Skin Irritation Validation Study
Phase II: Analysis of the primary endpoint MTT and the secondary endpoint IL1-α

Report from the study biostatistician to the management team

Sebastian Hoffmann
European Commission, JRC (Joint Research Centre)
IHCP - Institute of Health and Consumer Protection
ECVAM
Via E. Fermi 1
21020 Ispra (VA), Italy

Table of contents

Table of tables

Table of Figures

Annexes

1 Introduction

Two alternative methods for skin irritation testing proceeded after showing promising results in a Phase I to a complete formal ECVAM validation study. The aim of this study is to validate *in vitro* skin irritation tests in a collaborative study, in order to replace the Draize skin irritation test performed on rabbits according to Method B.4 of Annex V to *Directive 67/548/EEC* or OECD TG 404.[1] The primary goal of this validation study is the scientific evaluation of the ability of the *in vitro* tests to reliably discriminate skin irritants (I) from non-irritants (NI), as defined with EU risk phrases (R38; no label) according to the Dangerous Substances Directive, *67/548/EEC*.[2] A secondary goal of this study is to retrospectively analyse the data to assess if the *in vitro* tests reliably discriminate between strong, mild and non-irritants, as defined by the 'Globally Harmonised System (GHS)' for classification and labelling, adopted by the United Nations.[3]

The Phase I was conducted in early 2004 in order to preliminary assess the performance regarding the within-laboratory variability and predictive capacity of the three tests EpiDerm, EPISKIN and SIFT (skin integrity function test). Twenty blinded chemicals were tested with each of the tests in the lead laboratories only running three independent runs, i.e. experiments. EpiDerm and EPISKIN measure cell viability via MTT, where the SIFT measures two endpoints (trans-epithelial water loss and electrical resistance). The results of this phase (Annex I) guided the decision of the Management Team (MT) whether a test should proceed to a formal collaborative validation study: The two commercially available reconstituted human epidermis models EpiDerm and EPISKIN showed promising results in terms of within-laboratory variability and predictive capacity so that they proceeded to Phase II. However, the third test SIFT (skin integrity function test) performed in a way that the MT decided that this test needs further optimisation and thus did not enter Phase II.

The objective of this report is to summarise and present a complete, objective and transparent analysis of within-laboratory and between-laboratory reproducibility as well as predictive capacity based on the submitted test data.

2 Material and Methods

2.1 Tests

Both the EpiDerm and the EPISKIN are commercially available reconstituted human epidermis models and the endpoint measured in these assays is cell viability via MTT. Thus identical analysis could be applied to both test data, which is presented in Chapter 3.

In addition, IL-1α was introduced as a second endpoint, which was performed with EpiDerm at ZEBET and at the three EPISKIN-laboratories to assess its usefulness in an exploratory manner. The analysis of this second endpoint is presented in Chapter 4.

2.2 Study design

After successful transfer and training to two additional laboratories per test, both EpiDerm and EPISKIN were performed in three laboratories (Table 1).

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

Table 1: Participating laboratories in Phase II

Each laboratory tested the same set of sixty chemicals in three runs each. These chemicals were coded and distributed by RCC-CCR (Rossdorf, Germany), a subcontracted laboratory. Thus chemicals were tested blind. The blinding code was provided to RCC-CCR by ECVAM. Contact between the laboratories during the testing was not allowed in order to safeguard the blinding.

2.3 Test chemicals

A Chemical Selection Sub-Committee (CSSC) was established to choose the test chemicals. This committee comprised regulators from the European Chemical Bureau (ECB) and the competent authorities of the UK and Germany as well as ECVAM staff. The criteria for the chemical selection for Phase I were in general maintained in Phase II. A summary of these criteria is given in Annex II. A detailed report will be provided by the CSSC. However, minor changes regarding the year of notification of new chemicals and the commercially availability were done.

For two of the 60 chemicals confidentiality issues with the respective producer/notifier could not be resolved, so that in Table 2 information on only 58 chemicals is given. Besides the substance name, the CAS-number and the database source, a number is assigned, which is used throughout this document as the chemical identifier. The chemicals were recruited from three different data bases: 33 chemicals were taken from the New Chemicals Database (NCD) of the European Chemicals Bureau (ECB), 19 from an ECETOC database[4] and six from the TSCA (Toxic Substances Control Act) database. All subsequent analyses are based on these 58 chemicals. Annexes III, IV, and V contain further information on the chemicals: Physical-chemical properties, the chemicals' skin irritation classifications and the blinding codes are given in Annex III, where the molecular structures are presented in Annex IV. Annex V contains the rabbit data, on which the classifications were based in case of the ECTOC and the TSCA databases. It does not contain the in vivo data of the NCD chemicals due to confidentiality issues.

chemical number	substance name	CAS-number	source
1	2-chloromethyl-3,5-dimethyl-4-methoxypyridine hydrochloride	86604-75-3	NCD
2	1-bromo-4-chlorobutane	6940-78-9	ECETOC
3	1-bromohexane	111-25-1	ECETOC
4	1-decanol	112-30-1	ECETOC
5	3-chloro-4-fluoronitrobenzene	350-30-1	ECETOC
6	3-diethylaminopropionitrile	5351-04-2	ECETOC
7	3-mercaptohexanol	51755-83-0	NCD
8	4-methylthio-benzaldehyde	3446-89-7	ECETOC
9	2,6-dimethyl-4-nitrobenzeneamine	16947-63-0	NCD
10	allyl heptanoate	142-19-8	ECETOC
11	allyl phenoxyacetate	7493-74-5	ECETOC
12	2-ethylhexyl 4-aminobenzoate	26218-04-2	NCD
13	1-[4-(2-dimethylaminoethoxy)phenyl]-2-phenylbutan-1-one	68047-07-4	NCD
14*			
15	a-terpineol	98-55-5	ECETOC
16	capryl-isostearate	209802-43-7	NCD
17	2-methyl-3-[(1,7,7-trimethylbicyclo[2.2.1]hept-2-yl)oxy]-1-propanol, bornyl isomer	128119-70-0	NCD
18	butyl methacrylate	97-88-1	TSCA
19	2,5-dimethyl-4-oxo-4,5-dihydrofuran-3-yl acetate	4166-20-5	NCD
20	cyclamen aldehyde	103-95-7	ECETOC
21	A mixture of: 5-exo-decylbicyclo[2.2.1]hept-2-ene; 5-endo-decylbicyclo[2.2.1]hept-2-ene	22094-85-5	NCD
22	diethyl phthalate	84-66-2	ECETOC
23	di-n-propyl disulphide	629-19-6	ECETOC
24	di-propylene glycol	25265-71-8	ECETOC
25	dipropylene glycol monobutyl ether	29911-28-2	TSCA
26	3,4-dimethyl-1H-pyrazole	2820-37-3	NCD
27	2-isopropyl-2-isobutyl-1,3-dimethoxypropane	129228-21-3	NCD
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
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
30	Mixture of: diethyl cis-1,4-cyclohexanedicarboxylate; diethyl trans-1,4-cyclohexanedicarboxylate	0072903-27-6	NCD
31	A 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 (mix).	NCD
32	2S-(2-furyl)-5R-hydroxy-4R-(1R,2-dihydroxy)ethyl-6S-hydroxymethyl-1,3-dioxane	7089-59-0	NCD
33	heptyl butyrate	5870-93-9	ECETOC
34	hexyl salicylate	6259-76-3	ECETOC
35	cyclohexadecanone	2550-52-9	NCD
36	isopropanol	67-63-0	ECETOC
37	[2-(cyclopentyloxy)ethyl]benzene(cyclopentyl 2-phenylethyl ether)	not allocated	NCD
38*			
39	methyl stearate	112-61-8	ECETOC
40	1-methyl-3-phenyl-1-piperazine	5271-27-2	NCD
41	naphthalene acetic acid	86-87-3	TSCA
42	disodium 2,2'-(1,4-phenylene)bis-(1H-benzimidazole-4,6-disulfonic acid or monosulfonic acid, monosulfonate or disulfonate	180898-37-7	NCD
43	A 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
44	phenylethylalcohol	60-12-8	ECETOC
45	(+/-) trans-3,3-dimethyl-5-(2,2,3-trimethyl-cyclopent-3-en-1-yl)-pent-4-en-2-ol	107898-54-4	NCD
46	4-methyl-8-methylenetricyclo[3.3.1.1(3,7)]decan-2-ol	122760-84-3	NCD
47	4-methyl-8-methylenetricyclo[3.3.1.1(3,7)]dec-2-yl acetate	122760-85-4	NCD

48	2-(formylamino)-3-thiophenecarboxylic acid	43028-69-9	NCD
49	isostearic acid monoisopropanolamide	152848-22-1	NCD
50	2-phenylhexanenitrile	3508-98-3	NCD

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
		(mix.)
			NCD

52	propyl (2S)-2-(1,1-dimethylpropoxy)-propanoate	0319002-92-1	NCD
53	silane A-1430	2530-87-2	TSCA

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

55	terpinyl acetate	80-26-2	ECETOC
56	benzenethiol, 5-(1,1-dimethylethyl)-2-methyl *(NB: CAS name from company)*	7340-90-1	NCD
57	triethylene glycol	112-27-6	TSCA
58	tri-isobutyl phosphate	126-71-6	TSCA
59	(E,E)-3,7,11-trimethyldodeca-1,4,6,10-tetraen-3-ol	125474-34-2	NCD
60	bis[(1-methylimidazol)-(2-ethyl-hexanoate)], zinc complex	not allocated	NCD

Table 2: List of 58 Phase II chemicals
(* confidential chemicals)

2.4 Data submission

A data submission template in EXCEL was developed in a collaborative effort between the leading laboratories and ECVAM for both tests. Agreeing on a final version, this version was password-protected by ECVAM and then provided to the leading laboratories, which passed them on to their participating laboratories. The spreadsheet containing the test data had to be returned to the biostatistician of the MT only. The data were excepted if the password-protection was still in place.

2.5 Quality criteria

Although the quality criteria controlling the consistency and interpretability of the MTT measurements are defined in the SOP of the two tests, they are presented here as the level of compliance might reveal useful information. The quality criteria addressed the responses and variability of the negative (NC) and the positive control (PC) as well as the variability of the test samples (Table 3). The variability criterion was established by an analysis of the Phase I data (Annex VI) and adjusted following a blinded interim analysis by the MT (Annex VII). The threshold value for the standard deviation of 18 can be interpreted in the way, that replicate measurements should always cover less than a third of the possible response scale of cell viability, i.e. 0 to about 100%,

The subsequent analyses were performed once including all test sample data and once excluding those samples, which did not produce three runs passing the test sample variability criterion.

Quality criteria	EpiDerm	EPISKIN
NC: absolute response	> 1.0 OD	> 0.6 OD
PC: mean viability	≤ 20%	≤ 40%
NC: variability	SD < 18	SD < 18
PC: variability	SD < 18	SD < 18
test samples: variability	SD < 18	SD < 18

Table 3: Quality criteria for EpiDerm and EPISKIN according to their SOPs

2.6 Data sets to be analysed

In the interpretation of the results of a validation study several aspects, e.g. focusing on biases or on balance in the data set to be analysed, have to be considered. This might lead to different views on how calculation of reproducibility and predictive capacity for a given test method should be performed. To highlight this problem, four potential options and some of their advantages and disadvantages in the context of this biostatistical report were considered.

1. The first option would be to analyse all reported data without excluding any information, even if it would be justified by quality criteria. In the scope of this study the importance of the quality criteria regarding the controls were never questioned. However, omitting the test sample criterion (Table 3), which controlled variability, would have benefits, as this would result in a not perfectly, but highly balanced data set. The consequences would be that the necessity of the criterion might be questioned and that highly problematic/variable data would enter the analysis. With an increasing amount of the latter interpretation of results would become more and more difficult.
2. The second option would be to analyse all valid runs, i.e. runs, which met the variability criterion of SD < 18. This would result in the largest data set fulfilling all variability criteria. However, some imbalance are likely to be introduced, e.g. one chemical might in total have had two or three valid runs only, while another did not reveal any problems in any laboratory and might enter the dataset with a total of nine valid runs.
3. The third option would be to analyse data of those chemicals, which had three valid runs in a given laboratory. In this approach the dataset would be further reduced, as here e.g. all data of a chemical with two valid and two non-valid runs in a given laboratory would be excluded. However, a chemical might have three valid runs in one laboratory, but not in another. In this way, the imbalance of the dataset would further increase. Assuming that chemicals might tend to be either problematic or unproblematic over all laboratories, options two and three would give very similar results.
4. The last option would be to consider only data from chemicals, which had three valid runs in all three laboratories. This would result in a perfectly balanced data set. However, the number of chemical considered might be severely reduced. Furthermore, a bias resulting in overestimation of

reproducibility and predictive capacity is very likely as all chemicals, which were problematic in one of the three laboratories, would be excluded.

For this report, initially the first three options were chosen before the results were known. In a Management Team meeting in February 2006 it was agreed to include only the results calculated using the first option, i.e. considering all data, and the third option, i.e. considering all data of chemicals, for which three valid runs (in terms of the variability criterion) were available. Any post-hoc rationalisation in this choice was avoided. For completeness, it has to be pointed out that in the few cases, when more than three valid runs were available, the first three were considered.

Finally, the Management Team agreed in the final meeting in May 2006 that when evaluating the overall outcome for each test the main emphasis would be put on the analysis undertaken for each chemical with three valid runs.

2.7 Within-laboratory variability

The within-laboratory variability of the primary cell viability endpoint of each of the participating six laboratories was analysed with four statistical techniques. Here as well as in all further analyses the run is considered as the experimental unit, which better reflects current use and potentially applications in the future. These range from very rigorous, i.e. aiming to detect optimal reproducibility, to less demanding approaches aiming to provide a detailed and complete evaluation of the within-laboratory variability. The most rigorous technique applied was a 1-way ANOVA comparing the data of the three runs for each single chemical, where a significance level of 1% was chosen. Second, the within-laboratory standard deviation (s_R), a measure of repeatability according to ISO standards[5], was determined for each chemical over the runs. It has to be noted that this measure is not completely transferable from the ISO guidance as variable aspects of testing, e.g. the operator and material used, were not systematically included in the study design. In a third step, the correlation according to Bravais-Pearson was calculated to compare all three pairs of runs. Finally, the predicted classification resulting from the prediction models (PM) were compared between the runs by a simple measure of similarity, i.e. the proportion of identical predictions.

In general, these parameters were calculated considering all available experimental runs - allowing direct comparisons of all laboratories - and considering the runs of those chemicals, for which three runs met the variability criterion for test samples, i.e. a SD < 18. The latter approach can be expected to result in unbalanced data sets over the laboratories, as one laboratory might not have produced three valid runs for a given chemical, while the other laboratories have. For reasons of simplicity and clarity, an analysis of all valid runs was excluded based on a Management Team decision taken at a meeting in Berlin (February 2006). The respective results of such an approach can be expected to range between the two chosen approaches.

In addition, the results of 18 chemicals, which were tested in the lead laboratories in both phases, were compared. Depending on the particular test method, chemical and run, these phases were between five and twelve month apart, where the same operator performed the experiments in the two phases for each of the test methods. Regarding this aspect, a t-test (significance level of 1%) per substance using the raw data was applied comparing the results of both phases. Furthermore, a paired t-test was calculated with the mean run results of the phases.

The within-laboratory variability of the secondary IL1-α endpoint was evaluated similarly. Omitting the calculation of correlations, the coefficient of variation (CV) was additionally introduced. The comparison of the two phases could be performed with eleven, respectively twelve chemicals.

2.8 Between-laboratory variability

The variability between the three laboratories of the primary endpoint was assessed with three statistical techniques. First, a 1-way ANOVA/t-test comparing the data of the three/two laboratories for each single chemical (significance level of 1%) was applied, where the mean values of the triplicates were used. Second, taking the run mean per laboratory, the standard deviation of these three means was calculated. Third, the proportion of identical run classifications and identical median run classifications over the three laboratories was evaluated.

As for within-laboratory reproducibility, these parameters were calculated considering all available experimental runs and considering the runs of those chemical, for which three runs met the variability criterion for test samples, i.e. a SD < 18. For reasons of simplicity and clarity, an analysis of all valid runs was excluded. The respective results can be expected to range between the two chosen approaches.

Additionally, the chemicals posing difficulties in the single laboratories regarding the quality criterion were compared.

The between-laboratory variability of the secondary endpoint was evaluated for the controls employing a 1-way ANOVA with a Bonferroni post-test comparing the laboratories pair-wise (significance level of 0.01). Furthermore, the standard deviation and CV of the mean IL1-α release of the three laboratories were calculated.

2.9 The reference test

As currently the Draize rabbit test for skin irritation[6] is the most frequent test foreseen in regulations, the MT chose this test as the reference test for comparison. The respective rabbit dominant median, a concept developed by Hoffmann et. al,[7] of the 58 chemicals are presented in Table 4 together with their classification and dominant endpoint. The dominant median is a concept developed to simplify the interpretation of 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, i.e., dominant one. This median allows classification of a chemical according to both the European classification scheme (ECS)[2] and the Globally Harmonised System (GHS)[3] by comparison with the classification cut-off points, i.e., 2 in the case of ECS or 1.5 and 2.3 in the case of GHS.

	classification		dominant	dominant
chemical	ECS	GHS	median	endpoint
1	R38	I	2.7	B
2	no label	NI	0.0	B
3	R38	I	2.7	E
4	R38	I	2.3	E
5	no label	NI	1.0	E
6	no label	NI	0.0	B
7	no label	NI	0.0	B
8	no label	NI	1.0	E
9	no label	NI	0.3	E
10	no label	MI	1.7	E
11	no label	NI	0.3	E
12	no label	NI	0.7	E
13	R38	MI	2.0	E
14*				
15	R38	I	2.7	O
16	no label	NI	1.0	E
17	no label	MI	1.7	E
18	R38	I	3.0	E
19	no label	NI	0.0	B
20	R38	I	2.3	O
21	no label	MI	1.7	E
22	no label	NI	0.0	E
23	R38	I	3.0	E
24	no label	NI	0.0	E
25	no label	NI	0.0	E
26	no label	NI	0.0	B
27	R38	I	4.0	E
28	no label	NI	0.0	B
29	R38	MI	2.0	B
30	no label	NI	1.3	E
31	R38	MI	2.0	O
32	no label	NI	0.0	B
33	no label	MI	1.7	E
34	R38	MI	2.0	B
35	no label	NI	0.0	B
36	no label	NI	0.3	E
37	R38	I	3.0	E
38*				
39	no label	NI	1.0	E
40	R38	I	3.3	E
41	no label	NI	0.0	B
42	no label	NI	0.0	B
43	R38	MI	2.0	B
44	no label	NI	1.0	E
45	R38	I	2.7	E
46	R38	MI	2.0	B
47	R38	MI	2.0	B
48	no label	NI	0.0	B
49	R38	MI	2.0	E
50	no label	MI	1.7	E
51	R38	MI	2.0	E
52	no label	NI	0.7	E
53	no label	NI	0.0	B
54	no label	NI	1.3	E
55	R38	MI	2.0	B
56	R38	I	3.3	O
57	no label	NI	0.0	B
58	R38	MI	2.0	E
59	R38	I	4.0	E
60	R38	MI	2.0	E

Table 4: Classification of the 60 chemicals according to European system and to the GHS together with the dominating median score and the dominating endpoints
(NI: non irritant; MI: mild irritant; I: Irritant; E: eryhtema; O: oedema; B: both; *confidential chemicals)

Annex V contains the rabbit data, from which, in case of the ECETOC and the TSCA databases, the classifications were derived by applying the respective classification scheme. It does not contain the in vivo data of the NCD chemicals due to confidentiality issues. However, the official European classifications of the chosen NCD chemicals were in line with the in vivo data. In addition, these chemicals were classified according to the globally harmonised system (GHS).

2.10 Predictive capacity

As the test systems were designed to predict the EU risk phrases, i.e. R38 for skin irritants and no label for non-irritants, the predictions and the respective European classification of the chemicals were combined in 2x2 contingency tables. From these tables the predictive capacity was calculated in terms of sensitivity, specificity, accuracy and positive and negative predictive value (PPV, NPV). These parameters were determined for the endpoint MTT with the cell viability of samples relative to the respective negative control. For the endpoint IL1-α, the fold-increase in comparison to the negative control and the total IL1-α amount released were analysed in this way. In general, the predictive parameters were calculated considering all available experimental runs and considering the runs of those chemical, for which three runs met the variability criterion for test samples, i.e. a SD < 18. For reasons of simplicity and clarity, an analysis of all valid runs was excluded. The respective results can be expected to range between the two chosen approaches.

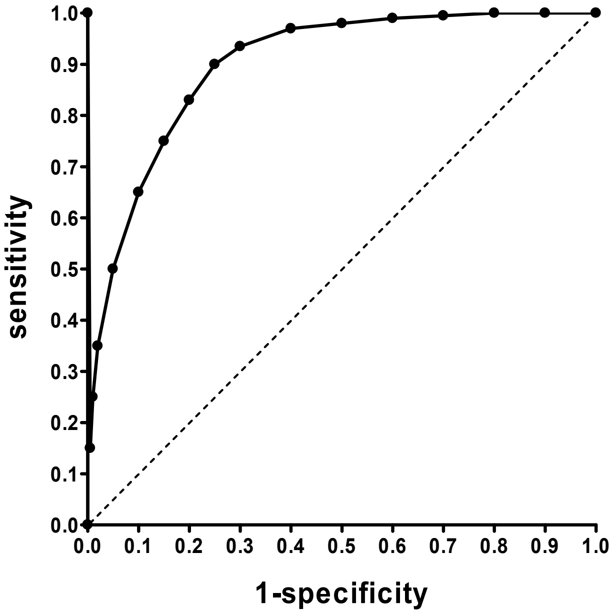


Figure 1: Example of a receiver operation curve (ROC), where the dotted line of identity indicates the ROC of a useless test

Receiver operation curve (ROC) analysis was performed to check how shifting of the PM-thresholds of the test systems to discriminate irritants from non-irritants affects the predictive capacity. This approach is the most common way to assess diagnostic test in medicine.[8,9] The parameters sensitivity and (1-specificity) are calculated for each observed data value and plotted against each other, as exemplarily shown in Figure 1. Thus, a test with a curve close to the line of identity would be useless, while a test with a curve approaching the upper left corner of the plot has increasing merit. The sum of sensitivity and specificity, which weighs the two parameters equally, was chosen to assess the ROC.

Additionally, the *in vivo* test data, which were used to classify the employed chemicals, were correlated with the endpoints of the new test systems. Therefore, the concept of the dominating median[7] was applied in order to reduce the *in vivo* data to a one-dimensional measure while the loss of information was minimized. Extracting the median for each of the endpoints of the *in vivo* experiment, i.e. erythema and oedema, and choosing the larger one results in the dominating median of a given chemical.

In addition, a further PM for the endpoint IL1-α was based on comparison of a samples response and the negative control response. Therefore, a 1-way ANOVA with Dunnett's post test, which was designed for such type of comparison,[10] was applied.

The secondary aim, the assessment of the test systems performance in terms of the Globally Harmonised System (GHS) was done in a post-hoc analysis. As the results from Phase I did not allow to define a PM, two thresholds maximising the accuracy were chosen for each test method. Disregarding aspects of reproducibility, the median classifications for chemicals with three valid runs were chosen for this analysis.

2.11 Statistics

All calculations were either performed in Microsoft EXCEL 2002, Graphpad Prism 4.02 or S-Plus 6.2. Regarding the reproducibility aspects, 1-way ANOVA techniques (confidence level of 1%) and descriptive measures, such as standard deviation, coefficient of variation, correlation and similarity measures were applied. In terms of predictive capacity, contingency tables were applied, where, when appropriate, confidence intervals for the estimated parameters are reported. ROC-analysis was used to describe the predictive capacities in a more complete manner.

3 Results MTT

3.1 EpiDerm

3.1.1 Data submission

ZEBET, the lead laboratory for the EpiDerm assay, submitted the data to ECVAM on 05.06.2005. One operator tested all sixty chemicals between the 24.10.2004 and the 23.05.2005, where 15 chemicals were tested per run. In the provided spreadsheet no remarks were reported.

The data from IIVS were received in the agreed format on 07.07.2005. Up to 34 chemicals were tested in one run. As the provided spreadsheet had a maximum capacity for 30 chemicals, in a few cases data of one run had to be submitted in two separate files. Two operators performed the experiments between the 28.01.2005 and the 29.07.2005 and no remarks were made in the spreadsheet. As the data for four tests did not fulfil the quality criteria, these cases were, according to a decision of the MT, retested and their data submitted on the 01.08.2005.

BASF submitted the data on the 13.06.2005 and a missing part, whose absence was only recognised later, on the 06.07.2005. Two operators tested either 15 or 30 chemicals per run between the 24.11.2004 and the 13.07.2005, where no remarks in the spreadsheets were reported. However, as the data for eight tests did not fulfil the quality criteria, these cases were, according to a decision of the MT, retested and their data submitted on the 18.07.2005. Furthermore, the MT agreed that a run, in which the negative control did not met the variability criterion of a SD < 18, does not have to be repeated. It was decided to exclude the one replicate causing the high variation. However, classification of the chemical tested in this run was not effected by this exclusion.

3.1.2 Analysis of quality criteria

In total, five data related quality criteria were included in the validation SOP of EpiDerm, where the first four addressed the controls and the fifth the tested sample. The first one demanded a mean response (in OD) of the negative control larger than 1.0 OD. As shown in Figure 2, this was always the case. To allow a comparison with Phase I, the respective data were added. The flexible experimental test set-up (in terms of amount of chemicals tested per sub-set) caused the different sample sizes per laboratory. While there was no significant difference (1-way ANOVA with Bonferroni post test) between the two ZEBET data sets and between IIVS and BASF, the negative controls at ZEBET were significantly smaller than at IIVS and BASF.

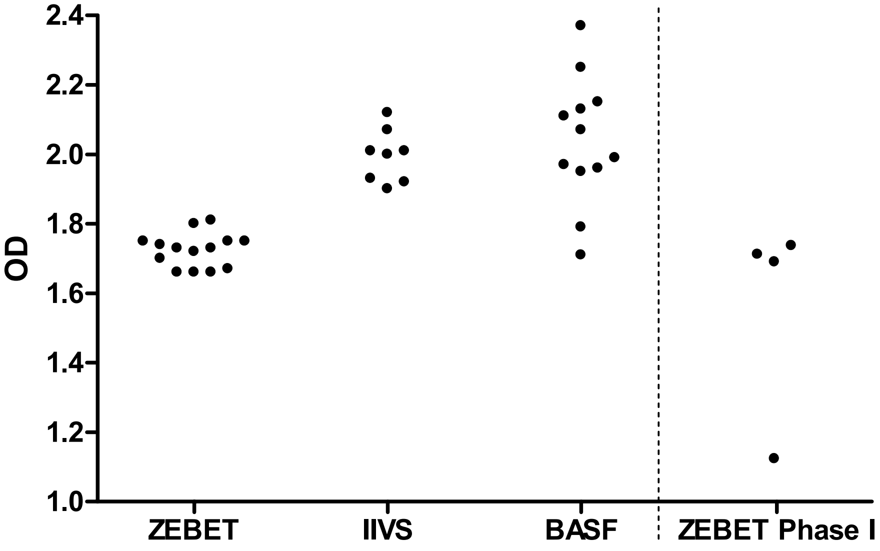


Figure 2: Response of the negative controls in the three laboratories in Phase II and the lead laboratory in Phase I

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. In Figure 3, the data for all laboratories and both phases show that this criterion was always met. However, a more variable response of the positive control could be observed for ZEBET in Phase II, which was also significantly larger than the respective responses at IIVS and BASF.

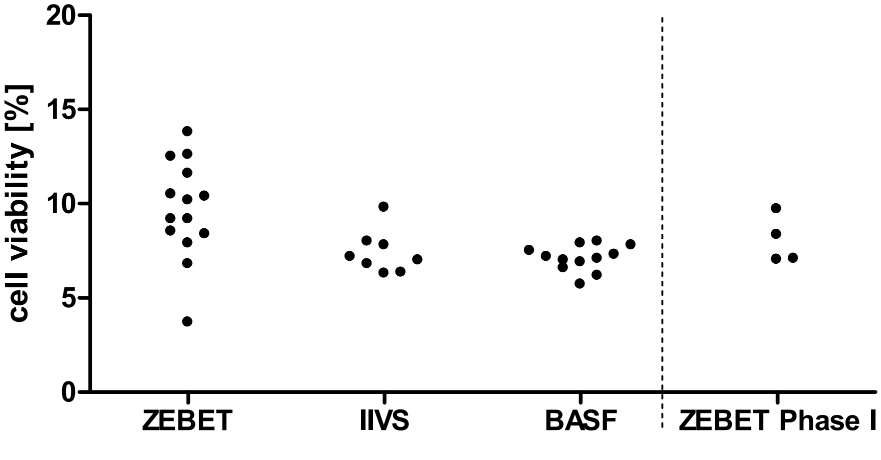


Figure 3: Relative response of the positive controls in the three laboratories in Phase II and the lead laboratory in Phase I

The third criterion demanded that the variability in terms of standard deviation (SD) of the negative control replicates was smaller than 18. The fourth criterion was identical, but referred to the variability of the positive control. The data for both controls are shown in Figure 4, where one negative control at BASF did not fulfil the criterion. However, the aberrant replicate causing this variability was excluded according to a MT decision.

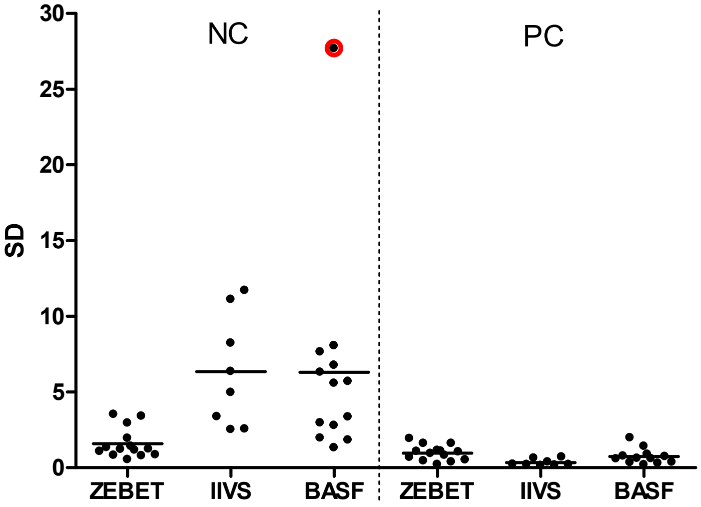


Figure 4: Variability measure as standard deviation (SD) of the negative (NC) and positive controls (PC) in the three laboratories in Phase II
(The red marked data point was due to one outlying replicate, which was excluded from analysis resulting in an SD of the respective run of 1.96.)

Table 5 clearly shows that in the lead laboratory ZEBET less test substances were retested. In total, with the 58 chemicals 193 tests were performed, of which 18 tests did not meet the variability criterion. As the variability criterion before the interim analysis was fixed at a standard deviation of 11, some chemicals were retested according to this criterion. Chemical 7 was retested, because it had physical-chemical properties in run 1, which did not met those as described. The first run of chemical 1 it was reported that a wrong chemical was tested so that these data were not considered at all. At IIVS, 196 tests were carried out, where 32 of these did not meet the SD-criterion. BASF performed 200 tests, 36 of which had an unacceptable variability. In this laboratory all tests of the fourth run were triggered by failure of the SD-criterion.

chemical number	ZEBET						IIVS					BASF			
	run						run					run			
	1	2	3	4	5	6	1	2	3	4	5	1	2	3	4
1	0.9	0.2	0.4				1.0	0.4	1.3			0.1	0.6	0.1	
2	0.5	6.1	**18.2**	11.2			**35.3**	0.9	**22.5**	11.5		11.8	1.5	4.1	
3	2.3	3.7	4.5				10.0	1.4	7.2			3.4	2.3	8.5	
4	0.5	0.5	0.8				0.6	0.4	0.2			15.2	**18.5**	**43.4**	9.0
5	1.8	4.3	2.2				**20.2**	**50.9**	**43.7**	12.9		17.3	13.7	9.2	
6	2.5	8.4	**31.5**	11.8			10.0	**22.6**	13.4	**19.0**		**21.0**	6.3	6.4	7.7
7	8.3	3.8	29.0	31.3			**18.5**	6.6	**19.4**	**23.7**		**20.1**	11.3	**25.9**	**27.3**
8	4.0	1.6	2.1				10.1	**55.9**	5.7	5.4		9.1	8.4	5.5	**51.7**
9	5.7	2.5	3.8				4.1	11.7	6.6			11.7	5.5	**35.6**	10.6
10	3.0	4.6	3.1				3.2	5.0	3.8			14.8	4.2	11.9	
11	3.2	4.9	3.2				9.4	6.8	4.9			**18.1**	6.3	5.9	0.9
12	2.9	1.3	3.2				4.4	4.2	0.6			0.6	8.4	3.0	
13	4.0	17.8	15.3	7.5	12.7		**18.1**	**36.5**	**28.4**	38.2		**46.1**	**44.8**	**31.8**	**43.0**
14*															
15	**52.9**	**33.3**	**29.6**	15.6	**24.2**		0.1	**27.9**	0.4	**47.1**		12.5	**41.4**	**37.9**	9.6
16	1.2	0.6	0.7				5.2	3.8	4.4			5.0	1.5	3.5	
17	1.5	1.3	1.0				0.8	0.5	0.6			0.8	12.3	1.0	
18	3.0	2.9	1.1				8.9	6.1	7.2			4.2	8.5	**23.4**	**35.2**
19	7.1	9.9	5.7				7.5	11.1	9.2			5.6	0.9	8.7	
20	2.5	2.0	16.0	3.8			3.8	0.2	7.7			**18.2**	14.7	4.3	14.1
21	4.4	1.8	2.5				5.5	3.5	7.2			11.2	4.9	2.5	
22	1.8	5.8	3.7				7.2	7.1	4.4			5.7	4.3	2.7	
23	1.6	5.3	0.3				3.5	0.9	2.8			14.1	9.3	5.7	
24	6.6	1.0	0.3				5.1	0.9	5.2			7.7	1.3	4.6	
25	8.4	3.4	1.4				4.4	4.9	5.0			12.4	3.7	**24.0**	4.6
26	0.8	1.6	0.7				7.1	8.2	10.2			0.2	1.3	1.3	
27	4.7	2.1	2.3				6.2	7.0	4.8	2.4		11.6	8.8	2.1	
28	0.7	3.9	0.3				5.8	4.5	11.7	7.1		6.9	2.6	4.0	
29	0.6	0.2	2.3				0.8	0.2	0.2			0.5	0.5	**41.6**	0.3
30	5.9	1.2	2.8				2.7	1.5	0.7			3.7	5.2	4.2	
31	**28.5**	4.3	1.4	2.9			**45.8**	15.4	**22.1**	3.7		0.2	**40.0**	0.1	6.1
32	1.0	2.6	3.0				14.9	13.2	2.8	3.2		5.2	2.0	4.7	
33	0.5	3.4	2.8				3.9	**19.8**	7.2	13.5		17.5	10.8	3.1	
34	1.8	2.8	1.7				3.2	1.5	1.1			16.3	7.2	1.9	4.7
35	2.2	4.3	2.0				4.5	8.0	3.8			0.7	0.6	0.6	
36	4.9	3.9	2.7				11.1	10.1	11.7			11.0	4.1	5.8	
37	4.1	3.2	1.0				12.9	7.4	13.8			13.3	14.7	7.0	6.1
38*															
39	1.0	3.0	2.4				3.4	0.9	5.9			2.8	5.8	4.3	
40	5.6	6.4	3.8				6.3	4.2	6.1	1.5	14.1	16.7	**49.6**	16.8	0.3
41	6.5	1.9	3.1				3.9	2.3	3.4			3.0	2.4	6.7	
42	1.8	3.4	2.1				1.7	3.7	10.1			4.9	10.0	4.4	
43	3.2	1.1	0.7				3.4	11.5	6.5			14.4	4.2	**20.1**	12.1
44	**31.4**	**36.1**	5.0	**43.9**	16.1	**49.8**	**38.2**	0.9	**43.7**	**49.5**		**23.5**	2.8	4.9	5.4
45	0.6	1.2	0.5				0.6	0.4	0.6			8.0	17.9	1.0	
46	0.9	2.7	0.4				6.5	8.1	9.2			0.6	0.7	0.4	
47	2.0	17.7	17.7	0.6			1.0	**37.1**	16.6	**27.0**		**36.7**	0.3	0.9	0.1
48	5.9	2.1	0.6				3.6	2.9	3.9			14.4	1.2	0.3	
49	3.0	2.9	1.9				**18.5**	1.9	3.3	6.7		1.8	6.3	0.5	
50	7.2	17.9	12.4	13.0			7.9	4.9	**22.3**	1.1		2.7	5.3	17.2	
51	0.3	1.6	1.8				6.1	0.5	0.3			8.1	8.5	9.5	
52	2.3	6.6	0.9				9.5	**20.1**	3.1	1.9		7.3	**40.0**	2.2	**24.2**
53	8.7	1.0	3.3	1.7			8.7	9.3	3.6			14.3	14.7	**41.0**	**33.0**
54	11.6	14.1	8.2	6.4			4.3	4.8	3.8	11.5		16.1	15.3	1.6	15.8
55	**23.8**	7.2	**28.4**	**18.2**			0.4	0.2	9.2			**22.3**	15.3	6.5	**25.7**
56	1.9	0.3	1.7				1.0	17.4	0.3			1.0	1.2	2.3	
57	3.9	5.9	1.4				5.4	5.4	1.9			7.8	5.7	5.1	
58	**36.5**	**45.1**	5.7	1.1	4.5		0.2	**49.9**	**23.2**	**19.6**	**53.4**	0.7	**41.9**	**43.5**	10.3
59	0.8	1.3	1.0				0.7	1.8	0.0			**18.3**	16.2	**36.8**	7.1
60	1.4	0.9	0.8				0.8	0.4	0.6			4.0	12.0	**21.7**	0.4

Table 5: Standard deviations of all runs and substances in the three laboratories (grey cells: SD > 18; *: confidential chemicals)

3.1.3 Within-laboratory variability

3.1.3.1 ZEBET

3.1.3.1.1 1-way ANOVA

To compare the independent experiments of a given chemical within a laboratory, first a 1-way ANOVA was applied. If there was a significant difference between any pair of experiments to the level of 1%, the chemical was considered not to be reproducible in terms of this measure. The ANOVA was calculated once for all available runs of a given chemical and once for those chemicals, which had three runs meeting the variability acceptance criterion. If in the later case more than three acceptable runs were available, the first three were considered.

