Optimisation of the EpiDerm Test Protocol for the Upcoming ECVAM Validation Study on *In Vitro* Skin Irritation Tests

Helena Kandárová, Manfred Liebsch, Elke Genschow, Ingrid Gerner, Dieter Traue, Birgitta Slawik and Horst Spielmann

Centre for Documentation and Evaluation of Alternative Methods to Animal Experiments (ZEBET) at the BfR (Federal Institute for Risk Assessment), D-Berlin

Summary

An ECVAM-funded prevalidation study (PV) was conducted during 1999 and 2000 to identify in vitro tests capable of reliably distinguishing between skin irritants (I) and non-irritants (NI) according to European Union risk phrases ("R38" or no classification). The tests evaluated were EpiDerm™, EPISKIN™, PREDISKIN™, the non-perfused pig ear method, and the mouse skin integrity function test (SIFT). Whereas reproducibility of the two human skin model tests and SIFT was acceptable, none of the methods was deemed ready to enter a formal validation study due to their low predictivity. The ECVAM Skin Irritation Task Force therefore suggested improvements of protocols and prediction models for these tests. Furthermore, it was agreed that experience gained with the two human-skin models be shared, and a common protocol should be developed for EpiDerm and EPISKIN (Zuang et al., 2002). When we applied an improved EPISKIN protocol (Portes et al., 2002) to the EpiDerm model, an acceptable specificity (80%) was achieved, whereas the sensitivity (60%) was far too low.

In 2003, the EPISKIN protocol was further refined by extension of the post-incubation period following chemical exposure. In the current study, we evaluated this EPISKIN refinement by applying it to EpiDerm. In addition, we developed technical improvements for the application of the test chemicals and rinsing procedure, which reduced the variability of results and increased the percentage of correct predictions. A set of twenty non-coded reference substances from the ECVAM prevalidation study phase III (Fentem et al., 2001) was tested with the final protocol in three independent runs. Both high sensitivity (80%) and high specificity (78%) were achieved, and the statistical probability of correct classifications was high, so that the test is now regarded ready for formal validation.

Zusammenfassung: Optimierung des EpiDerm Tests für die bevorstehende ECVAM Validierungsstudie von *in vitro* Tests zur Vorhersage hautreizender Stoffe

Von 1999 bis 2000 wurde in einer ECVAM Prävalidierungsstudie geprüft, inwieweit in vitro Methoden hautreizende Stoffe und solche, die die Haut nicht reizen, verlässlich nach der Europäischen Gefahrstoffklassifikation „R38" (hautreizend) oder „keine Klassifizierung" unterscheiden können. Die bewerteten Methoden waren EpiDerm™, EPISKIN™, der PREDISKIN™ Test, das nicht-perfundierte Schweineohr und der „Mäuse-Hautintegritäts-Funktionstest" (SIFT). Obwohl die beiden menschlichen Hautmodelle und der SIFT gut reproduzierbare Ergebnisse lieferten, versagten alle Tests hinsichtlich der korrekten Vorhersage hautreizender Stoffe, so dass keine Methode für eine formale Validierungsstudie in Frage kam. Die ECVAM Skin Irritation Task Force schlug daher eine Verbesserung der Protokolle und der Prädiktionsmodelle vor. Weiterhin wurde vereinbart, dass die mit den Hautmodellen EpiDerm und EPISKIN gemachten Erfahrungen ausgetauscht werden sollten, um ein für beide Hautmodelle gültiges Testprotokoll zu entwickeln (Zuang et al., 2002). Als wir das erste, verbesserte EPISKIN Protokoll (Portes et al., 2002) auf das EpiDerm Modell anwandten, erzielte der Test zwar eine hinreichende Spezifität (80%), aber die Sensitivität (60%) war unzureichend.

Im Jahr 2003 wurde das EPISKIN Protokoll durch eine Verlängerung der Inkubationszeit nach der Behandlung mit den Chemikalien noch einmal verbessert. In der vorliegenden Studie haben wir dieses Protokoll auf EpiDerm übertragen. Zusätzlich haben wir technische Verbesserungen für die Applikation und vollständige Entfernung der Testsubstanzen eingeführt, die nicht nur die Variabilität der Ergebnisse verringerten, sondern auch die Vorhersageleistungen des Tests verbesserten. Mit diesem neuen Protokoll wurden die zwanzig Referenzstoffe aus der ECVAM Prävalidierungsstudie Phase III (Fentem et al., 2001) in drei unabhängigen Testläufen getestet. Die Vorhersageleistungen waren mit jeweils 80% Sensitivität und 78% Spezifität balanciert und ausreichend hoch. Die statistische Wahrscheinlichkeit der korrekten Klassifizierungen war ebenfalls hoch, so dass der Test in der gegenwärtigen Form in einer formalen Validierungsstudie berücksichtigt werden kann.

Keywords: skin irritation, EpiDerm, alternative methods, validation, prevalidation, ECVAM, common protocol

1 Introduction

The determination of skin irritation potential is an international regulatory requirement for the testing of chemicals. To replace the Draize skin irritation test in rabbits and to avoid dangerous human volunteer experiments, several *in vitro* test systems have been developed. Promising results were obtained with three-dimensional human skin models, which are already accepted for the prediction of skin corrosion potential of chemicals. Therefore, in the ECVAM funded pre-validation study on *in vitro* skin irritation tests, two human epidermal models EPISKIN™ (EPISKIN SNC, Lyon, France) and EpiDerm™ (MatTek, Ashland, USA) were evaluated amongst other systems.

