5.3 SKIN IRRITATION STUDIES

5.3.1 RESULTS OBTAINED WITH EPIDERM MODEL

The overall aim of the present study was to prepare the EpiDerm skin irritation test for the upcoming ECVAM skin irritation validation study. As described in the Introduction (section 3.3.2), the study was divided into the two phases:

1. transfer of the refined EPISKIN test protocol (Portes *et al*, 2002 and Cotovio, 2003 personal communication) to EpiDerm model, and optimisation of the assay with the reference set of chemicals from the non-successful ECVAM pre-validation study (Phase I),
2. evaluation of the predictive performance (sensitivity, specificity and accuracy) of the new EpiDerm skin irritation test applying a different set of the test chemicals as those used for optimisation (Phase II).

Performance of the EpiDerm skin irritation test in Phase I

The refined EPISKIN protocol when applied on EpiDerm model showed promissing performance. However, the proposed application volume of 16 µl / 16 mg (equivalent to 10 µl / mg for EPISKIN model) had to be increased for EpiDerm to 25 µl / 25 mg, as too many results close to the classification border were initially obtained for irritants (data not shown). In addition, some technical improvements, e.g. application technique for liquids using the nylon mesh and more efficient washing procedure were implemented into the EpiDerm experimental design (for details see Materials and Methods). This improvements helped to reduce variability and thus increased the robustness of the method. The optimised assay was finally evaluated in three independent experimental runs using twenty chemicals from the third phase of the ECVAM skin irritation prevalidation study (Fentem *et al*., 2001). The experimental outcome is summarised in Table 29.

The balanced overall prediction (accuracy of 79%) obtained with the new protocol was promising and met the acceptance criteria defined by the management team of the prevalidation study (Fentem *et al*., 2001). In addition, comparison with EPISKIN results obtained with the same set of test chemicals (Portes *et al*., 2002; Cotovio *et al*., 2005) showed that the use of a "common skin model" test protocol and prediction model seems to be possible.

Table 29. Results obtained with the final EpiDerm protocol in three independent experiments with 20 pre-validation chemicals.

Chemical	Relative cell viability %						Mean of 3 runs ;CI		
	Run 1		Run 2		Run 3		Mean	95 % CI boundaries	
	mean	SD	mean	SD	mean	SD		lower	upper
Sodium lauryl sulphate (50%)	14.5	2.9	10.8	0.5	9.6	0.1	11.9	9.6	14.2
1,1,1-Trichloroethane	19.6	2.1	12.5	0.3	16.6	5.2	16.2	13.0	19.4
Potassium hydroxide (5%)	11.5	1.0	9.9	0.3	10.2	0.4	10.5	9.8	11.3
Heptanal	12.8	1.6	9.1	0.5	10.4	0.2	10.8	9.4	12.2
Methyl palmitate	107.5	1.8	81.0	5.5	98.6	3.2	95.7	86.4	105
Lilestralis/Lilial	11.8	2.6	11.4	0.6	12.8	0.4	12.0	10.8	13.2
1-Bromopentane	88.2	13.4	67.8	25.1	89.2	5.4	81.7	79.1	94.7
dl-Citronellol	12.3	0.5	9.7	0.5	11.4	0.4	11.1	10.2	12.1
d-Limonene	15.0	2.7	23.6	11.9	10.4	0.7	16.3	9.9	22.8
10-Undecenoic acid	17.5	6.7	12.3	5.1	10.0	0.7	13.2	9.1	17.8
Dimethyl disulphide	15.9	1.4	13.9	0.7	19.5	3.6	16.4	14.0	18.9
Soap 20/80 coconut oil/tallow	102.9	5.7	86.1	2.3	100.8	5.6	96.6	89.7	104
Cis-Cyclooctene	97.3	15.2	84.1	1.9	71.8	30.2	84.4	81.8	98.9
2-Methyl-4-phenyl-2-butanol	14.5	5.9	9.7	0.9	66.3	47.4	30.2	2.4	58.0
2,4-Xylidine	14.1	1.2	12.5	0.9	13.3	0.4	13.3	12.5	14.1
Hydroxycitronellal	25.6	18.8	100.3	2.7	91.4	9.4	72.5	44.7	100.8
3,3'-Dithiodipropionic acid	112.1	3.5	80.8	1.0	94.2	6.7	95.7	84.9	107
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
4-Amino-1.2.4-triazole	94.3	10.6	82.8	2.1	99.1	2.9	92.1	85.0	99.1
3-Chloronitrobenzene	100.5	8.0	87.9	3.2	102.3	0.6	96.9	90.7	103.1

SD = standard deviation, 95 % CI = 95 % confidence interval.

Table 30. Performance of the optimised test.

Contingency table Statistics	Results obtained with 20 Chemicals (Fentem *et al*. 2001) and protocol performed according to Portes *et al*. 2003 (1)	Results obtained with 20 chemicals (Fentem *et al*. 2001 and optimized protocol (2)	Results obtained with 19 chemicals (Fentem *et al*. 2001 (3)
Sensitivity (%)	60	80	80%
Specificity (%)	80	70	78%
Positive prediction (%)	75	73	80%
Negative prediction (%)	67	78	79%
Accuracy (%)			
Prevalence of test set	1	1	1.11

(1) Experiment performed with EpiDerm model according to EPISKIN refined protocol published by Portes et al., in 2002. The protocol is based on 15 min exposure and 18 hours postincubation. Application volume was 10 µl (liquids) or 10 mg (solids) (data presented at the 3 Management team meeting of the Skin Irritation Validation Study. Liebsch, 2002).

(2) Optimized EpiDerm protocol - based on investigations of Cotovio, 2003.

Both protocol are based on 15 min exposure and 42 hours postincubation. Application volume was increased for EpiDerm model to 25 µl (liquids) and 25mg (solids).

(3) Change in predictive performance after excluding Dimethyl disulphide from experimental set due to the insufficient data for clear classification (for details see part Materials and Methods).

Performance of the EpiDerm skin irritation test in Phase II

To exclude the possibility that the methodological refinements were only valid for this specific set of twenty test chemicals, it was decided to verify the protocol improvements in additional study with different set of test chemicals. This chemicals were selected from the ECETOC database No. 66 (ECETOC, 1995) and their classification according to EU and GHS rules was performed by an BfR expert (for detailed information see Materials and Methods). Experimental outcome of Phase II is summarised in Table 31.

Table 31. Results obtained with the final EpiDerm protocol in three independent experiments with 26 new chemicals.