Table 6 shows that eight chemicals had non-acceptable runs. Chemical numbers 2, 6 and 31 had only one non-valid, but three valid runs. Chemical numbers 7, 15, 44, 55 and 58 had more than one non-valid run. For these three valid runs were not available with the exception of chemical number 58. Here, the MT decided to consider the three valid out of the total of five runs. Twenty-two out of the 58 chemicals had a p-value smaller than 0.01. The same 22 chemicals still had a p-value smaller than 0.01 when only three valid runs were considered. The four chemicals, for which less than three valid runs were available, were very likely to be reproducible in terms of the ANOVA, as data with high variability enter the comparison. Consequently, for those chemicals with individual runs showing a small variability small difference between runs can result in a significant result, which might not have any relevance. In fact, ZEBET had the lowest within-run variability (see Chapter 3.1.3.4).

chemical	number of runs	number of non-valid runs	p-value	p-value of three valid runs
1	3	0	0.0014	0.0014
2	4	**1**	0.8943	0.7357
3	3	0	0.4701	0.4701
4	3	0	0.0057	0.0057
5	3	0	0.0064	0.0064
6	4	**1**	0.0252	0.3388
7	4	**2**	0.2386	-
8	3	0	0.0035	0.0035
9	3	0	0.0568	0.0568
10	3	0	0.0620	0.0620
11	3	0	0.0053	0.0053
12	3	0	0.7175	0.7175
13	5	0	<0.0001	0.0009
14*				
15	5	**2**	0.6848	-
16	3	0	0.0077	0.0077
17	3	0	0.0046	0.0046
18	3	0	0.2736	0.2736
19	3	0	0.7017	0.7017
20	4	0	0.037	0.0564
21	3	0	0.0282	0.0282
22	3	0	0.0212	0.0212
23	3	0	0.0028	0.0028
24	3	0	0.2541	0.2541
25	3	0	0.0604	0.0604
26	3	0	0.0091	0.0091
27	3	0	0.4859	0.4859
28	3	0	0.0155	0.0155
29	3	0	0.1120	0.1120
30	3	0	0.4090	0.4090
31	4	**1**	0.5441	0.1236
32	3	0	0.0005	0.0005
33	3	0	0.0144	0.0144
34	3	0	0.0043	0.0043
35	3	0	0.0034	0.0034
36	3	0	0.2436	0.2436
37	3	0	0.0394	0.0394
38*				
39	3	0	0.1036	0.1036
40	3	0	0.0108	0.0108
41	3	0	0.1935	0.1935
42	3	0	0.0059	0.0059
43	3	0	0.0242	0.0242
44	6	**4**	0.6749	-
45	3	0	0.0006	0.0006
46	3	0	0.0019	0.0019
47	4	0	0.1546	0.2690
48	3	0	0.3740	0.3740
49	3	0	0.3073	0.3073
50	4	0	0.0979	0.0693
51	3	0	0.0002	0.0002
52	3	0	0.0002	0.0002
53	4	0	0.0014	0.0042
54	4	0	0.0052	0.0093
55	4	**3**	0.4206	-
56	3	0	0.0085	0.0085
57	3	0	0.0143	0.0143
58	5	**2**	0.084	0.3765
59	3	0	0.0008	0.0008
60	3	0	0.1866	0.1866

Table 6: ZEBET within-laboratory reproducibility: 1-way ANOVA p-values (*confidential chemicals)

3.1.3.1.2 Within-laboratory standard deviation s_R

Also the within-laboratory standard deviation was calculated for all available runs per chemical and for the first three valid runs per chemical. The data for all chemicals are displayed in Table 7. Transferring the value of 18 from the variability criterion to this type of standard deviation, four chemicals (number 6, 13, 54, 58) showed a $s_R > 18$ when considering all runs. Focusing on the three valid runs, only two chemicals (number 13, 58) had a $s_R > 18$. This can be interpreted as evidence that the variability criterion of SD > 18 supports the reproducibility of the test by identifying highly variable runs, which tend to be aberrant. Although only descriptive, this measure of within-laboratory variability is highly informative and well interpretable. The distribution of s_R in the three laboratories is compared in Chapter 3.1.3.4.

chemical	number of runs	run 1	run 2	run 3	run 4	run 5	run 6	s_R	
								all runs	three valid runs
1	3	5.57	5.35	8.24				1.61	1.61
2	4	86.98	90.56	91.95	85.93			2.86	2.43
3	3	94.53	96.25	98.38				1.93	1.93
4	3	11.12	13.50	13.35				1.33	1.33
5	3	96.65	99.65	108.70				6.27	6.27
6	4	94.12	82.94	42.48	87.65			23.34	5.62
7	*4*	*74.82*	*49.58*	*45.23*	*36.14*			*16.56*	-
8	3	95.06	103.16	90.15				6.57	6.57
9	3	93.76	103.38	94.48				5.36	5.36
10	3	97.02	100.29	105.90				4.49	4.49
11	3	87.18	103.92	95.27				8.37	8.37
12	3	104.11	105.57	103.96				0.89	0.89
13	5	102.65	75.40	20.30	96.24	46.69		34.57	41.95
14*									
15	*5*	*72.50*	*43.27*	*71.30*	*81.04*	*58.60*		*14.71*	-
16	3	102.93	99.66	100.61				1.68	1.68
17	3	11.72	14.20	8.36				2.93	2.93
18	3	98.14	96.28	99.92				1.82	1.82
19	3	86.87	91.07	92.03				2.75	2.75
20	4	16.48	16.49	37.25	16.17			10.44	11.99
21	3	94.72	107.15	101.84				6.24	6.24
22	3	106.51	93.55	97.61				6.63	6.63
23	3	90.97	91.34	104.73				7.84	7.84
24	3	97.30	100.57	103.16				2.94	2.94
25	3	92.04	97.34	105.19				6.61	6.61
26	3	9.26	10.54	6.28				2.18	2.18
27	3	103.33	101.40	99.93				1.70	1.70
28	3	90.80	98.72	95.72				4.00	4.00
29	3	9.81	11.80	12.61				1.44	1.44
30	3	97.99	95.40	99.86				2.24	2.24
31	4	28.59	15.17	11.57	17.76			7.33	3.11
32	3	91.06	106.14	103.38				8.03	8.03
33	3	97.20	105.59	98.66				4.48	4.48
34	3	97.03	106.68	102.82				4.86	4.86
35	3	87.17	101.38	95.92				7.17	7.17
36	3	97.54	95.01	101.10				3.06	3.06
37	3	44.34	37.73	45.72				4.27	4.27
38*									
39	3	100.21	95.91	96.15				2.42	2.42
40	3	32.18	21.78	11.93				10.12	10.12
41	3	94.29	99.14	101.51				3.68	3.68
42	3	95.73	102.24	106.49				5.42	5.42
43	3	100.41	104.11	106.63				3.13	3.13
44	*6*	*46.32*	*67.75*	*88.90*	*70.23*	*87.83*	*67.15*	*15.73*	-
45	3	13.02	12.11	7.93				2.72	2.72
46	3	9.89	16.26	7.86				4.38	4.38
47	4	9.39	30.66	22.66	7.68			10.98	10.74
48	3	101.99	97.85	101.47				2.26	2.26
49	3	98.56	94.88	96.35				1.85	1.85
50	4	56.83	80.80	86.76	71.09			13.06	15.84
51	3	100.97	110.23	111.65				5.80	5.80
52	3	78.36	100.75	111.24				16.80	16.80
53	4	84.67	109.36	98.20	102.26			10.38	12.36
54	4	90.37	97.66	55.71	73.71			18.68	22.41
55	*4*	*73.25*	*95.60*	*88.15*	*69.14*			*12.43*	-
56	3	12.60	13.11	18.01				2.99	2.99
57	3	86.80	96.94	101.10				7.35	7.35
58	5	62.22	52.57	13.45	8.66	12.96		25.33	2.64
59	3	20.67	14.10	17.60				3.29	3.29
60	3	9.24	7.77	9.44				0.91	0.91

Table 7: ZEBET within-laboratory standard deviation s_R
(light grey cells: runs, which were not considered for three valid runs; dark grey cells: chemicals with $s_R > 18$; italic: chemicals without three valid runs;
* confidential chemicals)

3.1.3.1.3 Correlation

The third measure of reproducibility within a laboratory was the Bravais-Pearson correlation coefficient r. It was applied to correlate the results of two complete runs. However, as already seen in Phase I, the value of this measure is limited due to the fact that the test protocol was designed to separate irritants from non-irritants. Thus, the test did not produce viabilities evenly distributed over the whole response range, but lumped together at both ends. This property restricted the usefulness of the correlation coefficient as it can be expected to be high. Calculating the correlation of the mean cell viability for all three pairs of runs for the 54 chemicals with three valid runs confirmed the expectation: The correlation was in all three instances larger than 0.9 (Table 8). Considering the first three runs for all 60 chemicals resulted in correlations of the runs between 0.8689 and 0.9507.

	correlation r
Run 1 – Run 2	0.9704
Run 1 – Run 3	0.9200
Run 2 – Run 3	0.9585

Table 8: ZEBET run correlations

3.1.3.1.4 Proportion of identically classified chemicals

The crudest measure of within-laboratory reproducibility is the proportion of identically classified chemicals. The classification were derived by the Prediction Model (PM) of the SOP, i.e. a chemical, considered independently per run, would be classified as irritant when the mean viability was below 50% and as non-irritant otherwise. This measure was first applied to the 54 chemicals with three valid runs. The classifications, which can easily be derived from Table 7, are identical for 53 chemicals. Only chemical number 13 was classified non-consistently. When considering all runs, 52 out of 58 chemicals were consistently classified. In comparison, the EPISKIN test, which was the only test with three runs in the ECVAM Validation Study of in vitro test for skin corrosion, classified none to nine (depending on the laboratory) of 60 chemicals inconsistently.[11] For clarification, it has to be stressed explicitly that this measure did not take the correctness of the classifications into account, i.e. a chemical can be within-laboratory reproducible, but consistently wrongly classified.

3.1.3.2 IIVS

3.1.3.2.1 1-way ANOVA

To compare the independent experiments within a laboratory, again a 1-way ANOVA (significance level of 1%) was applied to the data of each test chemical. The ANOVA was calculated for all available runs per chemical and additionally for those chemicals, which had three runs meeting the variability acceptance criterion. If in the later case more than three acceptable runs were available, the first three were considered.

chemical	number of runs	number of non-valid runs	p-value	p-value of three valid runs
1	3	0	0.0699	0.0699
2	4	2	0.4340	-
3	3	0	0.7105	0.7105
4	3	0	0.0171	0.0171
5	4	3	0.5553	-
6	4	2	0.7754	-
7	4	3	0.1985	-
8	4	1	0.2220	0.9007
9	3	0	0.6346	0.6346
10	3	0	0.0046	0.0046
11	3	0	0.4212	0.4212
12	3	0	0.0017	0.0017
13	4	0	0.5069	-
14*				
15	4	2	0.6024	-
16	3	0	0.0020	0.0020
17	3	0	0.0033	0.0033
18	3	0	0.2148	0.2148
19	3	0	0.5821	0.5821
20	3	0	0.6711	0.6711
21	3	0	0.0909	0.0909
22	3	0	0.4535	0.4535
23	3	0	0.0002	0.0002
24	3	0	0.0184	0.0184
25	3	0	0.9756	0.9756
26	3	0	0.6718	0.6718
27	4	0	<0.0001	0.0336
28	4	0	0.2977	0.2296
29	3	0	0.0135	0.0135
30	3	0	0.1255	0.1255
31	4	2	0.0468	-
32	4	0	0.0002	0.0007
33	4	1	0.7043	0.9276
34	3	0	<0.0001	<0.0001
35	3	0	0.1693	0.1693
36	3	0	0.7922	0.7922
37	3	0	0.9463	0.9463
38*				
39	3	0	0.0773	0.0773
40	5	0	<0.0001	0.0004
41	3	0	0.0793	0.0793
42	3	0	0.2741	0.2741
43	3	0	0.3976	0.3976
44	4	3	0.2896	-
45	3	0	0.0043	0.0043
46	3	0	0.6667	0.6667
47	4	2	0.7105	-
48	3	0	0.7424	0.7424
49	4	1	0.6643	0.6113
50	4	1	0.2687	0.0083
51	3	0	0.0730	0.0730
52	4	1	0.8897	0.8365
53	3	0	0.0555	0.0555
54	4	0	0.0070	0.0322
55	3	0	0.3022	0.3022
56	3	0	0.3121	0.3121
57	3	0	0.5739	0.5739
58	5	4	0.1374	-
59	3	0	0.0319	0.0319
60	3	0	0.0023	0.0023

Table 9: IIVS within-laboratory reproducibility: 1-way ANOVA p-values

Table 9 shows that 12 of the 58 chemicals and 11 of the 48 chemicals with three valid runs were not reproducible in terms of the 1-way ANOVA.

3.1.3.2.2 Within-laboratory standard deviation s_R

Also the within-laboratory standard deviation was calculated for all available runs per chemical and for the first three qualifying runs per chemical. The data for all substances are displayed in Table 10. Transferring the value of 18 from the variability criterion to this type of standard deviation, seven chemicals (numbers 8, 27, 31, 32, 40, 44, 58) showed a $s_R > 18$ when considering all runs. Focusing on the three valid runs, only two chemicals (numbers 32 and 40) had a $s_R > 18$. This can be interpreted as evidence that the variability criterion of SD > 18 supports the reproducibility of the test by identifying highly variable runs, which tend to be aberrant. The distribution of s_R in the three laboratories is compared in Chapter 3.1.3.4.

chemical	number of runs	run 1	run 2	run 3	run 4	run 5		s_R
							all runs	three valid runs
1	3	6.60	4.32	5.90			1.17	1.17
2	*4*	*43.53*	*54.97*	*61.50*	*32.89*		*12.64*	-
3	3	98.68	103.26	99.29			2.49	2.49
4	3	7.39	7.66	6.26			0.74	0.74
5	*4*	*30.22*	*71.60*	*49.69*	*61.25*		*17.74*	-
6	*4*	*22.80*	*32.04*	*19.39*	*19.39*		*5.98*	-
7	*4*	*49.73*	*60.35*	*28.44*	*33.57*		*14.69*	-
8	4	94.12	50.43	96.58	94.26		22.31	1.39
9	3	103.75	108.25	110.14			3.28	3.28
10	3	96.01	112.52	97.74			9.07	9.07
11	3	98.29	104.36	96.33			4.19	4.19
12	3	101.17	92.46	82.19			9.50	9.50
13	*4*	*56.29*	*32.42*	*25.77*	*58.61*		*16.62*	-
14*								
15	*4*	*8.05*	*23.46*	*6.67*	*33.26*		*12.78*	-
16	3	94.31	115.68	114.09			11.91	11.91
17	3	9.65	6.58	7.48			1.58	1.58
18	3	97.66	94.45	85.81			6.13	6.13
19	3	84.97	76.92	79.02			4.17	4.17
20	3	7.65	7.54	10.83			1.87	1.87
21	3	102.27	111.23	114.17			6.20	6.20
22	3	95.34	102.31	99.06			3.49	3.49
23	3	112.26	99.29	90.78			10.81	10.81
24	3	98.52	92.39	84.46			7.05	7.05
25	3	102.52	102.12	102.99			0.44	0.44
26	3	15.67	10.35	16.20			3.23	3.23
27	4	49.56	45.31	62.27	102.93		26.28	8.83
28	4	95.00	90.57	82.46	86.29		5.41	6.36
29	3	8.80	9.20	7.59			0.84	0.84
30	3	101.52	100.66	98.04			1.81	1.81
31	*4*	*57.94*	*18.57*	*63.41*	*95.22*		*31.44*	-
32	4	54.46	117.26	119.16	99.77		30.09	36.82
33	4	96.10	98.17	94.22	86.52		5.08	5.07
34	3	92.51	109.45	88.02			11.30	11.30
35	3	119.54	110.96	110.35			5.14	5.14
36	3	81.69	85.66	87.86			3.13	3.13
37	3	47.29	48.87	45.68			1.60	1.60
38*								
39	3	104.29	104.19	96.23			4.62	4.62
40	5	88.14	88.00	53.89	17.85	27.45	32.89	19.73
41	3	96.81	102.97	96.17			3.75	3.75
42	3	111.46	104.35	102.82			4.61	4.61
43	3	106.75	101.02	97.37			4.73	4.73
44	*4*	*50.03*	*7.68*	*54.22*	*70.80*		*26.87*	-
45	3	8.88	6.48	8.06			1.22	1.22
46	3	15.83	14.93	20.59			3.04	3.04
47	*4*	*7.12*	*30.28*	*18.16*	*22.50*		*9.67*	-
48	3	98.44	98.47	96.50			1.13	1.13
49	4	91.12	98.49	90.04	97.29		4.27	4.57
50	4	71.41	82.48	81.79	92.87		8.77	10.73
51	3	116.03	120.56	112.17			4.20	4.20
52	4	99.65	103.40	97.00	97.31		2.95	1.45
53	3	89.06	108.48	98.12			9.72	9.72
54	4	46.26	52.20	39.60	66.77		11.58	6.30
55	3	7.69	8.35	14.45			3.73	3.73
56	3	9.44	20.05	7.07			6.91	6.91
57	3	92.68	95.68	96.61			2.05	2.05
58	*5*	*5.27*	*64.96*	*67.10*	*17.78*	*68.06*	*30.57*	
59	3	7.24	9.56	6.48			1.61	1.61
60	3	8.28	5.04	6.34			1.63	1.63

Table 10: IIVS within-laboratory standard deviation s_R
(light grey cells: runs, which were not considered for three valid runs; dark grey cells: chemicals with $s_R > 18$; italic: chemicals without three valid runs;
* confidential chemicals)

3.1.3.2.3 Correlation

The third measure of reproducibility within a laboratory was Bravais-Pearson correlation coefficient r. It was applied to correlate the results of two complete runs. However, as already seen in Phase I, the value of this measure is limited due to the fact that the test protocol was designed to separate irritants from non-irritants. Calculating the correlation of the mean cell viability for all three pairs of runs for the 48 chemicals with three valid runs resulted in correlation coefficients larger than 0.9 (Table 11). Also when considering the first three runs for 58 chemicals the corresponding correlations were above 0.9 (data not shown).

	correlation r
Run 1 – Run 2	0.9592
Run 1 – Run 3	0.9431
Run 2 – Run 3	0.9810

Table 11: IIVS run correlations

3.1.3.2.4 Proportion of identically classified chemicals

The crudest measure for within-laboratory reproducibility was the proportion of identically classified chemicals. The classifications were derived by Prediction Model (PM) of the SOP. This measure was again first applied to the 48 chemicals with three valid runs. The classifications, which can easily be derived from Table 10, are identical in the three valid runs for 46 chemicals. Only chemical numbers 27 and 54 would be classified non-consistently. When considering all runs, ten of 58 chemicals showed different classifications between runs.

3.1.3.3 BASF

3.1.3.3.1 1-way ANOVA

To compare the independent experiments within a laboratory, first a 1-way ANOVA (significance level of 1%) was applied to the data of each test compound. The ANOVA was calculated for all available runs per chemical and additionally for those chemicals, which had three runs meeting the variability acceptance criterion. If in the later case more than three acceptable runs were available, the first three were considered. Table 12 shows that eleven of the 58 chemicals and ten of the 48 chemicals with three valid runs were not reproducible in terms of the 1-way ANOVA. The increase in the number of chemicals is caused by higher variability of the additional runs, which then results in non-significant differences between the runs.

3.1.3.3.2 Within-laboratory standard deviation s_R

Also the within-laboratory standard deviation was calculated for all available runs per chemical and for the first three qualifying runs per chemical. The data for all substances are displayed in Table 13. Transferring the value of 18 from the variability criterion to this type of standard deviation, twelve chemicals (numbers 4, 6, 8, 13, 15, 31, 34, 40, 52, 53, 54, 59) showed a $s_R > 18$ when considering all runs. Focusing on the three valid runs, only four chemicals (numbers 6, 8, 40, 54) had a $s_R > 18$. This can be interpreted as evidence that the variability criterion of SD > 18 supports the reproducibility of the test by identifying highly variable runs, which tend to be aberrant. The distribution of s_R in the three laboratories is compared in Chapter 3.1.3.4.

chemical number	number of runs	number of non-valid runs	p-value	p-value of three valid runs
1	3	0	0.2343	0.2343
2	3	0	0.2830	0.2830
3	3	0	0.0730	0.0730
4	4	2	0.0747	-
5	3	0	0.1759	0.1759
6	4	1	0.0023	0.0001
7	4	3	0.2106	-
8	4	1	0.0196	<0.0001
9	4	1	0.2513	0.1266
10	3	0	0.1485	0.1485
11	4	1	0.0884	0.1897
12	3	0	0.0011	0.0011
13	4	4	0.6388	-
14*				
15	4	2	0.2652	-
16	3	0	0.0039	0.0039
17	3	0	0.4530	0.4530
18	4	2	0.3780	-
19	3	0	0.5085	0.5085
20	4	1	0.4504	0.7941
21	3	0	0.1479	0.1479
22	3	0	0.2700	0.2700
23	3	0	0.0680	0.0680
24	3	0	0.3885	0.3885
25	4	1	0.0775	0.1714
26	3	0	0.2762	0.2762
27	3	0	0.2932	0.2932
28	3	0	0.0004	0.0004
29	4	1	0.4339	0.0027
30	3	0	0.1361	0.1361
31	4	1	0.0414	0.3033
32	3	0	0.0829	0.0829
33	3	0	0.5016	0.5016
34	4	0	0.0006	0.0194
35	3	0	0.5221	0.5221
36	3	0	0.1231	0.1231
37	4	0	0.0758	0.0950
38*				
39	3	0	0.0097	0.0097
40	4	1	0.1497	0.0064
41	3	0	0.9757	0.9757
42	3	0	0.4361	0.4361
43	4	1	0.6243	0.5145
44	4	1	0.1409	0.1341
45	3	0	0.4413	0.4413
46	3	0	0.0043	0.0043
47	4	1	0.2726	0.0218
48	3	0	0.6349	0.6349
49	3	0	0.3348	0.3348
50	3	0	0.0939	0.0939
51	3	0	0.2055	0.2055
52	4	2	0.1575	-
53	4	2	0.0573	-
54	4	0	0.0069	0.0049
55	4	2	0.4365	-
56	3	0	0.3136	0.3136
57	3	0	0.1122	0.1122
58	4	2	0.7447	-
59	4	2	0.1562	-
60	4	1	0.5180	0.6784

Table 12: BASF within-laboratory reproducibility: 1-way ANOVA p-values

chemical number	number of runs	run 1	run 2	run 3	run 4	s_R	
						all runs	three valid runs
1	3	6.22	6.36	6.73		0.27	0.27
2	4	22.57	18.27	28.75		5.26	5.26
3	3	97.40	101.51	88.86		6.45	6.45
4	*3*	*24.11*	*88.84*	*56.94*	*64.72*	*26.73*	-
5	3	72.85	97.26	84.36		12.21	12.21
6	3	33.87	9.18	63.15	15.12	24.28	29.59
7	*3*	*82.60*	*70.67*	*58.61*	*41.67*	*17.48*	-
8	3	17.90	24.80	95.92	72.41	37.62	43.19
9	3	101.70	102.20	82.75	118.56	14.64	9.60
10	3	85.14	105.50	100.40		10.59	10.59
11	4	86.43	101.16	105.46	110.10	10.24	4.47
12	4	90.10	113.50	118.42		15.13	15.13
13	*4*	*73.89*	*36.10*	*35.91*	*38.72*	*18.53*	-
14*							
15	*3*	*50.00*	*83.47*	*55.67*	*94.33*	*21.41*	-
16	4	89.99	105.95	102.43		8.39	8.39
17	3	8.11	15.08	8.53		3.91	3.91
18	4	112.12	104.29	87.88	83.24	13.59	-
19	3	89.95	89.01	94.64		3.02	3.02
20	3	33.06	25.63	31.30	15.57	7.88	7.97
21	4	113.34	98.75	101.74		7.71	7.71
22	4	98.41	103.80	104.25		3.25	3.25
23	3	100.46	76.33	99.50		13.66	13.66
24	3	103.81	97.66	97.55		3.58	3.58
25	3	117.50	105.75	88.55	85.55	15.03	16.16
26	3	7.73	8.27	6.75		0.77	0.77
27	3	96.83	105.38	108.52		6.05	6.05
28	4	82.93	100.46	65.46		17.50	17.50
29	4	9.42	8.68	32.58	7.21	12.11	1.13
30	3	101.81	109.70	108.51		4.26	4.26
31	4	7.08	57.09	6.41	10.99	24.55	2.47
32	4	110.37	108.61	99.40		5.89	5.89
33	4	93.38	105.34	97.27		6.10	6.10
34	3	61.51	86.31	94.28	116.77	22.80	17.09
35	4	8.16	7.54	7.86		0.31	0.31
36	3	97.33	100.22	84.27		8.50	8.50
37	4	44.39	18.54	36.30	24.78	11.57	13.22
38*							
39	3	113.66	96.59	101.76		8.75	8.75
40	3	64.66	49.52	37.96	7.38	24.27	28.66
41	4	104.02	103.23	103.81		0.41	0.41
42	3	88.96	96.77	92.88		3.90	3.90
43	3	95.67	107.76	93.65	103.40	6.60	6.12
44	4	77.16	99.83	101.64	93.01	11.14	4.55
45	3	13.35	20.92	7.95		6.52	6.52
46	3	9.77	7.57	7.41		1.32	1.32
47	3	32.58	6.82	6.43	5.08	13.25	0.91
48	4	110.20	103.05	103.85		3.92	3.92
49	3	97.75	102.30	98.24		2.50	2.50
50	4	73.60	94.18	74.99		11.50	11.50
51	4	111.46	113.09	124.80		7.28	7.28
52	*3*	*104.46*	*69.56*	*104.96*	*68.52*	*20.60*	-
53	*3*	*17.96*	*82.08*	*58.49*	*88.34*	*31.88*	-
54	4	55.42	105.48	103.71	86.00	23.22	28.40
55	*4*	*74.99*	*70.13*	*95.48*	*81.18*	*10.99*	-
56	4	9.90	8.78	10.97		1.09	1.09
57	3	102.72	90.52	92.64		6.52	6.52
58	*3*	*7.29*	*30.61*	*31.20*	*28.22*	*11.43*	-
59	*4*	*27.03*	*50.60*	*62.42*	*21.04*	*19.52*	-
60	4	10.11	17.81	20.48	6.39	6.56	5.83

Table 13: BASF within-laboratory standard deviation s_R
(light grey cells: runs, which were not considered for three valid runs; dark grey cells: chemicals with s_R > 18; italic: chemicals without three valid runs;
* confidential chemicals)

3.1.3.3.3 Correlation

The third measure of reproducibility within a laboratory was Bravais-Pearson correlation coefficient r. It was applied to correlate the results of two complete runs. However, as already seen in Phase I, the value of this measure is limited due to the fact that the test protocol was designed to separate irritants from non-irritants. Calculating the correlation of the mean cell viability for all three pairs of runs for the 48 chemicals with three valid runs resulted in correlation coefficients around 0.9 (Table 14). Considering the first three runs for all chemicals resulted in correlations above 0.85 (data not shown).

	correlation r
Run 1 – Run 2	0.9285
Run 1 – Run 3	0.8972
Run 2 – Run 3	0.9221

Table 14: BASF run correlations

3.1.3.3.4 Proportion of identically classified chemicals

The crudest measure for within-laboratory reproducibility was the proportion of identically classified chemicals. The classifications were derived from the Prediction Model (PM) of the SOP. This proportion was again only applied to the 48 chemicals with three valid runs. The classifications, which can easily be derived from Table 13, are identical in the three valid runs for 45 chemicals. Only chemical numbers 6, 8 and 40 would be classified non-consistently. Considering all runs, ten out of 58 chemicals were classified non-consistently.

3.1.3.4 Summary within-laboratory variability results

The results of the within-laboratory variability of all applied measures are summarised for the three laboratories in Table 15. Regarding the sample size, it is obvious that in the lead laboratory ZEBET more chemicals had three valid runs than in the additional laboratories. Furthermore, the number of chemicals, which gave significant ANOVA results, differs between laboratories as it depends on the within-assay variability. Comparing this variability via the relative cumulative distribution of the standard deviations of all tests revealed that these were smallest at ZEBET, followed by IIVS and BASF (Figure 5). Same ranking of laboratories is reflected when focusing on the variability criterion, which considers $s_R > 18$ as unacceptable. Having in mind the problems of the further two measures (correlation and proportion of identically runs), the most informative measure for within-laboratory variability was, although only a descriptive measure, the within-laboratory standard deviation s_R. However, the number of chemicals not consistently classified in the runs in a given laboratory, i.e. one to ten, is comparable to the respective number for the EPISKIN test in the ECVAM Validation Study of in vitro test for skin corrosion, i.e. none to nine of 60 chemicals.[11]

Variability measure	ZEBET		IIVS		BASF	
	all runs	three valid runs	all runs	three valid runs	all runs	three valid runs
sample size	58	54	58	48	58	48
ANOVA: number of chemicals with significant run differences	22	22	12	11	11	10
number of chemicals with s_R >18	4	2	7	2	12	4
mean correlation of runs	0.9177	0.9496	0.9180	0.9611	0.8590	0.9159
proportion of identically classified chemicals	52/58	53/54	48/58	46/48	48/58	45/48

Table 15: Summary of within-laboratory variability evaluation of EpiDerm

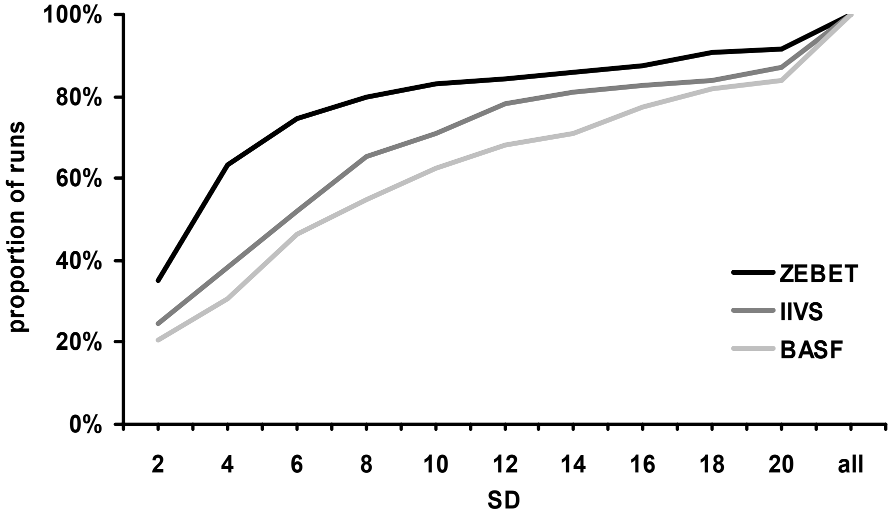


Figure 5: Relative cumulative distribution of the standard deviations of all tests in the three EpiDerm-laboratories.

Overall, the results on the within-laboratory variability are promising, where a ranking of laboratories was nevertheless obvious, i.e. the lead laboratory showed overall the smallest variability.

3.1.4 Within-laboratory variability between phases

The comparison of the results of the 18 chemicals tested in both phases in the validation study was performed with the three valid runs only. The resulting information adds to the assessment of the within-laboratory variability, as up to twelve month lay between Phase I Run 1 and Phase II Run3 and at least 5 month lay between Phase I Run 3 and Phase II Run 1. All tests were carried out by the same ZEBET-operator, but for each phase a new samples of the respective chemicals were provided. When applying a t-test with a significance level of 1% to each of the run data of each of the chemicals, only one substance gave significantly different results in the two phases II. However, when having a closer look to chemical 12 in Table 16, it becomes evident that this was caused by highly reproducible results within the phases.

chemical number	Phase I			Phase II			t-test p-value
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	
1	5.58	7.38	6.48	8.24	5.57	5.35	0.9351
9	101.77	98.27	107.34	93.76	103.38	94.48	0.266
12	96.58	100.46	99.02	104.11	105.57	103.96	**0.0092**
13	79.31	74.06	86.94	102.65	75.40	20.30	0.5987
16	103.56	115.10	115.19	102.93	99.66	100.61	0.0622
17	9.22	10.45	9.23	11.72	14.20	8.36	0.3616
28	99.92	98.65	95.00	90.80	98.72	95.72	0.368
29	10.59	9.71	9.05	9.81	11.80	12.61	0.1604
30	102.12	106.60	105.49	97.99	95.40	99.86	0.0201
32	95.55	100.76	94.96	91.06	106.14	103.38	0.5673
35	106.33	107.71	101.74	87.17	101.38	95.92	0.0819
37	21.16	42.84	42.43	44.34	37.73	45.72	0.4004
40	31.14	6.90	23.86	32.18	21.78	11.93	0.8926
42	100.21	101.34	102.11	95.73	102.24	106.49	0.9371
49	105.49	99.79	105.96	98.56	94.88	96.35	0.0337
51	102.72	116.15	109.75	100.97	110.23	111.65	0.7266
52	84.41	100.29	97.08	78.36	100.75	111.24	0.8051
59	9.35	13.09	10.28	20.67	14.10	17.60	0.0411

Table 16: Cell viability and comparison of the 18 chemicals tested in both phases at ZEBET with EpiDerm

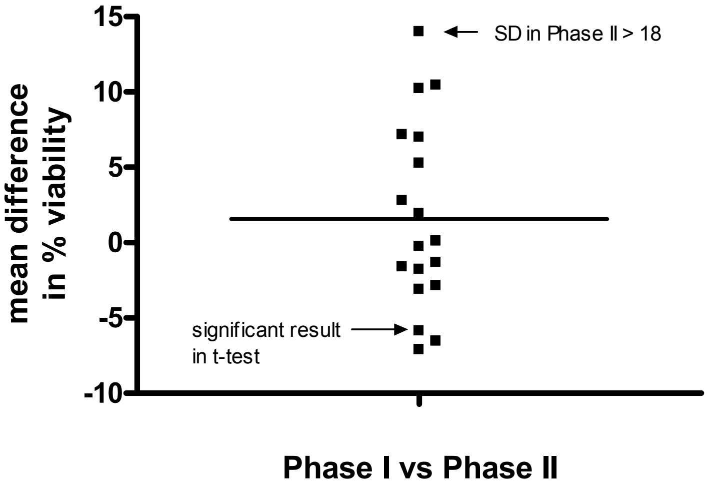


Figure 6: Differences in mean viability of the 18 chemicals tested in both phases at ZEBET with EpiDerm

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, indicating good within-laboratory reproducibility. The differences between these values are presented in Figure 6.

3.1.5 Between-laboratory variability

The between-laboratory variability was first assessed with a 1-way ANOVA (significance level of 1%). Taking the run means of a chemical, the data of the three laboratories were compared once taking all runs of all chemicals into account and once taking only those chemicals with at least three valid runs in two laboratories into account. In those cases were only two laboratories had three valid runs, these were compared by a t-test (significance level of 1%).

In Table 17, the ANOVA/t-test p-values and the ANOVA sum of squares are given for both approaches. Considering all runs, six chemicals (numbers 2, 5, 6, 15, 35, 55) gave significantly different results. As their respective sum of squares were larger than 5000, those chemicals with a sum of squares of > 5000 are marked grey, where ten had a p-value > 0.01. This is an indicator that the variability within laboratories resulted in non-significant results, although there are substantial differences between the laboratories. Indeed, of the 16 chemicals, which were not reproducible in at least one laboratory in terms of the within-laboratory standard deviation s_R, twelve had a sum of squares > 5000. The sum of squares of the other four substances ranged between 2134 and 3148. Six chemicals (numbers 7, 13, 15, 44, 55, 58) did not have three valid runs in at least two laboratories. It is obvious that these chemicals posed problems in EpiDerm. Ten of the remaining 52 chemicals had three valid runs in two laboratories only. In total, four chemicals (numbers 2, 4, 27, 35) had an ANOVA/t-test p-value below 0.01 and five chemicals (numbers 8, 27, 35, 40, 54) had an ANOVA sum of squares larger than 5000. Summarising, 36 chemicals were, according to the here applied measure, reproducible between the three laboratories, where one of these (number 34) was not reproducible within one laboratory when considering all runs. The sums of squares were, with the exception of chemical 34, below 1200. Another eight chemicals (numbers 5, 6, 18, 31, 47, 52, 53, 59) were reproducible between two laboratories. Chemical number 4, which gave similar response in two laboratories, resulted in a significant t-test because of the high within-laboratory reproducibility in both laboratories. Seven chemicals (numbers 2, 8, 27, 35, 40, 54, 55) were not reproducible between-laboratories, where four of these were not within-laboratory reproducible, i.e. $s_R > 18$, in at least one laboratory.

	all runs				three valid runs per laboratory			
chemical number	n	laboratories with s_R >18	ANOVA p-value	ANOVA sum of squares	n	laboratories with s_R >18	ANOVA/t-test p-value	ANOVA sum of squares
1	9		0.6369	9	9		0.6369	9
2	11		**<0.0001**	8347	6		**<0.0001**	
3	9		0.4024	140	9		0.4024	140
4	10	1	0.0105	7904	6		**0.0032**	
5	10		**0.0064**	5597	6		0.1008	
6	12	2	**0.0080**	10250	6	1	0.0273	
7	12		0.2581	3224				
8	11	2	0.1446	9447	9	1	0.0849	8694
9	10		0.5011	880	9		0.1728	472
10	9		0.7499	473	9		0.7499	473
11	10		0.7050	542	9		0.1965	370
12	9		0.2304	1044	9		0.2304	1044
13	13	2	0.3158	8363				
14*								
15	13	1	**0.0018**	9705				
16	9		0.4664	554	9		0.4664	554
17	9		0.3771	73	9		0.3771	73
18	10		0.7660	686	6		0.2123	
19	9		0.0141	282	9		0.0141	282
20	11		0.0541	1078	9		0.1101	879
21	9		0.4040	370	9		0.4040	370
22	9		0.6673	153	9		0.6673	153
23	9		0.6468	844	9		0.6468	844
24	9		0.1340	278	9		0.1340	278
25	10		0.8704	796	9		0.7407	566
26	9		0.0283	104	9		0.0283	104
27	10	1	0.0365	5541	9		**<0.0001**	5278
28	10		0.3971	954	9		0.4504	946
29	10		0.6376	506	9		0.0331	25
30	9		0.0247	181	9		0.0247	181
31	12	2	0.0639	9096	6		0.0437	
32	10	1	0.8629	3041	9	1	0.8774	3039
33	10		0.2731	279	9		0.2146	278
34	10	1	0.6213	2134	9		0.1601	1633
35	9		**<0.0001**	19260	9		**<0.0001**	19260
36	9		0.0711	441	9		0.0711	441
37	10		0.0713	942	9		0.1705	706
38*								
39	9		0.4347	274	9		0.4347	274
40	12	2	0.2777	8375	9	2	0.0441	7437
41	9		0.1294	110	9		0.1294	110
42	9		0.0341	406	9		0.0341	406
43	10		0.6860	217	9		0.8738	146
44	14	1	0.0136	8253				
45	9		0.2563	162	9		0.2563	162
46	9		0.0366	182	9		0.0366	182
47	12		0.6963	1267	6		0.0763	
48	9		0.0296	140	9		0.0296	140
49	10		0.1828	120	9		0.3331	88
50	11		0.5635	1163	9		0.7612	1092
51	9		0.1924	361	9		0.1924	361
52	11	1	0.5114	2204	6		0.9076	
53	11	1	0.0661	7022	6		0.9059	
54	12	2	0.0493	5986	9	1	0.1024	5767
55	11		**<0.0001**	11800				
56	9		0.4713	149	9		0.4713	149
57	9		0.9968	202	9		0.9968	202
58	14	2	0.4591	7715				
59	10	1	0.0313	3148	6		0.0101	
60	10		0.1582	231	9		0.3092	111

Table 17: EpiDerm between-laboratory reproducibility: ANOVA
(bold type: significant ANOVA; grey cells: ANOVA sum of squares > 5000;
* confidential chemicals)

chemical number	all runs				three valid runs			
	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
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

Table 18: EpiDerm between-laboratory variability: the standard deviation of laboratory run means
(bold: SD > 18; * confidential chemicals)

The second measure of between-laboratory variability was the standard deviation of the means of the runs per laboratory (Table 18). Transferring the value of 18 from the variability criterion for test sample showed in the analysis of all runs that 13 chemicals (2, 4, 5, 6, 8, 15, 27, 31, 35, 44, 53, 54, 55) had a standard deviation when comparing laboratories larger 18. All of these were also not reproducible according to the respective analysis with the ANOVA above. Focusing on the 52 chemical with three valid runs, seven (numbers 2, 6, 8, 27, 35, 40, 54) showed an SD > 18, where chemical number 6 was, due to large within laboratory variability, not significant in the respective ANOVA analysis. Chemical numbers 2 and 6 had three valid runs in two laboratories only. Executing the same analysis with all valid runs, even if there were only one or two in a laboratory, gave similar results to the analysis of all runs presented here. An additional chemical (number 40) had an SD >18 (data not shown).

Applying the third measure – the proportion of identical classified chemicals taking the median classification per laboratory into account – to those 52 chemicals with three valid runs in at least two laboratories, 47 chemicals were identically classified. Six of these had three valid runs in two laboratories only. The not reproducible substances according to this measure are the same as in the respective analysis of the between-laboratory standard deviation. Considering all runs, 15 out of 58 chemicals were not consistently classified (Table 19). In a pair-wise comparison of the laboratories the concordance of classifications was 47/58 = 81.0% for ZEBET-IIVS, 51/58 = 87.9% for ZEBET-BASF and 46/58 = 79.3% for IIVS-BASF.

chemical number	EU classification	median classification			between-laboratory reproducible
		ZEBET	IIVS	BASF	
2	no label	0	1	1	-
5	no label	0	1	0	-
6	no label	0	1	1	-
7	no label	1	1	0	-
8	no label	0	0	1	-
9	no label	0	0	0	+
10	no label	0	0	0	+
11	no label	0	0	0	+
12	no label	0	0	0	+
16	no label	0	0	0	+
17	no label	1	1	1	+
19	no label	0	0	0	+
21	no label	0	0	0	+
22	no label	0	0	0	+
24	no label	0	0	0	+
25	no label	0	0	0	+
26	no label	1	1	1	+
28	no label	0	0	0	+
30	no label	0	0	0	+
32	no label	0	0	0	+
33	no label	0	0	0	+
35	no label	0	0	1	-
36	no label	0	0	0	+
39	no label	0	0	0	+
41	no label	0	0	0	+
42	no label	0	0	0	+
44	no label	0	0	0	+
48	no label	0	0	0	+
50	no label	0	0	0	+
52	no label	0	0	0	+
53	no label	0	0	0	+
54	no label	0	1	0	-
57	no label	0	0	0	+
1	R38	1	1	1	+
3	R38	0	0	0	+
4	R38	1	1	0	-
13	R38	0	1	1	-
15	R38	0	1	0	-
18	R38	0	0	0	+
20	R38	1	1	1	+
23	R38	0	0	0	+
27	R38	0	1	0	-
29	R38	1	1	1	+
31	R38	1	0	1	-
34	R38	0	0	0	+
37	R38	1	1	1	+
40	R38	1	0	1	-
43	R38	0	0	0	+
45	R38	1	1	1	+
46	R38	1	1	1	+
47	R38	1	1	1	+
49	R38	0	0	0	+
51	R38	0	0	0	+
55	R38	0	1	0	-
56	R38	1	1	1	+
58	R38	1	0	1	-
59	R38	1	1	1	+
60	R38	1	1	1	+
		reproducible non-labeled chemicals			78.8%
		reproducible R38-labeled chemicals			68.0%
		overall reproducibility			74.1%

Table 19: Between-laboratory reproducibility of EpiDerm in terms of identical median classifications (‘0’: no label/non-irritant; ‘1’: R38/irritant) between the laboratories when considering all runs
(‘-‘ indicates a non-reproducible chemical; ‘+’ indicates a reproducible chemical; grey cells highlight the inconclusive cases, i.e. those with equal numbers of negative and positive classifications in the individual runs, which were conservatively considered as skin irritants)

Finally those chemicals, which did not have three valid runs in at least one of the laboratories, are compared in Table 20. The chemicals number 7 and 15 were problematic in all three laboratories. Four chemicals (numbers 13, 44, 55 and 58) did not have three valid runs in two laboratories, where chemical number 58 was considered to have three valid runs at ZEBET according to a MT decision. The remaining substances were problematic in one of the two additional laboratories. This might indicate a higher level of routine handling of EpiDerm in the lead laboratory.

chemical number	ZEBET	IIVS	BASF
2	3/4	2/4	3/3
4	3/3	3/3	2/4
5	3/3	1/4	3/3
6	3/4	2/4	3/4
7	2/4	1/4	1/4
13	5/5	1/4	0/4
15	2/5	2/4	2/4
18	3/3	3/3	2/4
31	3/4	2/4	3/4
44	2/6	1/4	3/4
47	3/4	2/4	3/4
52	3/3	3/4	2/4
53	3/4	3/3	2/4
55	1/4	3/3	2/4
58	3/5	1/5	2/4
59	3/3	3/3	2/4

Table 20: Chemicals without three valid runs in at least one EpiDerm-laboratory (indicated by grey cells)

3.1.6 Predictive Capacity

3.1.6.1 ZEBET

The prediction model of EpiDerm was designed to predict the current European classifications for skin irritation, i.e. the label R38 (skin irritant) versus no label: a test substance in an experiment was predicted to be a skin irritant if it reduced in average the relative cell viability below 50% compared to the mean cell viability of the negative control. If the mean cell viability was above 50%, it was considered to be a not skin irritating in terms of the European classification system.