Based on former studies performed with human skin models, it was suggested that the performance of well-developed human skin models is comparable and that common protocols could be used for different skin models (Liebsch et al., 1997; Liebsch et al., 2000; Liebsch et al., 2004). A retrospective analysis of EpiDerm and EPISKIN results from the pre-validation revealed comparable sensitivity of both skin models. However, the predictive performance of both models was not sufficient (Zuang et al., 2002), so that further investigations were necessary to improve the tests. The ECVAM Skin Irritation Task Force suggested collaboration between L'ORÉAL and ZEBET to develop a refined common protocol applicable to EPISKIN and EpiDerm (Zuang et al., 2002).

Later, the EPISKIN protocol was refined by reducing the test substance exposure time to 15 minutes, followed by a post-incubation period of 18 hours (Portes et al., 2002). With this protocol the specificity of the EPISKIN test was significantly increased (from 40% to 80%). When this protocol was applied by ZEBET to the EpiDerm skin model, the specificity was improved to 80%, though the sensitivity was only 60% (Liebsch, 2002).

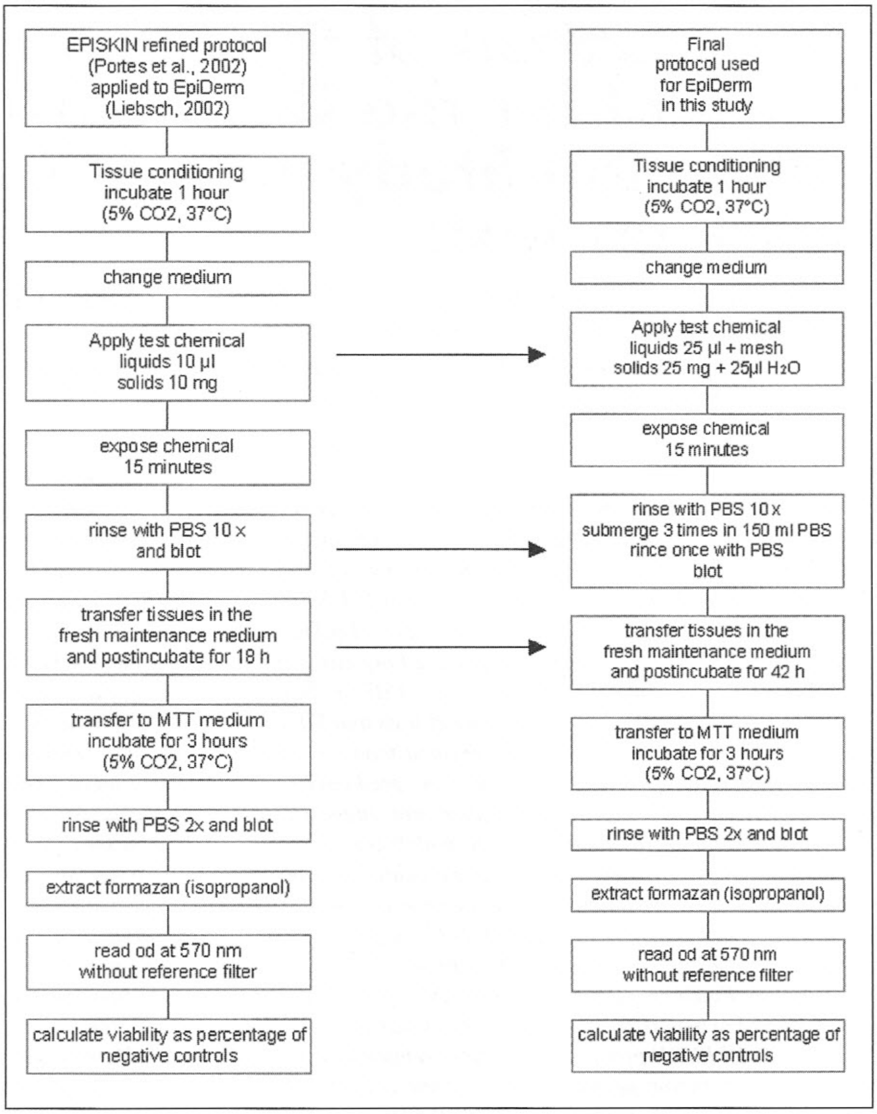
EPISKIN refined protocol (Portes et al., 2002) applied to EpiDerm (Liebsch, 2002)
Tissue conditioning incubate 1 hour (5% CO2, 37°C)
change medium
Apply test chemical liquids 10 µl solids 10 mg
expose chemical 15 minutes
rinse with PBS 10 x and blot
transfer tissues in the fresh maintenance medium and postincubate for 18 h
transfer to MTT medium incubate for 3 hours (5% CO2, 37°C)
rinse with PBS 2x and blot
extract formazan (isopropanol)
read od at 570 nm without reference filter
calculate viability as percentage of negative controls
Final protocol used for EpiDerm in this study
Tissue conditioning incubate 1 hour (5% CO2, 37°C)
change medium
Apply test chemical liquids 25 µl + mesh solids 25 mg + 25µl H2O
expose chemical 15 minutes
rinse with PBS 10 x submerge 3 times in 150 ml PBS rince once with PBS blot
transfer tissues in the fresh maintenance medium and postincubate for 42 h
transfer to MTT medium incubate for 3 hours (5% CO2, 37°C)
rinse with PBS 2x and blot
extract formazan (isopropanol)
read od at 570 nm without reference filter
calculate viability as percentage of negative controls

Fig. 1: Protocol improvements. Comparison of the first refinement of the EPISKIN protocol (Portes et al., 2002) (left side) and the improved final protocol (right side) applied to EpiDerm resulting from refinements developed at L'ORÉAL (post-incubation time) and ZEBET (application and rinsing techniques). Changes are marked with arrows between the charts.