Chemical	**In vivo class**	**Relative cell viability % n = 3 single tissues**						**Mean of 3 runs ;CI**			**In vitro class**
		Run 1		**Run 2**		**Run 3**		**Mean**	**95 % CI boundaries**		
		mean	SD	mean	SD	mean	SD		lower	upper	
SLS (20% aq.)	I	7.7	0.6	10.8	1.5	21.2	1.0	13.2	8.5	18.0	I
Tetrachloroethylene	I	11.3	0.7	15.1	5.4	15.7	0.6	14.0	11.4	16.7	I
alpha-Terpinelol	I	10.5	0.2	10.6	0.3	9.5	0.4	10.2	9.7	10.7	I
Tallow propylene polyamine	I	16.4	0.8	28.5	8.8	16.9	2.9	20.6	14.8	26.4	I
1-Bromohexane	I	96.5	19.2	84.1	15.4	100.0	5.6	93.5	82.4	104.7	NI
Methyl laurate	I	107.4	13.2	100.2	2.5	102.5	8.0	103.3	96.9	109.8	NI
Cinnamaldehyde	I	11.3	0.8	12.4	1.0	10.5	0.4	11.4	10.5	12.2	I
Linalyl acetate	I	90.0	10.7	97.8	5.4	97.1	3.4	95.0	89.4	100.5	NI
Eugenol	I	12.1	0.2	12.6	1.2	10.8	0.4	11.8	11.0	12.6	I
Linalol	I	79.3	10.5	71.5	4.6	68.5	17.0	73.1	64.4	81.8	NI
Methylstearate	NI	105.9	7.4	94.3	0.5	102.9	1.1	101.0	96.1	105.9	NI
Benzylalcohol	NI	75.1	24.5	93.8	7.8	98.6	6.0	89.2	76.1	102.3	NI
2-ethoxy ethyl methacrylate	NI	104.8	2.8	99.5	3.3	92.5	7.9	99.0	93.6	104.3	NI
Benzyl benzoate	NI	89.4	4.8	101.3	2.3	109.8	20.4	100.1	89.5	110.8	NI
Benzyl acetate	NI	103.5	2.7	91.3	9.1	113.6	10.3	102.8	93.6	112.0	NI
Isopropyl palmitate	NI	122.3	15.0	103.6	3.7	97.4	1.8	107.8	97.3	118.3	NI
Isopropyl myristate	NI	103.7	4.4	92.6	1.7	106.2	1.0	100.8	95.7	106.0	NI
n-Butyl propionate	NI	97.9	1.2	95.6	4.4	132.5	2.2	108.7	94.8	122.6	NI
Sodium bisulphite	NI	91.1	10.9	100.2	0.7	89.1	11.1	93.5	86.3	100.6	NI
1,6 – Dibromohexane	NI	11.6	0.7	15.1	1.9	9.5	1.2	12.1	10.0	14.1	I
Isopropanol	NI	109.2	2.8	97.1	3.4	93.4	3.3	99.9	94.1	108.8	NI
Benzyl salicylate	NI	101.0	2.3	87.8	3.0	103.2	26.1	97.3	85.8	108.9	NI
Lauric acid	NI	105.7	1.0	98.5	1.0	99.8	4.3	101.4	98.3	104.4	NI
Dipropylene glycol	NI	116.4	20.9	101.3	3.5	104.6	3.1	107.4	97.7	117.2	NI
Sodium bicarbonate	NI	96.9	2.5	105.5	3.6	99.4	7.0	100.6	96.2	104.9	NI
Erucamide	NI	86.7	7.1	100.0	3.9	124.0	15.1	103.6	89.4	117.8	NI

SD = standard deviation; 95 % CI = 95 % confidence interval

With the new test set of 26 chemicals, a high specificity (94%) was obtained, while at 60%, the sensitivity was at the border of acceptance. However, it is important to note that the test set was not balanced. The prevalence of 0.63 indicates that only 10 chemicals were *in vivo* irritants and the rests were non-irritating chemicals. Moreover, three, with regard to correct classification, disputable chemicals (methyl laurate, linalol and linalyl acetate) contributed significantly to the low sensitivity.

For the full set of 45 chemicals a sensitivity of 70% and specificity of 88% were obtained. The positive predictive value was 82%, the negative predictive value was 79%, and the resulting accuracy was 80% (Table 32).

Table 32. Comparison of statistical performance measures.

Contingency table Statistics	Results obtained with 26 new chemicals (1)	Predictivity for all 45 chemicals tested with final protocol (2)
Sensitivity (%)	60%	70%
Specificity (%)	94%	88%
Positive prediction (%)	86%	82%
Negative prediction (%)	79%	79%
Accuracy (%)	81%	80%
Prevalence of the test set	0.63	0.8

(1) Final protocol evaluated on set of new 26 chemical (10 irritants. 16 non-irritants).
(2) Predictivity shown for all 45 chemicals tested with final protocol on EpiDerm (20 in vivo irritants and 25 in vivo non-irritants).

Assay variability and correct prediction

In all experiments, the variability (expressed by the standard deviations) between individual tissues treated identically within a single test run was low (see Tables 29 and 31). Similarly the variability of results between independent experiments expressed by the 95% confidence interval was low, except for two chemicals: 2-methyl-4-phenyl-2-butanol (#2) and hydroxycitronellal (#17). Variable results for these two substances have also been found in all previous studies with EPISKIN and EpiDerm.

Of the 20 chemicals that were classified as *in vivo* skin irritants (based on test in rabbits), methyl palmitate (#33), methyl laurate (#43), 1-bromopentane (#28), 1-bromohexane (#37), linalol (#41) and linalyl acetate (#42) were predicted to be non-irritant in the EpiDerm test.

Of the 25 chemicals that were classified *in vivo* as nonirritants, 2,4-xylidine (#1), 2-methyl-4-phenyl-2-butanol (#2) and 1,6-dibromohexane (#9) were predicted to be irritants in the EpiDerm test.

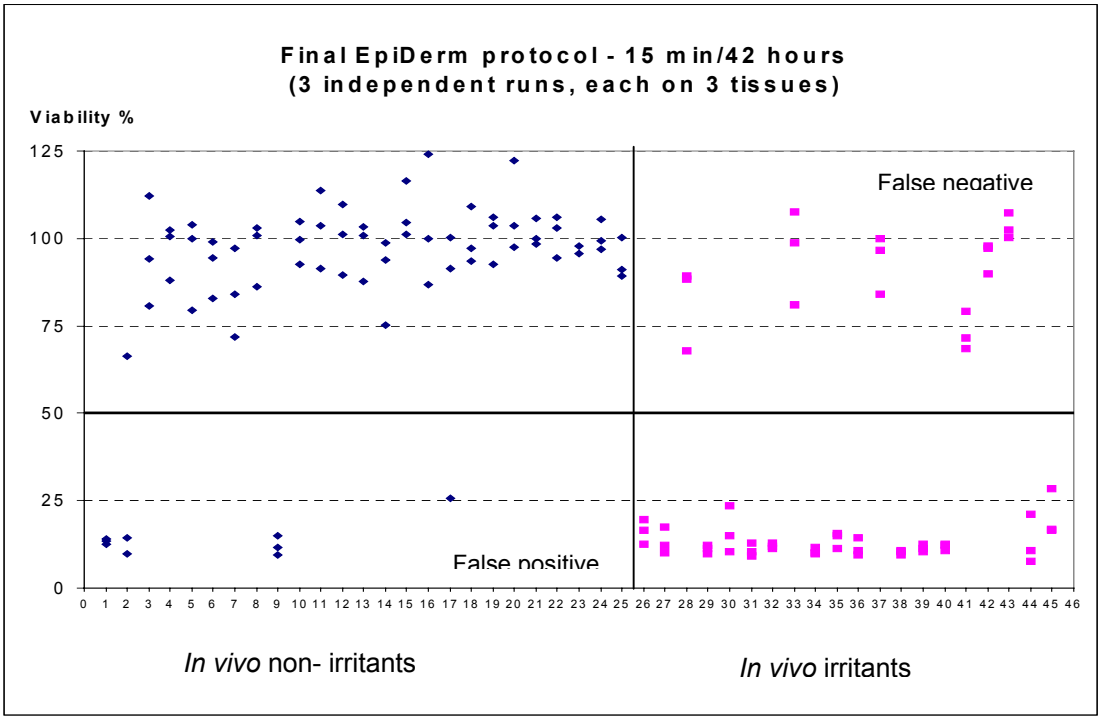

Figure 31. Distribution of relative tissue viabilities obtained with EpiDerm protocol: 45 chemicals.
Distribution of 135 test results (45 chemicals tested 3 times) expressed as % tissue viability of negative controls is given. Chemicals predicted as irritant are well separated from chemicals predicted as non-irritant, since none of values are close to the 50 % viability cut-off, which separates the two groups.

In summary, 22 out of 25 rabbit non-irritants and 14 out of 20 rabbit skin irritants were classified correctly, resulting in sensitivity of 70 % and specificity of 88% of the EpiDerm test. The overall accuracy of the assay with 45 chemicals was 80 % (see Table 32).

Probability of prediction

An *in vitro* toxicity test is characterised by a combination of a biological test system and a specific PM. The test system provides a means of generating *in vitro* data for chemical properties of interest, whereas the PM is an unambiguous algorithm for converting these data into predictions of specific toxicological endpoints (Archer *et al.*, 1997). In validation studies, both the test system and PM must undergo independent assessment (Worth and Balls, 2001).