	EU	run						median approach	
chemical	classification	1	2	3	4	5	6	all runs	three valid runs
2	no label	0	0	0	0			0	0
5	no label	0	0	0				0	0
6	no label	0	0	**1**	0			0	0
7	no label	**1**	**1**	**1**	0			**1**	-
8	no label	0	0	0				0	0
9	no label	0	0	0				0	0
10	no label	0	0	0				0	0
11	no label	0	0	0				0	0
12	no label	0	0	0				0	0
16	no label	0	0	0				0	0
17	no label	**1**	**1**	**1**				**1**	**1**
19	no label	0	0	0				0	0
21	no label	0	0	0				0	0
22	no label	0	0	0				0	0
24	no label	0	0	0				0	0
25	no label	0	0	0				0	0
26	no label	**1**	**1**	**1**				**1**	**1**
28	no label	0	0	0				0	0
30	no label	0	0	0				0	0
32	no label	0	0	0				0	0
33	no label	0	0	0				0	0
35	no label	0	0	0				0	0
36	no label	0	0	0				0	0
39	no label	0	0	0				0	0
41	no label	0	0	0				0	0
42	no label	0	0	0				0	0
44	no label	**1**	0	0	0	0	0	0	-
48	no label	0	0	0				0	0
50	no label	0	0	0	0			0	0
52	no label	0	0	0				0	0
53	no label	0	0	0	0			0	0
54	no label	0	0	0	0			0	0
57	no label	0	0	0				0	0
1	R38	1	1	1				1	1
3	R38	**0**	**0**	**0**				**0**	**0**
4	R38	1	1	1				1	1
13	R38	**0**	**0**	1	**0**	1		**0**	**0**
15	R38	**0**	1	**0**	**0**	**0**		**0**	**-**
18	R38	**0**	**0**	**0**				**0**	**0**
20	R38	1	1	1	1			1	1
23	R38	**0**	**0**	**0**				**0**	**0**
27	R38	**0**	**0**	**0**				**0**	**0**
29	R38	1	1	1				1	1
31	R38	1	1	1	1			1	1
34	R38	**0**	**0**	**0**				**0**	**0**
37	R38	1	1	1				1	1
40	R38	1	1	1				1	1
43	R38	**0**	**0**	**0**				**0**	**0**
45	R38	1	1	1				1	1
46	R38	1	1	1				1	1
47	R38	1	1	1	1			1	1
49	R38	**0**	**0**	**0**				**0**	**0**
51	R38	**0**	**0**	**0**				**0**	**0**
55	R38	**0**	**0**	**0**	**0**			**0**	**-**
56	R38	1	1	1				1	1
58	R38	**0**	**0**	1	1	1		1	1
59	R38	1	1	1				1	1
60	R38	1	1	1				1	1

Table 21: EpiDerm-Classification of the 58 chemicals according to the prediction model at ZEBET
(0: non irritant (no label); 1: irritant (R38); bold: misclassifications; grey cells: not valid or not necessary runs)

In Table 21, the predictions are presented for all test chemicals, where misclassifications are bold. To summarise the prediction over the runs, two approaches based on the median classification were applied: the median of all available runs and of three valid runs, when available, where in cases with more than three valid runs the first three were considered. However, differences between these were minor. In Table 22, the parameters specificity and sensitivity calculated from Table 21 are presented for each single run and for the summarising approaches. Besides the sample sizes for both parameters, also the exact lower 5%-confidence bounds are given.

Regardless the way of analysis, the specificity was always around 90%. Nevertheless, it can be seen that the analysis of valid runs, i.e. those meeting the quality criterion of a SD < 18, have a slightly increased specificity. A similar effect was observed for sensitivity, which, depending on the analysis ranged between 52.0 and 65.2%. Summarising the data in a conservative way, i.e. classifying a chemical as irritant when it was classified as irritant in at least one run, produced very similar results (data not shown).

		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

Table 22: Predictive Capacity of EpiDerm at ZEBET in terms of specificity and sensitivity with the respective sample sizes and 5% lower confidence bounds for each run and summaries over all runs

In order to investigate how the balance of specificity and sensitivity depend on the PM-threshold we submitted the ZEBET data of all valid runs (n = 175) to a ROC-analysis (Figure 7). This approach was chosen as a compromise between the ‘all runs’-approach, which would be more unbalanced with regard to the number of runs per chemical, and the ‘three valid runs’-approach, by which chemicals would be excluded. This analysis revealed a steep curve for specificities above 90%, which then became relative flat for specificities below 90%. This angular shape indicated that an optimal balance of the performance parameters is reached around the angle. Rendering the test either more sensitive or more specific can only be achieved by extreme trade-off.

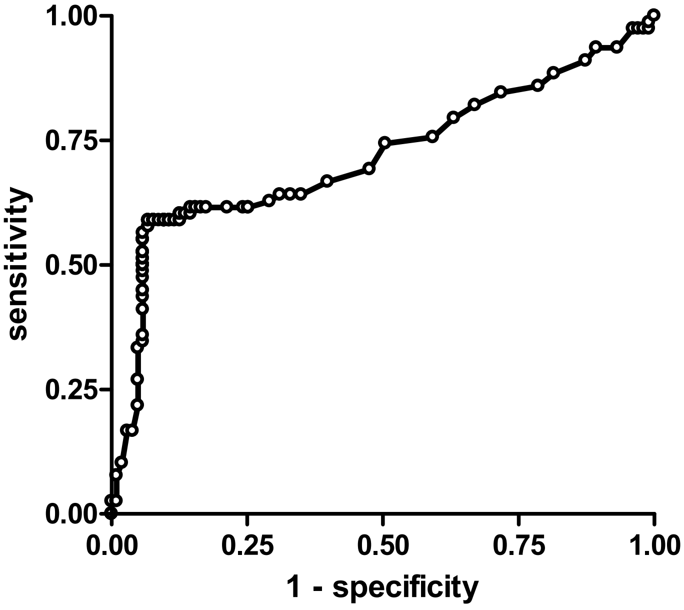


Figure 7: Receiver operation curve of all valid runs from ZEBET

To get more insight into the predictive capacity and its threshold dependence, we plotted the sensitivity, the specificity and the sum of these, which allows optimising the PM-threshold choice when weighing sensitivity and specificity equally. In Figure 8A, which still considered the chemicals classifications according to the European system, it is obvious that the sum of sensitivity and specificity did not change over a broad spectrum of thresholds ranging from about 38% to about 71%. This insensitivity to threshold changes in the middle cell viability response range clearly reflected the optimization efforts with EpiDerm towards a two-class system. Furthermore, it confirmed that the prediction model of 50% was an appropriate choice.

Although moving the in vitro threshold has little influence, we also modeled to move the in vivo threshold. In the European system, this threshold is equivalent the dominant median score of 2. We performed the same analysis as done for this in vivo threshold of 2, when moving it downwards to 1.7 and upwards to 2.3. The results are shown in Figure 8B and C. As basically the shapes of all curves were similar to those in Figure 8A, EpiDerm predictive capacity could not be improved when moving the in vivo classification threshold.

Furthermore, the mean viability of all runs, included in Table 18, was plotted against the dominant median of the in vivo data, which is included in Table 4. Figure 9 allows a more detailed evaluation of the severity of the misclassification. For example, six of the in this graph eleven false negative classified chemicals had a dominant median of 2.0, i.e. at the classification threshold of the European classification system.

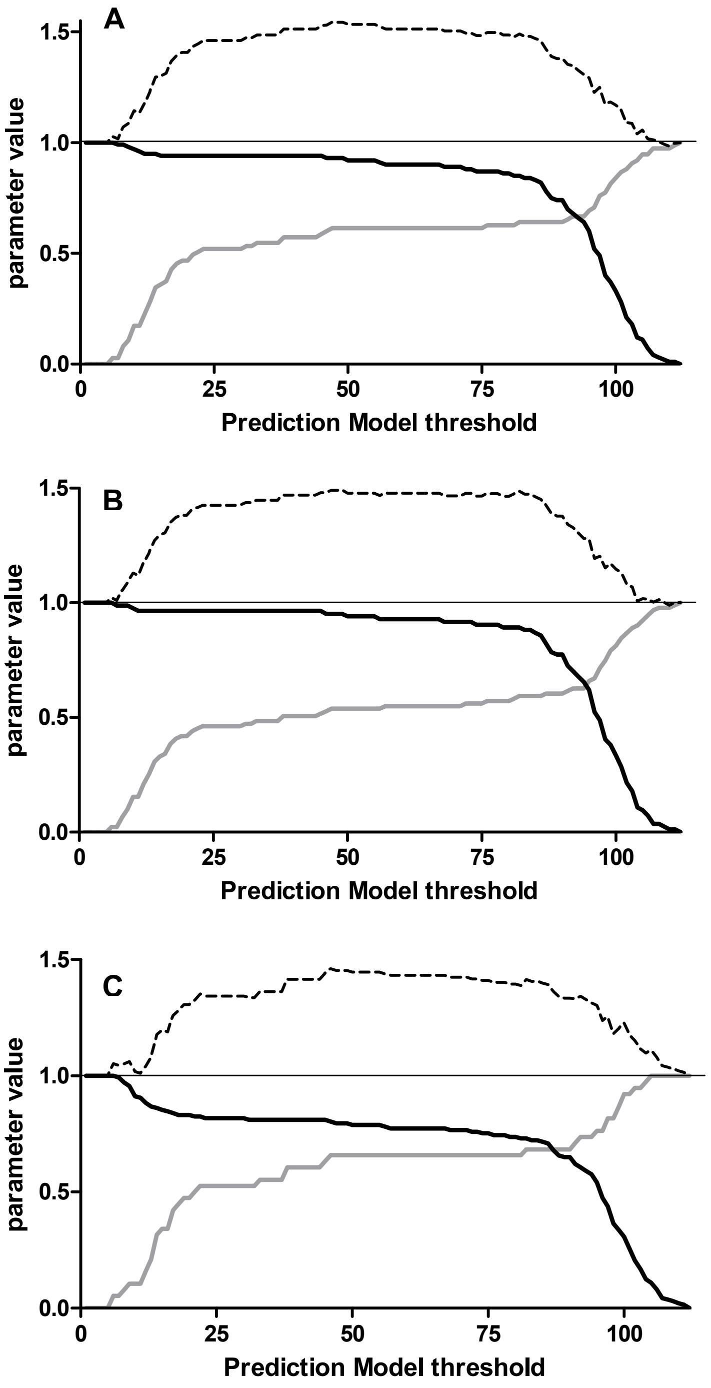


Figure 8: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold [%] when considering all valid ZEBET-runs. A: Classification of the in vivo data according to the European classification system, i.e. a threshold of 2. B: In vivo classification threshold of 1.7. C: In vivo classification threshold of 2.3.
(black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity).

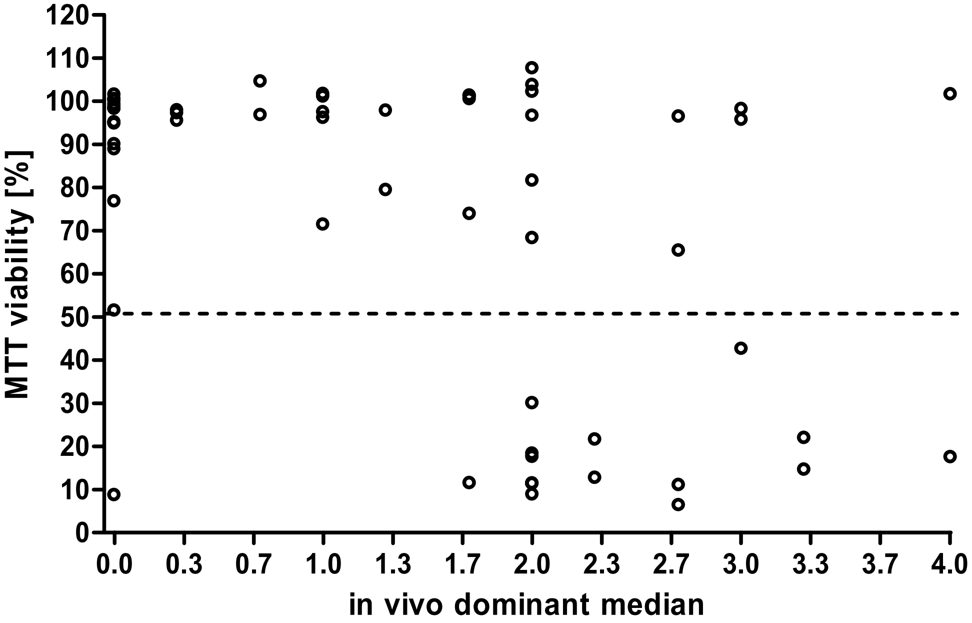


Figure 9: Correlation of the in vivo dominant median with the mean viability of all available runs from ZEBET with EpiDerm, where the dotted line indicates the PM-threshold

3.1.6.2 IIVS

The predictions are presented for all test chemicals (Table 23). In order to summarise the prediction over the runs, two approaches based on the median classification were applied: the median of all available runs and of three valid runs, when available. In Table 24, the parameters specificity and sensitivity calculated from Table 23 are presented for each single run and for the summarising approaches. Besides the sample sizes for both parameters, also the exact lower 5%-confidence bound are given. For simplicity, the inconclusive results, i.e. '2 vs 2' and '1 vs 1' were considered as an irritant summary classification. The specificity ranged between 79% and 89%. The maximal specificity is reached when considering only the 20 chemicals with three valid runs. The sensitivity was between 56 and 65%. Nevertheless, it can be seen that the analysis of three valid runs had an increased specificity. Summarising the data in a conservative way, i.e. classifying a chemical as irritant when it was classified as irritant in at least one run, produced very similar results (data not shown).

chemical	EU	run					median approach	
number	classification	1	2	3	4	5	all runs	three valid runs
2	no label	**1**	0	0	**1**		2 vs 2	-
5	no label	**1**	0	**1**	0		2 vs 2	-
6	no label	**1**	**1**	**1**	**1**		**1**	-
7	no label	**1**	0	**1**	**1**		**1**	-
8	no label	0	0	0	0		0	0
9	no label	0	0	0			0	0
10	no label	0	0	0			0	0
11	no label	0	0	0			0	0
12	no label	0	0	0			0	0
16	no label	0	0	0			0	0
17	no label	**1**	**1**	**1**			**1**	**1**
19	no label	0	0	0			0	0
21	no label	0	0	0			0	0
22	no label	0	0	0			0	0
24	no label	0	0	0			0	0
25	no label	0	0	0			0	0
26	no label	**1**	**1**	**1**			**1**	**1**
28	no label	0	0	0	0		0	0
30	no label	0	0	0			0	0
32	no label	0	0	0	0		0	0
33	no label	0	0	0	0		0	0
35	no label	0	0	0			0	0
36	no label	0	0	0			0	0
39	no label	0	0	0			0	0
41	no label	0	0	0			0	0
42	no label	0	0	0			0	0
44	no label	0	**1**	0	0		0	-
48	no label	0	0	0			0	0
50	no label	0	0	0	0		0	0
52	no label	0	0	0	0		0	0
53	no label	0	0	0			0	0
54	no label	**1**	0	**1**	0		2 vs 2	**1**
57	no label	0	0	0			0	0
1	R38	1	1	1			1	1
3	R38	**0**	**0**	**0**			**0**	**0**
4	R38	1	1	1			1	1
13	R38	**0**	1	1	**0**		2 vs 2	-
15	R38	1	1	1	1		1	-
18	R38	**0**	**0**	**0**			**0**	**0**
20	R38	1	1	1			1	1
23	R38	**0**	**0**	**0**			**0**	**0**
27	R38	1	1	**0**	**0**		2 vs 2	1
29	R38	1	1	1			1	1
31	R38	**0**	1	**0**	**0**		**0**	-
34	R38	**0**	**0**	**0**			**0**	**0**
37	R38	1	1	1			1	1
40	R38	**0**	**0**	**0**	1	1	**0**	**0**
43	R38	**0**	**0**	**0**			**0**	**0**
45	R38	1	1	1			1	1
46	R38	1	1	1			1	1
47	R38	1	1	1	1		1	-
49	R38	**0**	**0**	**0**	**0**		**0**	**0**
51	R38	**0**	**0**	**0**			**0**	**0**
55	R38	1	1	1			1	1
56	R38	1	1	1			1	1
58	R38	1	**0**	**0**	1	**0**	**0**	-
59	R38	1	1	1			1	1
60	R38	1	1	1			1	1

Table 23: EpiDerm-Classification of the 58 chemicals according to the prediction model at IIVS
(0: non irritant (no label); 1: irritant (R38); bold: misclassifications; grey cells: not valid or not necessary runs)

		specificity			sensitivity		
		n	%	LB-5%	n	%	LB-5%
Run 1	all	33	78.8	63.8	25	60.0	41.7
	valid	28	83.3	68.1	22	68.2	48.6
Run 2	all	33	87.9	74.4	25	64.0	45.6
	valid	27	88.9	73.7	21	61.9	41.7
Run 3	all	33	81.8	67.2	25	56.0	37.9
	valid	29	86.2	71.2	22	59.1	39.5
all runs (median)		33	78.8	63.8	25	60.0	41.7
three valid runs (median)		28	89.3	74.6	20	60.0	39.4

Table 24: Predictive Capacity of EpiDerm at IIVS in terms of specificity and sensitivity with the respective sample sizes and 5% lower confidence bounds for each run and summaries over all runs

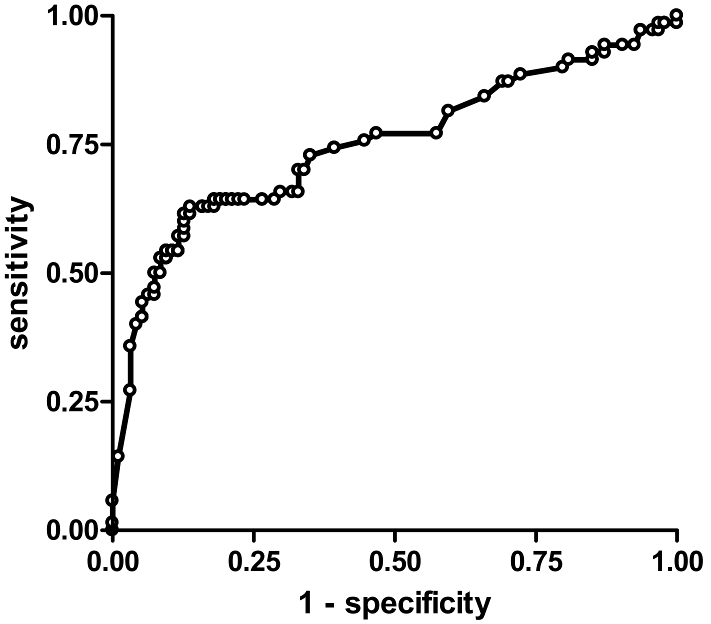


Figure 10: Receiver operation curve of all valid runs from IIVS

To investigate how the balance of specificity and sensitivity depend on the PM-threshold we submitted the IIVS data of all valid runs (n = 164) to a ROC-analysis (Figure 10). This analysis revealed a steep curve for specificities above 85%, which then became relative flat for specificities below 85%. This angular shape indicated that an optimal balance of the performance parameters was reached around the angle. Rendering the test either more sensitive or more specific could only be achieved by extreme trade-off.

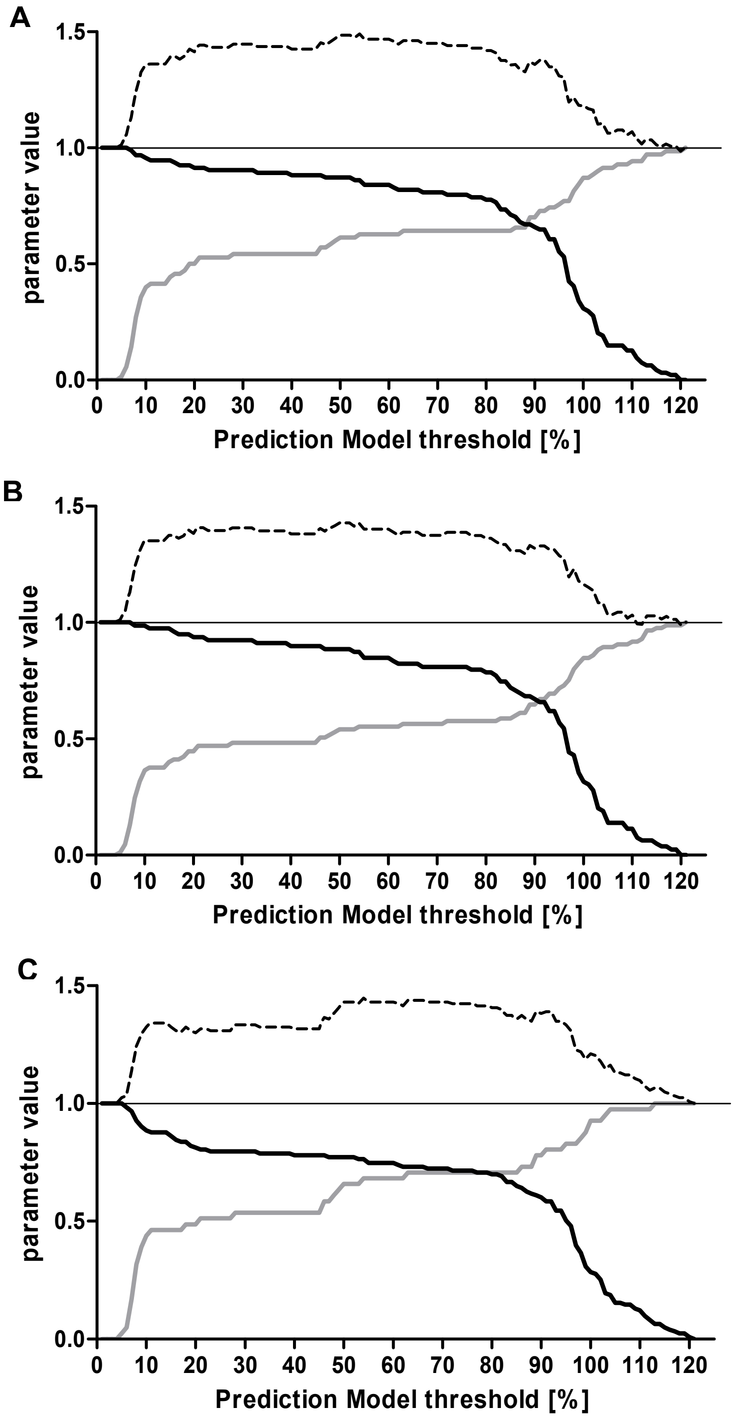


Figure 11: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold [%] when considering all valid IIVS-runs. A: Classification of the in vivo data according to the European classification system, i.e. a threshold of 2. B: In vivo classification threshold of 1.7. C: In vivo classification threshold of 2.3.
(black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity).

To get more insight into the predictive capacity and its threshold dependence, we plotted the sensitivity, the specificity and the sum of these, which allows optimising the PM-threshold choice when weighing sensitivity and specificity equally. In Figure 11A, which considered the chemicals classifications according to the European system, it is obvious that the sum of sensitivity and specificity did not change over a broad spectrum of thresholds ranging from about 25% to about 80%. This insensitivity to threshold changes in the middle cell viability response range clearly reflected the optimisation efforts with EpiDerm towards a two class system. Furthermore, it confirmed that the prediction model of 50% was an appropriate choice.

Although moving the in vitro threshold had little influence, we also modeled to move the in vivo threshold. In the European system, this threshold is equivalent the dominant median score of 2. We performed the same analysis as done for this in vivo threshold of 2, when moving it downwards to 1.7 and upwards to 2.3. The results are shown in Figure 11B and C. As basically the shapes of all curves are similar to those in Figure 11A, EpiDerm predictive capacity could not be improved when moving the in vivo classification threshold.

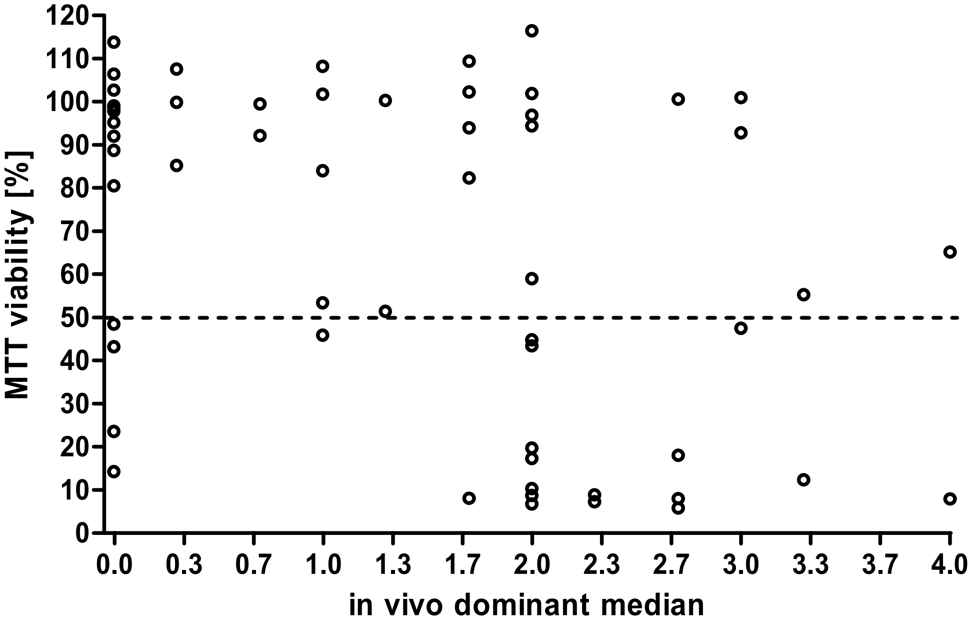


Figure 12: Correlation of the in vivo dominant median with the mean viability of all available runs from IIVS with EpiDerm, where the dotted line indicates the PM-threshold

Furthermore, the mean viability of all runs, included in Table 18, was plotted against the dominant median of the in vivo data, which is included in Table 4. Figure 12 allowed a more detailed evaluation of the severity of the misclassification. For example, half of the ten false negative classified chemicals had a dominant median of 2.0, i.e. the classification threshold of the European classification system.

3.1.6.3 BASF

In Table 25, the predictions are presented for all test chemicals. To summarise the prediction over the runs, two approaches based on the median classification were applied: the median of all available runs and of three valid runs, when available. In Table 26, the parameters specificity and sensitivity calculated from Table 25 are presented for each single run and for the summarising approaches. Besides the sample sizes for both parameters, also the exact lower 5%-confidence bound are given. For simplicity, the inconclusive results, i.e. ‘2 vs 2’ and ‘1 vs 1’ were considered as an irritant summary classification. The specificity ranged between 79% and 88%. The maximal specificity is reached when considering only the 50 chemicals with three valid runs. The sensitivity was between 47% and 61%.

In order to investigate how the balance of specificity and sensitivity depend on the PM-threshold we submitted the ZEBET data of all valid runs (n = 164) to a ROC-analysis (Figure 13). This analysis revealed a steeper curve for specificities above 80%, which then flattened.

chemical	EU		Run			median approach	
number	classification	1	2	3	4	all runs	three valid runs
2	no label	**1**	**1**	**1**		**1**	**1**
5	no label	0	0	0		0	0
6	no label	**1**	**1**	0	**1**	**1**	**1**
7	no label	0	0	0	**1**	0	-
8	no label	**1**	**1**	0	0	2 vs 2	**1**
9	no label	0	0	0	0	0	0
10	no label	0	0	0		0	0
11	no label	0	0	0	0	0	0
12	no label	0	0	0		0	0
16	no label	0	0	0		0	0
17	no label	**1**	**1**	**1**		**1**	**1**
19	no label	0	0	0		0	0
21	no label	0	0	0		0	0
22	no label	0	0	0		0	0
24	no label	0	0	0		0	0
25	no label	0	0	0	0	0	0
26	no label	**1**	**1**	**1**		**1**	**1**
28	no label	0	0	0		0	0
30	no label	0	0	0		0	0
32	no label	0	0	0		0	0
33	no label	0	0	0		0	0
35	no label	**1**	**1**	**1**		**1**	**1**
36	no label	0	0	0		0	0
39	no label	0	0	0		0	0
41	no label	0	0	0		0	0
42	no label	0	0	0		0	0
44	no label	0	0	0	0	0	0
48	no label	0	0	0		0	0
50	no label	0	0	0		0	0
52	no label	0	0	0	0	0	-
53	no label	**1**	0	0	0	0	-
54	no label	0	0	0	0	0	0
57	no label	0	0	0		0	0
1	R38	1	1	1		1	1
3	R38	**0**	**0**	**0**		**0**	**0**
4	R38	1	**0**	**0**	**0**	**0**	-
13	R38	**0**	1	1	1	1	-
15	R38	1	**0**	**0**	**0**	**0**	-
18	R38	**0**	**0**	**0**	**0**	**0**	-
20	R38	1	1	1	1	1	1
23	R38	**0**	**0**	**0**		**0**	**0**
27	R38	**0**	**0**	**0**		**0**	**0**
29	R38	1	1	1	1	1	1
31	R38	1	**0**	1	1	1	1
34	R38	**0**	**0**	**0**	**0**	**0**	**0**
37	R38	1	1	1	1	1	1
40	R38	**0**	1	1	1	1	1
43	R38	**0**	**0**	**0**	**0**	**0**	**0**
45	R38	1	1	1		1	1
46	R38	1	1	1		1	1
47	R38	1	1	1	1	1	1
49	R38	**0**	**0**	**0**		**0**	**0**
51	R38	**0**	**0**	**0**		**0**	**0**
55	R38	**0**	**0**	**0**	**0**	**0**	-
56	R38	1	1	1		1	1
58	R38	1	1	1	1	1	-
59	R38	1	**0**	**0**	1	2 vs 2	-
60	R38	1	1	1	1	1	1

Table 25: EpiDerm-Classification of the 58 chemicals according to the prediction model at BASF
(0: non irritant (no label); 1: irritant (R38); bold: misclassifications; grey cells: not valid or not necessary runs)

		specificity			sensitivity		
		n	%	LB-5%	n	%	LB-5%
Run 1	all	33	78.8	63.8	25	56.0	37.9
	valid	29	79.3	63.2	20	55.0	34.7
Run 2	all	33	81.8	67.2	25	48.0	30.5
	valid	32	81.3	66.3	19	47.4	27.4
Run 3	all	33	87.9	74.4	25	48.0	30.5
	valid	30	86.7	72.0	16	56.3	33.3
all runs (median)		33	87.8	67.2	25	56.0	37.9
three valid runs (median)		30	80.0	64.3	18	61.1	39.2

Table 26: Predictive Capacity of EpiDerm at BASF in terms of specificity and sensitivity with the respective sample sizes and 5% lower confidence bounds for each run and summaries over all runs

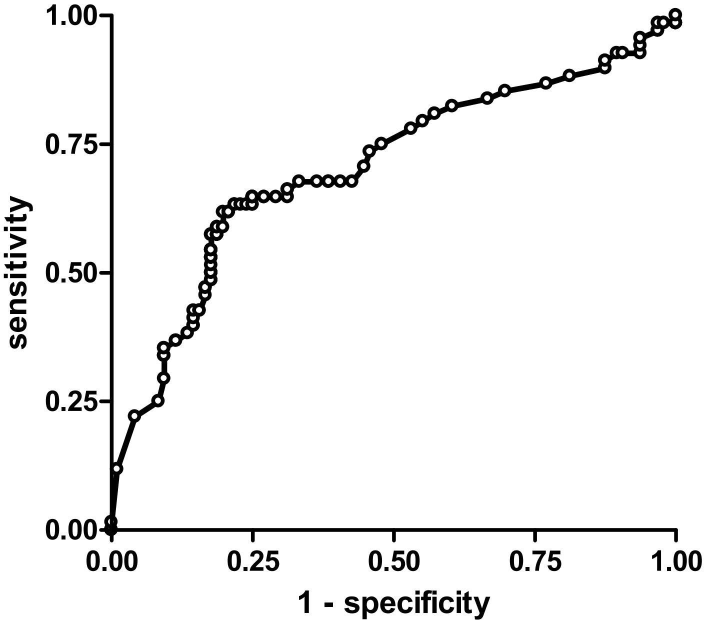


Figure 13: Receiver operation curve of all valid runs from BASF

To get more insight into the predictive capacity and its threshold dependence, we plotted the sensitivity, the specificity and the sum of these, which allows optimising the PM-threshold choice when weighing sensitivity and specificity equally. In Figure 14A, which still considered the chemicals classifications according to the European system, it is obvious that the sum of sensitivity and specificity did not change over a broad spectrum of thresholds ranging from about 38% to about 85%. This insensitivity to threshold changes in the middle cell viability response range clearly reflected the optimization efforts with EpiDerm towards a two class system. Furthermore, it confirmed that the prediction model of 50% was an appropriate choice.

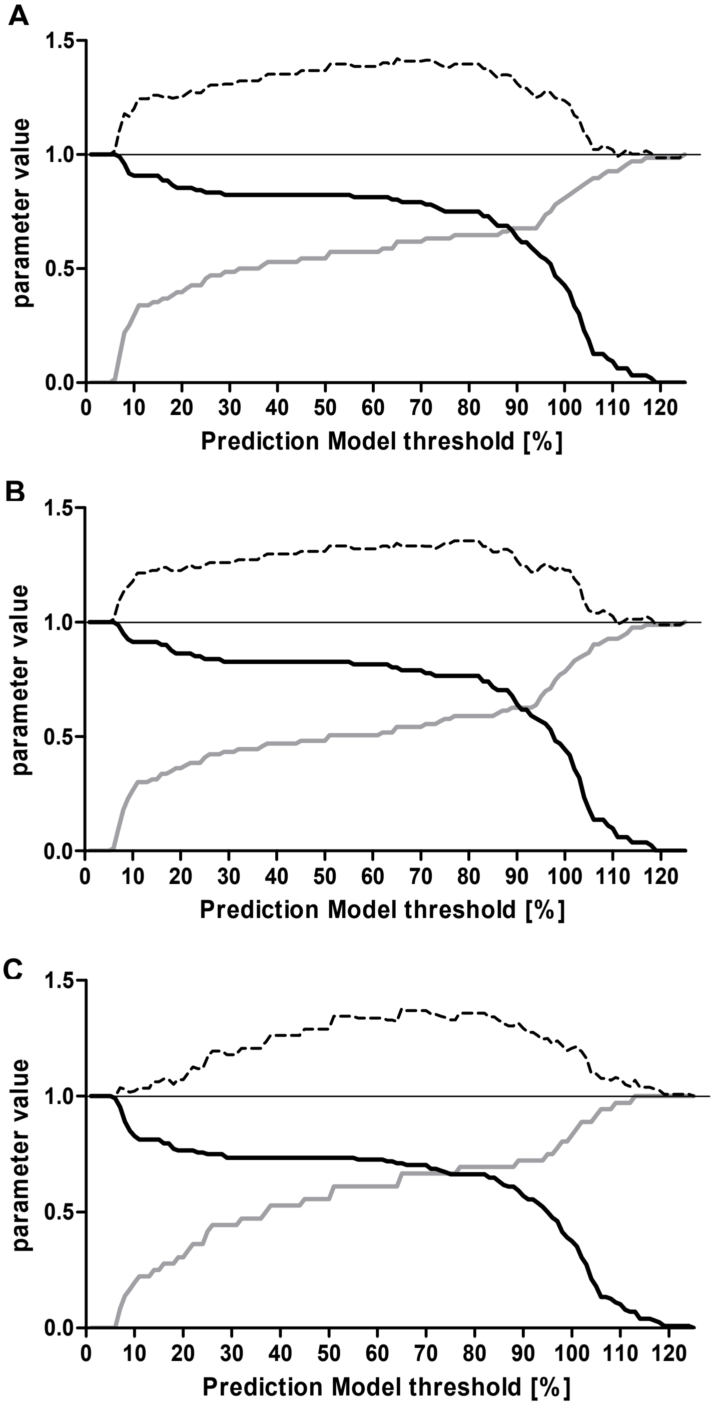


Figure 14: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold [%] when considering all valid BASF-runs. A: Classification of the in vivo data according to the European classification system, i.e. a threshold of 2. B: In vivo classification threshold of 1.7. C: In vivo classification threshold of 2.3.
(black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity).

Although moving the in vitro threshold has little influence, we also modeled to move the in vivo threshold. In the European system, this threshold is equivalent the dominant median score of 2. We performed the same analysis as done for this in vivo threshold of 2, when moving it downwards to 1.7 and upwards to 2.3. The results are shown in Figure 14B and C. As basically the shapes of all curves are similar to those in Figure 14A, EpiDerm predictive capacity could not be improved when moving the in vivo classification threshold.

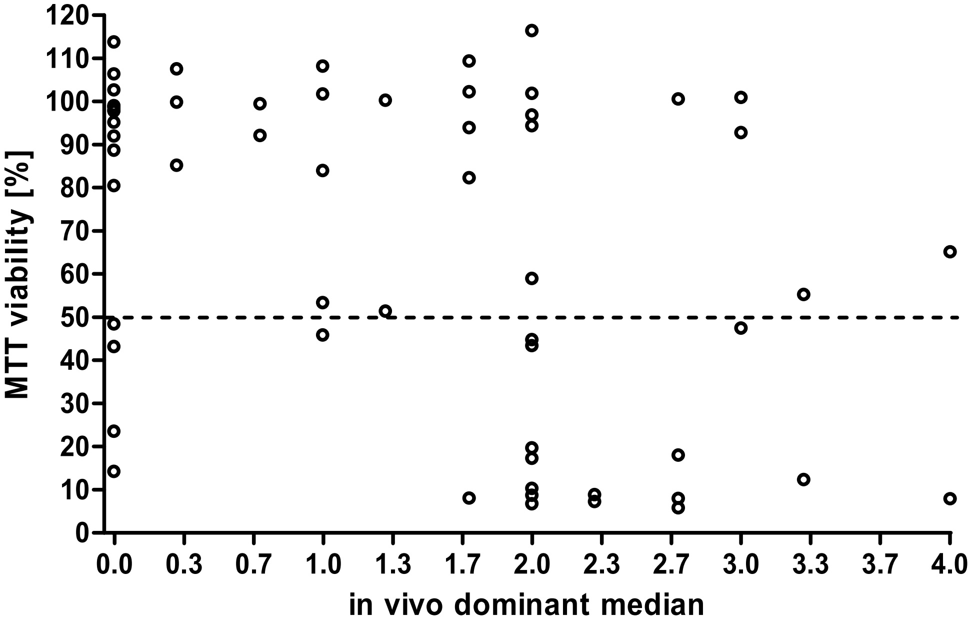


Figure 15: Correlation of the in vivo dominant median with the mean viability of all available runs from BASF with EpiDerm, where the dotted line indicates the PM-threshold

Furthermore, the mean viability of all runs, included in Table 18, was plotted against the dominant median of the in vivo data, which is included in Table 4. Figure 14 allowed a more detailed evaluation of the severity of the misclassification. For example, half of the ten false negative classified chemicals had a dominant median of 2.0, i.e. the classification threshold of the European classification system.

3.1.6.4 Misclassifications

To compare the misclassified chemicals of the three EpiDerm-laboratories, those chemicals, which were misclassified at least once in one of the laboratories, are summarised in Table 27. While only eleven of the 33 not labelled chemicals were misclassified at least once, 15 of the 25 R38-chemicals had at least one misclassification. Two non-irritants were consistently classified as irritant in all runs and all laboratories. Seven irritants were classified as non-irritants in all runs and all laboratories.

chem. no	EU class	dominant median	ZEBET run						IIVS run					BASF run				total number of runs	mis-classifying runs [%]
			1	2	3	4	5	6	1	2	3	4	5	1	2	3	4		
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				**1**	0	**1**	0		0	0	0		10	20.00
8	no label	1	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	**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**	10	100.00
49	R38	2	**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**		9	100.00
23	R38	3	**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**	10	100.00
18	R38	3	**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**				1	1	**0**	**0**		**0**	**0**	**0**		10	80.00
55	R38	2	**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**	1	**0**	1		**0**	1	1	**0**		**0**	1	1	1	13	46.15
58	R38	2	**0**	**0**	1	1	1		1	**0**	**0**	1	**0**	1	1	1	1	14	35.71
31	R38	2	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	1	1	1				1	1	1			1	**0**	**0**	1	10	20.00

Table 27: Summary of chemicals, which were misclassified at least once in one of the EpiDerm-laboratories
(bold type: misclassified runs; grey cells: SD > 18)

3.1.6.5 Summary predictive capacity results

First the specificity and sensitivity over all runs per laboratory from Tables 22, 24 and 26 according to the two approaches of analysis are presented in Table 28.

laboratory	**specificity [%]**		**sensitivity [%]**	
	all runs	three valid runs	all runs	three valid runs
ZEBET	90.9	93.6	56.0	60.9
IIVS	78.8	89.3	60.0	60.0
BASF	87.8	80.0	56.0	61.1

Table 28: Summary of the predictive capacity (specificity and sensitivity) in the three EpiDerm laboratories considering either chemical runs or only those chemicals, which had three valid (SD<18) runs

Second, from all runs in all laboratories we calculated the sample sizes and specificity and sensitivity, again for the two different approaches, either considering all individual classification or the median classification for a given chemical (Table 29). Here, it has to be kept in mind that there are minor imbalances in the data set regarding the numbers of classifications considered per laboratory, whose effect on the results was nevertheless negligible. Taking all individual classifications (n = 607) into account resulted in the lowest estimation for both parameters indicating that a large sample variability within a run (SD > 18) increased the chance of misclassification. Considering only those chemicals, for which three valid runs were available (n = 462), a specificity of 89% and a sensitivity of 60% was achieved. Similar parameter estimations and a similar pattern were present when summarising the median classifications. Due to the strong dependencies between the data in terms of reproducibility, no estimation of confidence bounds was performed. Considering only those chemicals, which had three valid runs in all three laboratories, i.e. 26 non-labelled and 16 labelled (R38) chemicals, resulted in a specificity of 89.3% and a sensitivity of 54.7% (data not shown).

	specificity		**sensitivity**	
	n	**%**	**n**	**%**
all runs (individually classification)	328	84.76	261	56.32
three valid runs (individually classification)	267	88.76	183	60.11
all runs (median classification)	99	83.83	75	57.33
three valid runs (median classification)	89	87.66	61	60.66

Table 29: Summary of EpiDerm specificity and sensitivity considering the two different approaches

Summarizing the receiver operation curve of the three laboratories, which were based on all valid runs, Figure 16 shows that the overall RO-curve averaged the three laboratory RO-curves. The dotted square indicates the area, where the individual curves differ most: At BASF the increase is less strong as the increase in sensitivity was traded-off by a more severe loss in specificity than in the other two laboratories. Again, the angled shape of the RO-curve reflected the effects of the test protocol optimization to separate the two irritation classes of chemicals as clear as possible.