Since with both skin models several test results were close to the classification cut-off (50% tissue viability), L'ORÉAL developed a further improvement: the post-incubation period was increased from 18 to 42 hours to allow manifestation of significant effects (Cotovio, 2003). In preparation for a common skin model protocol, we applied this refinement to EpiDerm. However, an increase in the sensitivity of the test, i.e. a reduction of false negative predictions was also required. To reduce the false negative predictions, uneven distribution of some liquids was improved with paper patches, and later with a nylon mesh (suggested by L'ORÉAL). Furthermore, the rinsing technique was improved to reduce variability of the test results. The development of the final protocol described here is shown in Figure 1.

2 Materials and methods

2.1 Epidermis model

The EpiDerm™ (EPI-200) skin model produced by MatTek Corporation (Ashland, MA, USA) consists of normal, hu-

Received 29 April 2004; received in final form and accepted for publication 16 July 2004

man epidermal keratinocytes (NHEK) cultured to form a multilayer, highly differentiated model of the human epidermis *in vitro* (Cannon et al., 1994; Earl et al., 1999). EpiDerm consists of organised basal, spinous, granular and cornified layers analogous to those found *in vivo*. The EpiDerm™ tissues (surface 0.63 cm^2) are cultured on specially prepared cell culture inserts and shipped as kits containing 24 tissues on agarose, culture medium and MTT assay components.

Each EpiDerm batch is controlled by the manufacturer. Tissues as well as the culture media are tested for viral, bacterial, fungal and mycoplasma contamination. MatTek also provides information on the ET 50 of the standard test chemical Triton X-100 and on tissue viability (MTT test) for each EpiDerm lot.

2.2 Test chemicals

20 non-coded chemicals from the ECVAM prevalidation study phase III were evaluated (Tab. 1). The set of 20 chemicals had been selected for the prevalidation study by an independent Chemical Selection Sub-Committee commissioned by the management team (MT) of the prevalidation study (Fentem et al., 2001). The chemicals were chosen primarily on the basis of skin irritation classifications derived from rabbit data included in the ECETOC database (ECETOC, 1995).

An expert of the BfR (Federal Institute for Risk Assessment, Berlin) conducted an additional evaluation of the classification according to EU classification rules (Tab. 2). This revision revealed that two of the *in vivo* experiments listed in the ECETOC database provided inconclusive results (#6, lilestralis/lilial, and #9, d-limonene). One test was not properly conducted (#11, dimethyl disulphide) so that this chemical should not have been chosen for the ECVAM prevalidation study, and cannot be recommended for further studies as a reference material.

2.3 Biostatistics, prediction model (PM)

The PM applied in the current study is based on a comparison of the cell viability of treated tissues with that of negative controls (treated with water) using the MTT assay (Mossman, 1983). If in this specific test design (15 minutes chemical exposure and 42 hours post-incubation) the mean cell viability of three treated tissues is reduced to less than 50% of control values, the chemical is classified as an "irritant".

The classification of each test chemical by the defined prediction model only provides information on whether the classification is false or correct but not on the likelihood of assigning each test chemical to one of the two classes. Therefore, the posterior probability for correct classification was estimated according to the Bayes' Rule for each of the 20 test chemicals. A case is classified in the group for which the posterior probability is the largest (Norušis, 1994; Backhaus et al., 1996; Bortz, 1993). The sum of the probabilities of the two classes is 100%, since each chemical must belong to one of the two classes. For each class, the location of the point that represents the mean of all variables can be determined. These points are called group centroids. For each chemical the distance of the respective chemical from each of the group centroids can be computed. A chemical is assigned to the specific class to which it is nearest. By using these distances for classification, probability values of correct classifications were derived (Norušis, 1994; Backhaus et al., 1996; Bortz, 1993).

2.4 Special technical improvements

2.4.1 Application technique for liquids

Different application volumes of liquid substances were evaluated (10, 15, 20 and 25 µl) and the volume was increased from 10 µl (EPISKIN protocol) to 25 µl. To reduce the incidence of false negative predictions due to uneven distribution of hydrophobic liquids on the hydrated EpiDerm surface we initially used Finn chamber paper patches (d = 8 mm) placed on the EpiDerm surface. This type of material is commonly used in *in vivo* experiments on humans. However, the disadvantage of this technique is the absorption of the test substance and enhancement of evaporation by the paper. This may cause under-prediction in some cases. At the technical meeting of the EpiDerm and EPISKIN lead laboratories at ZEBET it was suggested to replace paper patches by a nylon mesh. By this

Tab. 1: Test chemicals – specification

No.	Chemical	CAS No.	Purity (%)	Chemical Type	Solid/ Liquid
1	sodium lauryl sulphate (50 %)	151-21-3	94.8	soap/surfactant	liquid
2	1,1,1-trichloroethane	71-55-6	>99.95	chlorinated	liquid
3	potassium hydroxide (5 %)	1310-58-3	R.G.	alkali	liquid
4	heptanal	111-71-7	95	aldehyde	liquid
5	methyl palmitate	112-39-0	99	ester	liquid
6	lilestralis/lilial	80-54-6	97.8	aldehyde	liquid
7	1-bromopentane	110-53-2	99	brominated	liquid
8	*dl*-citronellol	106-22-9	98.7	alcohol	liquid
9	d-limonene	5989-27-5	98.8	miscellaneous	liquid
10	10-undecenoic acid	112-38-9	98.8	organic acid	liquid
11	dimethyl disulphide	624-92-0	99	sulphur containing	liquid
12	soap 20/80 coconut oil/tallow	-	-	soap/surfactant	solid
13	*cis*-cyclooctene	931-87-3	95	hydrocarbon (unsat.)	liquid
14	2-methyl-4-phenyl-2-butanol	103-05-9	100	alcohol	liquid
15	2,4-xylidine	95-68-1	98.1	amine	liquid
16	hydroxycitronellal	107-75-5	98.7	aldehyde	liquid
17	3,3'-dithiodipropionic acid	1119-62-6	99	sulphur containing	solid
18	4,4-methylene bis-(2,6-di -*tert*-butyl) phenol	118-82-1	98	phenolic	liquid
19	4-amino-1,2,4-triazole	584-13-4	96.7	miscellaneous	solid
20	3-chloronitrobenzene	121-73-3	99.6	halogenated aromatic	solid