In the current study, SPSS® analysis was performed to evaluate the predictive power of the method, since it is not sufficient to just assign the results according to a PM and to calculate the predictivity measures (*sensitivity*, *specificity*, *accuracy*, etc.) in a 2×2 contingency table. Therefore, the probability of correct classification was determined for each of the test chemicals in each experimental run. When a chemical has been correctly classified with a likelihood > 0.75 (= 75%), this can be regarded as sufficiently robust and

thus “relevant”. The testing of 25 *in vivo* non-irritating chemicals (upper part of Figure 32) provided 21 “clear-cut” correct negative predictions with a likelihood of > 75%. For chemical # 1 (2,4-xylidine) and #9 (1,6-dibromohexane), the likelihood of being classified false-positive was > 75%. In the group of 20 *in vivo* irritating chemicals (lower part of Figure 32), 14 chemicals were correctly classified with a likelihood of > 75%. Five chemicals, #28 (1-bromopentane), #37 (1-bromohexane), #33 (methyl palmitate), #43 (methyl laurate) and #42 (linalyl acetate) were classified as false-negative with a probability of more than 75%. Chemical #41 (linalol) was also classified as false-negative, but with a probability of less than 75%. For the majority of the chemicals tested, the probability for correct classification was high (> 75 %), which confirms the good test performance.

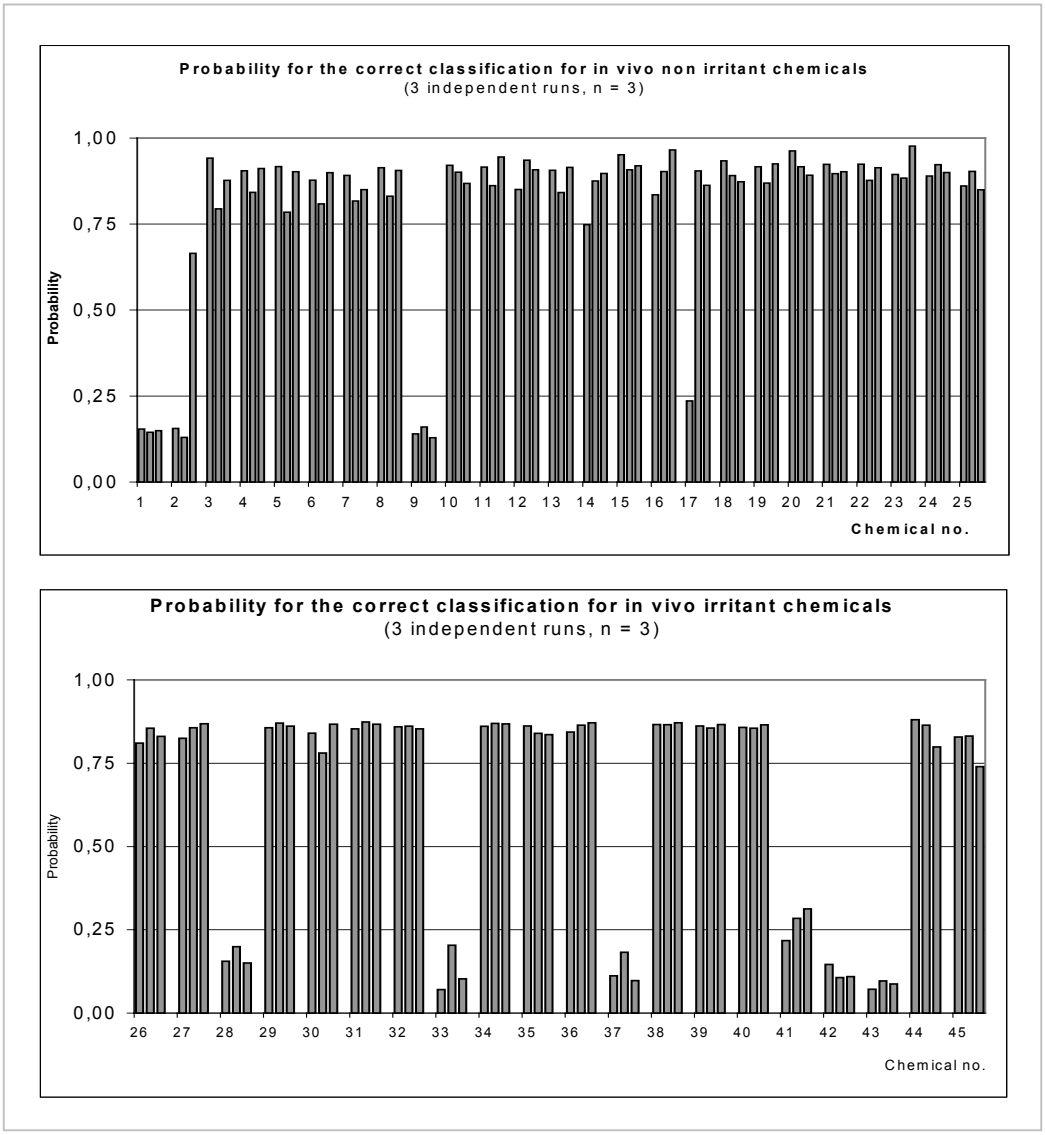

Figure 32. Probability for the correct classification for 45 test chemicals. The upper graph shows the *in vivo* non irritants # 1-25 the lower graph shows the *in vivo* irritating chemicals # 26-46. Values for the probability range between 0 and 1 correspond to 0 to 100%. A robust (relevant) test result should be backed by a classification probability of >75%.

5.3.2 DISCUSSION

The predictive value of the EpiDerm skin irritation test

The present study had two main objectives. One objective was to transfer the original EPISKIN protocol to the EpiDerm model and to optimise it (Phase I). The second objective was to confirm the results obtained in Phase I with a new set of chemicals, to evaluate the reliability and promising performance of the test.

Twenty *in vivo* (rabbit) irritants and 25 *in vivo* (rabbit) non-irritants from diverse chemical groups were tested by using the EpiDerm skin irritation assay. Six of the irritating chemicals were not classified in concordance with the *in vivo* classification, namely: methyl palmitate (#33), methyl laurate (#43), 1-bromopentane (#28), 1-bromohexane (#37), linalol (#41) and linalyl acetate (#42).

Methyl palmitate has provided a negative result in all *in vitro* models in which it has been tested to date (Fentem *et al*., 2001; Portes *et al*., 2002, Heylings *et al*., 2003; Cotovio *et al*. 2005). The same result was obtained also with methyl laurate. Both chemicals belong to the chemical group of fatty acid methylesters. Although they are predicted to be severe irritants in the rabbit skin test, in the human patch test they are either non-irritating or only very slightly irritating (Basketter *et al*., 1999; ECETOC, 2002; Basketter *et al*., 2004).

Linalol, an unsaturated tertiary alcohol, and linalyl esters (for example, linalyl acetate), are found in a variety of fruits, vegetables and spices. Both linalol and linalyl acetate are terpens and are used as fragrance ingredients in cosmetics, as well as in non-cosmetic products. Based on *in vivo* rabbit data in the *ECETOC Data Bank* (ECETOC, 1995), linalol and linalyl acetate were classified as irritants. Linalol, however, shows a large variability in response in single animals (see Table 12, Materials and Methods). This observation was confirmed in other rabbit and also guinea-pig studies, in which linalol and linalyl acetate induced very mild to severe erythema (Bickers *et al*., 2003; Letizia *et al*., 2003a; Letizia *et al*., 2003). In contrast to animal data, linalyl acetate induced no irritation in 30 of 31 volunteers in human patch tests (Basketter *et al*., 2004).

It has been reported that the irritating and sensitising effects of terpens depend very much on the amount of isomers and impurities in the sample and on the age of the sample. Important is also the test design and the conductance of the *in vivo* test. It is therefore not surprising, that a large variability of *in vivo* responses has been observed with terpens, for example, alfa-terpinelol and citronellol, which has resulted in different conclusions with respect to humans and rabbits (Basketter *et al*., 2004).