This effect can also be seen in Figure 17, where the sensitivity, specificity and their sum are displayed. In the threshold range between 45% and 73% both the sensitivity and specificity curves are almost flat, where the sum of both remains approximately constant, i.e. larger than 1.42. The maximum sum of 1.454 is reached at a threshold of 54% close to the predefined prediction model threshold of 50%. Similarly to the results in Figures 16 and 17, also the resulting curves when moving the in vivo threshold to 1.7 or 2.3 represented an average result (data not shown).

The negative and positive predictive values, which incorporate specificity and sensitivity as well as prevalence, i.e. the proportion of irritating chemicals in a defined population of chemicals, can be found in Annex VIII.

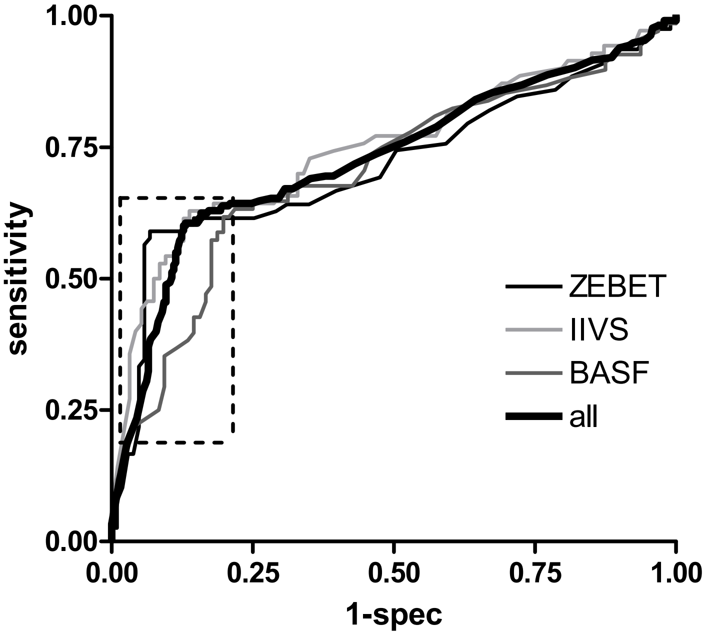


Figure 16: Receiver operation curves of all valid runs of all EpiDerm-laboratories.

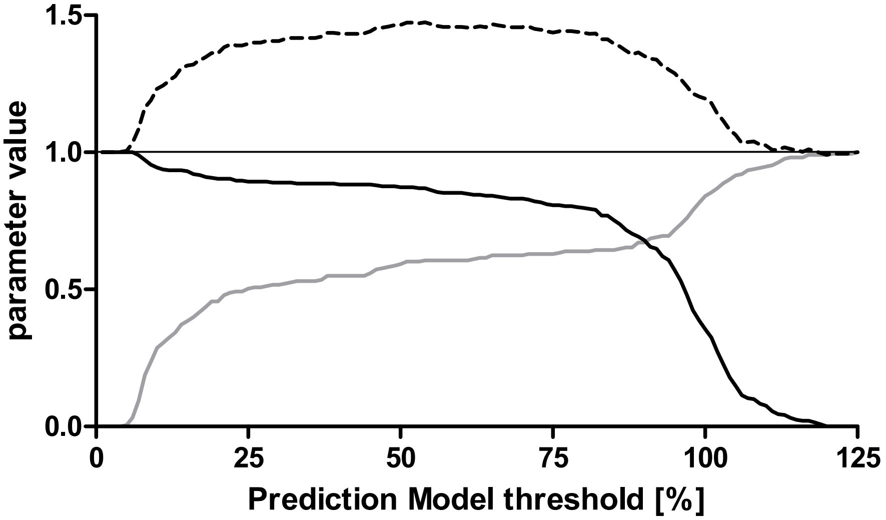


Figure 17: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold [%] when considering all valid EpiDerm-runs (black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity).

Although the B- and C-curves in Figures 8, 11 and 14 already suggest that it will be impossible to find a satisfactorily performing PM for the three GHS classes, performance of EpiDerm to predict the three classes of the GHS was analysed to confirm this expectation. To keep this analysis simple and disregarding reproducibility, only the median run classification of chemical with three valid runs for all laboratories were considered. This resulted in a dataset with a sample size of 150, where 33 entries were GHS-irritants, 43 GHS-mild irritants and 74 GHS-non irritants. As in Phase I (Annex I) no satisfactory PM could be identified, a post-hoc approach to construct a new PM was chosen. Therefore, the two thresholds of viability maximizing the sum of sensitivity and specificity in the ROC-analyses of discriminating GHS-non irritants from GHS-mild irritants and GHS irritants and of discriminating GHS-non and mild irritants from GHS irritants, respectively, were used. As here the two optimal thresholds were almost identical – by itself a strong indication confirming the expectation – one threshold was switched to the next highest value. The respective PM consisted of the threshold of 60% viability, below which chemicals would be classified as GHS-irritants, and of 81%, above which chemicals would be classified as GHS-non irritants, was constructed. Chemicals with viabilities between these two thresholds would be classified as GHS-mild irritants. Applying the PM resulted in the correct classification of 66.7% GHS-irritants, 9.3% GHS-mild irritants and 87.8% GHS-non irritants. It has to be noted that only six entries had viability between 60% and 81%. This confirms the results of Phase I that EpiDerm is not able to predict the three GHS-classes. The fact that GHS-mild irritants were either giving high or low viabilities reflects that the EpiDerm protocol was optimized for the European classification system.

3.2 EPISKIN

3.2.1 Data submission

L'Oréal, the lead laboratory for the EPISKIN assay, submitted the data to ECVAM on 08.06.2005. One operator tested all sixty chemicals between the 20.12.2004 and the 11.04.2005, where 20 chemicals were tested per run. In the provided spreadsheet, no remarks were given. As some chemicals interacted with the MTT, their data required for taking this interaction into account were submitted in adjusted spreadsheets.

The data from Unilever were received in the agreed format on 13.06.2005. Twenty chemicals were tested per run. The experiments were carries out between the 04.01.2005 and the 25.04.2005, where no remarks were provided in the spreadsheets. With one exception, the same operator performed all experiments. Data adjustment for MTT-interaction was provided on separate spreadsheets.

Sanofi submitted the data on the 09.06.2005. One operator tested 20 chemicals per run between the 10.12.2004 and the 18.04.2005, where no remarks in the spreadsheets were noted. Data adjustment for MTT-interaction was provided on separate spreadsheets.

3.2.2 Analysis of quality criteria

In total, five data related quality criteria were included in the validation SOP of EPISKIN, where the first four addressed the controls and the fifth the tested sample. The first one demands a mean response (in OD) of the negative control larger than 0.6 OD. As shown in Figure 18, this criterion was failed once at Sanofi, which triggered a repetition of the experiment. To allow a comparison with Phase I, the respective data were added. The different sample sizes per laboratory were caused by failed quality criteria triggering additional experiments.

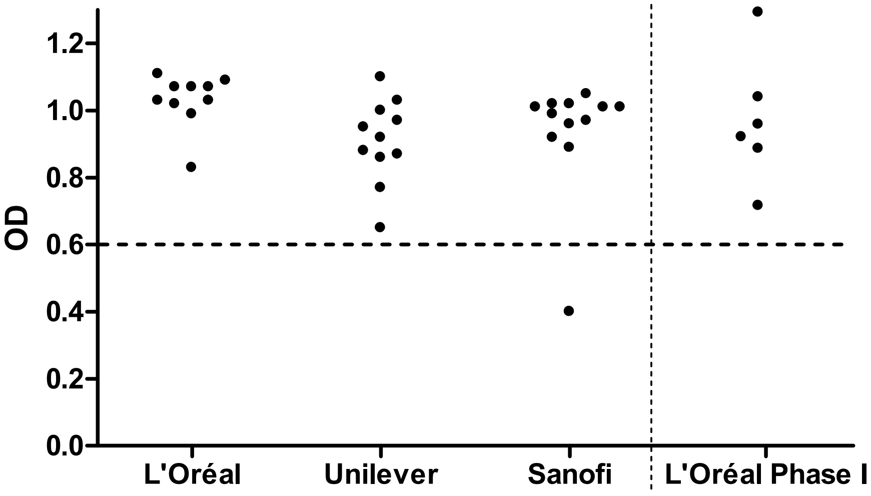


Figure 18: Response of the negative controls in the three EPISKIN-laboratories in Phase II and the lead laboratory in Phase I.

In addition, a run was only considered valid according to the SOP, when the mean relative viability of the positive control was below 40% of the viability of the negative control. The data for all laboratories and both phases show that this criterion was always met (Figure 19). However, a more variable response of the positive control could be observed for Sanofi. Furthermore, at L'Oréal the response of the negative control was decreased between the two phases.

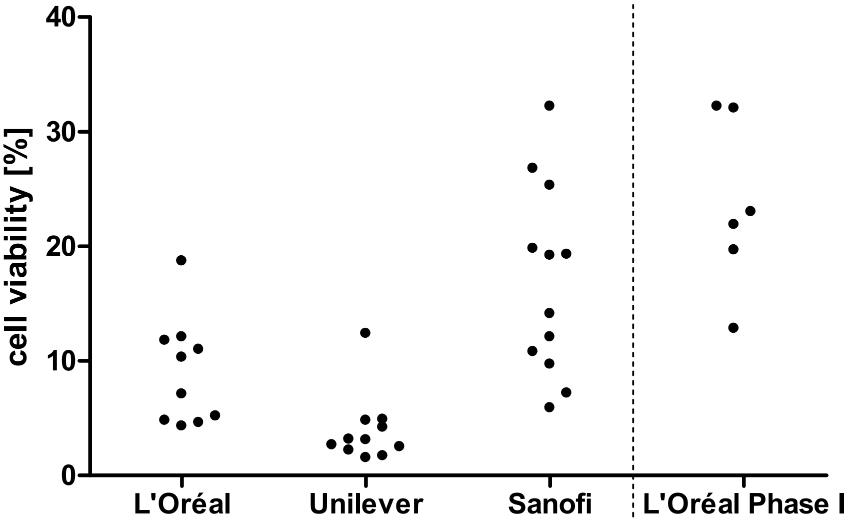


Figure 19: Relative response of the positive controls in the three EPISKIN-laboratories in Phase II and the lead laboratory in Phase I.

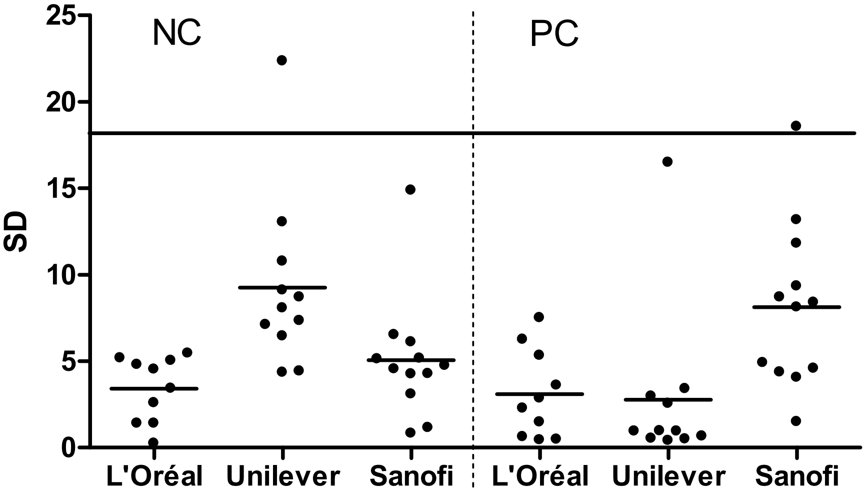


Figure 20: Variability measure as standard deviation (SD) of the negative (NC) and positive controls (PC) in the three EPISKIN-laboratories in Phase II.

The third criterion demanded that the variability in terms of standard deviation (SD) of the negative control replicates was smaller than 18. The fourth criterion was identical, but referred to the variability of the positive control. The data for both controls are shown in Figure 20, where one negative control at Unilever and one positive control at Sanofi did not fulfil the respective criterion. These two experiments were repeated.

The last criterion focused on the variability of a tested sample. In order to be interpretable, the variability criterion of a tested sample was set at a SD of 18. This means that a test sample, whose three replicates showed an SD > 18 had to be retested. However, a chemical could only be retested once.

At L'Oréal, in total 178 tests were carried out with the 58 chemicals, where four chemicals were tested four times (Table 30). Ten tests did not meet the variability criterion. Unilever performed 187 tests, where 13 of these did not the SD-criterion and thus 13 chemicals were retested once. At Sanofi, 182 tests were performed, eight of which had an unacceptable variability and triggered thus a retest.

chemical number	L'Oréal				Unilever				Sanofi			
	run				run				run			
	1	2	3	4	1	2	3	4	1	2	3	4
1	0.77	0.15	1.20		1.22	3.44	0.57		0.43	1.08	0.49	
2	0.79	0.71	0.56		0.29	0.93	0.19		0.26	0.69	0.32	
3	14.26	10.05	17.12		**34.14**	4.80	13.93	8.54	5.18	11.49	12.61	
4	1.83	2.76	1.26		1.11	2.48	1.81		0.18	2.20	0.83	
5	**23.09**	**23.51**	**28.72**		1.00	4.61	15.80		**26.73**	13.79	0.79	0.84
6	17.74	12.08	**18.49**	5.79	10.74	11.73	6.54		**27.65**	**25.19**	**29.11**	
7	3.94	5.51	13.88		15.52	5.08	2.27		7.68	13.11	1.86	
8	1.59	17.46	14.91		1.35	**19.81**	5.23	1.49	1.67	6.43	1.56	
9	4.48	6.59	5.56		8.89	**31.94**	4.26	5.67	2.75	8.96	4.41	
10	6.49	3.84	7.64		13.48	9.37	16.19		3.30	11.37	3.99	
11	2.15	7.08	7.24		11.99	9.69	7.71		3.42	5.99	4.11	
12	3.47	11.58	6.48		5.02	10.35	14.33		3.37	7.16	2.85	
13	0.71	1.39	0.35		0.55	0.32	2.48		4.45	0.39	17.52	
14												
15	11.69	7.84	**18.85**	6.37	0.83	0.40	0.25		1.13	15.88	0.19	
16	2.42	1.61	5.10		2.34	2.48	7.40		3.55	1.81	5.79	
17	2.71	0.61	1.87		0.96	2.62	0.05		2.62	0.66	0.70	
18	4.76	1.70	1.46		**25.23**	4.12	10.12	4.45	**44.18**	7.40	**26.47**	
19	5.20	1.53	2.45		10.59	4.86	6.78		3.54	**27.13**	4.70	5.44
20	11.36	2.89	13.13		0.60	3.89	2.20		3.52	**29.88**	6.17	**31.69**
21	3.04	7.26	3.92		10.28	4.61	7.29		2.65	1.72	7.62	
22	4.16	5.88	4.53		2.83	3.41	15.27		8.76	1.19	2.30	
23	**29.88**	11.68	6.66	**30.12**	1.59	21.64	3.27	11.19	2.16	**21.12**	6.55	10.49
24	11.58	5.09	7.29		4.36	9.53	8.02		**27.24**	6.38	2.83	3.99
25	3.38	6.60	1.18		2.02	3.28	9.97		13.10	10.08	1.68	
26	1.92	1.70	1.35		13.09	4.45	11.39		0.74	3.10	0.10	
27	14.90	17.82	13.22		**42.94**	2.13	13.72	0.94	7.51	7.93	5.75	
28	3.73	1.95	6.32		6.61	9.22	5.77		5.12	1.60	2.45	
29	2.28	1.32	0.38		0.54	1.22	1.33		0.44	0.55	1.10	
30	14.52	4.91	4.32		9.80	**21.52**	13.54	9.77	4.36	2.33	2.76	
31	1.32	2.90	1.03		0.35	16.36	2.92		2.00	0.89	0.92	
32	1.13	5.65	4.00		5.71	7.37	12.24		6.92	8.95	1.09	
33	1.98	2.77	7.27		6.46	13.06	5.76		7.14	6.45	5.07	
34	3.65	9.41	1.44		**28.67**	8.09	3.96	1.15	14.11	6.73	3.74	
35	4.85	3.99	2.25		7.67	14.06	7.44		9.55	9.10	4.88	
36	14.16	6.67	2.77		6.40	**21.80**	9.96	2.72	**27.26**	0.63	3.68	7.56
37	1.68	1.94	0.50		0.25	1.05	1.31		0.41	12.27	0.74	
38												
39	1.44	4.32	6.94		7.46	6.02	8.56		2.21	3.96	7.24	
40	1.26	7.75	6.08		11.42	3.92	11.15		**22.69**	0.81	6.62	0.1
41	2.10	3.86	2.01		4.54	6.69	3.71		7.99	3.33	6.13	
42	5.25	1.38	5.86		**44.07**	3.91	10.71	6.91	5.98	6.07	6.10	
43	3.17	15.36	5.13		2.43	2.90	4.13		**20.17**	16.80	7.59	16.92
44	5.23	5.13	5.00		14.97	16.43	**33.45**	10.04	10.19	5.19	12.70	
45	1.15	0.54	0.41		0.95	1.57	1.11		0.63	2.68	0.81	
46	3.41	9.73	4.73		1.33	0.66	3.48		0.30	2.24	0.75	
47	1.11	6.41	2.37		4.52	10.49	10.89		1.26	3.38	1.75	
48	10.98	5.31	2.42		4.10	5.76	5.54		9.24	6.11	2.62	
49	14.55	4.81	3.40		3.43	2.16	4.46		6.05	5.30	3.46	
50	5.77	5.68	3.98		2.87	**29.41**	9.65	7.72	7.22	3.73	5.02	
51	16.87	10.26	8.62		3.61	4.09	1.87		3.88	8.93	7.14	
52	**34.45**	5.08	9.21	5.32	15.30	10.20	**38.07**	16.14	3.11	6.64	17.22	
53	**28.20**	14.17	**24.75**		**21.20**	**22.61**	**18.91**		5.55	9.38	11.06	
54	8.61	7.49	5.31		2.25	17.55	9.84		5.77	4.54	10.69	
55	15.88	16.75	11.31		0.78	1.28	0.40		**20.99**	**23.03**	4.46	
56	2.06	1.44	1.54		**30.46**	10.88	**32.10**		8.40	2.69	2.35	
57	3.91	7.04	2.45		6.84	13.38	5.94		3.39	6.47	1.91	
58	0.47	0.33	1.72		0.49	0.93	0.44		0.74	1.18	0.36	
59	0.80	2.32	8.11		0.90	0.61	10.91		6.20	7.39	10.40	
60	2.86	1.54	4.42		1.98	0.78	0.64		3.70	5.93	17.62	

Table 30: Standard deviations of all EPISKIN-tests in the three laboratories of all 60 substances, where those with a SD > 18 are marked grey
(* confidential chemicals)

3.1.3 Within-laboratory variability

3.1.3.1 L'Oréal

3.2.3.1.1 1-way ANOVA

To compare the independent experiments within a laboratory, a 1-way ANOVA (significance level of 1%) was applied to the data of each test compound. The ANOVA was calculated for all available runs per chemical and for those chemicals, which had three runs meeting the variability acceptance criterion. Table 31 shows that same five chemicals were not reproducible in terms of the 1-way ANOVA for both approaches.

3.2.3.1.2 Within-laboratory standard deviation s_R

Also the within-laboratory standard deviation was calculated for all available runs per chemical and for the first three qualifying runs per chemical. The data for all substances are displayed in Table 32. Transferring the value of 18 from the variability criterion to this type of standard deviation, four chemicals (numbers 5, 43, 53, 55) showed a $s_R > 18$ when considering all runs. Focusing on the three valid runs, only two chemicals (numbers 43, 55) had a $s_R > 18$. This can be interpreted as evidence that the variability criterion of SD > 18 supports the reproducibility of the test by identifying highly variable runs, which tend to be aberrant. The distribution of s_R in the three laboratories is compared in Chapter 3.2.3.4.

chemical number	number of runs	number of non-valid runs	p-value of all runs	p-value of three valid runs
1	3		0.1995	0.1995
2	3		0.0065	0.0065
3	3		0.4370	0.4370
4	3		0.6467	0.6467
5	3	3	0.0785	-
6	4	1	0.4147	0.3547
7	3		0.4081	0.4081
8	3		0.0986	0.0986
9	3		0.1360	0.1360
10	3		0.1440	0.1440
11	3		0.0405	0.0405
12	3		0.1687	0.1687
13	3		0.8490	0.8490
14*	3		0.1223	0.1223
15	4	1	0.7475	0.3874
16	3		0.0675	0.0675
17	3		0.4154	0.4154
18	3		0.0575	0.0575
19	3		0.1895	0.1895
20	3		0.9150	0.9150
21	3		0.1220	0.1220
22	3		0.1139	0.1139
23	4	3	0.3606	-
24	3		0.3209	0.3209
25	3		0.3391	0.3391
26	3		0.2285	0.2285
27	3		0.3187	0.3187
28	3		0.1820	0.1820
29	3		0.3175	0.3175
30	3		0.1135	0.1135
31	3		0.0024	0.0024
32	3		0.0138	0.0138
33	3		0.0159	0.0159
34	3		0.0077	0.0077
35	3		0.5093	0.5093
36	3		0.8133	0.8133
37	3		0.1506	0.1506
38*	3		0.2053	0.2053
39	3		0.2839	0.2839
40	3		0.5433	0.5433
41	3		0.2926	0.2926
42	3		0.2645	0.2645
43	3		0.0005	0.0005
44	3		0.4282	0.4282
45	3		0.6197	0.6197
46	3		0.7687	0.7687
47	3		0.0234	0.0234
48	3		0.4851	0.4851
49	3		0.5063	0.5063
50	3		0.2736	0.2736
51	3		0.1202	0.1202
52	4	1	0.1069	0.2639
53	3	3	0.2358	-
54	3		0.0437	0.0437
55	3		0.0236	0.0236
56	3		0.0134	0.0134
57	3		0.0678	0.0678
58	3		0.1247	0.1247
59	3		0.0021	0.0021
60	3		0.0139	0.0139

Table 31: L’Oréal within-laboratory variability: 1-way ANOVA p-values (*confidential chemicals)

chemical number	number of runs	run 1	run 2	run 3	run 4	s_R	
						all runs	three valid runs
1	3	5.15	5.39	6.47		0.70	0.70
2	3	5.27	7.61	4.97		1.45	1.45
3	3	21.46	21.81	35.41		7.95	7.95
4	3	6.40	7.58	7.95		0.81	0.81
5	*3*	*88.35*	*34.81*	*41.53*		*29.16*	-
6	4	29.35	30.07	35.77	15.44	8.63	8.24
7	3	56.68	66.67	64.55		5.26	5.26
8	3	46.87	40.06	67.59		14.34	14.34
9	3	106.05	102.29	113.04		5.46	5.46
10	3	99.44	96.24	107.69		5.91	5.91
11	3	98.76	87.77	103.94		8.26	8.26
12	3	108.14	119.62	106.56		7.13	7.13
13	3	4.75	4.96	5.19		0.22	0.22
14*	3						
15	4	19.97	15.07	16.48	9.11	4.53	5.44
16	3	98.11	95.43	103.47		4.09	4.09
17	3	11.82	9.59	11.03		1.13	1.13
18	3	11.38	15.02	7.36		3.83	3.83
19	3	115.19	111.05	116.78		2.96	2.96
20	3	23.36	23.35	26.41		1.76	1.76
21	3	102.23	99.08	109.09		5.12	5.12
22	3	94.85	100.59	90.48		5.07	5.07
23	*4*	*62.51*	*55.20*	*59.31*	*31.15*	*14.25*	-
24	3	110.97	99.88	107.88		5.72	5.72
25	3	105.30	108.18	102.47		2.86	2.86
26	3	8.56	6.28	6.22		1.33	1.33
27	3	92.99	76.85	73.26		10.51	10.51
28	3	116.54	116.07	122.93		3.83	3.83
29	3	11.66	9.88	11.74		1.05	1.05
30	3	75.95	92.84	76.92		9.49	9.49
31	3	8.88	17.71	9.57		4.91	4.91
32	3	98.19	93.40	107.57		7.21	7.21
33	3	98.01	100.87	113.08		8.00	8.00
34	3	102.73	86.80	110.03		11.88	11.88
35	3	120.14	121.46	123.95		1.93	1.93
36	3	103.23	98.84	99.16		2.45	2.45
37	3	7.85	10.66	8.94		1.42	1.42
38*	3						
39	3	104.23	100.42	107.33		3.46	3.46
40	3	6.36	11.43	10.62		2.72	2.72
41	3	95.04	98.67	95.47		1.98	1.98
42	3	101.36	96.45	103.09		3.45	3.45
43	3	91.46	36.34	32.85		32.88	32.88
44	3	92.95	89.45	95.26		2.93	2.93
45	3	11.75	11.37	12.00		0.32	0.32
46	3	11.20	14.65	11.23		1.98	1.98
47	3	11.31	22.11	11.04		6.31	6.31
48	3	91.68	98.98	96.81		3.75	3.75
49	3	100.59	94.94	91.53		4.57	4.57
50	3	118.37	111.76	118.41		3.83	3.83
51	3	71.12	93.10	92.80		12.60	12.60
52	4	43.62	81.79	72.18	79.88	17.66	5.09
53	*3*	*57.34*	*59.47*	*89.89*		*18.21*	-
54	3	59.86	77.76	61.67		9.85	9.85
55	3	28.70	75.41	54.77		23.41	23.41
56	3	16.26	10.87	11.10		3.05	3.05
57	3	103.23	91.66	99.01		5.86	5.86
58	3	8.33	6.52	6.51		1.05	1.05
59	3	7.84	32.69	25.21		12.75	12.75
60	3	93.22	82.20	89.67		5.63	5.63

Table 32: L'Oréal within-laboratory standard deviation s_R
(light grey cells: runs, which were not considered for three valid runs; dark grey cells: chemicals with $s_R > 18$; italic: chemicals without three valid runs;
* confidential chemicals)

3.2.3.1.3 Correlation

The third measure of reproducibility within a laboratory was Bravais-Pearson correlation coefficient r. It was applied to correlate the results of two complete runs. However, as already seen in Phase I, the value of this measure is limited due to the fact that the test protocol was designed to separate irritants from non-irritants. Calculating the correlation of the mean cell viability for all three pairs of runs for the 55 chemicals with three valid runs resulted in correlation coefficients above 0.95 (Table 33). When considering the first three runs of all chemicals the correlation ranged between 0.9330 and 0.9738.

	correlation r
Run 1 – Run 2	0.9813
Run 1 – Run 3	0.9808
Run 2 – Run 3	0.9756

Table 33: L’Oréal run correlations

3.2.3.1.4 Proportion of identically classified chemicals

The crudest measure to for within-laboratory reproducibility was the proportion of identically classified chemicals. The classification was done according to the PM. This proportion was first applied to the 55 chemicals with three valid runs. The classifications, which can easily be derived from Table 32, are identical in the three valid runs for 52 substances. Only three chemicals (number 7, 43 and 55), the later two having a s_R >18, were classified non-consistently. When considering all runs, 52 of 58 chemicals were classified consistently.

3.1.3.2 Unilever

3.2.3.2.1 1-way ANOVA

To compare the independent experiments within a laboratory, a 1-way ANOVA (significance level of 1%) was applied to the data of each test compound. The ANOVA was calculated for all available runs per chemical and for those chemicals, which had three runs meeting the variability acceptance criterion. Table 34 shows that eleven chemicals were not reproducible in terms of the 1-way ANOVA for both sets and additional three chemicals were not reproducible in the set considering all runs.

chemical number	number of runs	number of non-valid runs	p-value of all runs	p-value of three valid runs
1	3		0.5263	0.5263
2	3		0.0237	0.0237
3	4	1	0.8547	0.2562
4	3		0.9264	0.9264
5	3		0.3578	0.3578
6	3		0.0209	0.0209
7	3		0.1021	0.1021
8	4	1	0.0076	0.3168
9	4	1	0.0561	0.6093
10	3		0.2602	0.2602
11	3		0.0029	0.0029
12	3		0.1174	0.1174
13	3		0.2373	0.2373
14*				
15	3		0.2451	0.2451
16	3		0.003	0.003
17	3		0.6272	0.6272
18	4	1	0.5024	0.013
19	3		0.3182	0.3182
20	3		0.3893	0.3893
21	3		0.0389	0.0389
22	3		0.0107	0.0107
23	4	1	0.7154	0.3874
24	3		0.0275	0.0275
25	3		0.0191	0.0191
26	3		0.9363	0.9363
27	4	1	0.2309	0.1209
28	3		0.0092	0.0092
29	3		0.9121	0.9121
30	4	1	0.0395	0.1527
31	3		0.3839	0.3839
32	3		0.0132	0.0132
33	3		0.1287	0.1287
34	4	1	0.0017	0.2107
35	3		0.01	0.01
36	4	1	0.4346	0.8259
37	3		0.0019	0.0019
38*				
39	3		0.2665	0.2665
40	3		0.1127	0.1127
41	3		0.0055	0.0055
42	4	1	0.0603	0.2264
43	3		0.6067	0.6067
44	4	1	0.1501	0.268
45	3		0.0312	0.0312
46	3		0.7736	0.7736
47	3		0.1473	0.1473
48	3		0.0312	0.0312
49	3		0.0017	0.0017
50	4	1	0.0028	<0.0001
51	3		0.029	0.029
52	4	1	0.0097	0.0484
53	3	3	0.1238	-
54	3		0.0041	0.0041
55	3		0.005	0.005
56	3	3	0.9687	-
57	3		0.0271	0.0271
58	3		0.0008	0.0008
59	3		0.0029	0.0029
60	3		0.044	0.044

Table 34: Unilever within-laboratory variability: 1-way ANOVA p-values (*confidential chemicals)

3.2.3.2.2 Within-laboratory standard deviation s_R

Also the within-laboratory standard deviation was calculated for all available runs per chemical and for the first three qualifying runs per chemical. The data for all substances are displayed in Table 35. Transferring the value of 18 from the variability criterion to this type of standard deviation, seven chemicals (numbers 11, 18, 27, 25, 50, 52, 54) showed a $s_R > 18$ when considering all runs. Focusing on the three valid runs, only five chemicals (numbers 11, 25, 50, 52, 54) had a $s_R > 18$. This can be interpreted as evidence that the variability criterion of SD > 18 supports the reproducibility of the test by identifying highly variable runs, which tend to be aberrant. The distribution of s_R in the three laboratories is compared in Chapter 3.2.3.4.

chemical number	number of runs	run 1	run 2	run 3	run 4	s_R	
						all runs	three valid runs
1	3	5.25	5.28	3.46		1.04	1.04
2	3	4.30	4.75	3.02		0.90	0.90
3	4	39.84	9.47	14.00	12.22	14.10	2.28
4	3	6.22	6.79	6.33		0.30	0.30
5	3	6.90	8.02	17.96		6.09	6.09
6	3	33.10	63.96	56.54		16.11	16.11
7	3	27.22	10.46	8.88		10.16	10.16
8	4	9.73	20.68	18.93	6.30	6.99	6.53
9	4	107.71	106.37	96.67	91.37	7.84	8.34
10	3	101.61	108.35	88.64		10.02	10.02
11	3	98.37	71.08	119.71		24.38	24.38
12	3	94.58	99.36	78.67		10.83	10.83
13	3	4.26	4.10	6.18		1.16	1.16
14*							
15	3	2.62	3.22	2.40		0.43	0.43
16	3	83.89	106.75	97.21		11.48	11.48
17	3	4.09	5.21	4.03		0.66	0.66
18	4	52.56	11.61	13.52	6.81	**21.14**	**3.46**
19	3	105.37	107.60	115.47		5.30	5.30
20	3	8.66	6.88	10.04		1.58	1.58
21	3	111.34	94.69	114.98		10.82	10.82
22	3	56.86	91.26	77.00		17.28	17.28
23	4	5.40	22.23	7.05	10.01	7.62	2.33
24	3	81.48	104.59	94.60		11.59	11.59
25	3	84.96	104.66	90.09		10.22	10.22
26	3	29.74	32.82	31.33		1.54	1.54
27	4	52.27	9.72	18.20	5.71	**21.18**	**6.38**
28	3	107.92	132.57	107.82		14.26	14.26
29	3	6.05	6.42	6.12		0.19	0.19
30	4	70.52	68.67	50.93	81.65	12.71	15.55
31	3	6.48	17.93	9.87		5.88	5.88
32	3	101.56	103.61	130.20		15.98	15.98
33	3	94.70	112.19	100.08		8.96	8.96
34	4	106.03	103.53	112.92	85.02	11.92	14.20
35	3	96.97	134.22	105.93		19.44	19.44
36	4	76.16	84.93	81.26	83.97	3.93	3.97
37	3	9.35	7.15	12.35		2.61	2.61
38*							
39	3	85.59	96.40	89.01		5.52	5.52
40	3	22.76	6.07	23.37		9.82	9.82
41	3	75.83	95.85	94.10		11.09	11.09
42	4	69.43	110.12	102.92	91.13	17.80	9.59
43	3	32.59	29.86	31.50		1.38	1.38
44	4	88.74	81.55	98.72	63.10	15.03	13.22
45	3	8.17	11.77	9.44		1.82	1.82
46	3	7.16	7.19	8.31		0.65	0.65
47	3	22.73	39.94	31.82		8.61	8.61
48	3	80.67	95.70	90.69		7.65	7.65
49	3	74.52	93.07	86.98		9.46	9.46
50	4	71.49	65.34	133.23	76.60	31.38	34.27
51	3	89.98	81.34	81.29		5.00	5.00
52	4	83.96	77.88	36.11	34.28	26.53	27.10
53	*3*	*58.57*	*35.96*	*77.97*		*21.03*	-
54	3	9.60	52.90	58.70		26.83	26.83
55	3	5.36	4.90	8.52		1.97	1.97
56	*3*	*90.11*	*84.78*	*88.34*		*2.71*	-
57	3	94.15	116.88	90.66		14.24	14.24
58	3	4.36	5.07	8.25		2.07	2.07
59	3	6.74	4.87	32.67		15.54	15.54
60	3	8.33	5.11	5.60		1.73	1.73

Table 35: Unilever within-laboratory standard deviation s_R
(light grey cells: runs, which were not considered for three valid runs; dark grey cells: chemicals with s_R > 18; italic: chemicals without three valid runs;
* confidential chemicals)

3.2.3.2.3 Correlation

The third measure of reproducibility within a laboratory was Bravais-Pearson correlation coefficient r. It was applied to correlate the results of two complete runs. However, as already seen in Phase I, the value of this measure is limited due to the fact that the test protocol was designed to separate irritants from non-irritants. Calculating the correlation of the mean cell viability for all three pairs of runs for the 56 chemicals with three valid runs resulted in correlation coefficients between 0.93 and 0.94 (Table 36). When considering the first three runs of all chemicals, the correlation was always larger than 0.9.

	correlation r
Run 1 – Run 2	0.9382
Run 1 – Run 3	0.9386
Run 2 – Run 3	0.9304

Table 36: Unilever run correlations

3.2.3.2.4 Proportion of identically classified chemicals

The crudest measure to for within-laboratory reproducibility was the proportion of identically classified chemicals. The classification was done according to the PM. This proportion was first applied to the 56 chemicals with three valid runs. The classifications, which can easily be derived from Table 35, were identical in the three valid runs for 53 substances. Only chemical numbers 6, 52 and 54, the latter two having a s_R >18, were classified non-consistently. Considering all runs, 52 of 58 chemicals were not consistently classified.

3.1.3.3 Sanofi

3.2.3.3.1 1-way ANOVA

To compare the independent experiments within a laboratory, again a 1-way ANOVA (significance level of 1%) was applied to the data of each test compound. The ANOVA was calculated for all available runs per chemical and for those chemicals, which had three runs meeting the variability acceptance criterion. Table 37 shows that eight chemicals were not reproducible in terms of the 1-way ANOVA in each of both sets. Seven of these were the same.

chemical number	number of runs	number of non-valid runs	p-value	p-value of three valid runs
1	3		0.0295	0.0295
2	3		0.4384	0.4384
3	3		0.4288	0.4288
4	3		0.0923	0.0923
5	4	1	0.0092	0.0593
6	3	3	0.0484	-
7	3		0.6223	0.6223
8	3		0.0390	0.0390
9	3		0.1290	0.1290
10	3		0.2710	0.2710
11	3		0.0009	0.0009
12	3		0.0012	0.0012
13	3		0.0414	0.0414
14*				
15	3		0.4522	0.4522
16	3		0.0181	0.0181
17	3		0.0148	0.0148
18	3	3	0.6450	-
19	4	1	0.5166	0.1687
20	4	2	0.2709	-
21	3		0.6379	0.6379
22	3		0.2249	0.2249
23	4	1	0.8776	0.5374
24	4	1	0.0850	0.0460
25	3		0.1282	0.1282
26	3		0.3786	0.3786
27	3		0.1603	0.1603
28	3		0.0015	0.0015
29	3		0.5450	0.5450
30	3		0.2747	0.2747
31	3		0.1778	0.1778
32	3		0.5079	0.5079
33	3		0.0167	0.0167
34	3		0.6247	0.6247
35	3		0.0298	0.0298
36	4	1	0.0861	0.0013
37	3		0.0837	0.0837
38*				
39	3		0.0384	0.0384
40	4	1	0.0529	0.4129
41	3		0.6530	0.6530
42	3		0.1190	0.1190
43	4	1	0.3927	0.2202
44	3		0.2895	0.2895
45	3		0.4781	0.4781
46	3		0.0846	0.0846
47	3		0.0002	0.0002
48	3		0.0590	0.0590
49	3		0.2394	0.2394
50	3		0.5102	0.5102
51	3		0.0065	0.0065
52	3		0.1112	0.1112
53	3		0.0008	0.0008
54	3		0.0007	0.0007
55	3	3	0.0488	-
56	3		0.8974	0.8974
57	3		0.0638	0.0638
58	3		0.0900	0.0900
59	3		0.1796	0.1796
60	3		0.1462	0.1462

Table 37: Sanofi within-laboratory variability: 1-way ANOVA p-values (*confidential chemicals)

3.2.3.3.2 Within-laboratory standard deviation s_R

Also the within-laboratory standard deviation was calculated for all available runs per chemical and for the first three qualifying runs per chemical. The data for all substances are displayed in Table 38. Transferring the value of 18 from the variability criterion to this type of standard deviation, five chemicals (numbers 5, 6, 53, 54, 55) showed a s_R > 18 when considering all runs. Focusing on the three valid runs, only two chemicals (numbers 53, 54) had a s_R > 18. This can be interpreted as evidence that the variability criterion of SD > 18 supports the reproducibility of the test by identifying highly variable runs, which tend to be aberrant. The distribution of s_R in the three laboratories is compared in Chapter 3.2.3.4.

chemical number	number of runs	run 1	run 2	run 3	run 4	s_R	
						all runs	three valid runs
1	3	2.89	5.06	3.82		1.09	1.09
2	3	4.15	4.67	4.37		0.26	0.26
3	3	53.57	43.77	43.03		5.88	5.88
4	3	5.22	7.96	7.65		1.50	1.50
5	4	55.98	21.69	4.93	3.88	**24.31**	**9.99**
6	*3*	*92.28*	*74.27*	*22.67*		*36.13*	-
7	3	41.37	43.27	48.44		3.66	3.66
8	3	28.70	39.50	35.86		5.50	5.50
9	3	91.98	111.89	109.22		10.81	10.81
10	3	104.57	104.25	95.19		5.33	5.33
11	3	88.90	117.33	100.17		14.32	14.32
12	3	94.74	121.46	99.90		14.17	14.17
13	3	6.33	4.03	29.97		14.36	14.36
14*	3						
15	3	3.70	12.23	3.26		5.06	5.06
16	3	98.84	110.17	97.98		6.81	6.81
17	3	12.54	6.89	9.76		2.83	2.83
18	3	33.94	14.20	35.64		11.92	-
19	4	99.54	88.30	98.64	106.25	7.41	4.16
20	*4*	*18.56*	*51.05*	*50.67*	*49.66*	*15.96*	-
21	3	102.40	106.22	104.04		1.92	1.92
22	3	89.58	97.39	90.65		4.23	4.23
23	4	74.95	73.44	81.03	75.06	3.35	3.48
24	4	70.04	104.02	91.68	96.63	14.61	6.21
25	3	99.46	113.87	95.91		9.51	9.51
26	3	2.12	3.90	1.78		1.14	1.14
27	3	92.92	93.69	82.01		6.53	6.53
28	3	104.05	123.06	114.11		9.51	9.51
29	3	8.04	9.02	9.97		0.97	0.97
30	3	102.53	99.42	97.82		2.40	2.40
31	3	8.91	7.93	6.51		1.21	1.21
32	3	98.51	104.72	103.56		3.30	3.30
33	3	100.25	121.56	112.80		10.71	10.71
34	3	90.83	98.47	94.37		3.82	3.82
35	3	105.04	128.40	111.18		12.11	12.11
36	4	68.66	97.46	92.33	71.10	14.63	13.98
37	3	5.02	20.22	8.02		8.05	8.05
38*							
39	3	97.24	109.06	96.94		6.91	6.91
40	4	31.25	3.50	6.85	2.56	13.60	2.25
41	3	90.68	94.72	90.49		2.39	2.39
42	3	93.67	102.03	90.05		6.15	6.15
43	4	51.89	37.39	59.99	43.43	9.87	11.70
44	3	95.63	103.22	89.12		7.06	7.06
45	3	9.76	8.03	8.69		0.87	0.87
46	3	3.58	6.34	3.71		1.55	1.55
47	3	9.24	23.70	7.06		9.04	9.04
48	3	86.60	102.57	91.11		8.23	8.23
49	3	92.95	100.79	96.02		3.95	3.95
50	3	111.76	116.09	110.94		2.77	2.77
51	3	51.45	80.49	68.01		14.57	14.57
52	3	74.62	96.94	83.63		11.23	11.23
53	3	43.77	94.64	89.71		28.05	28.05
54	3	28.22	71.49	67.87		24.01	24.01
55	*3*	*30.02*	*77.92*	*53.38*		*23.95*	-
56	3	15.01	13.06	14.51		1.01	1.01
57	3	101.16	106.99	96.32		5.34	5.34
58	3	5.86	7.25	7.60		0.92	0.92
59	3	10.88	24.99	15.48		7.20	7.20
60	3	85.34	84.24	66.85		10.37	10.37

Table 38: Sanofi within-laboratory standard deviation s_R
(light grey cells: runs, which were not considered for three valid runs; dark grey cells: chemicals with s_R > 18; italic: chemicals without three valid runs;
* confidential chemicals)

3.2.3.3.3 Correlation

The third measure of reproducibility within a laboratory was Bravais-Pearson correlation coefficient r. It was applied to correlate the results of two complete runs. However, as already seen in Phase I, the value of this measure is limited due to the fact that the test protocol was designed to separate irritants from non-irritants. Calculating the correlation of the mean cell viability for all three pairs of runs for the 54 chemicals with three valid runs resulted in correlation coefficients of at least 0.95 (Table 39). When considering the first three runs of all chemicals, the correlation was larger than 0.9.

	correlation r
Run 1 – Run 2	0.9521
Run 1 – Run 3	0.9583
Run 2 – Run 3	0.9829

Table 39: Sanofi run correlations

3.2.3.3.4 Proportion of identically classified chemicals

The crudest measure to for within-laboratory reproducibility was the proportion of identically classified chemicals. The classification was done according to the Prediction Model (PM) of the SOP. This proportion was first applied to the 55 chemicals with three valid runs. The classifications, which can easily be derived from Table 38, were identical in the three valid runs for 50 substances. Only chemical numbers 3, 43, 53 and 54, the later two having a s_R >18, were classified non-consistently. When considering all runs, 50 out of 58 chemicals had consistent classifications in all runs.