application technique, the chemical is applied directly to the tissue and a nylon mesh (8 mm diameter) is placed on the tissue surface. After several experiments, this refinement was implemented for the EpiDerm model and protocol. However, possible interactions of the nylon mesh with test substances must be ruled out for each test substance.

2.4.2 Application technique for solids

Solids were ground with a mortar and pestle, and the powder was applied to the tissue with a "sharp spoon" (Aesculap, #623) to give a constant bulk volume of about 25 ± 2 mg rather than constant weight. 25 µl sterile H_2O was then added to wet the test material. If necessary, the applied material was spread to match the size of the tissue. This application technique was also used successfully for EpiDerm in the ECVAM skin corrosion study (Liebsch et al., 2000).

2.4.3 Rinsing technique

To reduce the possibility of false positive predictions for chemicals that cannot easily be washed off the tissues and may thus influence the cells during the post-incubation period of 42 hours, we developed a better rinsing technique. This technique consists of 3 steps that enable complete removal of applied substances after the exposure time. Each tissue was rinsed 10 times in a soft stream of PBS from a washing bottle by filling and emptying the culture insert. Subsequently, the tissue was completely submerged 3 times in 150 ml PBS, and finally, rinsed once again with PBS in the stream from a washing bottle.

Tab. 2: Test chemicals – classification *in vivo*

No.	Chemical	Original classification based on ECETOC Technical Report No. 66 (Fentem et al., 2001)	Revised classification for each *in vivo* experiment listed in ECETOC Technical Report No.66		
		EU class	EU class		
			Exp.1	Exp.2	Exp.3
1	sodium lauryl sulphate (50 %)	R 38*	R 38*		
2	1,1,1-trichloroethane	R 38	R 38		
3	potassium hydroxide (5 %)	R 38*	R 38*		
4	heptanal	R 38	R 38		
5	methyl palmitate	R 38	R 38*		
6	lilestralis/lilial	R 38	R 38	NI	
7	1-bromopentane	R 38	R 38		
8	*dl*-citronellol	R 38	R 38	R 38	R 38
9	d-limonene	R 38	R 38	NI	
10	10-undecenoic acid	R 38	R 38		
11	dimethyl disulphide	NI	§		
12	soap 20/80 coconut oil/tallow	NI	NI		
13	*cis*-cyclooctene	NI	NI		
14	2-methyl-4-phenyl-2-butanol	NI	NI		
15	2,4-xylidine	NI	NI		
16	hydroxycitronellal	NI	NI	NI	
17	3,3'-dithiodipropionic acid	NI	NI		
18	4,4-methylene bis- (2,6-di -*tert*-butyl) phenol	NI	NI		
19	4-amino-1,2,4-triazole	NI	NI		
20	3-chloronitrobenzene	NI	NI		

Chemicals highlighted with a shadow should not have been selected for the ECVAM prevalidation study due to either inconclusive *in vivo* data (#6, #9), or improper conductance of the *in vivo* experiment (#11)
I = irritant, NI = non-irritant,
* = possibly corrosive (R34)
§ = Improperly conducted *in vivo* experiment.
3 of 6 animals demonstrated clear signs of irritation. Classification based on these 3 animals would lead to R38 classification. Experiment was terminated too early so that a possible irritation could not be properly developed by all 6 animals.

2.5 Protocol

EpiDerm™ kits were shipped from USA on Mondays and usually arrived in Berlin on Tuesday afternoons. The best reproducibility and results were obtained when the tissues were used on the day of arrival or following overnight storage at 4-6°C.

Initially, the ability of a test substance to directly reduce MTT was assessed by adding the test material to a 1.0 mg/ml MTT solution in Dulbecco's modified Eagle medium (DMEM). 25 µl of test liquid or 25 mg of test solid were added to the MTT solution (0.9 ml) and the mixtures were incubated in the dark at 37°C for 60 minutes. If the MTT solution significantly turned blue/purple, it was assumed that the test chemical had reduced the MTT (Liebsch et al., 2000). However, only those test materials that remain bound to the tissue after rinsing present a problem, giving a false MTT reduction. To evaluate, whether residual test material binds to the tissue, a functional check using freeze-killed control tissues (Liebsch et al., 2000) was performed. None of the MTT-reducing chemicals evaluated was significantly present in the tissues after the 42 h post-incubation period, therefore no respective corrections had to be performed with the 20 test chemicals.