Of the 25 *in vivo* rabbit non-irritatnts, three were overpredicted (classified as irritants) in the EpiDerm test. All three test chemicals, namely, 2,4-xylidine, 2-methyl-4-phenyl-2-

butanol and 1,6-dibromhexane, are well soluble in lipids or are lipid solvents. Therefore, these chemicals might readily affect the lipid barrier of the epidermis and penetrate into the deeper layers of the skin. The toxic effects of such substances *in vivo* and *in vitro* may depend on the level of the lipid barrier development. 2,4-xylidine, 2-methyl-4-phenyl-2-butanol were consistently over-predicted by all *in vitro* tests, regardless of the *in vitro* model used, the test design and the endpoints measured (Fentem *et al*., 2001; Portes *et al*., 2002, Heylings *et al*., 2003;Cotovio *et al*. 2005). Moreover, 2,4-xylidine has been overpredicted as corrosive substance in all validated human skin model *in vitro corrosivity* tests (Fentem *et al*. 1998; Liebsch *et al*., 2000).

It has been described above, that comparative studies in rabbits and in human volunteers often show differences in susceptibility to some classes of irritant substances and preparations. Human skin is in general less sensitive than rabbit skin to surfactants or cleaning products. These differences in sensitivity may be species specific (ECETOC 2002). The rabbit tests usually predict severe human skin irritants and non-irritants correctly, but fail to distinguish between mild and moderate skin irritants (Phillips *et al*., 1972). Therefore, skin patch testing in human volunteers would be important, particularly for substances at the lower end of the irritation scale. However, testing in human volunteers is permitted only with substances which are devoid of seriously hazardous properties (Robinson *et al*., 2001), so the skin irritation testing of chemicals in human volunteers is almost impossible. Nevertheless, there is good evidence from several chemical classes that rabbit and human data differ significantly. This has been reported mainly for fatty acids and acid blends (C8–C16), fatty alcohols (C8–C16), fatty acid methyl esters (C6–C16), detergents and cleaning products (Basketter *et al*., 1999; ECETOC, 2002; Basketter *et al*, 2004). These differences must be considered in the final classification of a chemical, if results from human patch studies are available.

Retrospective evaluation of PM

At the beginning of the present study, it was necessary to evaluate whether the prediction model defined for the EPISKIN assay (Portes *et al*., 2002; Cotovio, 2003) could be applied without alteration to the EpiDerm model. Identical chemicals as used for developing the EPISKIN test and prediction model were tested with EpiDerm model. Statistical analysis of EpiDerm results revealed, that the optimal classification border between irritants and non-irritants was 51% of relative viability. The result was in high concordance with previously developed EPISKIN prediction model (Portes *et al*., 2002; Cotovio, 2003), where the same classification border was proposed.

An evaluation of the assay robustness performed with 26 new chemicals shifted the classification cut-off to 61% of relative viability. Considering all test results (for all 45 chemicals), a viability of 58% was the optimum classification border. Since all the irritants resulted in less than 25% relative viability, and the majority of non-irritants were located above 75% relative viability, no effect on the final classification was observed when applying the new computed borderlines. Therefore, it was decided that a PM with classification border of 50% relative viability could be used for the reliable classification.

Limitations of the EpiDerm skin irritation test

The new test was applicable for both liquid and solid chemicals. Even chemicals that are MTT reducers and may interfere with the MTT endpoint for tissue viability were compatible with the assay.

The testing of highly volatile chemicals was difficult. Due to the low volume applied, strong ventilation in the laminar hood, and high volatility of some substances (for example, 1,1,1-trichloroethane), false-negative results were previously obtained. This technical problem was now solved by covering the plates with gas non-permeable film during the exposure period. However, the testing of high volatile substances may still be technically challenging.

The GHS classification of chemicals into three groups (irritants [I], mild irritants [SLI] and nonirritants [NI]) is an additional obstacle to the regulatory acceptance of the new method. The EpiDerm test was originally designed, developed and optimised in order to distinguish between two classes of chemicals, irritating and non-irritating, according to the EU classification system (EU, 1967). The test is therefore not designed to identify the third group: slight (SLI) or so-called “mild irritants”. In fact, no clear borderline (as in the case of *in vitro* skin irritation tests) seems to exist for defining the two (NI, I) or three (NI, SLI, I) irritation groups (see Figure 33). In addition, large variations in *in vivo* responses exist for many chemicals (for example, see chemical #45, in Figure 33). Moreover, if the available human data were to be taken into account, the classification of some irritants and many SLI chemicals, based on rabbit data, would have to be changed.

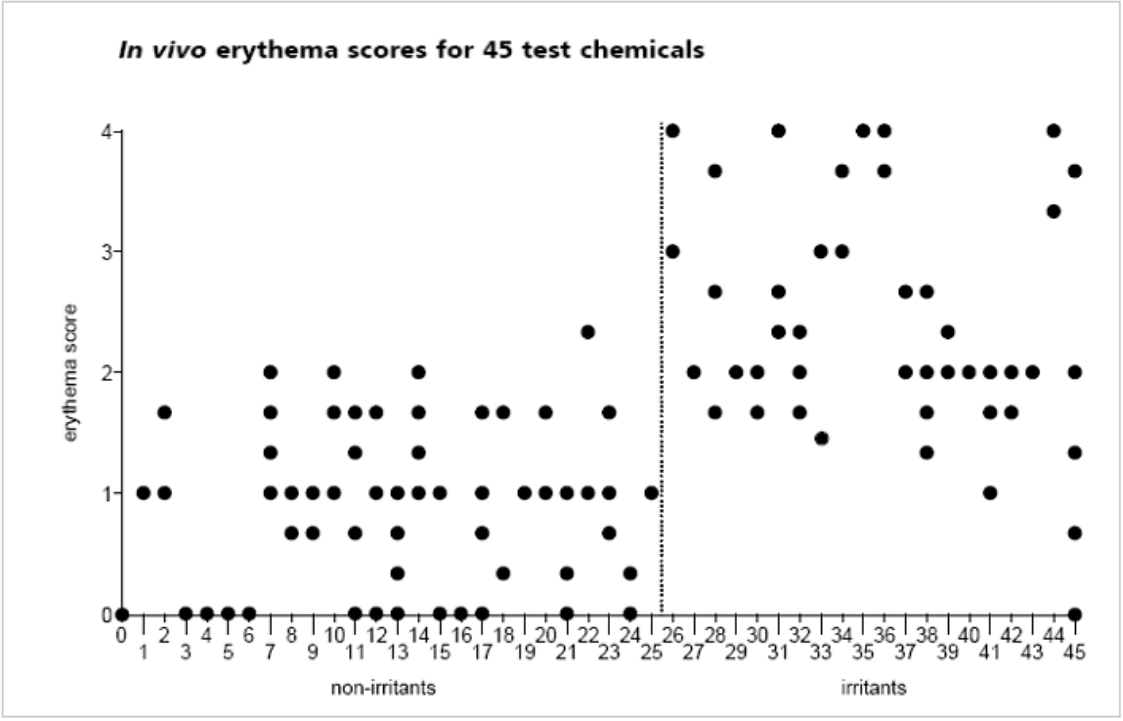

Figure 33. Erythema scores were calculated by using data from the ECETOC Data Bank 66 (18) for each *in vivo* experiment and test animal. In many cases (for example, chemicals 11, 28, 33 and 45), variability of *in vivo* responses to a single test chemical can be observed (for homogenous responses, only one or two points are displayed in the graph).

5.3.3 RESULTS OBTAINED WITH SKINETHIC MODEL

The purpose of the present study was to evaluate if SkinEthic model will provide similar results in the skin irritation assay using the “common skin irritation protocol” developed for EPISKIN and EpiDerm model. The study was divided into the two phases. First phase aimed for adaptation of the protocol to SkinEthic model and evaluation of the predictivity of the assay. The second phase was designed as an interlaboratory trial to assess inter-laboratory reproducibility.

Phase I: Assessment of the SkinEthic model performance with the common skin irritation protocol

In phase one, twenty substances from the third Phase of the ECVAM skin irritation pre-validation study (Fentem *et al*, 2001) were tested using the SkinEthic RHE model and the "common skin irritation protocol".