3.2.3.4 Summary within-laboratory variability results

The results of the within-laboratory variability of all applied measures are summarised for the three laboratories in Table 40. Regarding the sample size, the amounts of samples with three valid runs are similar. The number of chemicals, which gave significant ANOVA results, differs to some extend between laboratories as it depends on the within-assay variability.

	L'Oréal		Unilever		Sanofi	
	all	three valid runs	all	three valid runs	all	three valid runs
sample size	58	55	58	56	58	54
ANOVA: number of chemicals with significant run differences	5	5	14	11	8	8
number of chemicals with s_R >18	4	2	7	5	5	2
mean correlation of runs	0.9513	0.9792	0.9144	0.9358	0.9330	0.9691
proportion of identically classified chemicals	52/58	52/55	52/58	53/56	50/58	50/54

Table 40: Summary of within-laboratory variability evaluation of EPISKIN

Comparing this variability via the relative cumulative distribution of the standard deviations of all tests revealed, however, no substantial differences between the laboratories (Figure 21). In all three laboratories more than 90% of the experimental runs had a standard deviation smaller than 18. Regarding the variability criterion, which considers a s_R > 18 as unacceptable, only minor differences between the laboratories became evident when assessing the amount of failing chemicals.

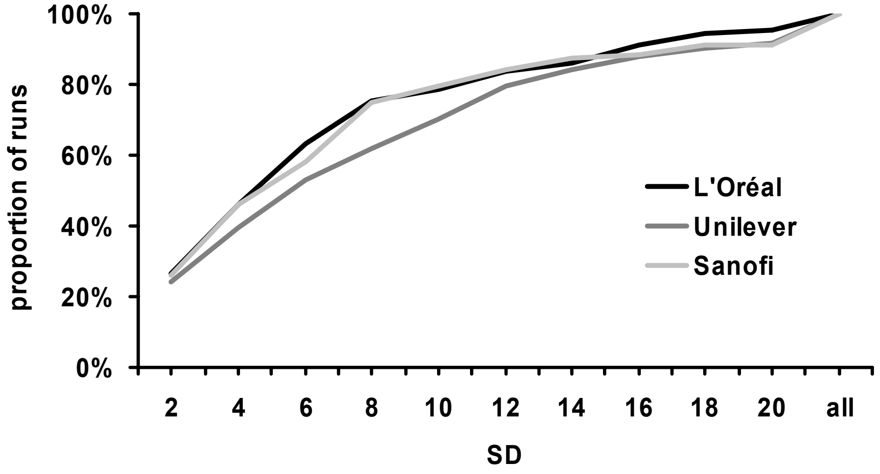


Figure 21: Relative cumulative distribution of the standard deviations of all tests in the three EPISKIN-laboratories

3.2.4 Within-laboratory variability between phases

The comparison the results of the 18 chemicals tested in both phases of the validation study was performed only with the three valid runs. This information added to the assessment of the within-laboratory variability, as up to twelve month lay between Phase I Run 1 and Phase II Run3 and at least 7 month lay between Phase I Run 3 and Phase II Run 1. All tests were carried out by the same operator, but for each phase a new sample of the respective chemicals were provided. When applying a t-test with a significance level of 1% to each of the run data of each of the chemicals, none of the substances gave significantly different results in the two phases II (Table 41).

chemical number	Phase I			Phase II			t-test p-value
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	
1	5.02	5.89	4.27	5.15	5.39	6.47	0.5316
9	100.04	103.86	105.48	106.05	102.29	113.04	0.2919
12	104.29	92.88	81.43	108.14	119.62	106.56	0.1281
13	7.34	5.54	9.91	4.75	4.96	5.19	0.1585
16	93.62	95.21	100.38	98.11	95.43	103.47	0.1743
17	13.18	12.76	8.48	11.82	9.59	11.03	0.7341
28	116.13	117.04	96.54	116.54	116.07	122.93	0.4354
29	13.57	10.75	9.42	11.66	9.88	11.74	0.9149
30	106.62	111.55	107.24	75.95	92.84	76.92	0.0212
32	95.90	93.30	107.05	98.19	93.40	107.57	0.2848
35	105.65	105.08	123.47	120.14	121.46	123.95	0.1726
37	7.37	7.91	6.93	7.85	10.66	8.94	0.1201
40	28.54	25.28	10.86	6.36	11.43	10.62	0.1992
42	95.76	103.91	102.37	101.36	96.45	103.09	0.9292
49	89.60	87.45	90.43	100.59	94.94	91.53	0.153
51	66.67	54.82	58.75	71.12	93.10	92.80	0.1379
52	90.02	95.25	70.42	79.88	81.79	72.18	0.2558
59	29.02	7.98	10.31	7.84	32.69	25.21	0.7028

Table 41: Cell viability and comparison of the 18 chemicals tested in both phases at L’Oréal with EPISKIN

Testing the mean viabilities of the phases for the 18 chemicals by a paired t-test resulted in a non-significant p-value of 0.458, which indicated good within-laboratory reproducibility. The differences between these values, which in two cases exceeded 20% points, are presented in Figure 22.

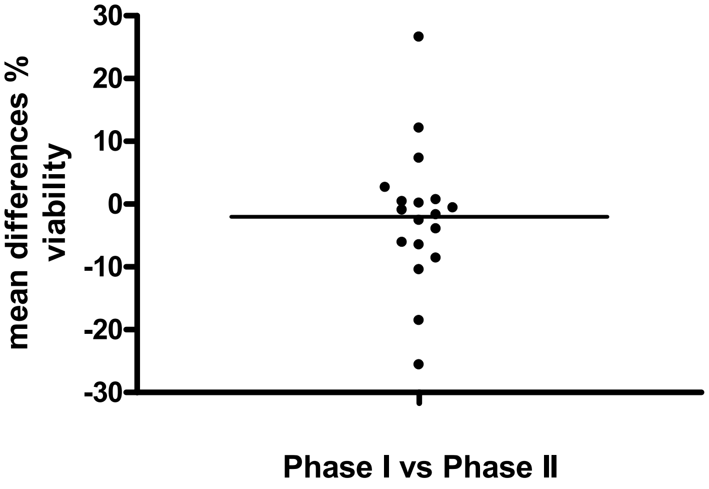


Figure 22: Differences in mean viability of the 18 chemicals tested in both phases at L'Oréal

3.2.5 Between-laboratory variability

The between-laboratory variability was first assessed with a 1-way ANOVA (significance level of 1%). Taking the run means of a chemical, the data of the three laboratories were compared once taking all runs of all chemicals into account and once taking only those chemicals with at least three valid runs in two laboratories into account. In those cases were only two laboratories had three valid runs, these were compared by a t-test (significance level of 1%).

In Table 42, the ANOVA/t-test p-values and the ANOVA sum of squares are given for both approaches. Considering all runs, eight chemicals (numbers 7, 8, 23, 26, 27, 46, 56, 60) gave significantly different results. Four of these (numbers 23, 27, 56, 60) also had a sum of squares > 5000, like another three chemicals (numbers 5, 6, 55).

chemical number	all tests				three valid runs per laboratory			
	n	laboratory with s_R >18	ANOVA p-value	ANOVA sum of squares	n	laboratory with s_R >18	ANOVA/t-test p-value	ANOVA sum of squares
1	9		0.1611	10	9		0.1611	10
2	9		0.1160	12	9		0.1160	12
3	10		0.0293	2172	9		**0.0010**	2052
4	9		0.5997	7	9		0.5997	7
5	10	2	0.1053	6751	6		0.9126	
6	10	1	0.1610	5651	6		0.0659	
7	9		**0.0005**	3674	9		**0.0005**	3674
8	10		**0.0036**	3087	9		**0.0067**	2958
9	10		0.5937	554	9		0.4952	546
10	9		0.9485	333	9		0.9485	333
11	9	1	0.9008	1796	9	1	0.9008	1796
12	9		0.1439	1408	9		0.1439	1408
13	9		0.4054	560	9		0.4054	560
14*								
15	10		0.0118	402	9		0.0356	336
16	9		0.6462	451	9		0.6462	451
17	9		0.0103	89	9		0.0103	89
18	9	1	0.4509	2078	6		0.8494	
19	10		0.0203	726	9		0.0269	361
20	10		0.0115	2779	6		**0.0003**	
21	9		0.8144	314	9		0.8144	314
22	9		0.1147	1409	9		0.1147	1409
23	12		**<0.0001**	9441	6		**0.0003**	
24	10		0.2651	1424	9		0.2367	664
25	9		0.2396	654	9		0.2396	654
26	9		**<0.0001**	1443	9		**<0.0001**	1443
27	10	1	**0.0011**	11530	9		**<0.0001**	11470
28	9		0.8508	651	9		0.8508	651
29	9		**0.0011**	40	9		**0.0011**	40
30	10		0.0114	2429	9		0.0274	2240
31	9		0.4935	152	9		0.4935	152
32	9		0.3793	879	9		0.3793	879
33	9		0.4775	662	9		0.4775	662
34	10		0.6561	831	9		0.7742	778
35	9	1	0.6804	1201	9	1	0.6804	1201
36	11		0.0556	1443	9		0.0697	1055
37	9		0.8849	153	9		0.8849	153
38*								
39	9		0.0496	490	9		0.0496	491
40	10		0.6288	870	9		0.0941	479
41	9		0.4099	357	9		0.4099	357
42	10		0.7634	1134	9		0.5407	348
43	10	1	0.3648	3278	9	1	0.4342	3221
44	10		0.3048	1116	9		0.0936	1027
45	9		0.0614	21	9		0.0614	21
46	9		**0.0020**	106	9		**0.0020**	107
47	9		0.0597	1002	9		0.0597	1002
48	9		0.5070	352	9		0.5070	352
49	9		0.1220	507	9		0.1220	508
50	10	1	0.1818	4881	9	1	0.3911	3272
51	9		0.1583	1465	9		0.1583	1465
52	11	1	0.2765	4550	9	1	0.4189	2370
53	9	3	0.6275	3646	6			
54	9	2	0.3897	3815	9	2	0.3897	3815
55	9	2	0.0381	6689	6	1	0.0262	
56	9		**<0.0001**	11070	9		0.4783	
57	9		0.8950	551	9		0.8950	551
58	9		0.5706	15	9		0.5706	15
59	9		0.7777	991	9		0.7777	991
60	9		**<0.0001**	12350	9		**<0.0001**	12350

Table 42: EPISKIN between-laboratory reproducibility: ANOVA (bold: significant ANOVA; grey cells: ANOVA sum of squares > 5000; * confidential chemicals)

Focusing on the chemicals with three valid runs in at least three laboratories, with the exception of chemical number 53 all chemicals did have three valid runs in at least two laboratories. Seven substances had three valid runs in two laboratories only. In total, ten chemicals (numbers 3, 7, 8, 20, 23, 26, 27, 29, 46, 60) had an ANOVA/t-test p-value below 0.01 and two chemicals (numbers 27, 60) had an ANOVA sum of squares larger than 5000. Summarising, 43 chemicals were according to the here applied measure reproducible between the three laboratories. As chemical numbers 29 and 46 gave significant results only due to high within-laboratory variability, finally nine chemicals (numbers 3, 7, 8, 20, 23, 26, 27, 56, 60) were not reproducible according to the ANOVA-analysis.

The second measure of between-laboratory variability was the standard deviation of the means of the runs per laboratory (Table 43). Transferring the value of 18 from the quality criterion for the standard deviation of the three replicates of a control or sample shows in the case of the analysis of all runs that nine chemicals (numbers 5, 6, 7, 8, 23, 27, 55, 56, 60) had a standard deviation when comparing laboratories larger 18. All of these were also not reproducible according to the respective analysis with the ANOVA above either in terms of the p-value or in terms of the sum of squares. Focusing on the chemical with three valid runs, chemical numbers 5 and 56 SD fell below 18, where chemical numbers 6 and 55 were not significant in the respective ANOVA/t-test analysis. Chemical numbers 6, 23 and 55 had three valid runs in two laboratories only. Executing the same analysis with all valid runs, even if there were only one or two in a laboratory, gave similar results to the analysis of all runs presented here. In addition, chemical number 40 would show an SD >18 (data not shown).

Applying the third measure – the proportion of identical classified chemicals taking into account the median classification per laboratory – 51 of 57 chemicals (chemical 53 excluded) were identically classified. Seven of these had three valid runs in two laboratories only. The not reproducible substances according to this measure were all not reproducible in terms of between-laboratory standard deviation. Considering all runs, 8 out of 58 chemicals were not consistently classified (Table 44). In a pair-wise comparison of the laboratories the concordance 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.

chemical	all runs				three valid runs			
number	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
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**

Table 43: EPISKIN between-laboratory variability: the standard deviation of laboratory run means
(bold: SD > 18; * confidential chemicals)

chemical	EU	median classification			between-laboratory
number	classification	L'Oréal	Unilever	Sanofi	reproducible
2	no label	**1**	**1**	**1**	+
5	no label	**1**	**1**	**1**	+
6	no label	**1**	0	0	-
7	no label	0	**1**	**1**	-
8	no label	**1**	**1**	**1**	+
9	no label	0	0	0	+
10	no label	0	0	0	+
11	no label	0	0	0	+
12	no label	0	0	0	+
16	no label	0	0	0	+
17	no label	**1**	**1**	**1**	+
19	no label	0	0	0	+
21	no label	0	0	0	+
22	no label	0	0	0	+
24	no label	0	0	0	+
25	no label	0	0	0	+
26	no label	**1**	**1**	**1**	+
28	no label	0	0	0	+
30	no label	0	0	0	+
32	no label	0	0	0	+
33	no label	0	0	0	+
35	no label	0	0	0	+
36	no label	0	0	0	+
39	no label	0	0	0	+
41	no label	0	0	0	+
42	no label	0	0	0	+
44	no label	0	0	0	+
48	no label	0	0	0	+
50	no label	0	0	0	+
52	no label	0	1	0	-
53	no label	0	0	0	+
54	no label	0	0	0	+
57	no label	0	0	0	+
1	R38	1	1	1	+
3	R38	1	1	1	+
4	R38	1	1	1	+
13	R38	1	1	1	+
15	R38	1	1	1	+
18	R38	1	1	1	+
20	R38	1	1	1	+
23	R38	**0**	1	**0**	-
27	R38	**0**	1	**0**	-
29	R38	1	1	1	+
31	R38	1	1	1	+
34	R38	**0**	**0**	**0**	+
37	R38	1	1	1	+
40	R38	1	1	1	+
43	R38	1	1	1	+
45	R38	1	1	1	+
46	R38	1	1	1	+
47	R38	1	1	1	+
49	R38	**0**	**0**	**0**	+
51	R38	**0**	**0**	**0**	+
55	R38	**0**	1	**0**	-
56	R38	1	**0**	1	-
58	R38	1	1	1	+
59	R38	1	1	1	+
60	R38	**0**	1	**0**	-
		reproducible non-labeled chemicals			90.9%
		reproducible R38-labeled chemicals			80.0%
		overall reproducibility			86.2%

Table 44: Between-laboratory reproducibility of EPISKIN in terms of identical median classifications ('0': no label/non-irritant; '1': R38/irritant) between the laboratories when considering all runs
('-' indicates a non-reproducible chemical; '+' indicates a reproducible chemical; grey cells highlight the inconclusive cases, i.e. those with equal numbers of negative and positive classifications in the individual runs, which were conservatively considered as skin irritants)

Finally those chemicals, which did not have three valid runs in at least one of the laboratories, are compared in Table 45. While chemical number 53 was problematic in two laboratories, the other chemicals caused problems in one laboratory only.

chemical number	L'Oréal	Unilever	Sanofi
5	0/3	3/3	3/4
6	3/4	3/3	0/3
18	3/3	3/4	0/3
20	3/3	3/3	2/4
23	0/3	3/4	3/4
53	0/3	0/3	3/3
55	3/3	3/3	0/3
56	3/3	0/3	3/3

Table 45: Chemicals without three valid runs in at least one EPISKIN-laboratory, indicated by grey cells

3.2.6 Predictive Capacity

3.2.6.1 L'Oréal

The prediction model of EPISKIN was designed in order to predict the current European classifications for skin irritation, i.e. the label R38 (skin irritant) versus no label, and was identical with the EpiDerm prediction model: a test substance in an experiment was predicted to be a skin irritant if it reduced in average the relative cell viability below 50% compared to the mean cell viability of the negative control. If the mean cell viability was above 50%, it was considered to be a not skin irritating in terms of the European classification system.

In Table 46, the predictions are presented for all test chemicals. To summarise the prediction over the runs, two approaches based on the median classification were applied: the median of all available runs and of three valid runs, when available. However, differences between these were minor. In Table 47, the parameters specificity and sensitivity calculated from Table 46 are presented for each single run and for the summarising approaches. Besides the sample sizes for both parameters, also the exact lower 5%-confidence bound are given.

chemical	EU	run				median approach	
number	classification	1	2	3	4	all runs	three valid runs
2	no label	**1**	**1**	**1**		**1**	**1**
5	no label	0	**1**	**1**		**1**	-
6	no label	**1**	**1**	**1**	**1**	**1**	**1**
7	no label	0	0	0		0	0
8	no label	**1**	**1**	0		**1**	**1**
9	no label	0	0	0		0	0
10	no label	0	0	0		0	0
11	no label	0	0	0		0	0
12	no label	0	0	0		0	0
16	no label	0	0	0		0	0
17	no label	**1**	**1**	**1**		**1**	**1**
19	no label	0	0	0		0	0
21	no label	0	0	0		0	0
22	no label	0	0	0		0	0
24	no label	0	0	0		0	0
25	no label	0	0	0		0	0
26	no label	**1**	**1**	**1**		**1**	**1**
28	no label	0	0	0		0	0
30	no label	0	0	0		0	0
32	no label	0	0	0		0	0
33	no label	0	0	0		0	0
35	no label	0	0	0		0	0
36	no label	0	0	0		0	0
39	no label	0	0	0		0	0
41	no label	0	0	0		0	0
42	no label	0	0	0		0	0
44	no label	0	0	0		0	0
48	no label	0	0	0		0	0
50	no label	0	0	0		0	0
52	no label	**1**	0	0	0	0	0
53	no label	0	0	0		0	-
54	no label	0	0	0		0	0
57	no label	0	0	0		0	0
1	R38	1	1	1		1	1
3	R38	1	1	1		1	1
4	R38	1	1	1		1	1
13	R38	1	1	1		1	1
15	R38	1	1	1	1	1	1
18	R38	1	1	1		1	1
20	R38	1	1	1		1	1
23	R38	**0**	**0**	**0**	1	**0**	-
27	R38	**0**	**0**	**0**		**0**	**0**
29	R38	1	1	1		1	1
31	R38	1	1	1		1	1
34	R38	**0**	**0**	**0**		**0**	**0**
37	R38	1	1	1		1	1
40	R38	1	1	1		1	1
43	R38	**0**	1	1		1	1
45	R38	1	1	1		1	1
46	R38	1	1	1		1	1
47	R38	1	1	1		1	1
49	R38	**0**	**0**	**0**		**0**	**0**
51	R38	**0**	**0**	**0**		**0**	**0**
55	R38	1	**0**	**0**		**0**	**0**
56	R38	1	1	1		1	1
58	R38	1	1	1		1	1
59	R38	1	1	1		1	1
60	R38	**0**	**0**	**0**		**0**	**0**

Table 46: EPISKIN-Classification of the 58 chemicals according to the prediction model at L'Oréal
(0: non irritant (no label); 1: irritant (R38); bold: misclassifications; grey cells: not valid runs)

Regardless the way of analysis, the specificity was always well above 80%. Nevertheless, it can be seen that the analysis of valid runs, i.e. those meeting the quality criterion of a SD < 18, have a slightly increased specificity. A similar effect was observed for sensitivity, which, depending on the analysis ranged between 72.0 and 76.0%. Summarising the data in a conservative way, i.e. classifying a chemical as irritant when it was classified as irritant in at least one run, produced very similar results (data not shown).

		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

Table 47: Predictive Capacity of EPISKIN at L'Oréal in terms of specificity and sensitivity with the respective sample sizes and 5% lower confidence bounds for each run and summaries over all runs

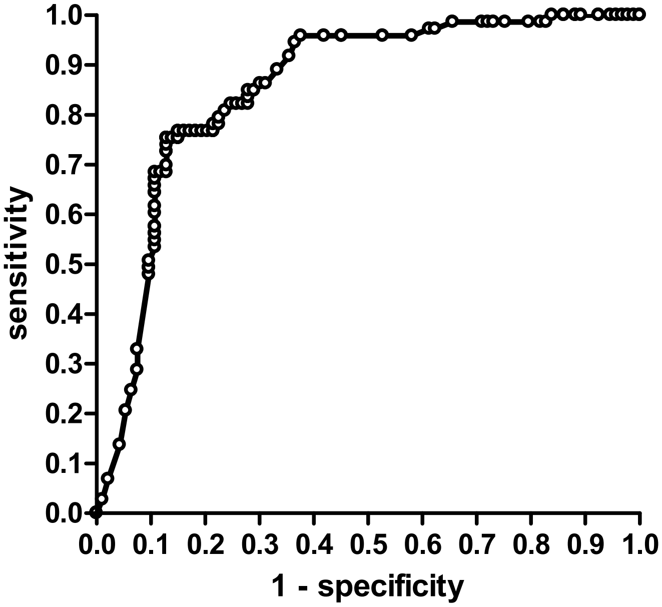


Figure 23: Receiver operation curve of all valid runs from L'Oréal

In order to investigate how the balance of specificity and sensitivity depend on the PM-threshold we submitted the L'Oréal data of all valid runs (n = 166) to a ROC-analysis. This analysis revealed a steep curve for sensitivity up to 75%, which becomes flatter for sensitivities between 75 and 95% and reaches almost a plateau above 95% (Figure 23). This angular shape indicates that an optimal balance of the performance parameters between the two angles. In this area the test can be rendered either more sensitive or more specific with minor trade-offs.

To get more insight into the predictive capacity and its threshold dependence, we plotted the sensitivity, the specificity and the sum of these, which allows optimising the PM-threshold choice when weighing sensitivity and specificity equally. In Figure 24A, which still considered the chemicals classifications according to the European system, it is obvious that the sum of sensitivity and specificity is larger than 1.60 over a wide spectrum of thresholds ranging from about 36% to about 59%. This insensitivity to threshold changes in the middle cell viability response range clearly reflected the optimization efforts with EPISKIN towards a two class system. Furthermore, it shows that the prediction model of 50% was an appropriate choice. However, moving the upper threshold up to 95% would, with one exception, still result in a sum larger than 1.55.

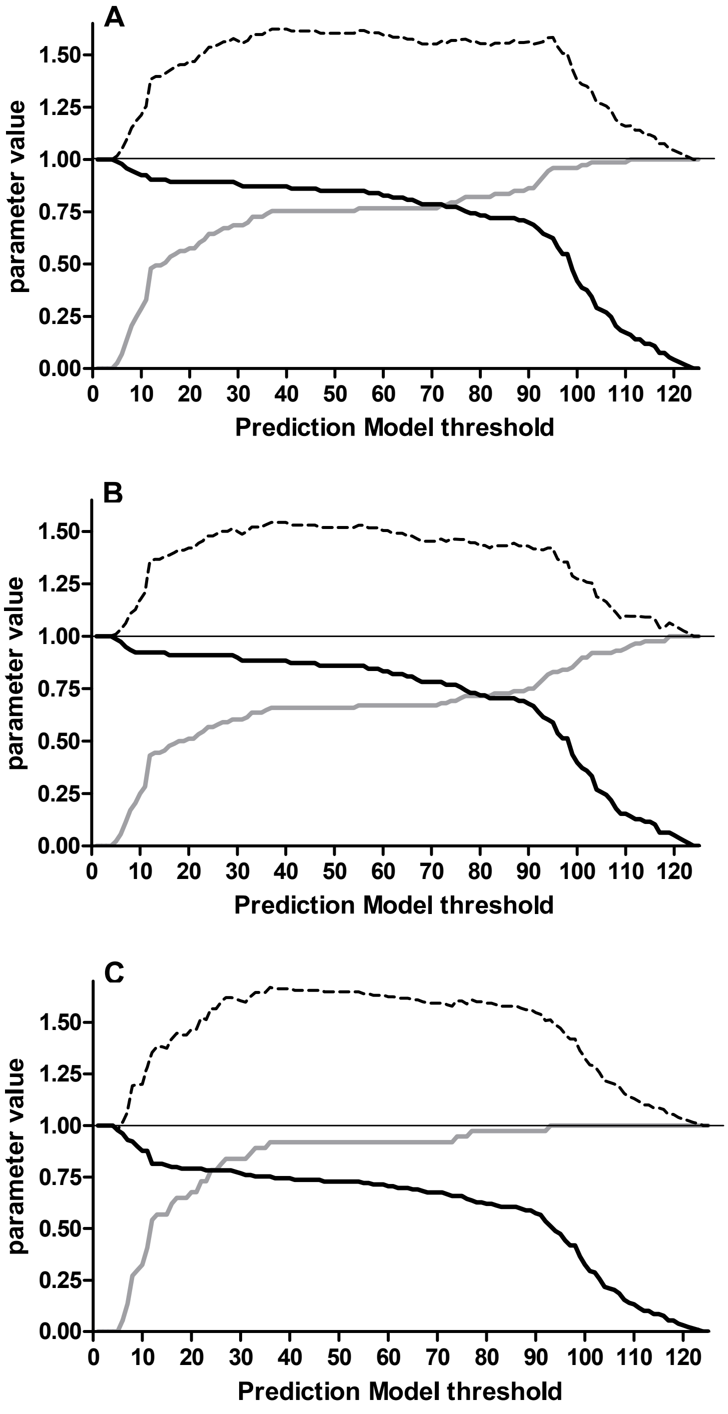


Figure 24: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold [%] when considering all valid L'Oréal-runs. A: Classification of the in vivo data according to the European classification system, i.e. a threshold of 2. B: In vivo classification threshold of 1.7. C: In vivo classification threshold of 2.3
(black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity).

Although moving the in vitro threshold has little influence, we also modeled to move the in vivo threshold. In the European system, this threshold is equivalent the dominant median score of 2. We performed the same analysis as done for this in vivo threshold of 2, when moving it downwards to 1.7 and upwards to 2.3. The results are shown in Figure 24B and C. As basically the shapes of all curves are similar to those in Figure 24A, EPISKIN predictive capacity at L'Oréal could not be improved when moving the in vivo classification threshold.

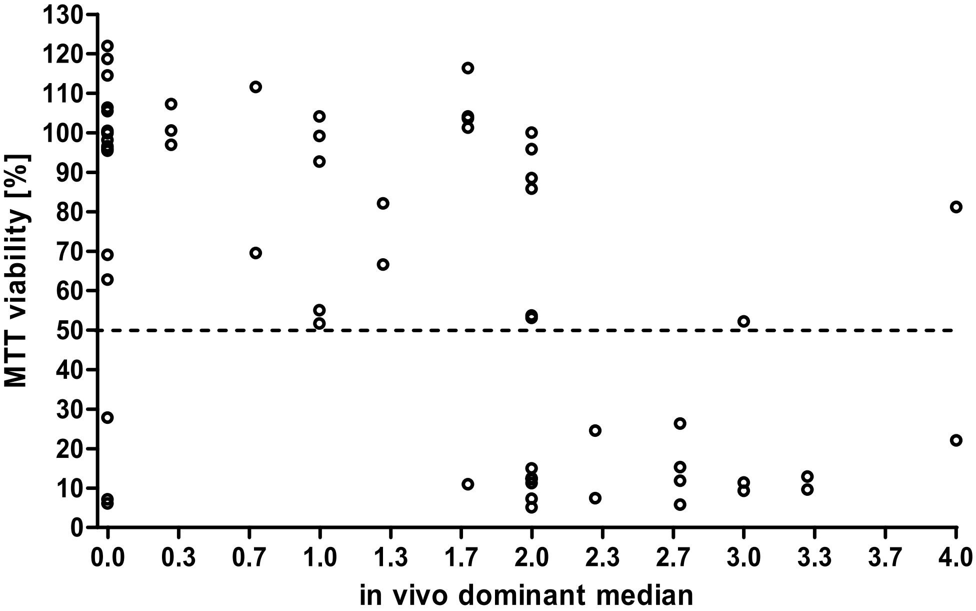


Figure 25: Correlation of the in vivo dominant median with the mean viability of all available runs from L'Oréal with EPISKIN, where the dotted line indicates the PM-threshold

Furthermore, the mean viability of all runs, included in Table 43, was plotted against the dominant median of the in vivo data, which is presented in Table 4. Figure 25 allowed a more detailed evaluation of the severity of the misclassification. For example, six of the eight false negative classified chemicals had a dominant median of 2.0, i.e. the classification threshold of the European classification system.

3.2.6.2 Unilever

In Table 48, the predictions are presented for all test chemicals. In Table 49, the parameters specificity and sensitivity calculated from Table 48 are presented for each single run and for the summarising approaches. Besides the sample sizes for both parameters, also the exact lower 5%-confidence bound are given.

chemical	EU	run				median approach	
number	classification	1	2	3	4	all runs	three valid runs
2	no label	**1**	**1**	**1**		**1**	**1**
5	no label	**1**	**1**	**1**		**1**	**1**
6	no label	**1**	0	0		0	0
7	no label	**1**	**1**	**1**		**1**	**1**
8	no label	**1**	**1**	**1**	**1**	**1**	**1**
9	no label	0	0	0	0	0	0
10	no label	0	0	0		0	0
11	no label	0	0	0		0	0
12	no label	0	0	0		0	0
16	no label	0	0	0		0	0
17	no label	**1**	**1**	**1**		**1**	**1**
19	no label	0	0	0		0	0
21	no label	0	0	0		0	0
22	no label	0	0	0		0	0
24	no label	0	0	0		0	0
25	no label	0	0	0		0	0
26	no label	**1**	**1**	**1**		**1**	**1**
28	no label	0	0	0		0	0
30	no label	0	0	0	0	0	0
32	no label	0	0	0		0	0
33	no label	0	0	0		0	0
35	no label	0	0	0		0	0
36	no label	0	0	0	0	0	0
39	no label	0	0	0		0	0
41	no label	0	0	0		0	0
42	no label	0	0	0	0	0	0
44	no label	0	0	0	0	0	0
48	no label	0	0	0		0	0
50	no label	0	0	0	0	0	0
52	no label	0	0	**1**	1	2 vs 2	0
53	no label	0	**1**	0		0	-
54	no label	1	0	0		0	0
57	no label	0	0	0		0	0
1	R38	1	1	1		1	1
3	R38	1	1	1	1	1	1
4	R38	1	1	1		1	1
13	R38	1	1	1		1	1
15	R38	1	1	1		1	1
18	R38	**0**	1	1	1	1	1
20	R38	1	1	1		1	1
23	R38	1	1	1	1	1	1
27	R38	**0**	1	1	1	1	1
29	R38	1	1	1		1	1
31	R38	1	1	1		1	1
34	R38	**0**	**0**	**0**	**0**	**0**	**0**
37	R38	1	1	1		1	1
40	R38	1	1	1		1	1
43	R38	1	1	1		1	1
45	R38	1	1	1		1	1
46	R38	1	1	1		1	1
47	R38	1	1	1		1	1
49	R38	**0**	**0**	**0**		**0**	**0**
51	R38	**0**	**0**	**0**		**0**	**0**
55	R38	1	1	1		1	1
56	R38	**0**	**0**	**0**		**0**	-
58	R38	1	1	1		1	1
59	R38	1	1	1		1	1
60	R38	1	1	1		1	1

Table 48: EPISKIN-Classification of the 58 chemicals according to the prediction model at Unilever
(0: non irritant (no label); 1: irritant (R38); bold: misclassifications; grey cells: not valid runs)

Regardless the way of analysis, the specificity was always around 80% except in the first run (75%). The sensitivity ranged, depending on the analysis, between 84.0 and 90.0% with the exception of the analysis of all runs in the first run. Less pronounced for the specificity, the analyses excluding data, which did not pass the variability quality criterion, resulted in higher sensitivities. Summarising the data in a conservative way, i.e. classifying a chemical as irritant when it was classified as irritant in at least one run, gave reduced specificities without affecting the sensitivities (data not shown).

		specificity			**sensitivity**		
		n	**%**	**LB-5%**	**n**	**%**	**LB-5%**
Run 1	all	33	75.8	60.5	25	76.0	76.9
	valid	31	74.2	58.2	20	90.0	71.7
Run 2	all	33	78.8	63.8	25	84.0	67.0
	valid	27	81.5	64.9	23	87.0	69.9
Run 3	all	33	78.8	63.8	25	84.0	67.0
	valid	30	80.0	64.3	24	87.5	70.8
all runs (median)		33	78.8	63.8	25	84.0	67.0
three valid runs (median)		32	81.3	66.3	24	87.5	70.8

Table 49: Predictive Capacity of EPISKIN at Unilever in terms of specificity and sensitivity with the respective sample sizes and 5% lower confidence bounds for each run and summaries over all runs

In order to investigate how the balance of specificity and sensitivity depend on the PM-threshold we submitted the Unilever data of all valid runs (n = 168) to a ROC-analysis. This analysis revealed a steep curve for sensitivity up to 90%, which then becomes flatter (Figure 26). This angular shape indicates that an optimal balance of the performance parameters is achieved at the angle and that rendering the test either more sensitive or specific can only be achieved by a substantial trade-off.

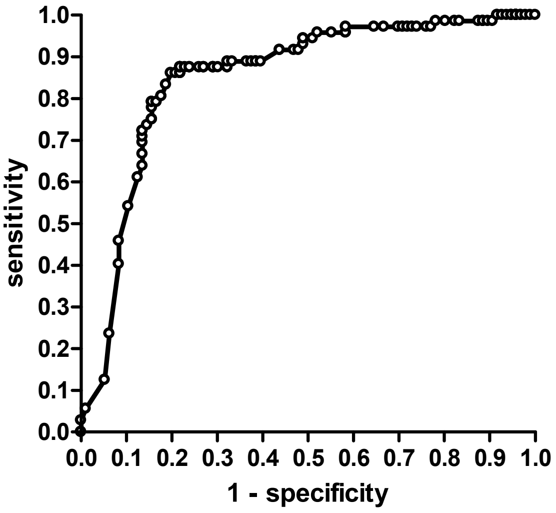


Figure 26: Receiver operation curve of all valid runs from Unilever for MTT

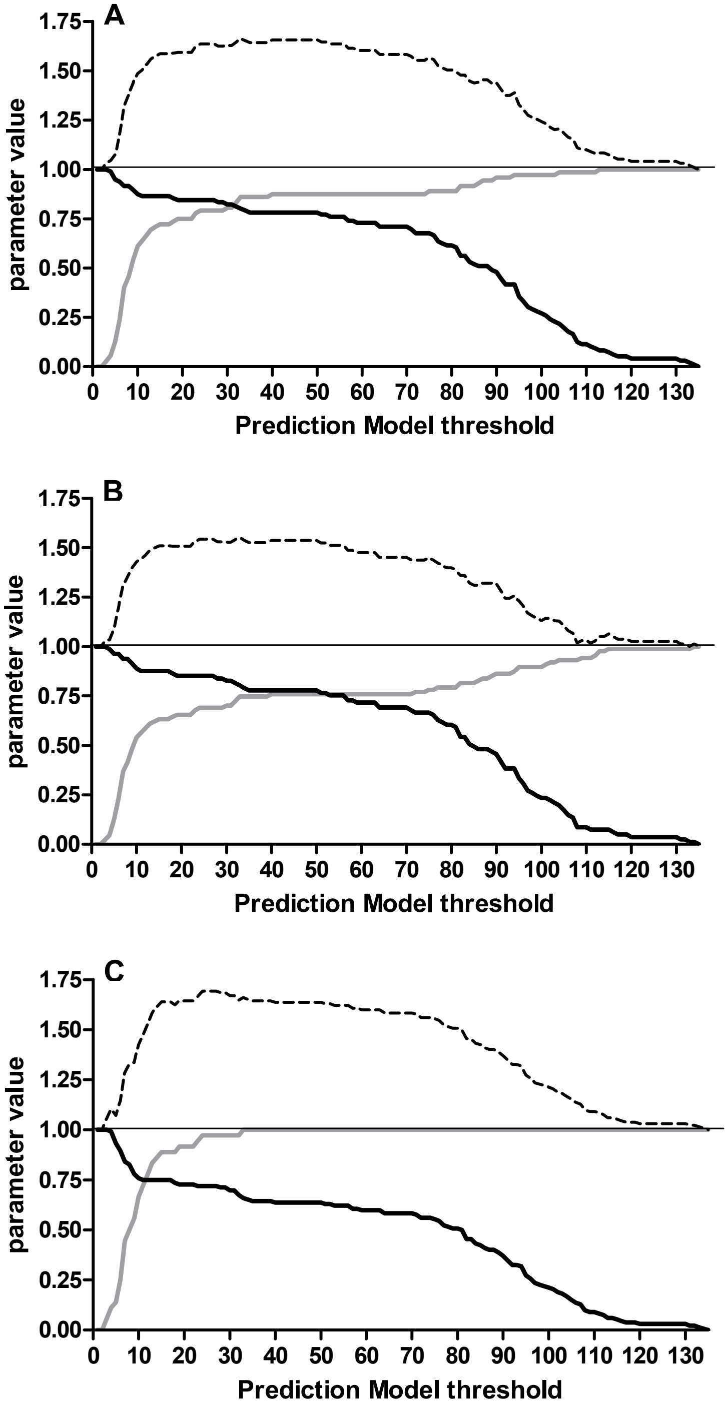


Figure 27: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold [%] when considering all valid Unilever-runs. A: Classification of the in vivo data according to the European classification system, i.e. a threshold of 2. B: In vivo classification threshold of 1.7. C: In vivo classification threshold of 2.3.
(black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity).

To get more insight into the predictive capacity and its threshold dependence, we plotted the sensitivity, the specificity and the sum of these, which allows optimising the PM-threshold choice when weighing sensitivity and specificity equally. In Figure 27A, which still considered the chemicals classifications according to the European system, it is obvious that the sum of sensitivity and specificity is larger than 1.60 over the area of thresholds ranging from about 26% to about 63% with a maximum of 1.66 at 33%. This insensitivity to threshold changes in the middle cell viability response range clearly reflects the optimization efforts with EPISKIN towards a two class system. Furthermore, it confirmed that the prediction model of 50% was an appropriate choice.

Although moving the in vitro threshold has had little influence, we also modeled to move the in vivo threshold. In the European system, this threshold is equivalent the dominant median score of 2. We performed the same analysis as done for this in vivo threshold of 2, when moving it downwards to 1.7 and upwards to 2.3. The results are shown in Figure 27B and C. As basically the shapes of all curves are similar to those in Figure 27A, EPISKIN predictive capacity at Unilever could not be improved when moving the in vivo classification threshold.

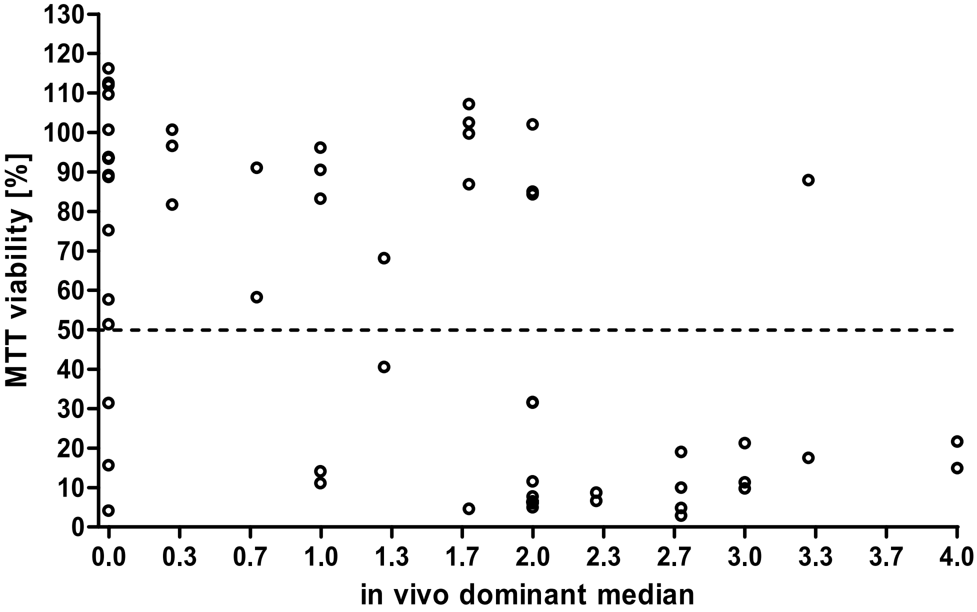


Figure 28: Correlation of the in vivo dominant median with the mean viability of all available runs from Unilever with EPISKIN, where the dotted line indicates the PM-threshold

Furthermore, the mean viability of all runs, included in Table 43, was plotted against the dominant median of the in vivo data, which is presented in Table 4. Figure 28 allowed a more detailed evaluation of the severity of the misclassification. For example, three of the four false negative classified chemicals had a dominant median of 2.0, i.e. the classification threshold of the European classification system.