On the day of the experiment, tissues were aseptically removed from the transport agarose and conditioned by a 1-hour incubation in 0.9 ml assay medium (5% CO_2, 37°C, saturated humidity) in 6-well plates to release transport stress-related compounds and debris accumulated during shipment. Then tissues were transferred to fresh assay medium and exposed topically to the test chemicals: liquids (25 ± 1 µl) were applied with a micropipette and a nylon mesh (diameter = 8 mm) was placed on the surface of the tissue. Solids were applied with a 25 mg ± 2 mg calibrated spoon and wet with 25 µl sterile water. If necessary, the mixture was gently spread on the surface of the epidermis with a microspatula. Waxy test materials were first applied to a stainless-steel disc and then placed on the tissue. Each test chemical was applied to three tissue samples. In addition, three tissues serving as negative controls were dosed with 25 µl sterile water, and three tissues serving as positive controls were dosed with 5% sodium lauryl sulphate (SLS).

To prevent chemical contamination across the wells of the 6-well plate, volatile substances were tested on separate plates. In addition, wells were covered with a membrane non-permeable to gas (NeoLab, #7-2220).

Dosing was performed consecutively at 60 second intervals (the time needed for the rinsing procedure). After 15 minutes of exposure each tissue was carefully rinsed with Ca^{2+}- and Mg^{2+}-free PBS using the new rinsing technique (see 2.4.3). Blotted inserts were then transferred to new 6-well plates containing 0.9 ml fresh maintenance medium and the surface of each tissue was dried with a sterile cotton tip. Tissues were post-incubated for 42 hours (5% CO_2, 37°C, saturated humidity) to allow development of cell damage (or cell recovery), which was subsequently assessed in the MTT assay (Mossman, 1983).

Blotted tissues were transferred to 24-well plates containing 0.3 ml freshly prepared MTT medium (1 mg/ml MTT) and incubated for 3 hours at 5% CO_2, 37°C and saturated humidity. Then, tissues were rinsed twice with PBS and transferred to new 24 well plates. Two ml isopropanol (analytical grade) were added to each well, completely immersing the inserts. Plates were sealed with parafilm and formazan extraction was performed at room temperature for 2 hours on a plate shaker. Afterwards, two aliquots (200 µl) per tissue of isopropanol extract were transferred to a 96-well plate. Optical density (OD) was measured at 570 nm using isopropanol as a blank.

The relative tissue viability was calculated as a percentage of the mean viability of the negative controls. The mean of the three values from identically treated replicate tissues was used to classify the chemical according to the PM.

3 Results and discussion

In addition to several experiments performed to evaluate the effect of method refinements (see 2.4), the final, optimised test protocol was evaluated in three independent experimental runs using three replicate tissues per treatment. The results are summarised in Table 3.

First, variability between single tissues treated identically within a single test run (expressed as standard deviations) was low. Second, variability of results between independent experiments (expressed by the 95% confidence interval) was low, with the exception of two chemicals: 2-methyl-4-phenyl-2-butanol (#14) revealed false positive results (10-15% tissue viability) in two runs and a correct negative result (66% tissue viability) in one run. Hydroxycitronellal (#16) revealed a false positive result (26% tissue viability) in one run and correct negative results (91-100% tissue viability) in two runs. Consequently, the upper and lower boundaries of the confidence interval cover the range from “irri-

Tab. 3: Results obtained with the final EpiDerm protocol in three independent experiments.

No.	Chemical	Relative cell viability % (mean ± SD) n = 3 single tissues						Mean of 3 runs ± CI		
		run 1		run 2		run 3		mean	95% CI boundaries	
		mean	± SD	mean	± SD	mean	± SD		lower	upper
1	sodium lauryl sulphate (50 %)	14,5	2,9	10,8	0,5	9,6	0,1	11,9	9,6	14,2
2	1,1,1-trichloroethane	19,6	2,1	12,5	0,3	16,6	5,2	16,2	13,0	19,4
3	potassium hydroxide (5 %)	11,5	1,0	9,9	0,3	10,2	0,4	10,5	9,8	11,3
4	heptanal	12,8	1,6	9,1	0,5	10,4	0,2	10,8	9,4	12,2
5	methyl palmitate	107,5	1,8	81,0	5,5	98,6	3,2	95,7	86,4	105
6	lilestralis/lilial	11,8	2,6	11,4	0,6	12,8	0,4	12,0	10,8	13,2
7	1-bromopentane	88,2	13,4	67,8	25,1	89,2	5,4	81,7	79,1	94,7
8	*dl*-citronellol	12,3	0,5	9,7	0,5	11,4	0,4	11,1	10,2	12,1
9	d-limonene	15,0	2,7	23,6	11,9	10,4	0,7	16,3	9,9	22,8
10	10-undecenoic acid	17,5	6,7	12,3	5,1	10,0	0,7	13,2	9,1	17,8
11	dimethyl disulphide	15,9	1,4	13,9	0,7	19,5	3,6	16,4	14,0	18,9
12	soap 20/80 coconut oil/tallow	102,9	5,7	86,1	2,3	100,8	5,6	96,6	89,7	104
13	*cis*-cyclooctene	97,3	15,2	84,1	1,9	71,8	30,2	84,4	81,8	98,9
14	2-methyl-4-phenyl-2-butanol	14,5	5,9	9,7	0,9	66,3	47,4	30,2	2,4	58,0
15	2,4-xylidine	14,1	1,2	12,5	0,9	13,3	0,4	13,3	12,5	14,1
16	hydroxycitronellal	25,6	18,8	100,3	2,7	91,4	9,4	72,5	44,7	100,8
17	3,3‘-dithiodipropionic acid	112,1	3,5	80,8	1,0	94,2	6,7	95,7	84,9	107
18	4,4-methylene bis-(2,6-di -tert-butyl) phenol	103,8	2,3	79,6	4,5	99,9	5,5	94,4	85,3	104
19	4-amino-1,2,4-triazole	94,3	10,6	82,8	2,1	99,1	2,9	92,1	85,0	99,1
20	3-chloronitrobenzene	100,5	8,0	87,9	3,2	102,3	0,6	96,9	90,7	103,1

CI = confidence interval, SD = standard deviation tant" (<50% tissue viability) to non-irritant (>50% tissue viability).