Initially, optimisation of the test protocol had to be accomplished, as the SkinEthic model showed slightly higher sensitivity to test substances when the original EpiDerm protocol was applied without changes (data not shown). Although the size of the SkinEthic

model was the same as size of EpiDerm cultures (0.63 cm^2), a dose of 25 µl / 25 mg (typical for EpiDerm) had to be reduced to 20 µl / mg, as too many results for non-irritation chemicals were close to the classification border. In fact, with this reduction of volume, the SkinEthic protocol stands in between EpiDerm and EPISKIN protocols. The main differences between the EpiDerm and SkinEthic protocol are shown in Figure 34 and are highlighted with grey colour.

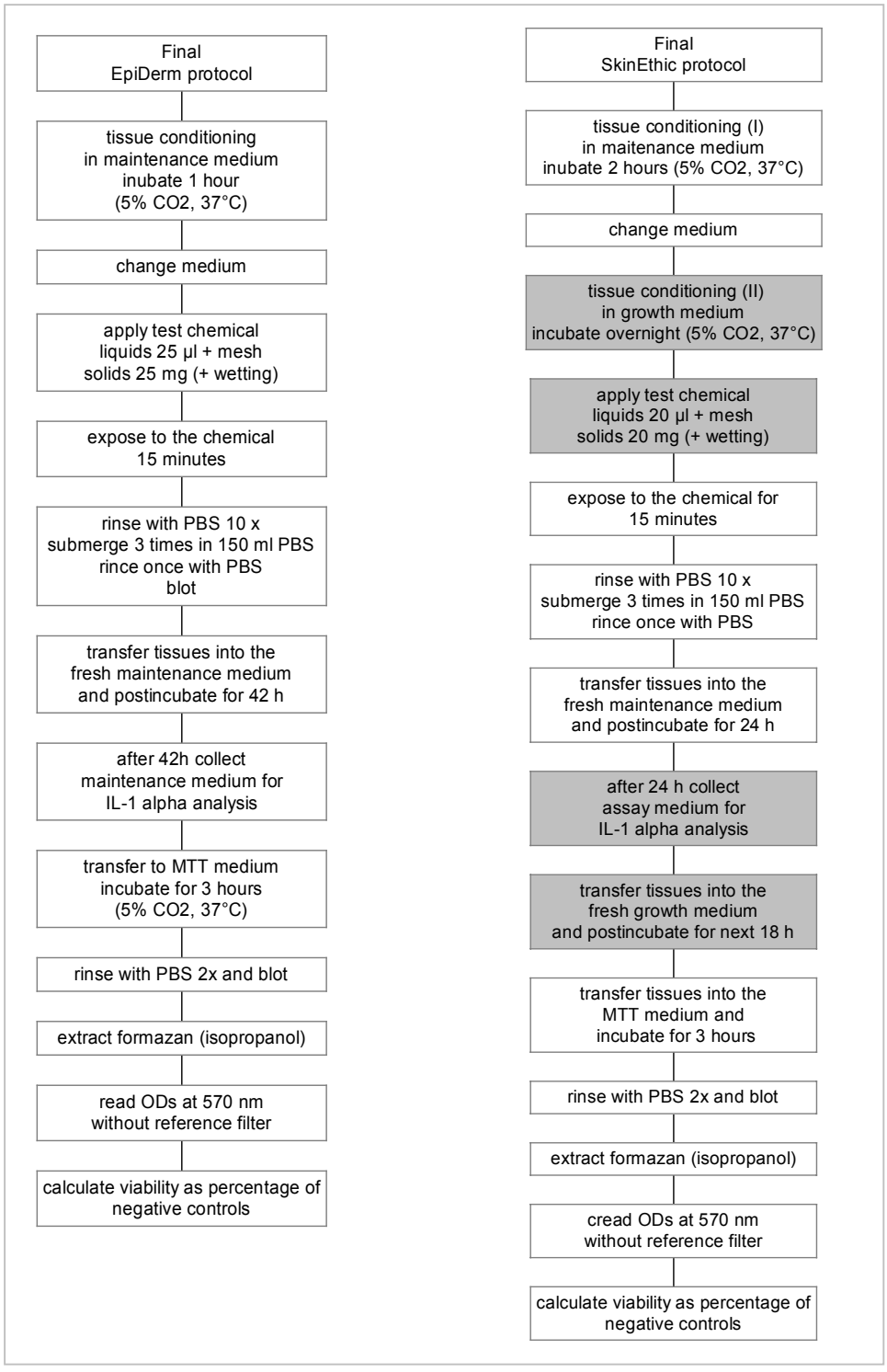

Figure 34. Comparison of the final EpiDerm protocol (left side) and the adapted SkinEthic protocol (right side). The differences between the two protocols are marked with grey colour.

As in the EpiDerm study, after the protocol optimisation, three independent runs were performed with three different batches of the SkinEthic model. The results are summarised in Table 33.

Table 33. Results obtained in Phase I with the SkinEthic model.

No.	Chemical	Relative cell viability % 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%)	1.4	1.50	0.6	0.28	2.8	0.31	1.6	0.6	2.5
2	1,1,1-Trichloroethane	12.5	4.89	2.9	0.57	2.9	1.33	6.1	1.9	10.3
3	Potassium hydroxide (5%)	0.8	0.26	0.6	0.31	2.6	0.43	1.3	0.6	2.1
4	Heptanal	1.6	0.30	1.5	0.11	1.4	0.64	1.5	1.2	1.8
5	Methyl palmitate	122.7	2.19	109.9	0.49	105.7	8.51	112.8	106.0	119.5
6	Lilestralis/Lilial	1.3	0.28	2.3	0.54	2.6	0.68	2.1	1.5	2.6
7	1-Bromopentane	1.3	0.40	1.8	0.85	2.9	0.27	2.0	1.4	2.7
8	*dl*-Citronellol	1.0	0.15	0.3	0.42	0.7	0.14	0.7	0.4	1.0
9	d-Limonene	1.6	1.01	0.7	0.16	1.0	0.31	1.1	0.6	1.6
10	10-Undecenoic acid	14.1	12.07	2.7	0.28	4.4	1.95	7.0	0.8	13.3
11	Dimethyl disulphide	4.4	2.46	2.4	3.26	6.4	1.83	4.4	2.2	6.6
12	Soap 20/80 coconut oil/tallow	90.8	1.12	101.8	2.03	112.0	3.43	101.5	94.3	108.7
13	*Cis*-Cyclooctene	0.8	0.51	0.6	0.26	0.9	0.24	0.8	0.5	1.0
14	2-Methyl-4-phenyl-2-butanol	78.0	19.63	6.6	7.13	73.1	16.29	52.6	24.1	81.0
15	2,4-Xylidine	0.3	0.23	0.0	0.23	1.3	0.17	0.5	0.1	1.0
16	Hydroxycitronellal	60.3	3.34	116.0	2.06	74.2	11.27	83.5	63.7	103.3
17	3,3'-Dithiodipropionic acid	82.9	1.01	97.8	6.25	113.5	3.66	98.1	87.5	108.6
18	4.4-Methylene bis-(2.6-di *-tert-* butyl) phenol	90.6	3.30	92.8	1.39	115.1	4.24	99.5	90.2	108.7
19	4-Amino-1,2,4-triazole	80.2	0.69	118.5	1.68	111.1	2.84	103.2	89.7	116.8
20	3-Chloronitrobenzene	79.9	2.24	113.9	1.91	103.8	0.79	99.2	87.5	110.9

SD = standard deviation.
95 % CI = 95 % confidence interval (based on 9 single values - 3 tissues per chemical in 3 independent experiments)

In the group of 10 irritants, one chemical (#5; methyl palmitate) was not predicted in concordance with the *in vivo* rabbit test. Although this chemical causes severe irritation to rabbits, the human patch test shows negative results (Basketter *et al*., 1999; ECETOC 2002; Basketter *et al*., 2004). In addition, all reconstructed human skin models predicted this chemical as being non-irritant (Fentem *et al*., 2001; Portes *et al*., 2002, Heylings *et al*., 2003; Kandárová *et al*., 2004, Cotovio *et al*., 2005).