3.2.6.3 Sanofi

In Table 50, the predictions are presented for all test chemicals. In Table 51, the parameters specificity and sensitivity calculated from Table 50 are presented for each single run and for the summarising approaches. Besides the sample sizes for both parameters, also the exact lower 5%-confidence bound are given.

chemical	EU	run				median approach	
number	classification	1	2	3	4	all runs	three valid runs
2	no label	**1**	**1**	**1**		**1**	**1**
5	no label	0	**1**	**1**	**1**	**1**	**1**
6	no label	0	0	**1**		0	-
7	no label	**1**	**1**	**1**		**1**	**1**
8	no label	**1**	**1**	**1**		**1**	**1**
9	no label	0	0	0		0	0
10	no label	0	0	0		0	0
11	no label	0	0	0		0	0
12	no label	0	0	0		0	0
16	no label	0	0	0		0	0
17	no label	**1**	**1**	**1**		**1**	**1**
19	no label	0	0	0	0	0	0
21	no label	0	0	0		0	0
22	no label	0	0	0		0	0
24	no label	0	0	0	0	0	0
25	no label	0	0	0		0	0
26	no label	**1**	**1**	**1**		**1**	**1**
28	no label	0	0	0		0	0
30	no label	0	0	0		0	0
32	no label	0	0	0		0	0
33	no label	0	0	0		0	0
35	no label	0	0	0		0	0
36	no label	0	0	0	0	0	0
39	no label	0	0	0		0	0
41	no label	0	0	0		0	0
42	no label	0	0	0		0	0
44	no label	0	0	0		0	0
48	no label	0	0	0		0	0
50	no label	0	0	0		0	0
52	no label	0	0	0		0	0
53	no label	**1**	0	0		0	0
54	no label	**1**	0	0		0	0
57	no label	0	0	0		0	0
1	R38	1	1	1		1	1
3	R38	**0**	1	1		1	1
4	R38	1	1	1		1	1
13	R38	1	1	1		1	1
15	R38	1	1	1		1	1
18	R38	1	1	1		1	-
20	R38	1	**0**	**0**	1	1	-
23	R38	**0**	**0**	**0**	**0**	**0**	**0**
27	R38	**0**	**0**	**0**		**0**	**0**
29	R38	1	1	1		1	1
31	R38	1	1	1		1	1
34	R38	**0**	**0**	**0**		**0**	**0**
37	R38	1	1	1		1	1
40	R38	1	1	1	1	1	1
43	R38	**0**	1	**0**	1	1	1
45	R38	1	1	1		1	1
46	R38	1	1	1		1	1
47	R38	1	1	1		1	1
49	R38	**0**	**0**	**0**		**0**	**0**
51	R38	**0**	**0**	**0**		**0**	**0**
55	R38	1	**0**	**0**		**0**	-
56	R38	1	1	1		1	1
58	R38	1	1	1		1	1
59	R38	1	1	1		1	1
60	R38	**0**	**0**	**0**		**0**	**0**

Table 50: EPISKIN-Classification of the 58 chemicals according to the prediction model at Sanofi
(0: non irritant (no label); 1: irritant (R38); bold: misclassifications; grey cells: not valid runs)

		specificity			sensitivity		
		n	%	LB-5%	n	%	LB-5%
Run 1	all	33	78.8	63.8	25	68.0	49.6
	valid	29	75.9	59.4	21	66.7	46.4
Run 2	all	33	81.8	67.2	25	68.0	49.6
	valid	31	80.6	65.3	21	76.2	56.3
Run 3	all	33	78.8	63.8	25	64.0	45.6
	valid	32	81.3	66.3	23	65.2	46.0
all runs (median)		33	81.8	67.2	25	68.0	49.6
three valid runs (median)		32	81.3	66.3	22	72.7	53.2

Table 51: Predictive Capacity of EPISKIN at Sanofi in terms of specificity and sensitivity with the respective sample sizes and 5% lower confidence bounds for each run and summaries over all runs

Regardless the way of analysis, the specificity was always around 80% except in the first run, when considering only the valid runs (75%). The sensitivity ranged, depending on the analysis, between 64.0% and 76.2%. Taking the valid runs into account resulted only for the sensitivity in the overall analyses in higher values, i.e. 68% for all runs and 73.9% for all valid runs. Summarising the data in a conservative way, i.e. classifying a chemical as irritant when it was classified as irritant in at least one run, gave reduced specificities without affecting the sensitivities (data not shown).

To investigate how the balance of specificity and sensitivity depend on the PM-threshold we submitted the Sanofi data of all valid runs (n = 164) to a ROC-analysis (Figure 29). This analysis revealed a steep curve for sensitivity up to 75%, which then becomes flatter. This angular shape indicates that an optimal balance of the performance parameters is achieved at the angle and that rendering the test either more sensitive or specific can only be achieved by a substantial trade-off.

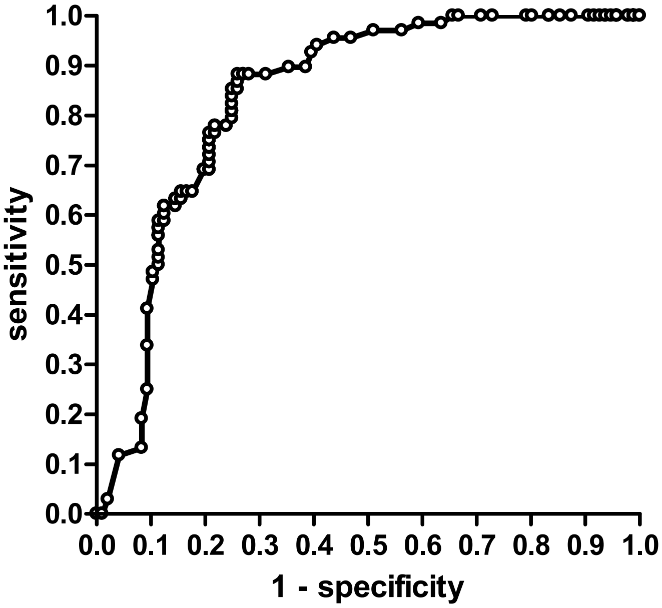


Figure 29: Receiver operation curve of all valid runs from Sanofi

To get more insight into the predictive capacity and its threshold dependence, we plotted the sensitivity, the specificity and the sum of these, which allows optimising the PM-threshold choice when weighing sensitivity and specificity equally.

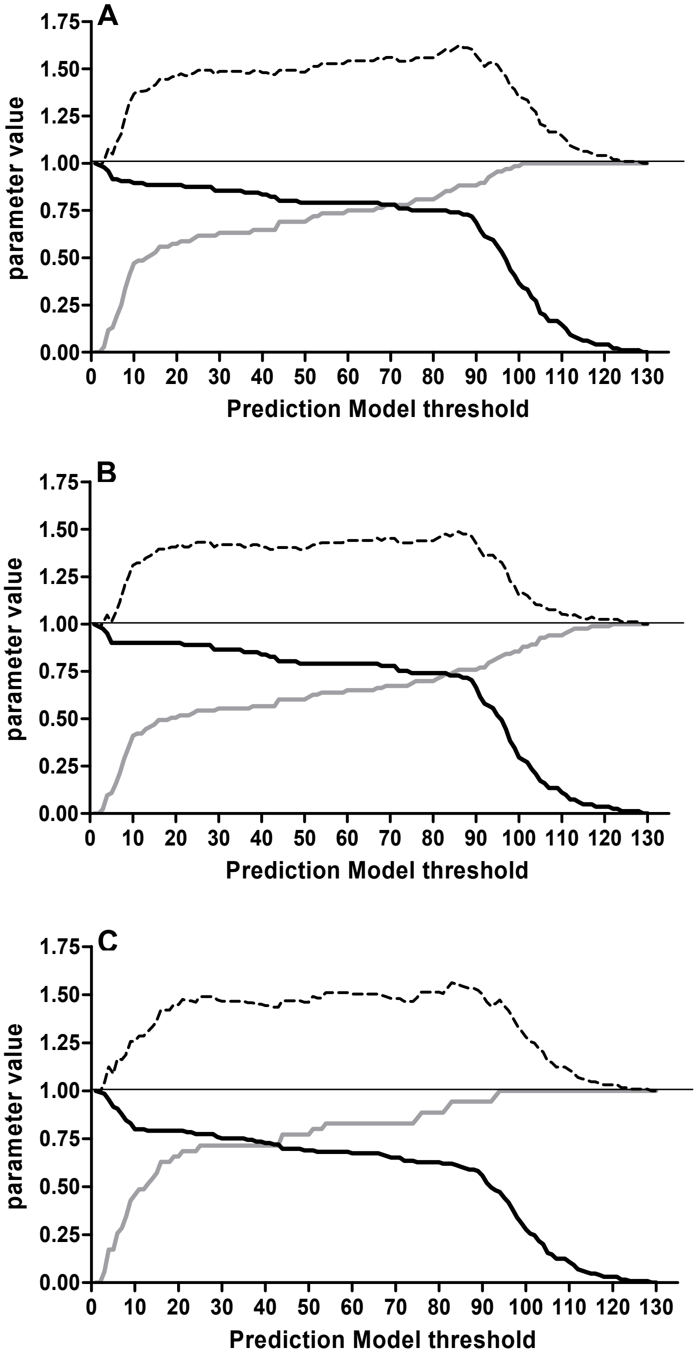


Figure 30: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold [%] when considering all valid Sanofi-runs. A: Classification of the in vivo data according to the European classification system, i.e. a threshold of 2. B: In vivo classification threshold of 1.7. C: In vivo classification threshold of 2.3.
(black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity).

In Figure 30A that still considers the chemicals classifications according to the European system a steadily increasing sum until a threshold of about 85% can be seen. However, already at a threshold of 25% a sum of 1.5 is approached. Considering the small peak at a maximum of 1.6 as negligible, the insensitivity to threshold changes in the middle cell viability response range clearly reflects the optimization efforts with EPISKIN towards a two class system. Furthermore, it confirmed that the prediction model of 50% was an appropriate choice.

Although moving the in vitro threshold has little influence, we also modeled to move the in vivo threshold. In the European system, this threshold is equivalent the dominant median score of 2. We performed the same analysis as done for this in vivo threshold of 2, when moving it downwards to 1.7 and upwards to 2.3. The results are shown in Figure 30B and C. As basically the shapes of all curves are similar to or lower than those in Figure 30A, EpiDerm predictive capacity is could not be improved when moving the in vivo classification threshold.

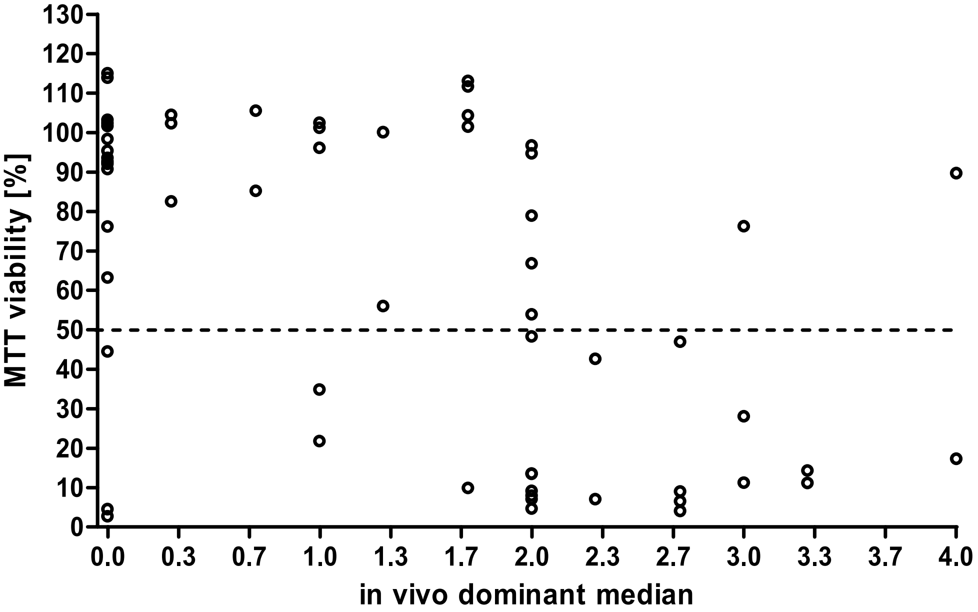


Figure 31: Correlation of the in vivo dominant median with the mean viability of all available runs from Sanofi with EPISKIN, where the dotted line indicates the PM-threshold.

Furthermore, the mean viability of all runs, included in Table 43, was plotted against the dominant median of the in vivo data, which is included in Table 4. Figure 31 allowed a more detailed evaluation of the severity of the misclassification. For example, five of the seven false negative classified chemicals had a dominant median of 2.0, i.e. the classification threshold of the European classification system.

3.2.6.4 Misclassifications

To compare the misclassified chemicals for the three EPISKIN-laboratories, those chemicals, which were misclassified at least once in one of the laboratories, are summarised in Table 52. While ten of the 33 not labelled chemicals were misclassified at least once, twelve of the 25 R38-chemicals had at least one misclassification. Three non-irritants were consistently classified as irritant in all runs and all laboratories. Three irritants were classified as non-irritants in all runs and all laboratories.

chem. no	EU class	dominant median	L'Oréal run				Unilever run				Sanofi run				total number of runs	misclassifying runs [%]
			1	2	3	4	1	2	3	4	1	2	3	4		
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	**1**	**1**	0		**1**	**1**	**1**	**1**	**1**	**1**	**1**		10	90.00
5	no label	1	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**		10	100.00
49	R38	2	**0**	**0**	**0**		**0**	**0**	**0**		**0**	**0**	**0**		9	100.00
51	R38	2	**0**	**0**	**0**		**0**	**0**	**0**		**0**	**0**	**0**		9	100.00
27	R38	4	**0**	**0**	**0**		**0**	1	1	1	**0**	**0**	**0**		10	70.00
60	R38	2	**0**	**0**	**0**		1	1	1		**0**	**0**	**0**		9	66.67
23	R38	3	**0**	**0**	**0**	1	1	1	1	1	**0**	**0**	**0**	**0**	12	58.33
55	R38	2	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**	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	1	1	1		**0**	1	1	1	1	1	1		10	10.00

Table 52: Summary of chemicals, which were misclassified at least once in one of the EPISKIN-laboratories
(bold type: misclassified runs; grey cells: SD > 18)

3.2.6.5 Summary predictive capacity results EPISKIN

To summarise, first the specificity and sensitivity over all runs per laboratory from Tables 47, 49 and 51 according to the two approaches of analysis are presented in Table 53.

laboratory	specificity [%]		sensitivity [%]	
	all runs	three valid runs	all runs	three valid runs
L’Oréal	81.8	83.9	72.0	75.0
Unilever	78.8	81.3	84.0	87.5
Sanofi	81.8	81.3	68.0	72.7

Table 53: Summary of the predictive capacity (specificity and sensitivity) in the three EPISKIN laboratories considering either all chemical runs or only those chemicals, which had three valid (SD<18) runs

Second, from the overall runs in all laboratories we calculated the sample sizes and specificity and sensitivity for the two different approaches, either considering all individual classifications or the median classification for a given chemical (Table 54). Taking all individual classifications (n = 547) into account resulted in the lowest estimation for both parameters indicating that a large replicate variability (SD > 18) increased the chance of misclassification. Considering only those chemicals, for which three valid runs were available (n = 495), a specificity of 81% and a sensitivity of 78% was achieved. Slightly increased parameter estimations and a similar pattern were obtained when summarising the median classifications. The third approach taking all valid runs into account gave similar performance parameters (data not shown). Due to the strong dependencies between the data in terms of reproducibility, confidence bounds were not calculated. Considering only those chemicals, which had three valid runs in all three laboratories, i.e. 30 non-labelled and 21 labelled (R38) chemicals, resulted in a specificity of 83.7% and a sensitivity of 73.0% (data not shown).

	specificity		sensitivity	
	n	%	n	%
all runs (individually classification)	311	79.74	236	73.73
three valid runs (individually classification)	285	80.70	210	77.62
all runs (median classification)	99	80.80	75	74.67
three valid runs (median classification)	95	82.15	70	78.56

Table 54: Summary of EPISKIN specificity and sensitivity considering three different approaches

Summarizing the receiver operation curve of the three laboratories, which were based on all valid runs, Figure 32 shows that the overall RO-curve averaged the three laboratory RO-curves. The dotted square indicates the area, where the individual curves differ most: Unilever performed better, where Sanofi predictive capacity was reduced. However, these differences are considered as minor and, if at all, should be discussed taking the specific problems encountered in the laboratories into account.

Finally, in Figure 33, the sensitivity, specificity and their sum are displayed. In the threshold range between 24% and 77% both the sensitivity and specificity curves were almost flat, where the sum of both remained approximately constant, i.e. larger than 1.55. The maximum sum of 1.593 is reached at 55% close to the predefined prediction model threshold of 50%. Similarly as in Figures 32 and 33, the resulting curves when moving the in vivo threshold to 1.7 or 2.3 represent an average (data not shown).

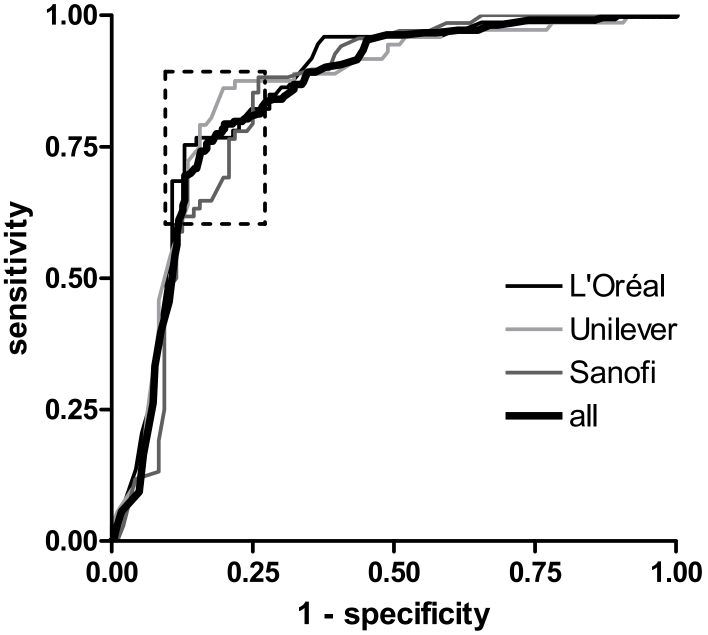


Figure 32: Receiver operation curves of all valid runs of all EPISKIN-laboratories.

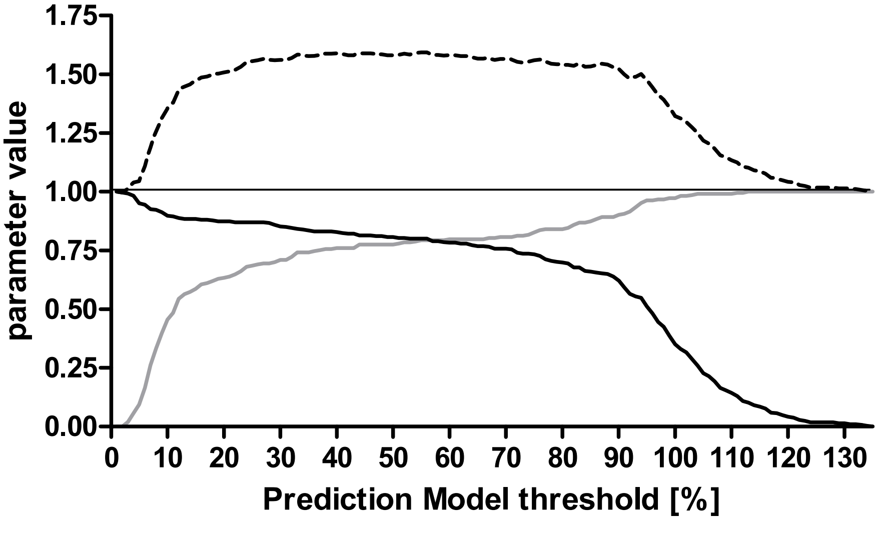


Figure 33: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold [%] when considering all valid EPISKIN-runs (black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity)

The negative and positive predictive values, which incorporate specificity and sensitivity as well as prevalence, i.e. the proportion of irritating chemicals in a defined population of chemicals, can be found in Annex VIII.

Although the B- and C-curves in Figures 24, 27 and 30 already suggest that it will be impossible to find a satisfactorily performing PM for the three GHS classes, performance of EPISKIN to predict the three classes of the GHS was analysed to confirm this expectation. To keep this analysis simple and disregarding reproducibility, only the median run classification of chemical with three valid runs for all laboratories were considered. This resulted in a dataset with a sample size of 165, where 35 entries were GHS-irritants, 50 GHS-mild irritants and 80 GHS-non irritants. As in Phase I (Annex I) no satisfactory PM could be identified, a post-hoc approach to construct a new PM was chosen. Therefore, the two thresholds of viability maximizing the sum of sensitivity and specificity in the ROC-analyses of discriminating GHS-non irritants from GHS-mild irritants and GHS irritants and of discriminating GHS-non and mild irritants from GHS irritants, respectively, were used. As here the two optimal thresholds were almost identical – by itself a strong indication confirming the expectation – one threshold was moved to the next highest value. The respective PM consisted of the threshold of 30% viability, below which chemicals would be classified as GHS-irritants, and of 50%, above which chemicals would be classified as GHS-non irritants, was constructed. Chemicals with viabilities between these two thresholds would be classified as GHS-mild irritants. Applying the PM resulted in the correct classification of 88.6% GHS-irritants, 6.0% GHS-mild irritants and 88.6% GHS-non irritants. It has to be noted that only eight entries had viability between 30% and 50%. This confirms the results of Phase I that EpiDerm is not able to predict the three GHS-classes. The fact that GHS-mild irritants were either giving high or low viabilities reflects that the EpiDerm protocol was optimized 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 when discriminating GHS-non irritants from the other two classes.

4 Results IL1-α

4.1 EpiDerm: ZEBET

4.1.1 Data submission

As IL1-α was established as a possible endpoint for skin irritation only in a later stage of test development, it was agreed to evaluate this second endpoint in a first step in the leading laboratories only. Because a highly irritating property of given substance might interact with IL1-α, the MT decided to measure this endpoint only for MTT-non irritant substances. Therefore the mean viability over all runs was considered. Additionally, chemicals, which showed high variability in the MTT-test (SD > 18), were included. ZEBET, the leading laboratory of EpiDerm, submitted the data on the 05.06.2006 for 45 chemicals. Forty-four of these triggered the IL1- by at least one of the agreed criteria, two of which were the confidential chemicals. One chemical was tested although it was reproducibly an MTT-irritant (mean viability of 43%) and it is not included here resulting in a total of 42 chemicals. The tests were performed between the 24.05.05 and the 04.06.05 with the stored supernatants from the MTT-tests, i.e. at least three runs each, by one operator, who also performed all MTT-tests. The two confidential and the additionally tested chemical(s) will be excluded from all analysis. The data processed in the following were the calculated as IL1-α in pg/ml and are shown in Figure 34. The chemicals were divided into four sub-sets.

Besides all positive controls, several chemicals (numbers 6, 7, 13, 15, 25, 44, 49, 54), had to be tested diluted. From the IL1-α content of the dilution the respective value of the undiluted samples was calculated. These calculated values are used in this report.

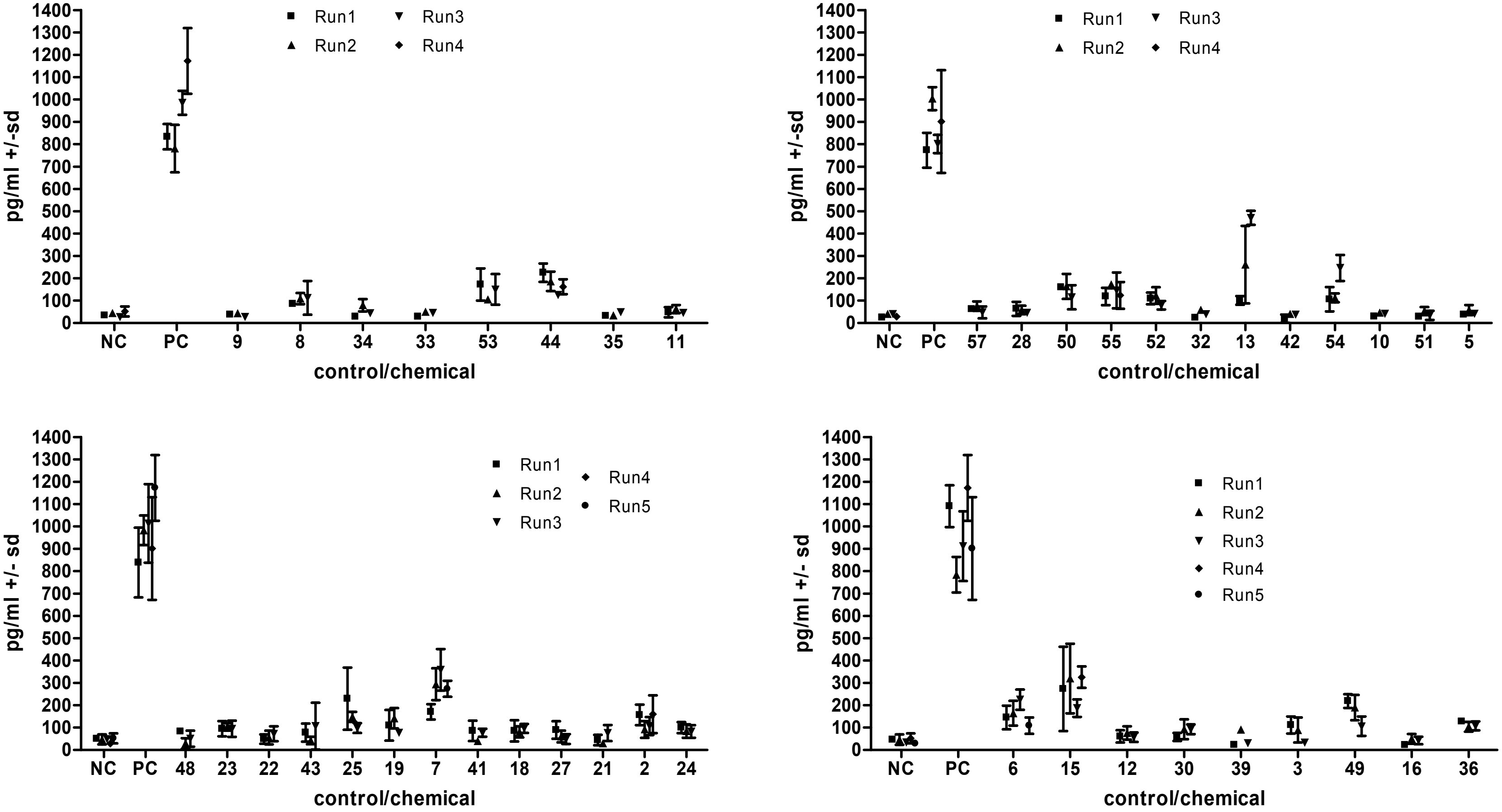


Figure 34: EpiDerm-ZEBET IL1-α data of all four sets of chemicals tested including the controls (NC: negative control; PC: positive control) expressed as mean and standard deviation

4.1.2 Intra-assay variability

Per run three replicates were measured. In order to assess the intra-assay variability, i.e. the variability between these replicates, the respective standard deviation and the coefficient of variation (CV) were calculated. To avoid graphical distortion, for the diluted samples and positives controls the respective values measure in the dilution were used. Regardless the measure, substantial intra-assay was observed (Figure 35). Both measures of variability depended on the response level, while the standard deviation increased in the same way the CV decreases with increasing response levels.

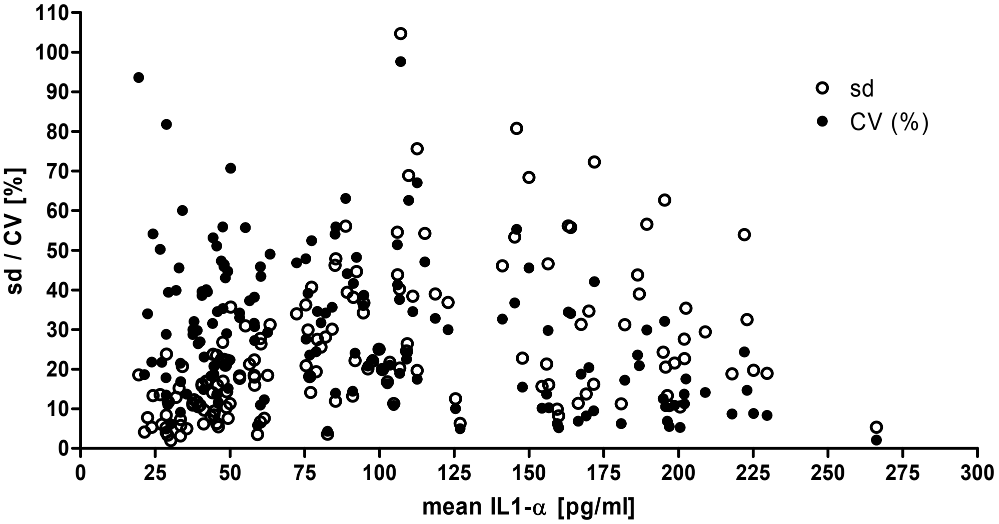


Figure 35: ZEBET IL1-α intra-assay variability expressed as standard deviation and coefficient of variation for all runs of all chemicals EpiDerm

4.1.3 Within-laboratory variability

The negative control gave in average a response of 40 pg/ml with a minimum of 24 and a maximum of 50 pg/ml (CV: 20.6%). The positive control induced a mean of 919 pg/ml IL1-α ranging from 773 to 1172 pg/ml.

The variability within the laboratory, i.e. between independent runs, was expressed by the standard deviation s_R and the coefficient of variation (Table 55). Regarding these descriptive measures, the CV is to be preferred as it is less dependent on the response range than the standard deviation. Furthermore, a 1-way ANOVA was calculated for each chemical with the raw and the logarithmically (natural) transformed data.

The mean variability in terms of CV between runs was 30.19% with a standard deviation of 15.46. When applying the ANOVA, no substantial differences between the two analyses were found. In total, two chemicals (numbers 15, 39) had p-values smaller than 0.01. Here again, the ANOVA was not appropriate for the assessment of within-laboratory reproducibility because of substantial variability within the runs. As the preliminary prediction model was based on the fold-increase of IL1-α release induced by a chemical in comparison to the respective negative control, the variability of this measure is included in Table 56. In this context, the CV is to be preferred, as it did not reveal any substantial response dependency (data not shown).

chemical number	chemical class (EU)	mean [pg/ml]			mean	s_R	CV [%]	1-way ANOVA	
		run 1	run 2	run 3				raw data	ln-transf.
2	no label	156.56	91.48	159.68	135.91	38.51	28.33	0.3530	0.3819
3	R38	111.37	88.93	33.14	77.81	40.28	51.77	0.1234	0.0559
5	no label	37.76	58.37	41.58	45.90	10.96	23.89	0.2941	0.3041
6	no label	145.43	164.19	109.07	139.56	28.02	20.08	0.4301	0.4120
7	no label	170.20	293.98	358.56	274.25	95.72	34.90	0.0214	0.0180
8	no label	85.48	109.22	112.79	102.50	14.84	14.48	0.7460	0.7658
9	no label	37.73	43.83	27.41	36.32	8.30	22.85	0.1320	0.1240
10	no label	29.40	46.02	41.39	38.94	8.58	22.03	0.1896	0.1627
11	no label	48.27	62.82	45.72	52.27	9.23	17.65	0.4625	0.4582
12	no label	60.39	76.29	56.74	64.47	10.39	16.12	0.6492	0.6691
13	R38	101.19	261.00	470.35	277.51	185.13	66.71	0.0129	**0.0076**
15	R38	272.81	319.81	187.14	259.92	67.27	25.88	0.5488	0.5523
16	no label	21.64	49.51	42.72	37.96	14.53	38.29	0.1695	0.0885
18	R38	85.61	80.63	96.32	87.52	8.02	9.16	0.8443	0.7842
19	no label	110.00	141.33	77.20	109.51	32.07	29.28	0.3357	0.3747
21	no label	44.52	29.60	75.60	49.91	23.47	47.02	0.1635	0.2078
22	no label	48.69	55.39	72.53	58.87	12.30	20.89	0.6098	0.7096
23	R38	94.81	102.86	94.92	97.53	4.62	4.73	0.9339	0.8760
24	no label	100.00	84.41	82.21	88.87	9.70	10.91	0.7065	0.6629
25	no label	228.94	148.00	97.90	158.28	66.12	41.78	0.2253	0.2449
27	R38	89.35	60.57	48.22	66.05	21.10	31.95	0.3009	0.2635
28	no label	63.64	58.31	46.29	56.08	8.89	15.85	0.6117	0.6382
30	no label	58.42	92.56	92.02	81.00	19.56	24.14	0.3554	0.2641
32	no label	24.01	58.45	39.51	40.66	17.25	42.43	0.0274	0.0135
33	no label	28.88	49.57	44.80	41.08	10.83	26.37	0.0446	0.0353
34	R38	28.91	79.47	44.38	50.92	25.91	50.88	0.0336	0.0247
35	no label	32.18	33.61	48.98	38.26	9.31	24.35	0.1369	0.1811
36	no label	127.17	103.64	106.95	112.59	12.74	11.31	0.2765	0.3102
39	no label	22.65	91.20	30.38	48.08	37.55	78.09	**0.0001**	**0.0003**
41	no label	85.37	39.21	76.89	67.16	24.57	36.59	0.2029	0.1013
42	no label	19.70	40.83	38.20	32.91	11.52	34.99	0.2776	0.1926
43	R38	77.53	42.28	107.28	75.70	32.54	42.99	0.5158	0.4379
44	no label	225.09	186.53	125.63	179.08	50.15	28.00	0.0366	0.0181
48	no label	82.80	28.97	50.41	54.06	27.10	50.13	0.0939	0.1079
49	R38	218.83	189.62	106.26	171.57	58.42	34.05	0.0513	0.0456
50	no label	159.68	163.38	115.39	146.15	26.70	18.27	0.4088	0.3562
51	R38	28.81	53.42	34.28	38.84	12.92	33.27	0.2170	0.3195
52	no label	109.49	123.14	79.00	103.88	22.60	21.76	0.2282	0.1716
53	no label	172.07	104.98	150.21	142.42	34.22	24.03	0.4049	0.3342
54	no label	106.20	112.84	246.52	155.19	79.17	51.01	0.0184	0.0326
55	R38	118.86	171.83	146.05	145.58	26.49	18.19	0.5062	0.4194
57	no label	61.70	75.64	47.77	61.70	13.94	22.58	0.3039	0.2483

Table 55: ZEBET Within-laboratory variability of IL1-a of EpiDerm (grey cells indicate runs with MTT-variability of SD >18)

Considering the predictions, which are based on the threshold value of 3-fold increase, eight (numbers 2, 8, 25, 36, 50, 52, 53, 54) of 42 chemicals were not reproducible.

As for the MTT-endpoint, a comparison of data of chemicals tested in both phases was carried out providing further information on the within-laboratory reproducibility. The mean IL1-α amount over the runs for the controls and the overlapping eleven chemicals are summarised in Table 56, whereas the mean individual run data are presented in Figure 36. The differences showed that the data of the two phases were very similar without any obvious trend. The larger difference for the positive control (PC) is acceptable for the high response level of about 1000 pg/ml. Indeed, a paired t-test excluding/including the positive control resulted in a non-significant p-value of 0.841/0.397.

chemical number	mean IL1-α [pg/ml] Phase 1	Phase 2	difference
NC	33.16	40.23	-7.08
PC	1110.42	919.32	191.09
9	37.16	36.32	0.84
12	61.33	64.47	-3.14
13	264.61	277.51	-12.90
16	34.21	37.96	-3.75
28	64.88	56.08	8.80
30	49.95	81.00	-31.05
32	42.47	40.66	1.81
35	41.66	38.26	3.40
42	57.61	32.91	24.70
49	140.81	171.57	-30.76
51	71.98	38.84	33.15
52	106.48	103.88	2.60

Table 56: Comparison of mean IL1-α amount in pg/ml of controls and twelve chemicals tested in both phases at ZEBET with EpiDerm

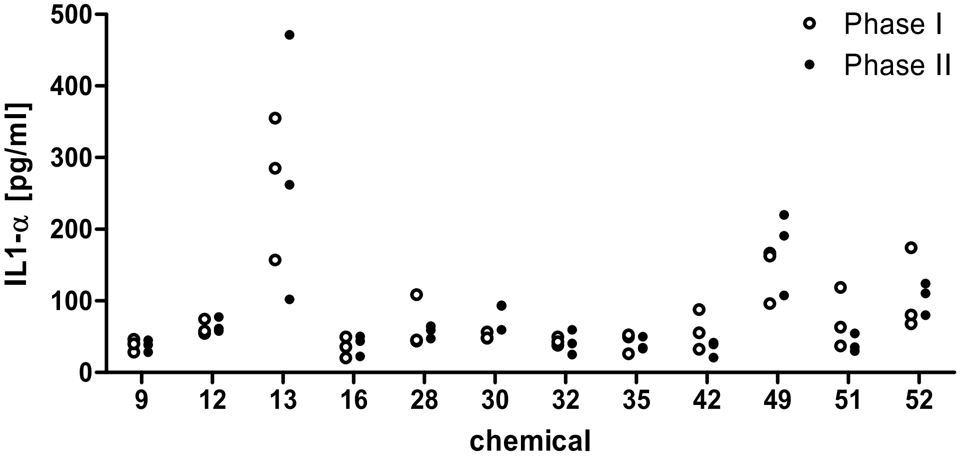


Figure 36: IL1-α amount in pg/ml of the individual runs of the twelve chemicals tested in both phases at ZEBET with EpiDerm

4.1.4 Predictive capacity

As IL1-α was considered from the very beginning in a strategic manner, 16 MTT-positive chemicals were not tested for this second endpoint. Of these 16, 14 were correctly and two wrongly classified as positives. Of the remaining 42 chemicals, eleven chemicals had a label (R38) and 31 had not. They included the four chemicals not having three acceptable MTT-runs (numbers 7, 15, 44, 55), two of which were irritants and two non-irritants.

chemical number	chemical class (EU)	fold increase			mean	sd	CV [%]	ANOVA p-value (log data)		
		1. run	2. run	3. run				1. run	2. run	3. run
2	no label	**3.12**	1.93	**5.52**	**3.52**	1.83	51.90	**< 0.05**	> 0.05	**< 0.05**
3	R38	2.43	1.94	0.99	1.79	0.73	40.98	**< 0.01**	> 0.05	> 0.05
5	no label	1.55	1.43	1.03	1.34	0.27	20.37	> 0.05	> 0.05	> 0.05
6	no label	**3.17**	**3.59**	**3.77**	**3.51**	0.31	8.77	**< 0.01**	**< 0.01**	**< 0.01**
7	no label	**3.39**	**6.22**	**7.30**	**5.64**	2.02	35.82	**< 0.05**	**< 0.01**	**< 0.01**
8	no label	2.53	2.44	**4.20**	**3.06**	0.99	32.43	**< 0.01**	**< 0.01**	**< 0.01**
9	no label	1.12	0.98	1.02	1.04	0.07	6.93	> 0.05	> 0.05	> 0.05
10	no label	1.20	1.13	1.03	1.12	0.09	7.63	> 0.05	> 0.05	> 0.05
11	no label	1.43	1.40	1.70	1.51	0.17	10.94	> 0.05	> 0.05	> 0.05
12	no label	1.32	1.67	1.69	1.56	0.21	13.34	> 0.05	> 0.05	> 0.05
13	R38	**4.15**	**6.39**	**11.7**	**7.41**	3.88	52.31	**< 0.01**	**< 0.05**	**< 0.01**
15	R38	**5.95**	**6.99**	**5.56**	**6.17**	0.74	11.99	**< 0.05**	**< 0.05**	**< 0.01**
16	no label	0.47	1.08	1.27	0.94	0.42	44.47	< 0.05	> 0.05	> 0.05
18	R38	1.70	1.70	1.96	1.79	0.15	8.40	> 0.05	> 0.05	> 0.05
19	no label	2.19	2.99	1.57	2.25	0.71	31.64	> 0.05	**< 0.05**	> 0.05
21	no label	0.89	0.63	1.54	1.02	0.47	45.95	> 0.05	> 0.05	> 0.05
22	no label	0.97	1.17	1.48	1.21	0.26	21.30	> 0.05	> 0.05	> 0.05
23	R38	1.89	2.17	1.93	2.00	0.15	7.58	> 0.05	> 0.05	> 0.05
24	no label	1.99	1.78	1.67	1.81	0.16	8.97	> 0.05	> 0.05	> 0.05
25	no label	**4.56**	**3.13**	1.99	**3.23**	1.29	39.91	**< 0.05**	**< 0.05**	> 0.05
27	R38	1.78	1.28	0.98	1.35	0.40	30.01	> 0.05	> 0.05	> 0.05
28	no label	2.61	1.43	1.15	1.73	0.77	44.79	**< 0.05**	> 0.05	> 0.05
30	no label	1.27	2.02	2.74	2.01	0.74	36.57	> 0.05	> 0.05	**< 0.01**
32	no label	0.98	1.43	0.98	1.13	0.26	22.99	> 0.05	> 0.05	> 0.05
33	no label	0.86	1.11	1.67	1.21	0.41	34.18	> 0.05	> 0.05	> 0.05
34	R38	0.86	1.77	1.65	1.43	0.49	34.65	> 0.05	**< 0.05**	> 0.05
35	no label	0.95	0.75	1.82	1.17	0.57	48.48	> 0.05	> 0.05	> 0.05
36	no label	2.77	2.26	**3.18**	2.74	0.46	16.84	**< 0.01**	> 0.05	**< 0.01**
37	no label	**5.02**	**5.03**	**6.74**	**5.60**	0.99	17.69	**< 0.01**	**< 0.01**	**< 0.01**
39	no label	0.49	1.99	0.90	1.13	0.78	68.81	< 0.05	> 0.05	> 0.05
41	no label	1.70	0.83	1.57	1.37	0.47	34.34	> 0.05	> 0.05	> 0.05
42	R38	0.81	1.00	0.95	0.92	0.10	10.71	> 0.05	> 0.05	> 0.05
43	no label	1.54	0.89	2.19	1.54	0.65	42.21	> 0.05	> 0.05	> 0.05
44	no label	**6.67**	**4.17**	**4.68**	**5.17**	1.32	25.53	**< 0.01**	**< 0.01**	**< 0.01**
48	R38	1.65	0.61	1.03	1.10	0.52	47.71	> 0.05	> 0.05	> 0.05
49	no label	**4.77**	**4.14**	**3.16**	**4.02**	0.81	20.17	**< 0.01**	**< 0.01**	**< 0.01**
50	R38	**6.54**	**4.00**	2.87	**4.47**	1.88	42.05	**< 0.01**	**< 0.01**	**< 0.05**
51	no label	1.18	1.31	0.85	1.11	0.24	21.30	> 0.05	> 0.05	> 0.05
52	no label	**4.49**	**3.02**	1.97	**3.16**	1.27	40.06	**< 0.01**	**< 0.01**	> 0.05
53	no label	**5.10**	2.34	**5.59**	**4.34**	1.75	40.34	**< 0.01**	**< 0.01**	**< 0.01**
54	R38	**4.35**	2.76	**6.13**	**4.41**	1.69	38.20	**< 0.01**	**< 0.01**	**< 0.01**
55	no label	**4.87**	**4.21**	**3.63**	**4.24**	0.62	14.64	**< 0.01**	**< 0.01**	**< 0.01**
57	no label	2.53	1.85	1.19	1.86	0.67	36.09	**< 0.05**	> 0.05	> 0.05

Table 57: Summary of IL1-α data of the tested chemicals and positive controls expressed as fold-increase relative to the respective negative controls together with 1-way ANOVA results of comparing the logarithmically transformed pg/ml-data with the respective negative control data using Dunnett's post test.