Of the ten *in vivo* rabbit skin irritants, chemical #5 (methyl palmitate) and chemical #7 (1-bromopentane) were under-predicted by the EpiDerm test as non-irritants. While the latter result cannot be easily explained, methyl palmitate (chemical #5) has so far been classified negative in all *in vitro* tests (Fentem et al., 2001; Portes et al., 2002; Heylings et al., 2003). Interestingly, whereas this chemical is predicted to be a severe irritant by the Draize rabbit skin test, it is either not or only very slightly irritating to the skin when tested in the human patch test (Basketter et al.,1997; ECETOC, 2002).

Of the ten chemicals that were classified as non-irritants based on the *in vivo* rabbit test, one chemical (#11, dimethyl disulphide) should have been excluded due to improper conductance of the Draize skin test (see 2.4 and Tab. 2). Two chemicals (#14, 2-methyl-4-phenyl-2-butanol and #15, 2,4-xylidine) were predicted false positive in the EpiDerm Test. This is not surprising, because both chemicals have been consistently over-predicted by all *in vitro* tests in the past regardless of the *in vitro* model, the test design and the endpoints measured (Fentem et al., 2001; Zuang et al., 2002; Portes et al., 2002; Heylings et al., 2003). Moreover, 2,4-xylidine, was even consistently predicted corrosive by the validated *in vitro* tests (TER, EPISKIN, and EpiDerm) in the ECVAM skin corrosion validation study and "catch-up" validation study (Fentem et al., 1998; Liebsch et al., 2000).

Figure 2 shows the distribution of the 60 test results (20 chemicals tested three times) expressed as % tissue viability of negative controls. Chemicals predicted irritant are perfectly separated from chemicals predicted *non-irritant*, since none of the values are close to the 50% viability, which separates the two groups. The fact that only 1 of 60 viability values is positioned within an area of 50 ± 25% suggests that the methodological improvements (increased post-incubation time and improved washing and application techniques) have aided to separate irritants from non-irritants, leading to clear-cut predictions. Regardless of whether some of the predictions are false negative, and some are false positive, only one of 60 values was close to the 50% cut-off line.

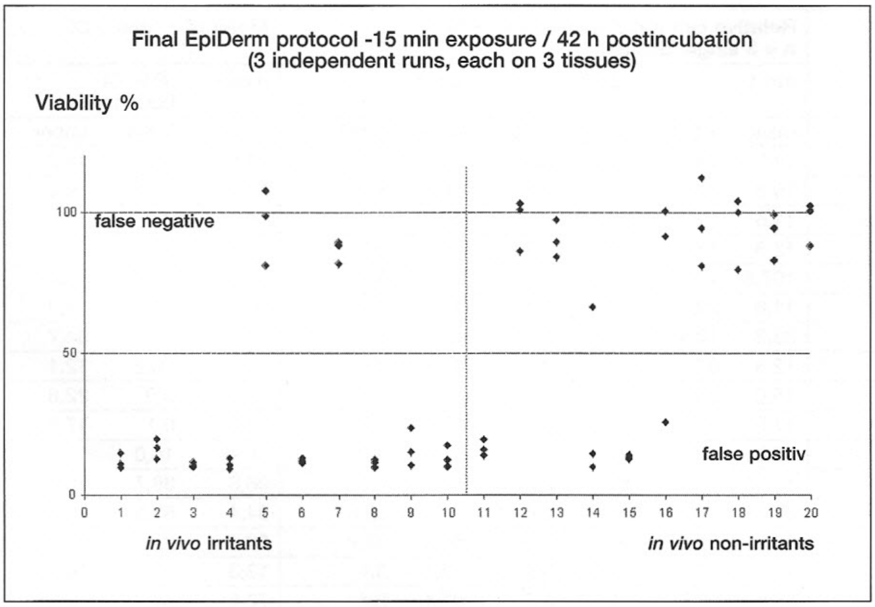

Fig. 2: Distribution of relative tissue viability obtained with the final protocol for 20 chemicals (3 independent runs for each chemical). Distribution of the 60 test results (20 chemicals tested three times) expressed as % tissue viability of negative controls. Chemicals predicted *irritant* are appropriately separated from chemicals predicted *non-irritant*, since none of the values are close to the 50% viability cut-off, which separates the two groups.

To evaluate the predictive power of a method, it is not sufficient to assign the results according to a pre-defined prediction model and to calculate the characteristic 2x2 table predictivity measures, *sensitivity, specificity, accuracy* etc. For example, a skin irritating chemical revealing 48%, 46%, and 49% tissue viability in three tests would be correctly classified in each run. However, the probability of being classified correctly every time is low. A more relevant assessment of test relevance can be made if the probability to be classified correctly is determined for each of the test chemicals (for details see 2.3). Figure 3 shows the outcome of these calculations.