In the group of 10 non-irritants, three chemicals, #11 (dimethyl disulphide), #13 (cis-cyclooctene) and #15 (2,4-xylidine), were classified as false positives by the SkinEthic model. Dimethyl disulphide and 2,4-xylidine provoked similar results in EPISKIN and EpiDerm models (Portes *et al*. 2002; Kandárová *et al*., 2004; Cotovio *et al*. 2005).

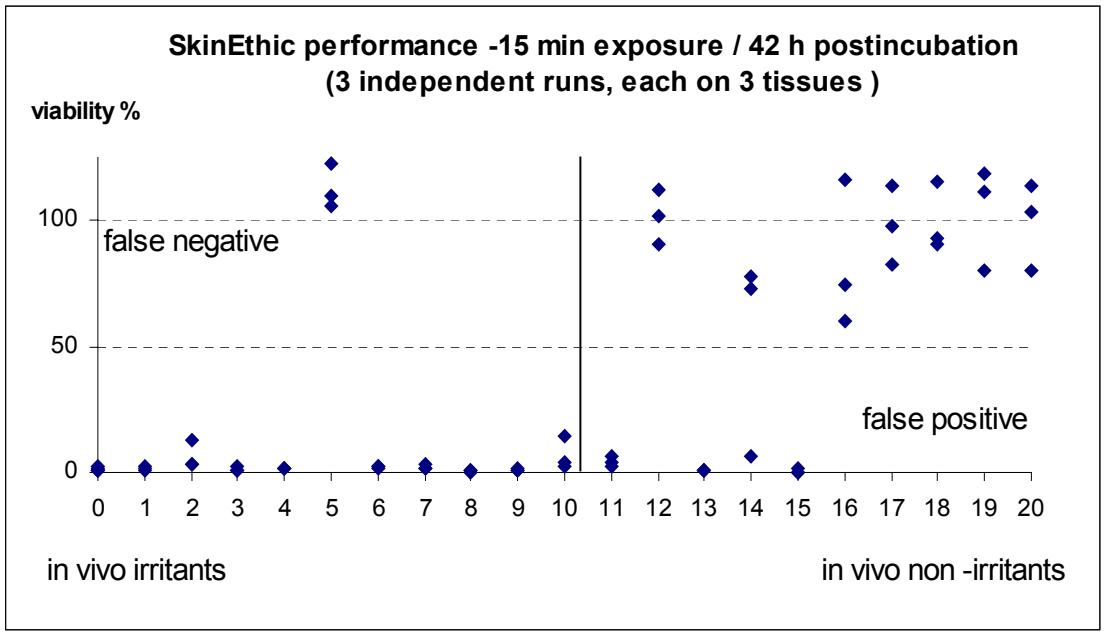

Figure 35. Performance of the SkinEthic RHE model with a "common protocol" and 20 samples from the ECVAM Skin irritation validation study.

With the chemicals tested, a sensitivity of 90 % and a specificity of 74 % were achieved with the SkinEthic model and the "common skin irritation protocol". The overall accuracy was 82% (Table 34). Chemical #11 (dimethyl disulphide) was excluded from these calculations due to the insufficient evidence of the *in vivo* classification.

Table 34. Contingency Table for Phase I.

		***In vivo* classification**		
		Irritant	Non-irritant	Unclear
	Irritant	27	7	0
	Non-irritant	3	20	0
***In vitro* prediction**				

Statistics for shadowed area of 2x2 table	
Sensitivity	90 %
Specificity	74 %
Positive predictive value	79%
Negative predictive value	87 %
Accuracy	82 %
Prevalence of test set	1.11

Assay variability

Evaluation of variability revealed following:

1) variability between single tissues treated identically within a single run was low.
2) variability between the three runs was low for most cases.

Significant variability within and between runs was observed only for chemical # 14 (2-methyl-4-phenyl 2-butanol) with a standard deviation (refereed to tissue viability) in two measurements above 15 %. Moreover, the predictions disagreed between runs. For chemical # 16 (hydroxycitronellal) significant differences in viability values were observed, however, correct prediction was obtained in all three runs. Overall, the variability of results in SkinEthic assay was low and comparable to the variability obtained with the EpiDerm assay (see Figure 35 and compare with Figure 31).

Probability of correct classification

The probability of correct classification was calculated separately for each chemical and run (see Figure 36). The grey columns show, for each chemical and run, the statistical likelihood of being classified correctly, whereas the dashed columns show the likelihood of being classified falsely. Any test result in which a test chemical has been classified correctly with a likelihood >75% can be regarded as sufficiently robust.

In the group of ten *in vivo* irritating chemicals (upper part of Figure 36), 9 chemicals were correctly classified with a likelihood of >75%. Only chemical #5 (methyl palmitate) was classified falsely negative (as non-irritant) with a likelihood of more than >75%. Of the nine *in vivo* non-irritating chemicals (dimethyl disulphide was excluded), five chemicals revealed "clear-cut" correct negative predictions with a likelihood of >75%. Chemicals #15 (2,4-xylidine) and #13 (cis-cyclooctene) revealed a clear likelihood of being classified falsely 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 one false positive prediction (mean tissue viability 6,6 %). The other chemical (#16, hydroxycitronellal) was classified correctly, but in one run with a likelihood of less than 75%.

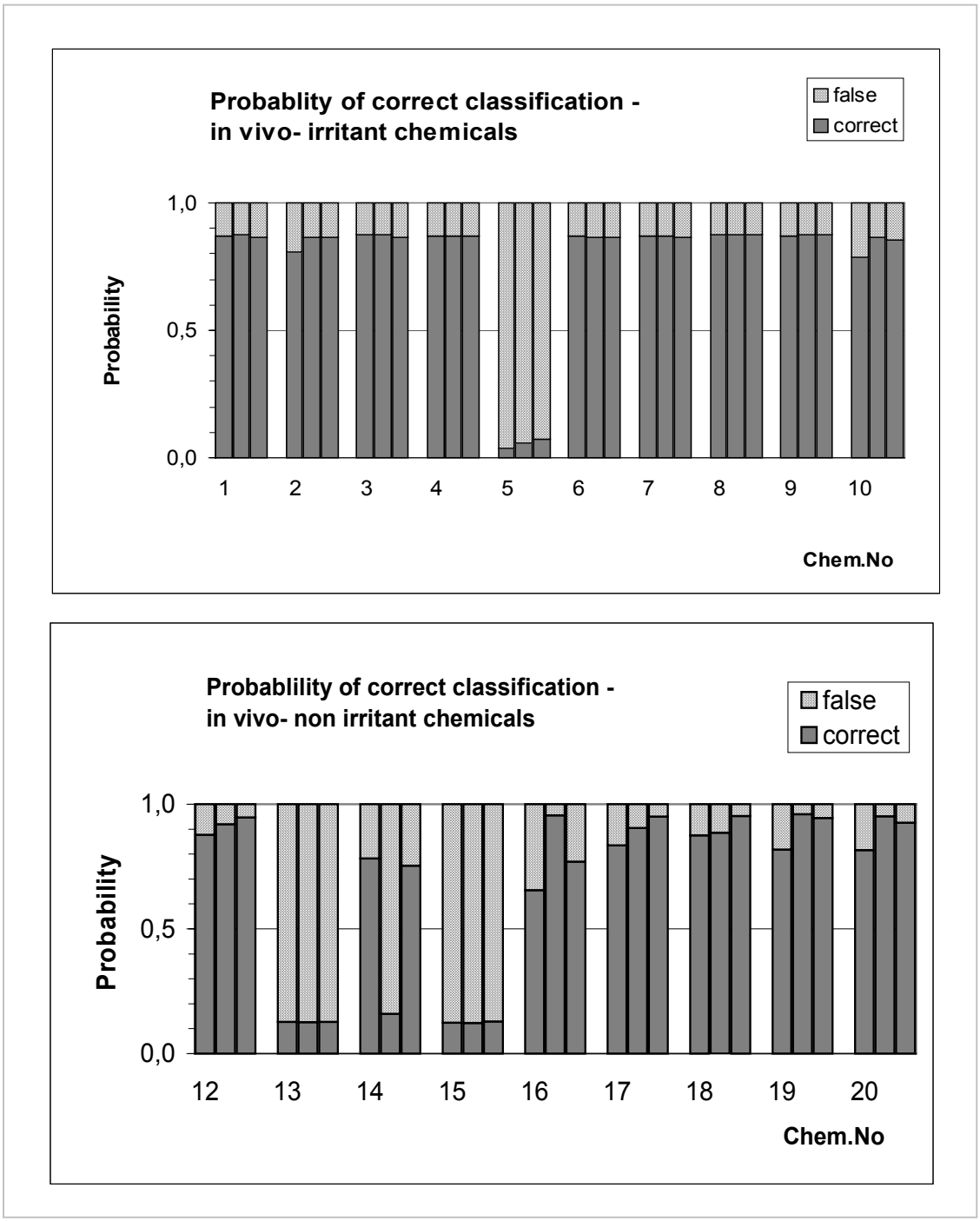

Figure 36. Probability of correct prediction for the 20 chemicals tested in Phase I.

