin Irritation Task Force Report 2. *ATLA* **30**, 109–129.
5. Cotovio, J., Grandidier, M-H., Portes, P., Roguet, R. & Rubinsteen, G. (2005). The *in vitro* acute skin irritation of chemicals: optimisation of the EPISKIN prediction model within the framework of the ECVAM validation process. *ATLA* **33**, 329–249.
6. Kandárová, H., Liebsch, M., Gerner, I., Schmidt, E., Genschow, E., Traue, D. & Spielmann, H. (2005). The EpiDerm test protocol for the upcoming ECVAM validation study on skin irritation tests — an assessment of the performance of the optimised test. *ATLA* **33**, 351–367.
7. Anon. (2001). Annex VI of *Directive 67/548/EEC*. General classification and labelling requirements for dangerous substances and preparations. *Official Journal of the European Communities* **L225**, 263–314.
8. Anon. (2002). *OECD Guideline for Testing of Chemicals No. 404: Acute Dermal Irritation/Corrosion*, 13pp. Paris, France: OECD.
9. Anon. (2003). Skin Corrosion/Irritation. In *UN Globally Harmonized System of Classification and Labelling of Chemicals*, pp. 123–135. New York, NY, USA and Geneva, Switzerland: United Nations Organisation.
10. Anon. (2005). *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment*. OECD Series on Testing and Assessment No. 34, ENV/JM/MONO(2005)14, 96pp. Paris, France: OECD. Available at http://appli1.oecd.org/olis/2005doc.nsf/linkto/env-jm-mono(2005)14 (Accessed 14.10.07).
11. R&D Systems (2007). Human IL-1α/IL-1F1 Immunoassay, Cat. No. DLA50, 16pp. Minneapolis, MN, USA: R&D Systems, Inc. Available at http://www.rndsystems.com/pdf/dla50.pdf (Accessed 04.04.07).
12. ECVAM (2007). ESAC statement on the validity of *in vitro* tests for skin irritation. *ATLA* **35**, 308–312. Also available at http://ecvam.jrc.it/index.htm (Accessed 14.10.07).
13. Mosman, T. (1983). Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *Journal of Immunological Methods* **65**, 55–63.
14. Eskes, C., Cole, T., Hoffmann, S., Worth, A., Cockshott, A., Gerner, I. & Zuang, V. (2007). The ECVAM international validation study on *in vitro* tests for acute skin irritation: selection of test chemicals. *ATLA* **35**, 603–619.
15. Anon. (1995). *Skin Irritation and Corrosion: Reference Chemicals Data Bank*. ECETOC Technical Report No. 66, 247pp. Brussels, Belgium: ECETOC.
16. Anon. (2003). *TSCA ITC (Interagency Testing Committee) Reports — Public Information Access Page*. Available at http://tsca-itc.syrres.com/ (Accessed 25.11.07).
17. Hoffmann, S., Cole, T. & Hartung, T. (2005). Skin irritation: prevalence, variability and regulatory classification of existing *in vivo* data from industrial chemicals. *Regulatory Toxicology & Pharmacology* **41**, 159–166.
18. Anon. (2000). Measurement methods and results: ISO 5725-1: Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions. In *Statistical Methods for Quality Control*, 5th edn, Vol. 2, pp. 13–35. Geneva, Switzerland: International Standards Organisation.
19. Altman, D.G. & Bland, J.M. (1994). Statistics notes: Diagnostic tests 3: receiver operating characteristic plots. *British Medical Journal* **309**, 188.
20. van der Schouw, Y.T., Verbeek, A.L. & Ruijs, S.H. (1995). Guidelines for the assessment of new diagnostic tests. *Investigative Radiology* **30**, 334–340.
21. Hoffmann, S. (2006). *ECVAM Skin Irritation Validation Study Phase 2: Analysis of the Primary Endpoint MTT and the Secondary Endpoint IL1-α,* 135pp. Will be available under *Downloads of study documents*, at http://ecvam.jrc.ec.europa.eu/index.htm.
22. Zuang, V., Eskes, C., Worth, A., Cole, T., Hoffmann, S., Gallegos Saliner, A., Netzeva, T., Patlewicz, G., Cockshott, A. & Gerner, I. (2006). *Report from the Chemicals Selection Sub-Committee (CSSC) to the Management Team on Potential Reasons for the Misclassification of Chemicals in the EPISKIN and EpiDerm Assays*, 30 + 24pp. Will be available under *Downloads of study documents*, at http://ecvam.jrc.ec.europa.eu/index.htm.
23. ECVAM (2007). *Performance Standards for Applying Human Skin Models to* In Vitro *Skin Irritation Testing*, 13pp. Will be available under *Downloads of study documents*, at http://ecvam.jrc.ec.europa.eu/index.htm.