The grey columns show, for each chemical, the statistical likelihood of being classified *non-irritating*, whereas the dashed columns show the likelihood of being classified irritating. Any test result in which a chemical has been correctly classified with a likelihood >75% can be regarded sufficiently robust (and thus "relevant"). Chemical #11 (dimethyl disulphide) was excluded from these calculations due to the insufficient evidence of the *in vivo* classification as a non-irritant (for details see Tab. 2). In the group of ten *in vivo* irritating chemicals (upper part of Fig. 3), eight chemicals were correctly classified with a likelihood of >75%, whereas for chemical #5 (methyl palmitate) and chemical #7 (1-bromopetane) the likelihood of being classified false negative (as non-irritant) was >75%. Of the nine *in vivo* non-irritating chemicals, six chemicals revealed "clear-cut" correct negative predictions with a likelihood of >75%, whereas one chemical (#15, 2,4-xylidine) revealed a clear likelihood of >75% of being classified false positive. For two further chemicals the likelihood of being classified correctly was below 75%, one of them (#14, 2-methyl-4-phenyl-2-butanol) revealed a false positive prediction (mean tissue viability 30%, see Tab. 3). The other chemical (#16, hydroxycitronellal) was classified negative correctly (mean tissue viability 73%, see Tab. 3), but only with a likelihood of 67%.

Table 4 gives an overview of the predictivity parameters (2x2 table statistics) obtained with the set of 20 chemicals from phase III of the ECVAM prevalidation study (Fentem et al., 2001) at the different stages of assay refinement. Column (1) of the table shows that when the EPISKIN prediction model used in phase III of the prevalidation study was applied to EpiDerm data, a comparably unbalanced result was obtained. Whereas the prediction of non-irritant chemicals was fairly acceptable, the prediction of irritants contained a high percentage of false positives. The EPISKIN method refinement (15 minutes chemical exposure and 18 hrs post-incubation) described by Portes et al. (2002) was then applied to EpiDerm, again with a similar outcome: while the specificity was significantly increased (from 40% to 80%), the sensitivity was reduced to 60% (column (2)). When the final common skin model protocol described here was used (column (3)), a sensitivity of 80% and specificity of 70% were achieved with the 20 test chemicals. Finally, column (4) shows that a very balanced predictivity of 80% sensitivity and 78% specificity was obtained when dimethyl disulphide was excluded from the data set. Whereas this chemical was selected and used in the ECVAM prevalidation study (Fentem et al., 2001), our (I.G.) analysis of the *in vivo* data published in ECETOC Report No. 66 revealed the chemical could not be classified based on these data. The animal experiments were terminated too early, and if the observation period had been extended, the probability of the substance being classified as an irritant would have been high.

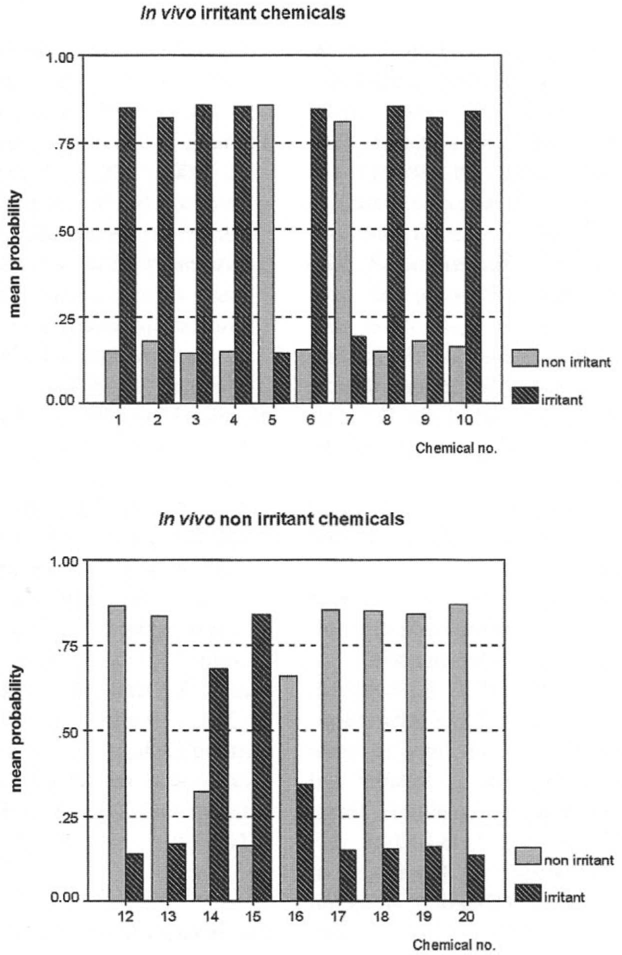

Fig. 3: Statistical probability of correct classification into two classes of skin irritancy obtained for each of 19 test chemicals. The mean probability values of three experiments for each test chemical of the set of 19 chemicals are given. Dimethyl disulphide (chemical #11) was excluded from the calculations due to insufficient *in vivo* data (for details see Tab. 2). The upper graph shows the *in vivo* irritant chemicals # 1-10, the lower graph shows the *in vivo* non-irritants # 12-20. Values for the probability range between 0 and 1 corresponding to 0 to 100%. Grey bars indicate the probability of being classified as non-irritant; dashed bars indicate the probability of being classified as irritant. A robust (relevant) test result should be backed by a classification probability of >75%.

Tab. 4: Comparison of statistical performance measures

Contingency table statistics	EPISKIN PM applied on EpiDerm (1)	Refined protocol (2)	Final protocol (3)	Final protocol Dimethyl disulphide excluded (4)
sensitivity (%)	90	60	80	80
specificity (%)	40	80	70	78
positive prediction (%)	60	75	73	80
negative prediction (%)	80	67	78	78
accuracy (%)	65	70	75	79

(1) The EPISKIN PM from the prevalidation study was applied to EpiDerm data obtained in phase III of prevalidation study. The comparison was made with the aim of evaluating whether the performance of both models is similar and whether a common protocol could be used for both models in the future (Zuang et al., 2002).