The probability for correct classification was calculated for each test chemical and run. The upper part of figure 1 (a) shows *in vivo* irritant chemicals (1-10), the lower part (b) shows *in vivo* non irritant chemicals (12-20). Values for the probability range between 0 and 1 correspond to 0 - 100%. The grey columns show for each chemical and for each run the statistical likelihood of the correct classification. A robust (relevant) test result should be backed by a correct classification probability of >75%.

Phase II: Inter-laboratory assessment of the protocol transferability and interleukin analysis

In phase two, six substances from the EpiDerm and EPISKIN optimisation studies (Cotovio *et al*., 2005 Kandárová *et al*., 2005) were selected to evaluate the protocol transferability between laboratories. In addition, investigations if the analysis of interleukin 1α release into the assay medium will contribute to the final prediction have been performed.

For the six coded substances, correct predictions were obtained in both laboratories in all experimental runs (see Table 35 and Figure 37).

Table 35. Relative tissue viability (%) obtained in SCHERING and ZEBET laboratories in three independent runs.

No.	Chemical name	SCHERING						ZEBET					
		Run 1		Run 2		Run 3		Run 1		Run 2		Run 3	
		mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
1	Lilestralis	3.4	0.4	0.7	0.1	1.4	0.2	3.2	0.2	0.9	0.0	0.9	0.0
2	Hydroxycitronellal	83.5	3.8	91.9	2.8	95.2	1.1	102.1	4.9	80.9	0.9	97.3	5.0
3	Isopropyl myristate	98.9	14.0	94.5	19.7	122.7	3.8	94.4	16.2	99.5	1.6	102.7	2.3
4	Sodium bicarbonate	85.1	4.8	78.0	23.5	105.5	3.5	100.5	8.2	84.6	1.4	89.7	0.6
5	Lauric acid	71.3	30.7	100.1	2.6	119.2	5.5	106.1	2.5	60.0	9.5	99.9	5.0
6	1-Bromohexane	3.1	0.4	0.7	0.1	1.2	0.2	2.7	0.4	0.8	0.0	0.9	0.1

SD = standard deviation.

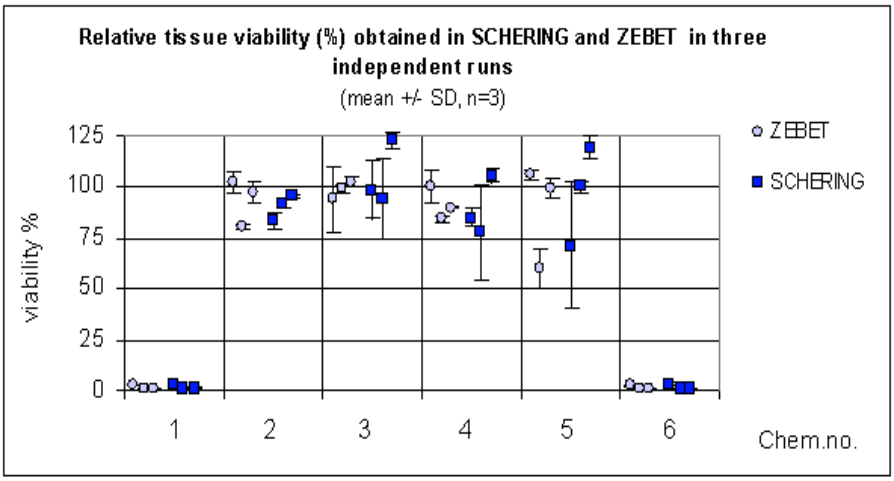

Figure 37. Interlaboratory study performed between ZEBET and SCHERING – MTT assay.

Variability between runs and between laboratories was very low for chemicals #1 (lilestralis), #2 (hydroxycitronellal) and #6 (1-bromohexane). For chemicals #3 (isopropyl myristate) and #4 (sodium bicarbonate) variability was in the acceptable range in both laboratories. Significant variation was observed only with chemical #5 (lauric acid) (see Figure 37).

Table 36. Statistical analysis of the inter-laboratory variability of the MTT assay.

Chem. no.		Variation	Intra-Batch	Between Batches	Total Total
1	Lilestralis	absolute	1.3	0.1	1.2
		relative (%)	76.7*	5.8	68.6*
2	Hydroxycitronellal	absolute	8.9	2.3	8.1
		relative (%)	9.7	2.5	8.8
3	Isopropyl myristate	absolute	11.1	4.6	10.2
		relative (%)	10.9	4.5	10.0
4	Sodium bicarbonate	absolute	11.6	1.5	10.4
		relative (%)	12.9	1.6	11.5
5	Lauric acid	absolute	24.6	5.8	22.1
		relative (%)	26.5	6.3	23.9
6	1-Bromohexane	absolute	1.2	0.2	1.1
		relative (%)	75.8*	10.0	68.0*

***** *in case of low viability, the relative variability provides disproportionately high values although the absolute differences are very low.*

The analysis of the culture medium for content of IL1-α supported the classifications based on MTT assay. Significant and constant interleukin release (above 70 pg/ml in all three runs) was observed only for chemicals #1 (lilestralis) and #6 (1-bromohexane) which also provided a strong positive result in the MTT assay. A slight increase of the interleukin level was noted for chemical # 5 (lauric acid) in the first run at Schering and the second run at ZEBET (Figure 38).

Although the results of interleukin assay support the classification based on results of the MTT assay, the variability of IL1-α assay was high and the inter-laboratory reproducibility was not optimal. More data should be generated to allow reliable estimation of the predictive power of this additional endpoint.

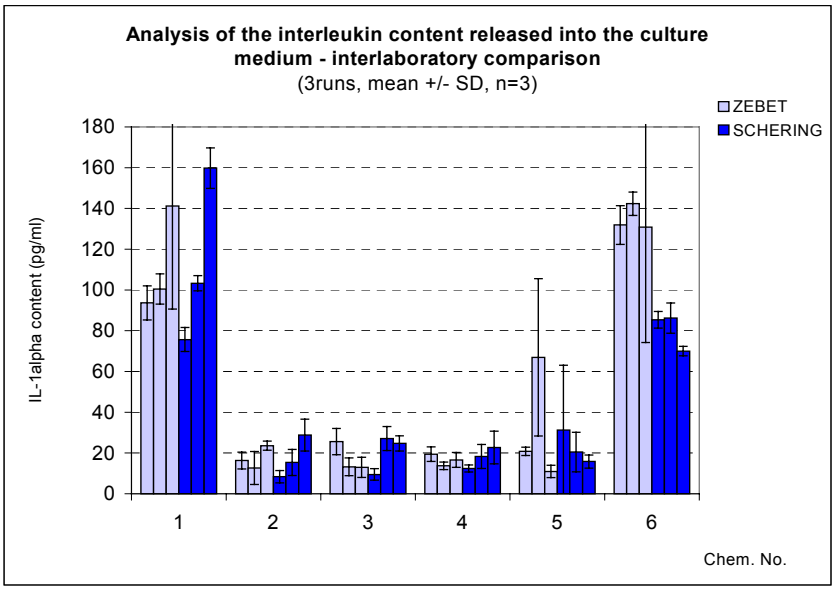

Figure 38. Interlaboratory study performed between ZEBET and SCHERING, IL1-α assay.