In Table 57, the data for two kinds of PM are summarised: the PM, which was based on the relative fold-increase compared to the negative control, and the results, i.e. p-values, of 1-way ANOVAs with a Dunnett post test comparing a given chemical's response to the response of the respective negative control. The later was applied to the raw data (data not shown) as well as to the log-transformed data. Although the differences were minor, the analysis of the log-transformed data is to be preferred as the variances were more homogenous over the response range. Simplifying the interpretation, the results over the runs were combined: A chemical was overall classified as irritant if the mean fold increase is larger than 3, or if a chemical gave in at least two runs a significantly ($p<0.05$) higher response than the negative control (log-transformed data). As the overall classification are identical for both prediction models (with the exception of chemical 36), the resulting predictive capacities are only presented for the prediction model based on the fold-increase. Of the 42 chemicals, four are correct positive, seven false negative, nine false positive and 22 correct negative. Including the 16 chemicals with MTT-data only, the strategy finally resulted in a specificity of 22/33=66.6% and a sensitivity of 18/15=72.0%. Compared to the predictive capacity of the MTT alone (specificity: 90.9%; sensitivity: 56.0%), the increase in sensitivity is achieved by a severe loss in specificity. Thus, the MT decided that IL1-α did not offer any advantages, so that the two additional laboratories were not asked to determine this second endpoint for their samples.

Although the negative controls were fairly reproducible in the 13 runs with a mean IL1-α of 40.23 pg/ml and a standard deviation of 9.14, a third PM based on the total amount of IL1-a released was explored. The IL1-α data reported in Table 55 were combined with the in vivo European classification and a ROC-analysis was performed (Figure 37). The flat curve, always close to a random process (indicated by the dotted line) shows that IL1-α as a stand-alone endpoint does not perform well. When considering the sum of both specificity and sensitivity, optimal performance for a stand-alone use of IL1-α is reached at a threshold of 77.0 pg/ml, i.e. a specificity of 52.7% and a sensitivity of 72.7%.

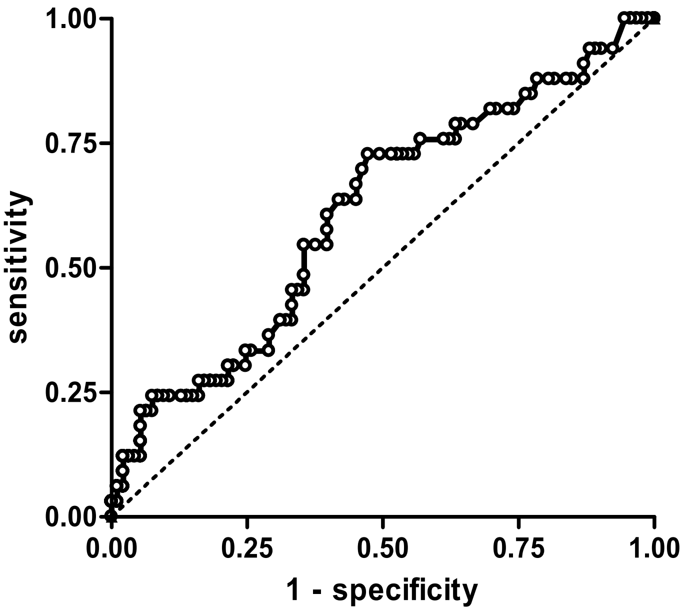


Figure 37: Receiver operation curve of the ZEBET IL1-α with EpiDerm, where the dotted line represents the curve of a useless test

Overall predictivity of the strategic combination of both endpoints was performed with the mean IL1-α amount (Table 55). The results are summarised in Figure 39, where for example the predictive capacity of MTT alone, i.e. specificity of 93.9% and sensitivity of 56.0%, is reached for thresholds above 278 pg/ml, i.e. the highest response of all runs. Choosing the best threshold from the analysis above considering IL1-α as a stand-alone endpoint, i.e. 75 pg/ml, resulted in a specificity of 17/33=51.5% and a sensitivity of 22/25=88.0%. In contrast, a PM-threshold of 170 pg/ml resulted in a specificity of 29/33=87.9% and a sensitivity of 17/25=68.0%, i.e. the overall maximum of the sum of both parameters (1.559). A stable, tough not optimal predictive capacity with both specificity and sensitivity around 70% was found for thresholds between 113 and 135 pg/ml. Thus, the second endpoint IL1-α could not substantially increase the overall predictive capacity in terms of specificity and sensitivity in comparison to MTT alone. In both cases the sum of the two parameters is about 150%. When interpreting these data, it has to be kept in mind that the IL1-α PM was optimised post-hoc, which usually results in an overestimation of predictive capacity. However, IL1-α could be used for shifting the balance between specificity and sensitivity.

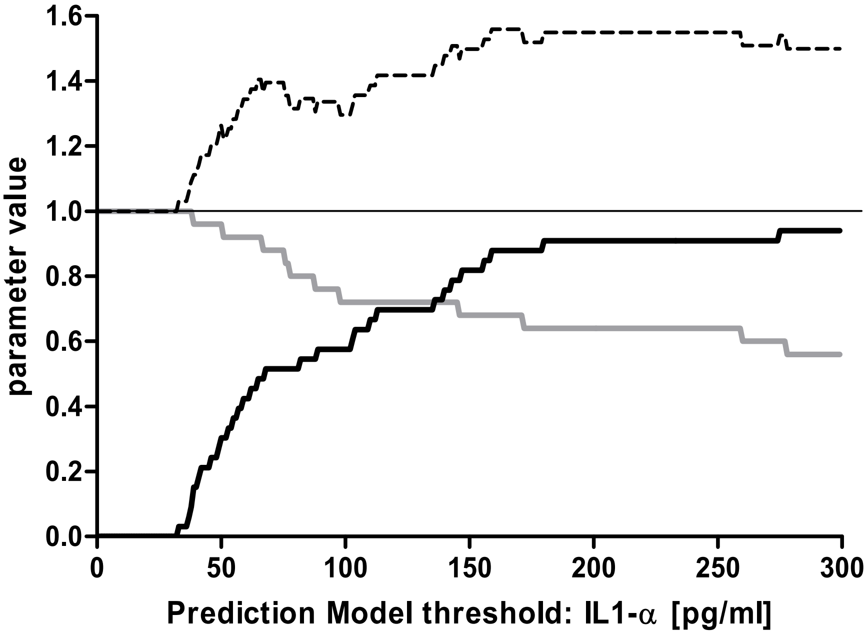


Figure 38: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold for IL1-α when considered in the proposed strategic manner together with MTT
(black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity).

Finally, it was attempted to construct post–hoc a PM for the GHS classification system. Due to restrictions in data availability, i.e. in amount and completeness, and the poor performance of IL1-α above, this was done empirically. To keep this analysis simple and disregarding reproducibility, for both endpoints only the mean over the respective runs was considered. Four chemicals were excluded, as they did not produce three valid runs for MTT. For 16 chemicals only viability data were available, because they showed viability smaller than 50%. Therefore, all of these had to be considered as GHS-irritants. The remaining 38 chemicals, of which four were GHS-irritants, nine GHS-mild irritants and 25 GHS-non irritants, were screened for obvious cut-off values in both endpoints: If for a given chemical viability was below 75% or IL1-α was larger than 160 pg/ml, then this would be classified as GHS-mild irritants. All the other cases would result in a classification as GHS-non irritant. This prediction model correctly classified for 66.7% of the GHS-irritants, for 25.0% of the GHS-mild irritants and for 96.1%. With the small amount of data available, no satisfactorily performing PM could be constructed. To evaluate this aspect in more detail and with adequate multivariate statistical tools a larger and more complete data set, i.e. with viability and IL1-α data, would be required.

4.2 EPISKIN: L'Oréal

4.2.1 Data submission

As IL1-α was established as a possible endpoint for skin irritation only in a later stage of test development, it was agreed to evaluate this second endpoint in a first step only in the leading laboratories. Because a highly irritating property of given substance might interact with IL1-α, the MT decided to measure this endpoint only for MTT-non irritant chemicals or chemicals, which showed high variability in the MTT-test. L'Oréal, the leading laboratory of EPISKIN, submitted the data on the 27.05.2006 for 39 chemicals. All chemicals, including the two confidential triggered the IL1-α by at least one of the agreed criteria. Two chemicals (numbers 8, 43) were already positive in two out of three MTT-runs, while the mean value was still larger than 50%. The tests were performed between the 12.04.05 and the 19.05.05 with the stored supernatants from the MTT-tests by one operator, who also performed all MTT-tests. For two chemicals four runs were available (numbers 23, 52), because the MTT-test was repeated once for each chemical due high variability. The two confidential chemicals were excluded from all analysis resulting in a total of 37 chemicals included in the analysis. The data processed in the following were the calculated IL1-α in pg/ml and are shown in Figure 39. The chemicals were divided in three sub-sets. It was remarked that some positives control would have had to be tested diluted. Although the respective data were not submitted, the effect is negligible as the response level would have been even higher.

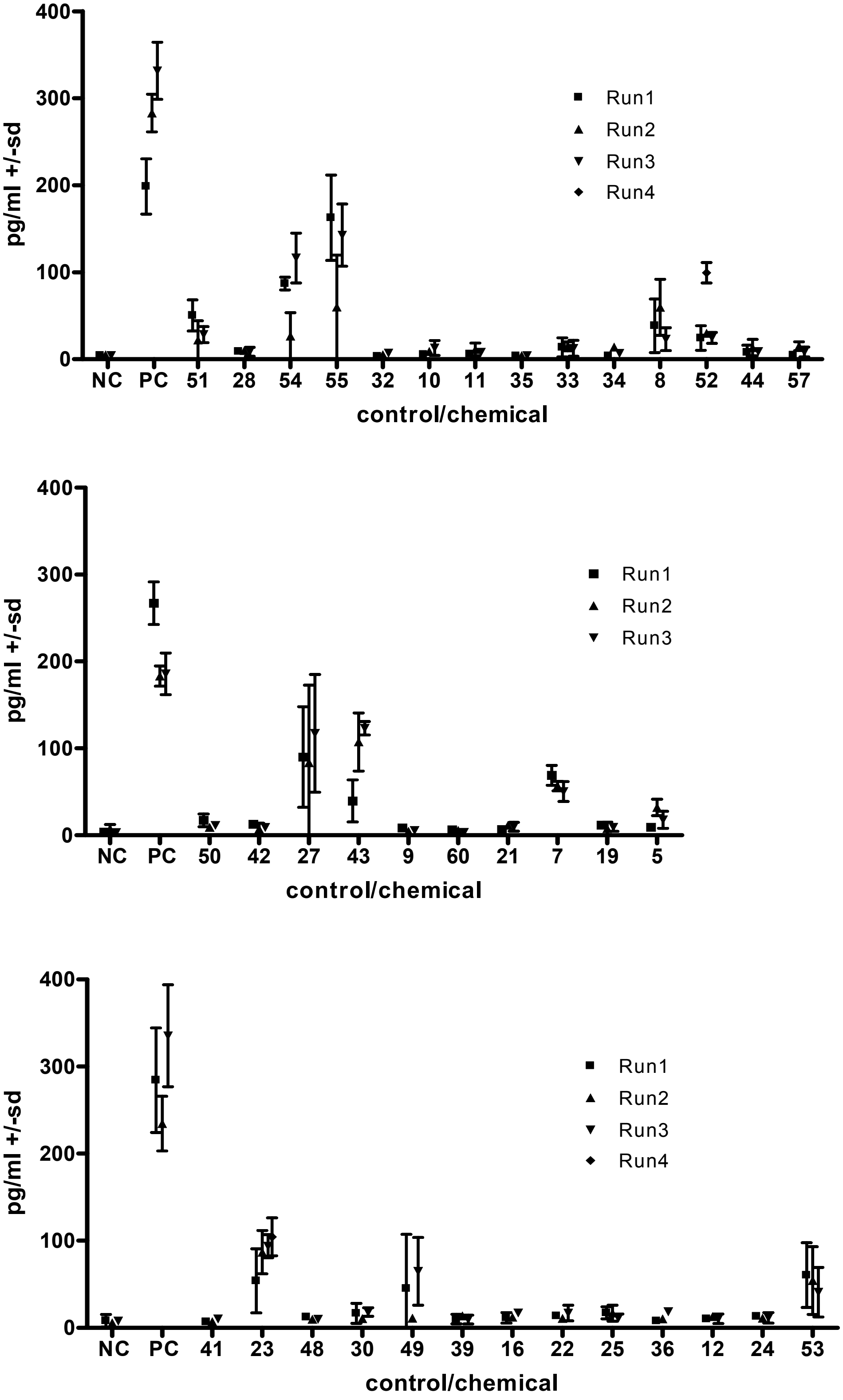


Figure 39: EPISKIN-L'Oréal IL1-α data of all four sets of chemicals tested including the controls (NC: negative control; PC: positive control) expressed as mean and standard deviation

4.2.2 Intra-assay variability

Per run three replicates were measured. In order to assess the intra-assay variability, i.e. the variability between these replicates, the respective standard deviation and the coefficient of variation were calculated. Regardless the measure, substantial intra-assay was observed (Figure 40). Both measures of variability depended on the response level, while the standard deviation increased in the same way the CV decreases with increasing response levels.

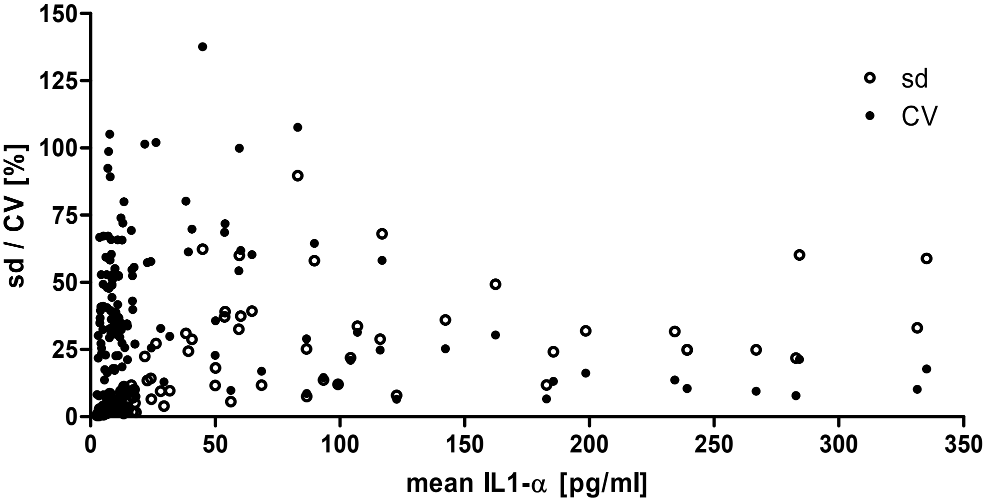


Figure 40: L'Oréal IL1-α intra-assay variability expressed as standard deviation and coefficient of variation for all runs of all chemicals EPISKIN

4.2.3 Within-laboratory variability

The negative control gave in average a response of 5.8 pg/ml with a minimum of 3.0 and a maximum of 8.6 pg/ml (CV: 34.9%). The positive control induced in average 254 pg/ml ranging from 183 to 335 pg/ml (CV: 21.9%).

The variability within the laboratory, i.e. between independent runs, was expressed by the standard deviation and the CV (Table 51). Furthermore, a 1-way ANOVA was calculated for each chemical with the logarithmically (natural) transformed data. Regarding the descriptive measures, the CV is to be preferred as it was less dependent on the response range than the standard deviation.

chemical number	chemical class	Run1	Run2	Run3	Run4	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

Table 58: L'Oréal within-laboratory variability of IL1-α of EPISKIN, where grey cells indicate runs with MTT-variability of SD >18

As the preliminary prediction model was based on the fold-increase of IL1-α release induced by a chemical in comparison with the respective negative control, the variability of this measure is included in Table 60.

Considering the predictions, which are based on the cut-off value of 5-fold increase, three (numbers 5, 49, 51) of 37 chemicals were not reproducible.

As for the MTT-endpoint, a comparison of data of chemicals tested in both phases was carried out providing further information on the within-laboratory reproducibility. The mean IL1-α amount over the runs for the controls and the overlapping eleven chemicals are summarised in Table 59, whereas the mean individual run data are presented in Figure 41. The differences show that the data of the two phases were very similar, whereas no obvious trend was present. The larger difference for the positive control (PC) is acceptable for the high response level of about 200 pg/ml. Indeed, a paired t-test including/excluding the positive control resulted in a non-significant p-value of 0.487/0.910.

chemical number	mean IL1-α [pg/ml] Phase 1	Phase 2	difference
NC	5.86	5.83	0.03
PC	192.33	254.33	-62.00
9	6.06	6.11	-0.05
12	11.68	11.16	0.52
16	8.37	13.59	-5.22
28	11.30	9.13	2.17
30	6.32	15.07	-8.75
32	5.64	5.18	0.46
35	7.59	4.11	3.48
42	5.36	9.52	-4.16
49	58.78	40.42	18.36
51	61.66	33.58	28.08
52	15.20	44.55	-29.35

Table 59: Comparison of mean IL1-α amount in pg/ml of controls and eleven chemicals tested in both phases at L’Oréal with EPISKIN

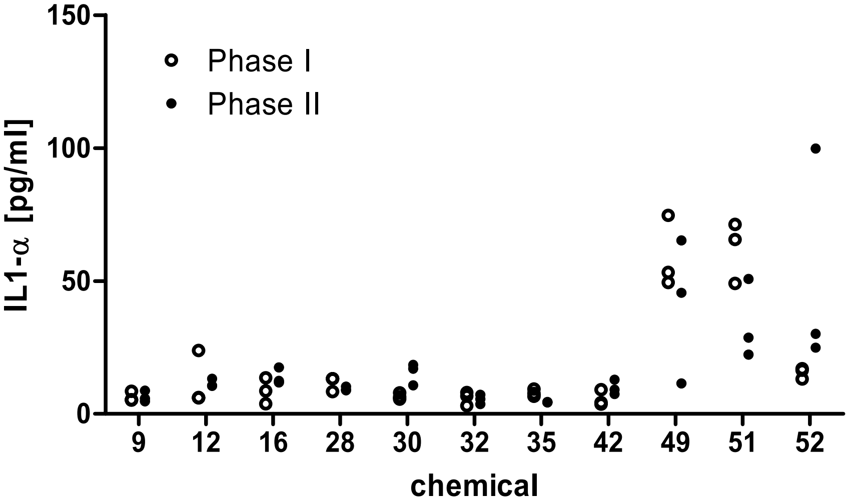


Figure 41: IL1-α amount in pg/ml of the individual runs of the twelve chemicals tested in both phases at L’Oréal with EPISKIN

4.2.4 Predictive capacity

As IL1-α was considered from the very beginning in a strategic manner, 21 MTT-positive chemicals were not tested for this second endpoint. Of these 21, 17 were correctly and four wrongly classified as positives. Of the remaining 37 chemicals, eight chemicals had a label (R38) and 29 had not. They included the three chemicals not having three acceptable MTT-runs (numbers 5, 23, 53), one of which was an irritant and two non-irritants. If available, the three valid runs per chemical were used. If not, the first three runs were considered for analysis.

chemical number	chemical class (EU)	fold increase			mean	sd	CV [%]	ANOVA p-value (log data)		
		1. run	2. run	3. run				1. run	2. run	3. run
5	no label	2.52	4.35	**6.02**	4.30	1.75	40.74	> 0.05	**< 0.05**	**< 0.01**
7	no label	**18.62**	**7.69**	**17.07**	**14.46**	5.91	40.90	**< 0.01**	**< 0.01**	**< 0.01**
8	no label	**8.51**	**12.27**	**5.22**	**8.67**	3.53	40.70	**< 0.01**	**< 0.01**	**< 0.01**
9	no label	2.29	0.60	1.85	1.58	0.88	55.49	> 0.05	> 0.05	> 0.05
10	no label	1.18	1.84	2.93	1.98	0.88	44.56	> 0.05	> 0.05	> 0.05
11	no label	1.34	2.30	1.87	1.84	0.48	26.18	> 0.05	> 0.05	> 0.05
12	no label	1.25	2.21	1.32	1.60	0.53	33.59	> 0.05	> 0.05	> 0.05
16	no label	1.41	2.07	2.19	1.89	0.42	22.22	> 0.05	> 0.05	> 0.05
19	no label	3.14	1.03	3.12	2.43	1.21	49.90	**< 0.05**	> 0.05	> 0.05
21	no label	1.73	1.54	3.33	2.20	0.98	44.69	> 0.05	> 0.05	> 0.05
22	no label	1.67	1.84	2.17	1.89	0.26	13.43	> 0.05	> 0.05	> 0.05
23	R38	**6.62**	**14.88**	**11.96**	**11.39**	3.45	37.56	**< 0.05**	**< 0.01**	**< 0.01**
24	no label	1.63	1.91	1.47	1.67	0.22	13.34	> 0.05	> 0.05	> 0.05
25	no label	2.11	2.89	1.43	2.14	0.73	34.09	> 0.05	**< 0.05**	> 0.05
27	R38	**24.35**	**11.33**	**39.79**	**25.15**	14.25	56.63	**< 0.01**	**< 0.01**	**< 0.01**
28	no label	1.95	2.05	1.95	1.98	0.06	2.91	> 0.05	> 0.05	> 0.05
30	no label	2.04	1.79	2.31	2.05	0.26	12.71	> 0.05	> 0.05	> 0.05
32	no label	0.73	1.08	1.58	1.13	0.42	37.81	> 0.05	> 0.05	> 0.05
33	no label	3.03	3.07	2.83	2.98	0.13	4.32	> 0.05	> 0.05	> 0.05
34	R38	0.86	2.88	1.52	1.75	1.03	58.75	> 0.05	> 0.05	> 0.05
35	no label	0.88	0.87	0.93	0.89	0.03	3.60	> 0.05	> 0.05	> 0.05
36	no label	0.96	1.78	2.37	1.70	0.71	41.57	> 0.05	> 0.05	> 0.05
39	no label	1.23	2.33	1.21	1.59	0.64	40.31	> 0.05	> 0.05	> 0.05
41	no label	0.84	1.19	1.32	1.12	0.25	22.23	> 0.05	> 0.05	> 0.05
42	no label	3.40	0.98	2.99	2.46	1.30	52.72	**< 0.05**	> 0.05	> 0.05
43	R38	**10.67**	**14.59**	**41.75**	**22.34**	16.93	75.78	**< 0.01**	**< 0.01**	**< 0.01**
44	no label	1.76	2.73	2.00	2.16	0.51	23.35	> 0.05	> 0.05	> 0.05
48	no label	1.53	1.68	1.27	1.49	0.20	13.89	> 0.05	> 0.05	> 0.05
49	R38	**5.54**	1.91	**8.29**	**5.25**	3.20	60.99	> 0.05	> 0.05	**< 0.01**
50	no label	4.63	1.22	3.86	3.24	1.79	55.25	**< 0.01**	> 0.05	**< 0.05**
51	R38	**11.15**	4.52	**6.42**	**7.36**	3.42	46.37	**< 0.01**	> 0.05	**< 0.01**
52	no label	**6.11**	**5.58**	**11.51**	**7.15**	2.92	42.43	**< 0.05**	**< 0.01**	**< 0.01**
53	no label	**7.41**	**9.29**	**5.23**	**7.31**	2.03	27.80	**< 0.01**	**< 0.01**	**< 0.01**
54	no label	**19.23**	**5.45**	**26.36**	**17.01**	10.63	62.48	**< 0.01**	> 0.05	**< 0.01**
55	R38	**35.98**	**12.31**	**32.30**	**26.87**	12.74	47.41	**< 0.01**	**< 0.01**	**< 0.01**
57	no label	1.04	3.11	1.93	2.02	1.04	51.24	> 0.05	> 0.05	> 0.05
60	R38	1.70	0.73	1.14	1.19	0.49	40.92	> 0.05	> 0.05	> 0.05

Table 60: Summary of L'Oréal IL1-α data of the tested chemicals and positive controls expressed as fold-increase relative to the respective negative controls together with 1-way ANOVA results of comparing the logarithmically transformed pg/ml-data with the respective negative control data (with Dunnett's post test)

In Table 60, the data for two kinds of PM are summarised: the PM, which bases on the relative increase compared to the negative control, and the results, i.e. p-values, of 1-way ANOVAs with a Dunnett post test comparing a given chemical's response to the response of the respective negative control. The later was applied to the raw data (data not shown) as well as to the logarithmically transformed data. Although the differences are minor, the analysis of the transformed data is to be preferred as the variances, i.e. the CVs, were more homogenous. To simplify the analysis, the results over the runs were combined: A chemical is overall classified as irritant if the mean fold increase is larger than 5, or if a chemical gave in at least two runs a significantly ($p<0.05$) higher response than the negative control (logarithmically transformed data).

As the overall classification are identical for both prediction models (preliminary and ANOVA-based on log-transformed data), the resulting predictive capacities are calculated for the prediction model based on the fold-increase only. Of the 37 chemicals, six are correct positive, two false negative, five false positive and 24 correct negative. The two chemicals, which were MTT-positive in two runs, were also IL1-α positive. Including the 21 chemicals with MTT-data only, the strategy finally resulted in a specificity of 24/33=72.7% and a sensitivity of 23/25=92.0%. Compared the predictive capacities of the MTT alone, the increase in sensitivity (20.0% points) is achieved by a smaller loss in specificity (11.7% points). Thus, IL1-α improved the predictive capacity of EPISKIN. Therefore the two additional laboratories were asked to determine this second endpoint for their samples.

4.3 EPISKIN: Unilever

4.3.1 Data submission

As IL1-α seemed to offer advantages, it was decided by the MT that the two additional laboratories determine this second endpoint. After training by L'Oréal, Unilever submitted the data on the 31.10.2006 for 32 chemicals in a summarised format, which was provided by the leading laboratory L'Oréal. All chemicals, including the two confidential, triggered the IL1-α by at least one of the agreed criteria. The summarising format did not include any information on the dates of testing or the operator. They included no specific remarks. For eight chemicals four runs were available, because the MTT-tests were repeated due to high variability. The two confidential chemicals will be excluded from all analysis resulting in a total of 30 chemicals, which were analysed here. The data processed in the following were the calculated IL1-α in pg/ml and are shown in Figure 42. The chemicals were tested in three sub-sets.

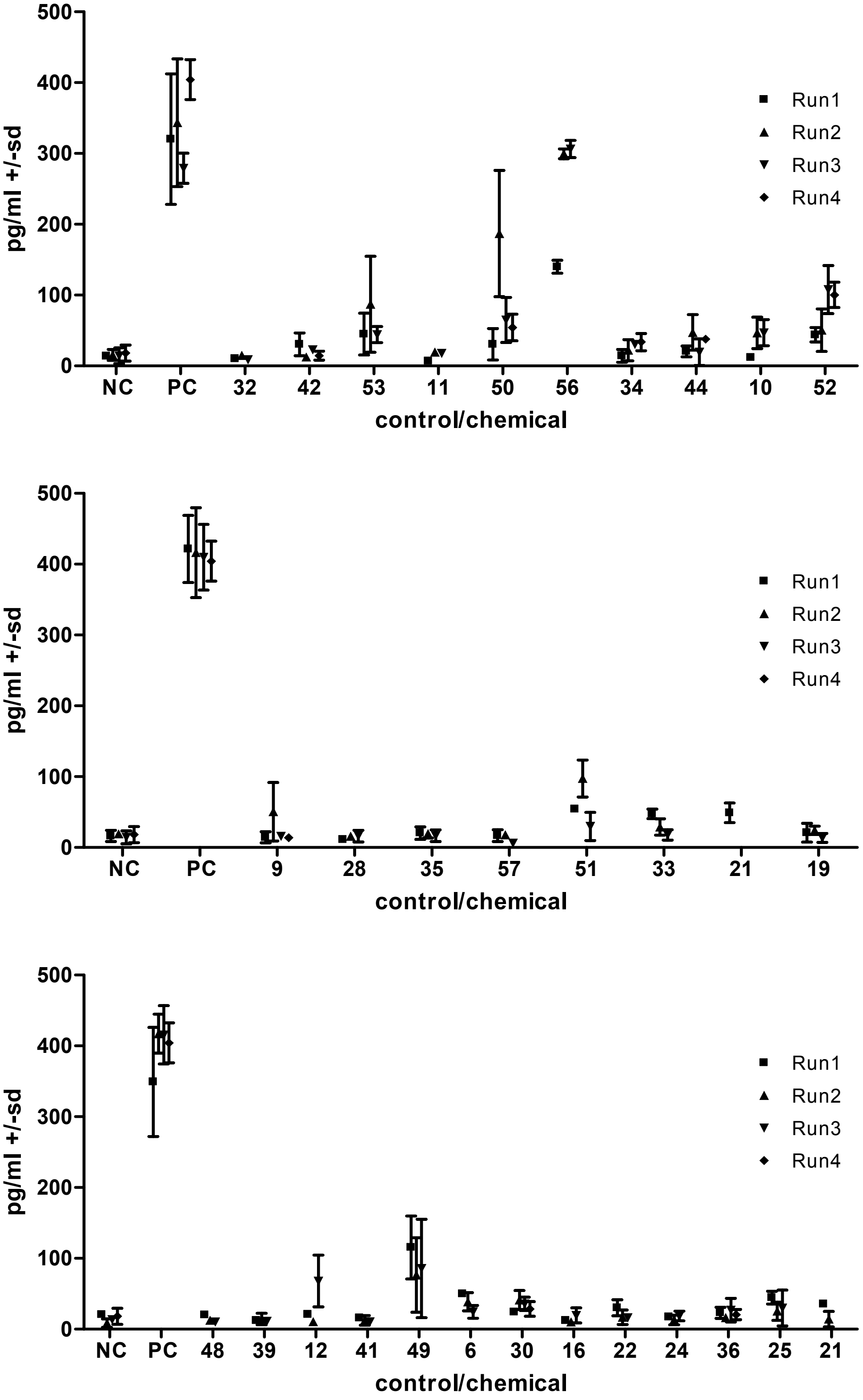


Figure 42: EPISKIN-Unilever IL1-α data of all four sets of chemicals tested including the controls (NC: negative control; PC: positive control) expressed as mean and standard deviation

4.3.2 Intra-assay variability

Per run three replicates were measured. In order to assess the intra-assay variability, i.e. the variability between these replicates, the respective standard deviation and the CV were calculated. Regardless the measure, substantial intra-assay was observed (Figure 43). Both measures of variability depended on the response level, while the standard deviation increased in the same way the CV decreases with increasing response levels.

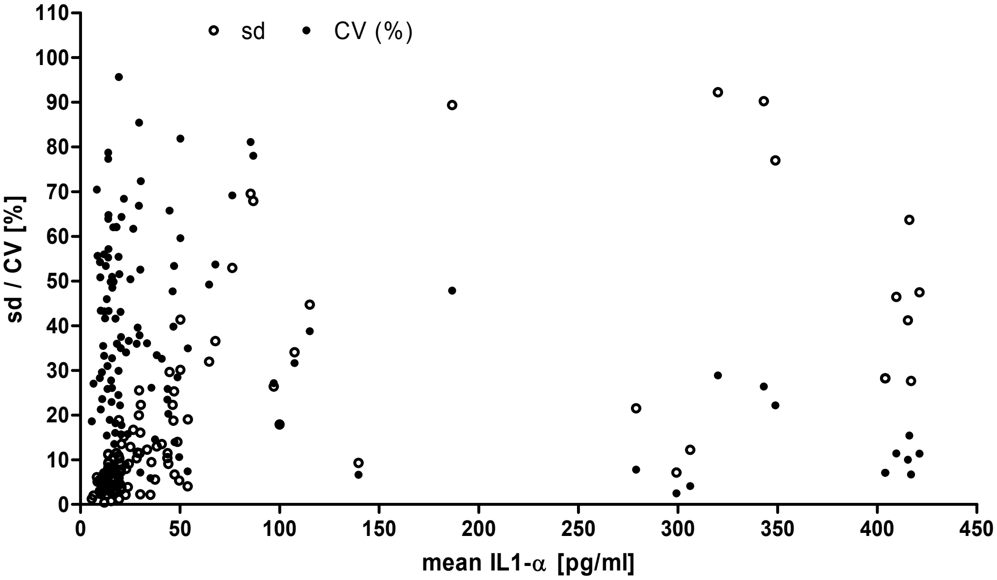


Figure 43: Unilever IL1-α intra-assay variability expressed as standard deviation and coefficient of variation for all runs of all chemicals EPISKIN

4.3.3 Within-laboratory variability

The negative control gave in average a response of 15.4 pg/ml with a minimum of 8.5 and a maximum of 20.6 pg/ml (CV: 22.4%). The positive control induced in average 378 pg/ml ranging from 279 to 422 pg/ml (CV: 13.6%).

The variability within the laboratory, i.e. between independent runs, was expressed by the standard deviation and the CV (Table 61). Furthermore, a 1-way ANOVA was calculated for each chemical with the raw and the logarithmically (natural) transformed data. Regarding the descriptive measures, the CV is to be preferred as it was less dependent on the response range than the standard deviation.

As the preliminary prediction model was based on the fold-increase of IL1-α release induced by a chemical in comparison to the respective negative control, the variability of this measure is included in Table 62.

Considering the predictions, which are based on the cut-off value of 5-fold increase, three (numbers 5, 49, 51) of 37 chemicals were not reproducible.

chemical number	chemical class (EU)	mean pg/ml				mean	sd	CV [%]
		Run1	Run2	Run3	Run4			
6	no label	49.87	38.54	24.50		37.64	12.71	33.76
9	no label	14.19	50.45	15.71	13.84	23.55	17.95	76.24
10	no label	11.70	46.61	47.00		35.10	20.27	57.73
11	no label	6.89	19.63	17.88		14.80	6.90	46.64
12	no label	20.95	10.53	68.00		33.16	30.62	92.33
16	no label	12.18	10.20	19.46		13.94	4.88	34.97
19	no label	20.89	24.20	13.53		18.06	3.96	21.95
21	no label	48.98	35.46	14.29		32.91	17.49	53.13
22	no label	30.03	16.83	16.15		21.00	7.83	37.26
24	no label	17.31	13.03	18.65		16.33	2.94	18.00
25	no label	44.45	25.37	29.77		33.20	9.99	30.10
28	no label	11.06	15.71	16.13		14.30	2.81	19.67
30	no label	24.00	41.13	35.90	28.50	32.39	7.62	23.53
32	no label	10.05	14.97	8.94		11.32	3.21	28.36
33	no label	47.53	29.00	17.89		31.47	14.98	47.59
34	R38	14.19	22.08	30.36	33.75	25.09	8.77	34.94
35	no label	20.37	19.40	16.63		18.80	1.94	10.32
36	no label	23.16	16.12	26.89	20.55	21.68	4.53	20.88
39	no label	12.29	14.36	11.24		12.63	1.59	12.56
41	no label	15.89	12.21	10.43		12.85	2.78	21.68
42	no label	30.41	12.34	22.92	14.48	20.04	8.29	41.36
44	no label	20.62	47.33	19.59	37.75	31.32	13.54	43.21
48	no label	20.13	12.66	10.06		14.28	5.23	36.59
49	R38	115.38	76.52	85.69		92.53	20.31	21.95
50	no label	30.63	186.94	64.89	54.22	84.17	69.99	83.15
51	R38	54.04	97.34	29.68		60.35	34.27	56.78
52	no label	43.93	50.45	107.77	100.30	75.61	33.07	43.73
53	no label	44.99	87.03	44.05		58.69	24.55	41.83
56	R38	139.87	299.50	306.32		248.56	94.19	37.89
57	no label	16.91	17.73	6.01		13.55	6.54	48.26

Table 61: Unilever within-laboratory variability of IL1-α of EPISKIN

4.3.4 Predictive capacity

As IL1-α was considered from the very beginning in a strategic manner, 28 MTT-positive chemicals were not tested for this second endpoint. Here, chemical number 54 was considered as positive as the mean viability value over all runs was below 50%, but in two of the three runs it was classified as negative. Thus, of these 28, 21 were correctly classified as positives and seven wrongly as positives (chemical numbers 2, 5, 7, 8, 17, 26, 54). Of the remaining 30 chemicals, four chemicals (numbers 34, 49, 51, 56) had a label (R38) and 26 had not. They included the three chemicals not having two acceptable MTT-runs (numbers 53, 56), one of which was an irritant and the other a non-irritants. In Table 62, the data for two kinds of PM are summarised: the PM, which was based on the relative increase compared to the negative control, and the results, i.e. p-values, of 1-way ANOVAs with a Dunnett-post test comparing a given chemical's response to the response of the respective negative control. The later was applied to the raw data (data not shown) as well as to the log-transformed data. The analysis of the transformed data is to be preferred as the variances were more homogenous over the entire response range. To simplify the analysis, the results over the runs were combined: A chemical is overall classified as irritant if the mean fold increase is larger than 5, or if a chemical gave in at least two runs a significantly (p<0.05) higher response than the negative control (logarithmically transformed data).

chemical number	chemical class (EU)	fold increase			mean	sd	CV [%]	ANOVA p-value (log data)		
		1. run	2. run	3. run				1. run	2. run	3. run
6	no label	2.42	4.54	1.82	2.927	1.429	48.83	**< 0.01**	**< 0.05**	> 0.05
9	no label	0.87	1.09	0.76	0.907	0.168	18.53	> 0.05	> 0.05	> 0.05
10	no label	0.84	3.02	3.31	2.390	1.350	56.49	> 0.05	> 0.05	**< 0.05**
11	no label	0.49	1.27	1.26	1.007	0.447	44.45	> 0.05	> 0.05	> 0.05
12	no label	1.02	1.24	**5.04**	2.433	2.260	92.88	> 0.05	> 0.05	**< 0.01**
16	no label	0.59	1.20	1.44	1.077	0.438	40.70	> 0.05	> 0.05	> 0.05
19	no label	1.28	1.30	0.94	1.173	0.202	17.24	> 0.05	> 0.05	> 0.05
21	no label	3.00	1.72	1.68	2.133	0.751	35.19	**< 0.05**	> 0.05	> 0.05
22	no label	1.46	1.98	1.20	1.547	0.397	25.68	> 0.05	> 0.05	> 0.05
24	no label	0.84	1.53	1.38	1.250	0.363	29.03	> 0.05	> 0.05	> 0.05
25	no label	2.16	2.99	2.21	2.453	0.465	18.97	**< 0.01**	> 0.05	> 0.05
28	no label	0.68	0.81	1.12	0.870	0.226	25.98	> 0.05	> 0.05	> 0.05
30	no label	1.17	2.66	1.56	1.797	0.773	43.01	> 0.05	**> 0.05**	> 0.05
32	no label	0.72	0.97	0.63	0.773	0.176	22.78	> 0.05	> 0.05	> 0.05
33	no label	2.92	1.50	1.25	1.890	0.901	47.66	**< 0.05**	> 0.05	> 0.05
34	R38	1.43	2.14	1.84	1.803	0.356	19.76	> 0.05	> 0.05	> 0.05
35	no label	1.25	1.00	1.16	1.137	0.127	11.14	> 0.05	> 0.05	> 0.05
36	no label	1.12	1.99	1.12	1.410	0.502	35.62	> 0.05	> 0.05	> 0.05
39	no label	0.60	1.69	0.83	1.040	0.575	55.24	> 0.05	> 0.05	> 0.05
41	no label	0.77	1.44	0.77	0.993	0.387	38.94	> 0.05	> 0.05	> 0.05
42	no label	0.80	1.61	0.79	1.067	0.471	44.12	> 0.05	> 0.05	> 0.05
44	no label	1.48	3.07	2.06	2.203	0.805	36.52	> 0.05	> 0.05	**< 0.05**
48	no label	0.98	1.49	0.75	1.073	0.379	35.29	> 0.05	> 0.05	> 0.05
49	R38	**5.60**	**9.01**	**6.35**	**6.987**	1.792	25.65	**< 0.01**	**< 0.01**	**< 0.01**
50	no label	2.20	4.56	2.96	3.240	1.205	37.18	> 0.05	**< 0.01**	**< 0.01**
51	R38	3.32	**5.04**	2.07	3.477	1.491	42.89	< 0.05	**< 0.01**	> 0.05
52	no label	3.15	3.27	**5.48**	3.967	1.312	33.07	> 0.05	> 0.05	**< 0.01**
53	no label	3.23	**5.65**	3.10	3.993	1.436	35.96	> 0.05	**< 0.05**	**< 0.05**
56	R38	**10.04**	**19.43**	**21.55**	**17.007**	6.126	36.02	**< 0.01**	**< 0.01**	**< 0.01**
57	no label	1.04	0.92	0.42	0.793	0.329	41.45	> 0.05	> 0.05	> 0.05

Table 62: Summary of Unilever IL1-α data of the tested chemicals and positive controls expressed as fold-increase relative to the respective negative controls together with 1-way ANOVA results of comparing the raw/log-transformed pg/ml-data with the respective negative control data (with Dunnett's post test)

Of the four R38-labelled chemicals, two had a mean fold-increase larger than 5. The two chemicals were also identified by the ANOVA-PM, which, in addition, also classified three non-irritant chemicals (numbers 6, 50, 53) as irritants. Thus the preferable PM based on the fold-increase, resulted when including the 28 chemicals with MTT-data only in the proposed strategic manner, in a specificity of 26/33=78.8% and a sensitivity of 23/25=92.0%. Compared to the predictive capacities of the MTT alone, the increase in sensitivity (8.0% points) is achieved by a smaller loss in specificity (3.0% points). This loss was caused by the two different approaches chemical 54 was considered, i.e. the median classification of the runs and the classification based on the mean viability over the runs.