(2) L'ORÉAL´s refined protocol (Portes et al., 2002) (15 min exposure/ 18 hours postincubation) was evaluated on EpiDerm – 1 experiment on 3 tissues per test chemical (Liebsch, 2002) .

(3) Final protocol (15 min exposure/ 42 hours postincubation) – 3 independent runs on 3 tissues per test substance.

(4) *In vivo* data for dimethyl disulphide based on ECETOC Technical Report are insufficient for clear classification (see Tab. 2).

4 Conclusion

The balanced overall prediction and reliability obtained with the new protocol is promising and meets the acceptance criteria defined by the MT of the prevalidation study (Fentem et al., 2001). Expectedly, variability between runs or even misclassifications were observed for a few chemicals. However, the results obtained with the new protocol design are sufficiently promising for the test to enter a formal validation study. In addition, comparison with EPISKIN results obtained with the same set of chemicals (Portes et al., 2002; Cotovio, 2003) shows that the use of a common "skin model" test protocol and prediction model is possible.

To exclude the possibility that the methodological refinements developed primarily by L'ORÉAL and partly by ZEBET are only valid for this specific set of test chemicals, both laboratories have decided to verify the improvements by applying the new method to a different set of chemicals. These studies are currently under way.

Acknowledgement
The authors would like to thank to Dr. José Cotovio, Dr. Roland Roguet and Ms. M. H. Grandidier for the fruitful scientific co-operation in developing and optimising a common protocol for the two commercial skin models EpiDerm and EPISKIN.

References

Backhaus, K., Erichson, B., Plinke, W. and Weiber, R. (1996). *Multivariate Analysemethoden* (591 pp.). 8th Edition, Berlin, Heidelberg, New York, London: Springer Verlag.

Basketter, D. A., Chamberlain, M., Griffiths, H. A. et al. (1997). The classification of skin irritants by human patch test. *Food Chemical Toxicology 35*, 845-852.

Bortz, J. (1993). *Statistik* (753 pp.). 3rd Edition, Berlin Heidelberg New York: Springer Verlag.

Botham, P. A., Earl, L. K., Fentem, J. H. et al. (1998). Alternative methods for skin irritation testing: the current status. ECVAM Skin Irritation Task Force report 1. *ATLA 26*, 195-211.

Cotovio, J. (2003). Personal communication. Presented at the Stakeholder Workshop of the Skin Irritation Validation Study. May 7-8th, Ispra, Italy.

ECETOC (1995). *Skin Irritation and Corrosion: Reference Chemicals Data Bank.* ECETOC Technical Report No. 66. European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels. 247 p.

ECETOC (2002). *Use of Human Data in Hazard Classification for Irritation and Sensitisation.*

ECETOC Monograph No. 32. European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels. 54p.

Fentem, J. H., Archer, G. E. B., Balls, M. et al. (1998). The ECVAM International Validation Study on In vitro Tests for Skin Corrosivity. 2. Results and Evaluation by the Management Team. *Toxicology in Vitro 12*, 483-524.

Fentem, J. H., Briggs, D., Chesné, C. et al. (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.

Heylings, J. R., Diot, S., Esdaile, D. J. et al. (2003). A prevalidation study on the in vitro skin irritation function test (SIFT) for prediction of acute skin irritation in vivo: results and evaluation of ECVAM Phase III. *Toxicology in Vitro 17*, 123-138.

Liebsch, M. (2002). Personal communication. Data presented at the ECVAM Skin Irritation Task Force meeting, 20-21 November 2002, Ispra, Italy.

Liebsch, M., Barrabas, Ch., Traue, D. und Spielmann, H. (1997). Entwicklung eines neuen in vitro Tests auf dermale Phototoxizität mit einem Modell menschlicher Epidermis (EpiDerm™). *ALTEX 14*, 165-174.

Liebsch, M., Kandárová, H., Spielmann, H. et al. (2004): Validation of the Human Epidermis Model SkinEthic for the Skin Corrosion Testing According to New OECD Test Guideline 431. 43rd Annual SOT meeting, March 21-25, 2004, Baltimore USA.

Liebsch, M., Traue, D., Barrabas, Ch. et al. (2000). The ECVAM prevalidation study on the use of EpiDerm for skin corrosivity testing. *ATLA 28*, 371-401.

Mossman, T. (1983). Rapid colorimetric assay for cell growth and survival-application to proliferation and cytotoxicity assays. *Journal of Immunological Methods 65*, 55-63.

Norušis, M. J. (1994). SPSS Professional Statistics 6.1. Chicago, Illinois: SPSS Inc.

Portes, P., Grandidier, M. H., Cohen, C. and Roguet, R. (2002). Refinement of the Episkin protocol for the assessment of acute skin irritation of chemicals: follow-up to the ECVAM prevalidation study. *Toxicogy in Vitro 16*, 765-770.

Zuang, V., Balls, M., Botham, P. et al. (2002). Follow-up to the ECVAM prevalidation study on in vitro tests for acute skin irritation. ECVAM skin irritation task force report 2. *ATLA 30*, 109-129.

Correspondence to
Ing. Helena Kandárová
ZEBET, Bundesinstitut für Risikobewertung (BfR)
Diedersdorfer Weg 1
D-12277 Berlin
Germany
e-mail: h.kandarova@bfr.bund.de

Abbreviations

CI, confidence interval; ECETOC, European Centre for Ecotoxicology and Toxicology of Chemicals; ECVAM, European Centre for the Validation of Alternative Methods; EU, European Union; I, irritant; MT, Management Team; NI, non-irritant; PM prediction model; R.G., reagent grade; SD, standard deviation; SLI, slight irritant; SLS, sodium lauryl sulphate