5.3.4 DISCUSSION

Adaptation of the protocol

As mentioned above, the SkinEthic skin irritation protocol is based on the optimised EpiDerm skin irritation assay. Overall, only minor adjustments in the protocol were needed to obtain results similar to EpiDerm and EPISKIN. The main difference between the EpiDerm and SkinEthic skin irritation protocols is the application volume of the test chemical and pre-incubation technique.

The EpiDerm and SkinEthic model have the same size (surface 0.63 cm^2), therefore the application volume (dose) should be the same in both models – 25 µl or 25 mg. However, for SkinEthic model the dose of 25 µl / mg was too high. Improved results were obtained when the application volume was decreased (to 20 µl for liquids and to 20 mg for solids. Due to the high hydrophobicity of the epidermal surface of SkinEthic model, the mesh application technique (used in EpiDerm model only for liquids) was extended also for solids that were completely dissolved after wetting with 20 µl of H_2O.

For the EpiDerm, EPISKIN and SkinEthic reconstructed human skin models different pre-incubation techniques are required. Therefore, this had to be considered during adaptation of the SOP. The use of growth medium during the second phase of the pre-incubation and during the entire post-incubation period was recommended by SkinEthic Laboratories.

According to the most recent modifications of the “common protocol” for the EPISKIN and EpiDerm models, the interleukin content is analysed after 42 hours of post-incubation. However, it seems that the main IL-1 α release occurs during the first 24 h, and the release during the next 18 h is only minor (experiments performed at Schering, data not shown). In addition, prolonged storage of the model in the medium (up to 42 h) at 37 °C may cause degradation of interleukins (Cotovio *et al*., 2005).

Due to the possible degradation of interleukins and also due to the cultivation conditions recommended by SkinEthic Laboratories, we decided to adjust the protocol and analyse the IL-1α content in the growth medium already after 24 h of post-incubation period. The medium was collected and kept frozen at –20 °C until analysis. The tissues were further post-incubated in fresh growth medium until the period of 42 hours was completed.

Assay performance and impact of specific factors on prediction of the skin irritation potential

The SkinEthic RHE model and its performance with the common skin irritation protocol was evaluated in two phases. First, the 20 chemicals used in Phase III of the ECVAM prevalidation study, and also during "follow-up" studies with EpiDerm and EPISKIN models (Portes *et al*., 2002; Cotovio *et al*., 2005; Kandárová *et al*. 2004, 2005), were used for the optimisation and evaluation of the predictive power. Next, the interlaboratory transferability of the protocol was assessed in Phase II by testing six coded chemicals in two laboratories (ZEBET and SCHERING).

Although the set of the “20 pre-validation” chemicals contains several difficult substances like 2,4-xylidine, dimethyl disulphide, 2-methyl-4-phenyl-2-butanol or *cis*-cyclooctene, testing of the complete set of the 20 chemicals was considered to be essential to allow direct comparison between performances of the SkinEthic, EpiDerm and EPISKIN RHE models.

2,4-xylidine and dimethyl disulphide where consistently over-predicted by all models and methods as irritants (Fentem *et al*., 2001; Portes *et al*., 2002, Heylings *et al*., 2003, Kandárová *et al*. 2004, Cotovio *et al*., 2005). Moreover, 2,4-xylidine was predicted in all *in vitro* skin corrosion studies with reconstructed human skin models as a corrosive chemical (Fentem *et al*., 1998, Liebsch *et al*., 2000; Kandárová *et al*., 2006). As expected, these two chemicals were predicted falsely in the SkinEthic skin irritation assay, too.

The SkinEthic RHE model over-predicted also *cis*-cyclooctene (chemical #13). Interestingly, whereas EPISKIN and SkinEthic RHE models consistently predict this substance as being irritating, the EpiDerm model classifies it as NI, in concordance with the EU classification. In the GHS classification systems, *cis*-Cyclooctene belongs to the group of mild irritants.

At the beginning of the study, several over-predictions for 3-chloronitrobenzene were obtained (data not shown). The reason was an insufficient washing procedure, which did not allow proper removal of the test substance from the SkinEthic surface. After evaluation of the washing procedure’s efficiency using light microscopy, it became clear that significant amounts of colourless crystals of the test substance remained at the surface. 3-chloronitrobenzene is poorly soluble in water (and PBS) and adheres strongly to the SkinEthic epidermal surface. During the long post-incubation time (42 hours) this toxic substance caused a significant decrease of tissue viability. When the remaining crystals were removed at the beginning of the post-incubation period with the cotton swab, the chemical was predicted correctly as non-irritant. In general, the washing procedure (= removal of the residues of the test chemicals) seems to be a critical issue not only for *in vitro* experiments,

but also for *in vivo* assays. As shown with 3-chloronitrobenzene, the efficiency of the washing procedure may have a significant impact on final results and consequently on the classifications.

One of the most interesting results was obtained for hydroxycitronellal. Three different samples of this chemical were tested, obtaining three different predictions. At ZEBET, in most cases, the original samples distributed by BIBRA TNO in Phase III of the pre-validation study were tested. SkinEthic Laboratories purchased the same chemicals and provided them to ZEBET for additional testing purposes. One of the hydroxycitronellal samples purchased by SkinEthic Laboratories was clearly predicted as irritant, the second sample revealed 50 % viability and the original sample from the pre-validation study (tested several times also on EpiDerm model) was predicated as non-irritant. The experiment was performed on the same tissue batch, therefore inter-batch variations can be excluded. It can be concluded, that although the purity of the samples was higher than 95 %, impurities and/or age of the different samples must have been responsible for the different results. These factors should be taken into account when controversial predictions are obtained in validation or follow–up studies.

Another interesting issue was to compare the prediction for 1-bromopentane and 1-bromohexane toxicity with different RHE models. Both chemicals were under-predicted as non-irritant by the EpiDerm model (Kandárová *et al*., 2004; 2005), whereas EPISKIN (Cotovio *et al*., 2005) and SkinEthic models classified these chemicals in concordance with *in vivo* rabbit data. One of the possible reasons for the different outcome might be different attributes of the stratum corneum (SC) of the three RHE models. As described by Ponec *et al*. (2000), the stratum corneum of SkinEthic and EPISKIN RHE models seems to be compact, while the EpiDerm model depicts a basket-wave pattern like *in vivo*. The higher number of corneocyte layers seems to be also a typical attribute of the SC of SkinEthic and EPISKIN models. Most probably, due to the specific shipping (4 °C) and storage conditions, the surface of EpiDerm cultures is quite hydrated and therefore the spreading of lipophilic chemicals requires to use the mesh. EPISKIN and SkinEthic epidermis have a rather dry and hydrophobic surface. It is almost impossible to spread aqueous solutions or water on the surface of SkinEthic epidermis without the mesh as spreading tool.

In the “common skin irritation protocol”, the surface of the RHE models is exposed to the test substance for 15 minutes. Then the chemical is rinsed off in a procedure which should assure the maximum achievable removal of the test substance. During the exposure time the test substance may penetrate into the SC, and depending on the characteristics of the SC and the chemical itself, it will reach different layers of the skin model. It may happen, that due to a thicker SC, a chemical is not properly removed from the tissue surface and residues form depot in the stratum corneum. During the post-exposure period, these

residues may reach the viable layers of the reconstructed human skin model, and cause toxicity as seen with 3-chloronitrobenzene and probably also *cis*-cyclooctene, 1-bromopentane and 1-bromohexane.

Not knowing the human responses to these chemicals, we have to rely on classifications based on *in vivo* rabbit data. However, more attention should be paid to the selection of test chemicals for future validation studies. For the development and validation of a new *in vitro* method for skin irritation, only those chemicals should be selected, for which well documented *in vivo* data are available. Moreover, when feasible, human patch tests should be performed. Only then it will become possible to assess the capacity of the reconstructed human skin models to predict the human response.