4.4 EPISKIN: Sanofi

4.4.1 Data submission

As IL1-α seemed to offer advantages, it was decided by the MT that the two additional laboratories determine this second endpoint. After training by L'Oréal, Sanofi submitted the data on the 04.11.2006 for 36 chemicals in a summarised format, which was provided by the leading laboratory L'Oréal. All chemicals, including the two confidential, triggered the IL1-α because they showed a mean viability in the MTT of larger than 50%. Two chemicals (numbers 18, 20), which had three MTT-runs not meeting the variability criterion (SD>18), were not tested for IL1-α. The submission did not include any information on the dates of testing or the operator or any specific remarks. For four chemicals four runs were available, because the MTT-tests were repeated due to high variability. The two confidential chemicals will be excluded from all analysis resulting in 34 chemicals included in this analysis. The data processed in the following were the calculated IL1-α in pg/ml and are shown in Figure 44. The chemicals were divided in three sub-sets.

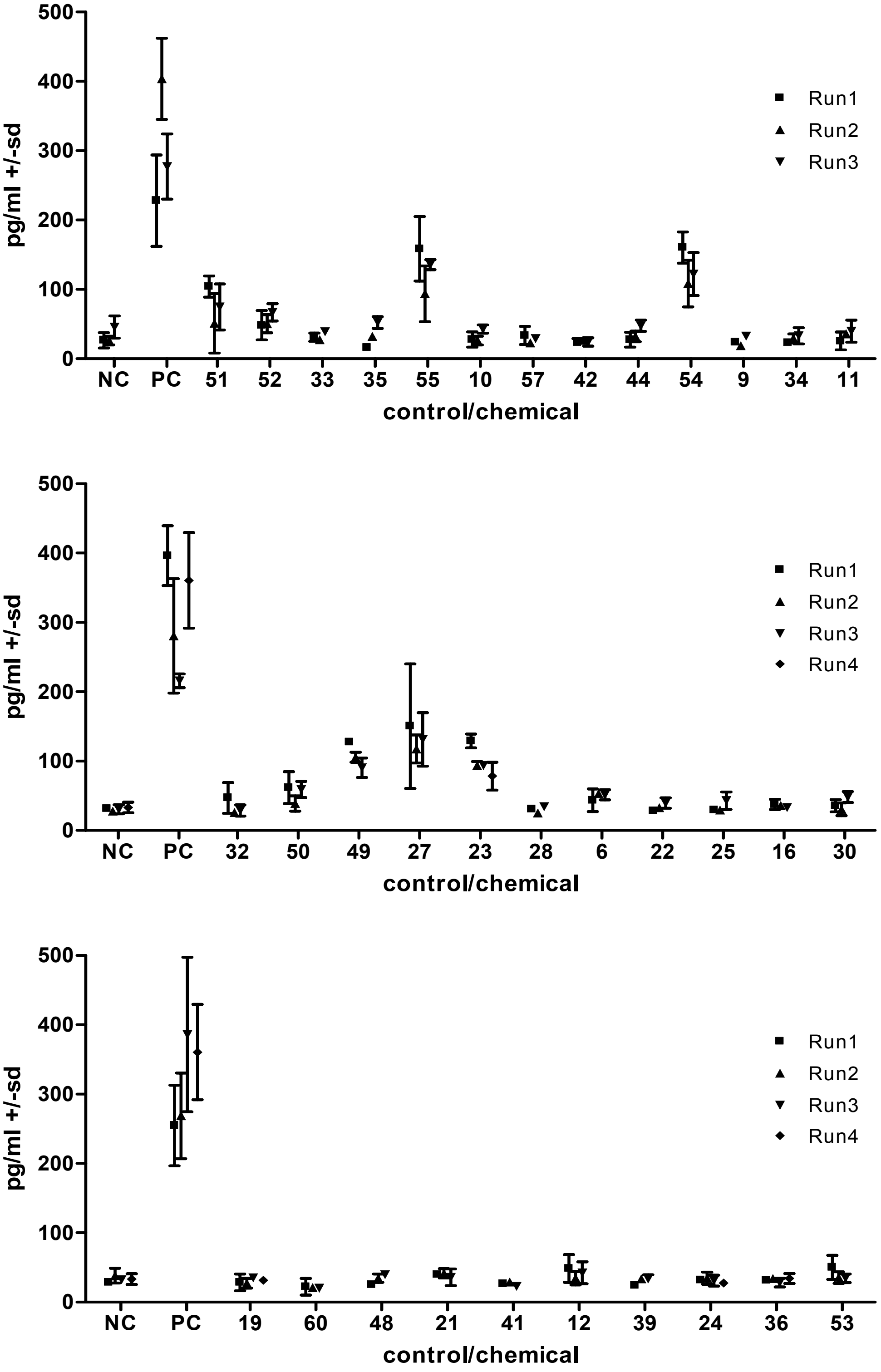


Figure 44: EPISKIN-Sanofi IL1-α data of all four sets of chemicals tested including the controls (NC: negative control; PC: positive control) expressed as mean and standard deviation

4.4.2 Intra-assay variability

Per run three replicates were measured. In order to assess the intra-assay variability, i.e. the variability between these replicates, the respective standard deviation and the coefficient of variation were calculated. Regardless the measure, less intra-assay was observed (Figure 45). While the standard deviation strongly depended on the response level, the CV was, with a mean of 21.7%, stable over the whole response range (data not shown).

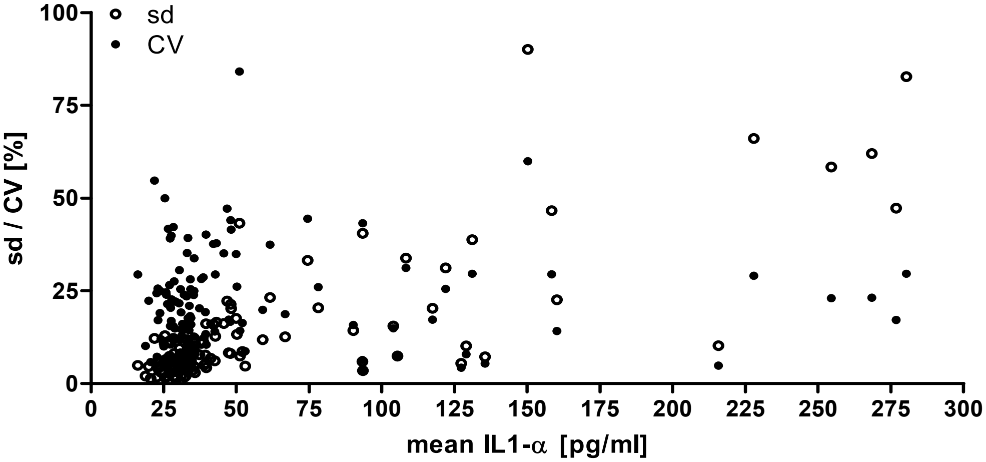


Figure 45: Sanofi IL1-α intra-assay variability expressed as standard deviation and coefficient of variation for all runs of all chemicals EPISKIN

4.4.3 Within-laboratory variability

The negative control gave in average a response of 32.1 pg/ml with a minimum of 26.1 and a maximum of 45.9 pg/ml (CV: 18.8%). The positive control induced in average 307 pg/ml ranging from 216 to 404 pg/ml (CV: 23.5%).

The variability within the laboratory, i.e. between independent runs, was expressed by the standard deviation and the CV (Table 63). Furthermore, a 1-way ANOVA was calculated for each chemical with the logarithmically (natural) transformed data. Regarding the descriptive measures, the CV is to be preferred as it was almost independent of the response (mean of 19.3%).

As the preliminary prediction model was based on the fold-increase of IL1-α release induced by a chemical in comparison with the respective negative control, the variability of this measure is included in Table 64.

Considering the predictions, which are based on the cut-off value of 5-fold increase, two (numbers 54, 55) of 34 chemicals were not reproducible.

chemical number	chemical Class (EU)	mean pg/ml				mean	sd	CV [%]
		Run 1	Run 2	Run 3	Run 4			
6	no label	43.40	53.22	51.45		49.36	5.24	10.61
9	no label	23.95	18.94	32.97		25.28	7.11	28.12
10	no label	27.86	26.12	42.76		32.25	9.15	28.36
11	no label	25.66	35.72	39.82		33.73	7.28	21.59
12	no label	48.40	34.43	42.32		41.71	7.01	16.80
16	no label	37.57	36.00	33.33		35.63	2.14	6.01
19	no label	28.55	27.29	35.30	31.42	30.64	3.55	11.59
21	no label	39.85	41.41	35.76		39.01	2.92	7.48
22	no label	28.26	32.80	39.57		33.54	5.69	16.98
23	R38	129.21	93.59	93.75	78.40	98.74	21.55	21.83
24	no label	31.68	34.51	31.02	27.52	31.18	2.87	9.21
25	no label	29.40	29.44	43.01		33.95	7.84	23.10
27	R38	150.38	117.63	131.35		133.12	16.45	12.35
28	no label	30.96	25.13	34.65		30.25	4.80	15.87
30	no label	35.60	30.66	48.15		38.14	9.02	23.64
32	no label	47.05	25.68	28.86		33.86	11.53	34.05
33	no label	31.25	27.65	39.17		32.69	5.89	18.03
34	R38	23.27	29.34	33.27		28.63	5.03	17.59
35	no label	16.30	32.22	52.34		33.62	18.06	53.73
36	no label	31.45	33.99	28.05	34.02	31.88	2.82	8.85
39	no label	24.27	33.62	33.54		30.48	5.37	17.63
41	no label	26.46	28.75	22.94		26.05	2.93	11.24
42	no label	23.25	26.49	24.05		24.60	1.69	6.86
44	no label	27.48	32.21	47.61		35.77	10.53	29.44
48	no label	25.30	34.58	39.72		33.20	7.31	22.01
49	R38	127.56	105.61	90.44		107.87	18.66	17.30
50	no label	61.84	38.85	59.30		53.33	12.61	23.64
51	R38	104.18	51.26	74.68		76.70	26.52	34.57
52	no label	48.26	50.47	66.97		55.23	10.23	18.52
53	no label	50.13	35.55	34.36		40.01	8.78	21.95
54	no label	160.40	108.53	122.21		130.38	26.88	20.62
55	R38	158.62	93.67	135.70		129.33	32.94	25.47
57	no label	33.58	22.83	29.01		28.47	5.40	18.96
60	R38	22.05	20.76	20.11		20.98	0.99	4.71

Table 63: Sanofi within-laboratory variability of IL1-α of EPISKIN

4.4.4 Predictive capacity

As IL1-α was considered from the very beginning in a strategic manner, 24 MTT-positive chemicals were not tested for this second endpoint. Two chemicals had failed the variability criterion for all three runs, which were, when disregarding the quality assurance aspect, correctly classified as positives. Of the 24 chemicals, 18 were correctly classified as positives and six wrongly as positives (numbers 2, 5, 7, 8, 17, 26). Of the remaining 34 chemicals, seven chemicals had a label (R38) and 27 had not. In Table 64, the data for two kinds of prediction model are summarised: the model, which bases on the relative increase compared to the negative control, i.e. 5-fold increase, and the results, i.e. p-values, of 1-way ANOVAs with a Dunnett-post test comparing a given chemical's response to the response of the respective negative control. The later was applied to the raw data (data not shown) as well as to the logarithmically transformed data. Although the differences are minor, the analysis of the transformed data is to be preferred as the variances were more homogenous over the response range. To simplify the analysis, the results over the runs were combined: A chemical is overall classified as irritant if the mean fold increase is larger than 5, or if a chemical gave in at least two runs a significantly (p<0.05) higher response than the negative control (logarithmically transformed data).

chemical number	chemical number	fold increase			mean	sd	CV [%]	ANOVA p-value (log data)		
		1. run	2. run	3. run				1. run	2. run	3. run
6	no label	1.38	1.91	1.68	1.66	0.266	16.04	> 0.05	**< 0.01**	**< 0.05**
9	no label	0.89	0.73	0.72	0.78	0.095	12.23	> 0.05	> 0.05	> 0.05
10	no label	1.04	1.00	0.93	0.99	0.056	5.62	> 0.05	> 0.05	> 0.05
11	no label	0.96	1.37	0.87	1.07	0.267	24.99	> 0.05	> 0.05	> 0.05
12	no label	1.69	0.90	1.31	1.30	0.395	30.39	> 0.05	> 0.05	> 0.05
16	no label	1.20	1.29	1.09	1.19	0.100	8.39	> 0.05	> 0.05	> 0.05
19	no label	1.00	1.09	0.95	1.01	0.071	7.00	> 0.05	> 0.05	> 0.05
21	no label	1.39	1.09	1.11	1.20	0.168	14.02	> 0.05	> 0.05	> 0.05
22	no label	0.90	1.18	1.29	1.12	0.201	17.90	> 0.05	> 0.05	> 0.05
23	R38	4.12	3.07	2.37	*3.19*	0.881	27.64	**< 0.01**	**< 0.01**	**< 0.01**
24	no label	0.90	0.96	0.83	0.90	0.065	7.26	> 0.05	> 0.05	> 0.05
25	no label	0.94	1.06	1.41	1.14	0.244	21.48	> 0.05	> 0.05	> 0.05
27	R38	4.79	3.15	4.30	*4.08*	0.842	20.63	**< 0.01**	**< 0.01**	**< 0.01**
28	no label	0.99	0.90	1.13	1.01	0.116	11.51	> 0.05	> 0.05	> 0.05
30	no label	1.13	1.10	1.58	1.27	0.269	21.17	> 0.05	> 0.05	> 0.05
32	no label	1.50	0.92	0.94	1.12	0.329	29.40	> 0.05	> 0.05	> 0.05
33	no label	1.17	1.06	0.85	1.03	0.163	15.84	> 0.05	> 0.05	> 0.05
34	R38	0.87	1.13	0.72	0.91	0.207	22.88	> 0.05	> 0.05	> 0.05
35	no label	0.61	1.24	1.14	1.00	0.339	33.97	> 0.05	> 0.05	> 0.05
36	no label	0.89	0.87	1.03	0.93	0.087	9.37	> 0.05	> 0.05	> 0.05
39	no label	0.85	0.88	1.04	0.92	0.102	11.06	> 0.05	> 0.05	> 0.05
41	no label	0.92	0.75	0.71	0.79	0.112	14.06	> 0.05	> 0.05	> 0.05
42	no label	0.87	1.02	0.52	0.80	0.257	31.94	> 0.05	> 0.05	> 0.05
44	no label	1.03	1.24	1.04	1.10	0.118	10.74	> 0.05	> 0.05	> 0.05
48	no label	0.88	0.91	1.23	1.01	0.194	19.27	> 0.05	> 0.05	> 0.05
49	R38	4.07	3.79	2.96	*3.61*	0.577	16.01	**< 0.01**	**< 0.01**	**< 0.01**
50	no label	1.97	1.39	1.94	1.77	0.327	18.48	> 0.05	> 0.05	**< 0.01**
51	R38	3.89	1.97	1.63	*2.50*	1.219	48.81	**< 0.01**	> 0.05	> 0.05
52	no label	1.80	1.94	1.46	1.73	0.247	14.24	> 0.05	> 0.05	> 0.05
53	no label	1.75	0.93	1.06	1.25	0.441	35.35	> 0.05	> 0.05	> 0.05
54	no label	**5.99**	4.17	2.66	*4.27*	1.667	39.02	**< 0.01**	**< 0.01**	**< 0.01**
55	R38	**5.93**	3.60	2.96	*4.16*	1.563	37.54	**< 0.01**	**< 0.01**	**< 0.01**
57	no label	1.25	0.88	0.63	0.92	0.312	33.91	> 0.05	> 0.05	> 0.05
60	R38	0.77	0.54	0.62	0.64	0.117	18.15	> 0.05	**< 0.05**	> 0.05

Table 64: Summary of Sanofi IL1-α data of the tested chemicals and positive controls expressed as fold-increase relative to the respective negative controls together with 1-way ANOVA results of comparing the logarithmically transformed pg/ml-data with the respective negative control data (with Dunnett's post test)

As no chemical induced a mean 5-fold increase of IL1-α release, this threshold was set to 2-fold. Considering this 2-fold PM and the ANOVA-PM, both models resulted in similar prediction with the exception of chemical 51, which was correctly classified as positive with the 2-fold PM. The resulting predictive capacities are only presented for the prediction model based on the fold-increase. Of the 34 chemicals, five are correct positive, two false negative, one false positive and 26 correct negative. Including the 24 chemicals with MTT-data only, the strategy finally resulted in a specificity of 26/33=78.8% and a sensitivity of 23/25=92.0%. Compared the predictive capacities of the MTT alone, the increase in sensitivity (24.0% points) is achieved by a smaller loss in specificity (3.0% points).

4.5 IL1-α: EPISKIN between-laboratory variability

4.5.1 Controls

A comparison of the response of the controls was performed not only to describe this aspect of variability but also to explore differences of predictions between the laboratories. Each laboratory provided ten data points for each control, which are summarised in Table 65 and Figure 46. While the responses of the positive control fell in the same range, the responses of the negative controls differed substantially between the laboratories. Transforming these data logarithmically and analysing both controls separately with a 1-way ANOVA and a Bonferroni post-test, resulted in significant differences ($p<0.001$) between all pairs of laboratories for the negative control. Regarding, the positive control, only L'Oréal and Unilever differed significantly ($p< 0.001$).

The mean response of the negative control being 32.1 pg/ml at Sanofi in contrast to 5.8 pg/ml (L'Oréal) and 15.4 pg/ml (Unilever), did not allow finding an appropriate common prediction model based on relative response to the negative control, i.e. fold-increase. Instead, a prediction model based on direct comparison of responses was proposed: By a 1-way ANOVA with a Dunnett post test applied to the logarithmically transformed IL1-α data, chemicals giving a significantly different/larger response than the respective negative control were identified. The respective p-values are included in Tables 60, 62 and 64.

	L'Oréal		Unilever		Sanofi	
	mean IL1-α [pg/ml]	SD	mean IL1-α [pg/ml]	SD	mean IL1-α [pg/ml]	SD
NC	4.52	2.38	13.93	4.30	26.76	11.12
	4.87	1.98	15.42	7.66	26.05	6.38
	4.41	1.19	14.22	11.17	45.90	16.03
	8.64	3.11	18.31	11.34	31.37	3.98
	3.70	0.28	16.30	7.89	27.86	4.65
	7.36	4.92	19.31	4.70	30.57	6.55
	2.95	0.24	14.36	9.29	33.08	7.73
	8.16	7.26	20.60	3.21	28.62	4.37
	5.84	0.78	8.49	5.97	38.16	10.68
	7.83	2.28	13.50	2.06	32.35	1.54
PC	198.76	31.73	320.29	92.14	228.07	65.91
	282.99	21.54	343.36	90.14	403.68	58.38
	331.67	32.84	279.26	21.39	277.11	47.08
	239.40	24.65	404.27	28.08	396.25	43.23
	267.09	24.63	421.47	47.37	280.56	82.53
	183.11	11.60	416.36	63.55	215.93	10.02
	185.82	23.96	409.78	46.32	360.63	68.76
	284.47	59.99	349.18	76.90	254.78	58.19
	234.51	31.49	417.25	27.48	268.71	61.77
	335.45	58.65	415.59	41.09	385.97	111.56

Table 65: Mean IL1-α amount in pg/ml and corresponding standard deviation (SD) for all negative and positive controls in the three EPISKIN-laboratories

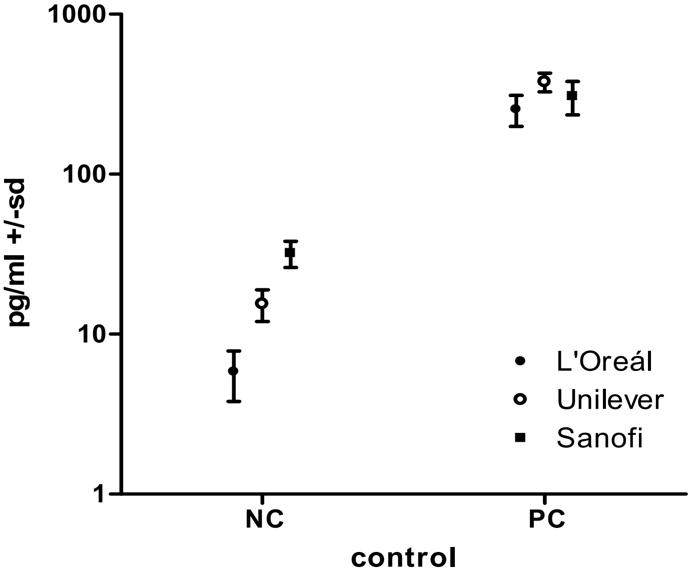


Figure 46: Summary of IL1-α data of the negative and positive controls in the three EPISKIN laboratories plotted on a logarithmic scale

4.5.2 Between-laboratory reproducibility

The reproducibility of test chemicals between laboratories will not be assessed in detail as substantial differences become evident from the preceding analysis. In Table 66 the chemicals, which were tested in at least two laboratories are listed. The reproducibility was assessed by the standard deviation and the CV of the mean IL1-α in the different laboratories. The mean standard deviation was 14.37, while the mean CV was about 50% (data not shown). As already seen in the analysis of within-laboratory reproducibility, the large between-laboratory reproducibility reflected lack of standardisation of this secondary endpoint.

chemical	chemical	mean IL1-α [pg/ml]			mean	sd	CV [%]
number	class (EU)	L'Oréal	Unilever	Sanofi			
6	no label	n.d.	37.64	49.36	43.50	8.29	19.06
9	no label	6.11	23.55	25.28	18.31	10.60	57.88
10	no label	9.06	35.1	32.25	25.47	14.28	56.07
11	no label	8.06	14.8	33.73	18.86	13.31	70.56
12	no label	11.16	33.16	41.71	28.68	15.76	54.96
16	no label	13.59	13.94	35.63	21.05	12.62	59.94
19	no label	9.46	18.06	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	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	15.1	33.2	33.95	27.42	10.67	38.92
28	no label	9.13	14.3	30.25	17.89	11.01	61.53
30	no label	15.07	32.39	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	13.7	31.47	32.69	25.95	10.63	40.96
35	no label	4.11	18.8	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	12.85	26.05	15.65	9.32	59.54
42	no label	9.52	20.04	24.6	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	33.2	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	103.50	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	8.2	25.09	28.63	20.64	10.92	52.91
49	R38	40.42	92.53	107.87	80.27	35.36	44.05
51	R38	33.58	60.35	76.7	56.88	21.77	38.28
55	R38	121.72	n.d.	129.33	125.53	5.38	4.29
60	R38	5.02	n.d.	20.98	13.00	11.29	86.85

Table 66: Between-laboratory variability of IL1-α of the three EPISKIN laboratories
(n.d.: not done)

4.6 Evaluation of a common EPISKIN prediction model

As obvious from the previous analyses, an optimal, generally applicable threshold of a PM based on the fold-increase could not be identified. The main reason for this was the different response levels of the negative controls in the laboratories, which indicate a need for further standardisation. The AOV-PM, which is based on the absolute difference of a sample's responses versus the response of the negative control, could be uniformly applied to the three laboratories. However, it performed not well in one of the laboratories, i.e. Unilever. Thus, a third PM based on the total IL1-α amount released (PM-total) was proposed. As this PM did not include the negative control, the difference in responses of the negative controls between the laboratories was circumvented.
In Figure 47 the absolute IL1-α amount released for all runs of all tested substances taken from Tables 58, 60 and 63 structured by laboratory and substance classification, i.e. non-irritant vs. irritant, are displayed.

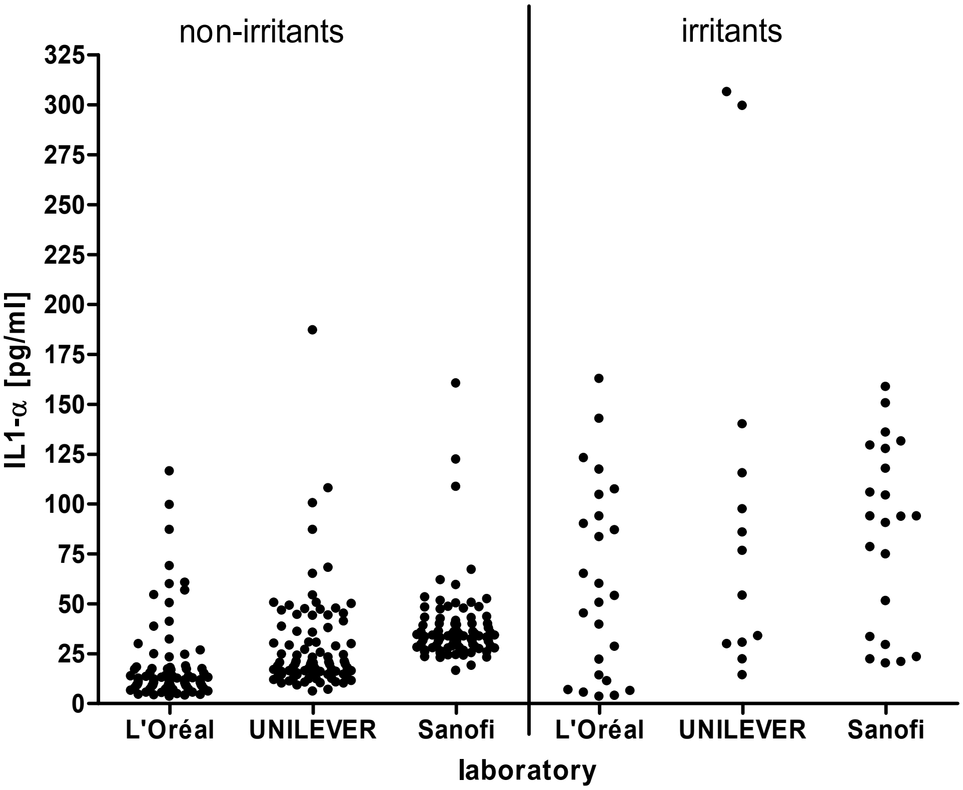


Figure 47: IL1-α release of all runs and substances tested in the three EPISKIN-laboratories

As irritant chemicals tended to induce higher IL1-α amount, the PM-total was evaluated in detail by a ROC-analysis to identify an optimal threshold value. Considering IL1-α as a stand-alone endpoint, thresholds between 48 and 54 pg/ml resulted in sums of specificity and sensitivity larger than 1.50, where the threshold of 51 pg/ml reached the maximum of 1.523 (Figure 48).

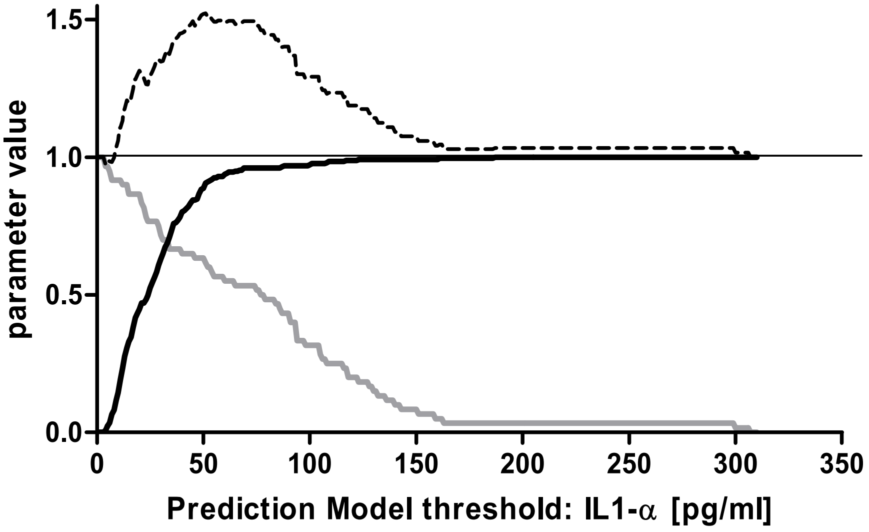


Figure 48: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold for IL1-α
(black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity).

When considering the mean IL1-α release of all runs on the level of the individual laboratories, this PM performed similar. At L'Oréal four additional correctly classified irritants were identified, while three non-irritants were falsely classified as positives (Table 67). At Unilever, three correct positives come along with three false positives, while at Sanofi five correct positives and three false positives were generated. In addition, Table 67 also demonstrates that variability issues persisted. The L'Oréal laboratory showed for some chemicals, i.e. chemical number 50, 49 and 51, aberrantly low IL1-α values. Furthermore, this PM gave good between-laboratory reproducibility of prediction for non-irritant chemicals: twenty-two of 25 chemicals tested in three laboratories were classified consistently. In contrast, only one of three irritant chemicals tested in three laboratories was classified consistently.

chemical number	chemical class (EU)	mean IL1-α of all runs [pg/ml]		
		L'Oréal	Unilever	Sanofi
5	no label	19.67	-	-
6	no label	-	37.64	49.36
7	no label	**58.56**	-	-
8	no label	40.42	-	-
9	no label	6.11	23.55	25.29
10	no label	9.06	35.10	32.25
11	no label	8.06	14.80	33.73
12	no label	11.16	33.16	41.72
16	no label	13.59	13.95	35.63
19	no label	9.46	19.54	30.64
21	no label	9.18	32.91	39.01
22	no label	13.79	21.00	33.54
24	no label	11.98	16.33	31.18
25	no label	15.10	33.20	33.95
28	no label	9.13	14.30	30.25
30	no label	15.07	32.38	38.14
32	no label	5.18	11.32	33.86
33	no label	13.70	31.47	32.69
35	no label	4.11	18.80	33.62
36	no label	12.27	21.68	31.88
39	no label	11.05	12.63	30.48
41	no label	8.06	12.84	26.05
42	no label	9.52	20.04	24.60
44	no label	10.01	31.32	35.77
48	no label	10.73	14.28	33.20
50	no label	12.49	**84.17**	**53.33**
52	no label	44.55	**75.61**	**55.23**
53	no label	**51.89**	**58.69**	40.01
54	no label	**76.61**	-	**130.38**
57	no label	9.44	13.55	28.47
23	R38	**84.76**	-	**98.74**
27	R38	**96.85**	-	**133.12**
34	R38	8.20	25.10	28.63
43	R38	**89.91**	-	-
49	R38	40.42	**92.53**	**107.87**
51	R38	33.58	**60.35**	**76.71**
55	R38	**121.72**	-	**129.33**
56	R38	-	**248.56**	-
60	R38	5.02	-	20.97

Table 67: Mean IL1-α of all available runs per laboratory, where the bold values indicate a classification as R38/irritant when considering 51 pg/ml (bold type) or 60 pg/ml (grey cells) as threshold of the PM-total

Overall predictivity of the strategic combination of both endpoints was performed with the mean IL1-α amount (Table 67). The results are summarised in Figure 49, where for example the predictive capacity of MTT alone, i.e. specificity of 82.8% and sensitivity of 74.7%, is displayed for thresholds above 248 pg/ml, i.e. the highest response of all runs. Choosing the best IL1-α threshold in terms of the sum of the predictive parameters of the stand-alone approach, 51 pg/ml was used resulting in a specificity of 73/99=73.7% and a sensitivity of 68/75=90.7%. Increasing the predictive capacity in terms of the parameter sum, the PM-threshold of 59 and 60 pg/ml resulted in a specificity of 78/99=78.8% and a sensitivity of 68/75=90.7%. Thus, the second endpoint IL1-α could increase the overall predictive capacity in terms of the sum of specificity and sensitivity in comparison to MTT alone. In addition, IL1-α could be used for shifting the balance between specificity and sensitivity. However, when interpreting these data, it has to be kept in mind that the IL1-α PM was optimised post-hoc, which usually results in an overestimation of predictive capacity.

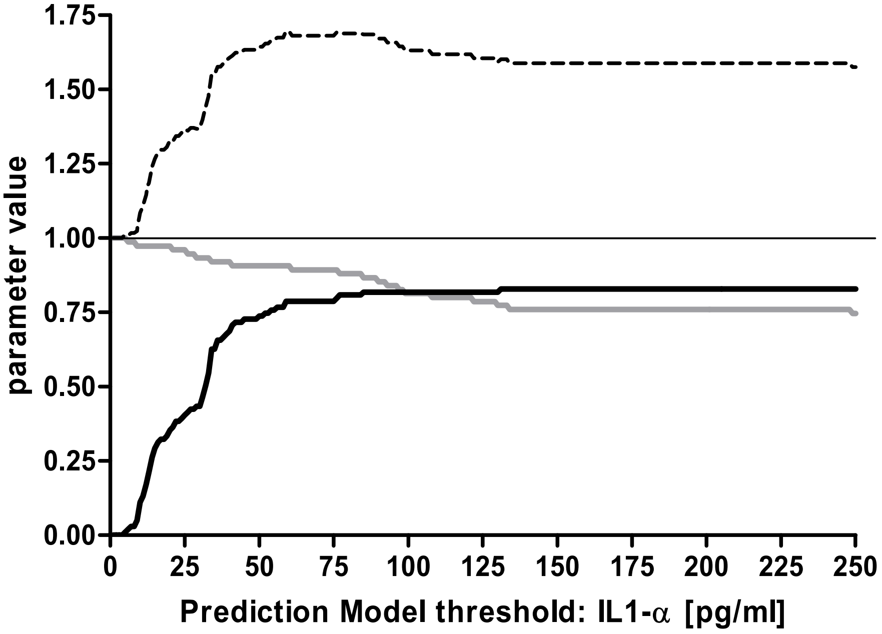


Figure 49: Curves of sensitivity, specificity and their sum depending on the in vitro Prediction Model threshold for IL1-α when considered in the proposed strategic manner together with MTT
(black line: specificity; grey line: sensitivity, dotted line: sum of sensitivity and specificity)

Finally, the predictive capacity of the strategy when moving the in vivo threshold for classification was analysed. When trying to discriminate GHS-irritants from GHS-mild and non-irritants the combination of endpoints resulted with an IL1-α prediction threshold of 60pg/ml in a specificity of 63.0% and a sensitivity of 100%. Moving the IL1-α to 80pg/ml increased the specificity to 65.9% maintaining a sensitivity of 100%. Discriminating GHS-irritants and GHS-mild irritants from GHS-non-irritants resulted for the IL1-α prediction threshold of 60pg/ml in a specificity of 79.8% and a sensitivity of 80.0%. This predictive capacity could not be improved by moving the IL1-α prediction.

In a last step, it was attempted to construct post–hoc a PM for the GHS classification system. Due to the poor performance of MTT to discriminate the three classes and of IL1-α above, which suggest already that EPISKIN cannot predict three classes, this was done in a simplified manner. A dataset was constructed with the data from all laboratories and had 162 entries, for all which three valid MTT runs were available. Seventy-one entries had viability smaller than 50%. To keep this analysis simple and disregarding reproducibility, for both endpoints only the mean over the respective runs was considered. Therefore, all of these had to be considered as GHS-irritants. The remaining 91 entries comprised three were GHS-irritants, 24 GHS-mild irritants and 64 GHS-non irritants. As seen in the analysis of the endpoint MTT, this viability did not allow discriminating GHS-non irritants from GHS-mild irritants. Therefore, the focus was put only on the endpoint IL1-α to discriminate these two classes. Subjecting all entries with viability above 50% and an IL1-α release below 60 pg/ml, i.e. the threshold correctly classifying all GHS-irritants, to a ROC-analysis resulted in a curve always close to or below the line of identity (data not shown). Thus, with the dataset available no satisfactorily performing PM for EPISKIN measuring both endpoints could be constructed.

4.7 Summary of EPISKIN predictive capacity

To summarise the predictive capacity of EPISKIN, both endpoints were considered in the strategic manner described above. The MTT prediction model was maintained, i.e. a cut-off value at 50% viability, and, as MTT was not able to distinguish the three GHS-classes, prediction of the European classification system assessed. However, for simplicity it was applied to the mean viability over all available runs for a given chemical, which was also the main criterion whether to test for IL1-α. For IL1-α, the prediction model based on the mean total IL1-α amount released over three runs was applied, because it could be applied to all data generating useful results. The classifications for both endpoints and all laboratories are summarised in Table 68.

chemical number	chemical	EU class	MTT classification (based on mean viability over all runs per laboratory)			IL1-α classification (Cut-Off: 60 mg/ml)			combined classification			combined classification (median over all laboratories)
			L'Oréal	Unilever	Sanofi	L'Oréal	Unilever	Sanofi	L'Oréal	Unilever	Sanofi	
1	2-chloromethyl-3,5-dimethyl-4-methoxypyridine hydrochloride	R38	1	1	1				1	1	1	1
2	1-bromo-4-chlorobutane	no label	1	1	1				1	1	1	1
3	1-bromohexane	R38	1	1	1				1	1	1	1
4	1-decanol	R38	1	1	1				1	1	1	1
5	3-chloro-4-fluoronitrobenzene	no label	0	1	1	0			0	1	1	1
6	3-diethylaminopropionitrile	no label	1	0	0		0	0	1	0	0	0
7	3-mercaptohexanol	no label	0	1	1	0			0	1	1	1
8	4-methylthio-benzaldehyde	no label	0	1	1	0			0	1	1	1
9	2,6-dimethyl-4-nitrobenzeneamine	no label	0	0	0	0	0	0	0	0	0	0
10	allyl heptanoate	no label	0	0	0	0	0	0	0	0	0	0
11	allyl phenoxyacetate	no label	0	0	0	0	0	0	0	0	0	0
12	2-ethylhexyl 4-aminobenzoate	no label	0	0	0	0	0	0	0	0	0	0
13	1-[4-(2-dimethylaminoethoxy)phenyl]-2-phenylbutan-1-one	R38	1	1	1				1	1	1	1
15	a-terpineol	R38	1	1	1				1	1	1	1
16	capryl-isostearate	no label	0	0	0	0	0	0	0	0	0	0
17	2-methyl-3-[(1,7,7-trimethylbicyclo[2.2.1]hept-2-yl)oxy]-1-propanol, bornyl isomer	no label	1	1	1				1	1	1	1
18	butyl methacrylate	R38	1	1	1				1	1	1	1
19	2,5-dimethyl-4-oxo-4,5-dihydrofuran-3-yl acetate	no label	0	0	0	0	0	0	0	0	0	0
20	cyclamen aldehyde	R38	1	1	1				1	1	1	1
21	A mixture of: 5-exo-decylbicyclo[2.2.1]hept-2-ene; 5-endo-decylbicyclo[2.2.1]hept-2-ene	no label	0	0	0	0	0	0	0	0	0	0
22	diethyl phthalate	no label	0	0	0	0	0	0	0	0	0	0
23	di-n-propyl disulphide	R38	0	1	0	1		1	1	1	1	1
24	di-propylene glycol	no label	0	0	0	0	0	0	0	0	0	0
25	dipropylene glycol monobutyl ether	no label	0	0	0	0	0	0	0	0	0	0
26	3,4-dimethyl-1H-pyrazole	no label	1	1	1				1	1	1	1
27	2-isopropyl-2-isobutyl-1,3-dimethoxypropane	R38	0	1	0	1		1	1	1	1	1

28	ethyl cis-4-[4-[[2-(2,4-dichlorophenyl)-2-(1H-imidazol-1-ylmethyl)-1,3-dioxolan-4-yl]methoxy]phenyl]piperazine-1-carboxylate	no label	0	0	0	0	0	0	0	0	0	0
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	R38	1	1	1				1	1	1	1
30	Mixture of: diethyl cis-1,4-cyclohexanedicarboxylate; diethyl trans-1,4-cyclohexanedicarboxylate	no label	0	0	0	0	0	0	0	0	0	0
31	A 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	R38	1	1	1				1	1	1	1
32	2S-(2-furyl)-5R-hydroxy-4R-(1R,2-dihydroxy)ethyl-6S-hydroxymethyl-1,3-dioxane	no label	0	0	0	0	0	0	0	0	0	0
33	heptyl butyrate	no label	0	0	0	0	0	0	0	0	0	0
34	hexyl salicylate	R38	0	0	0	0	0	0	0	0	0	0
35	cyclohexadecanone	no label	0	0	0	0	0	0	0	0	0	0
36	isopropanol	no label	0	0	0	0	0	0	0	0	0	0
37	[2-(cyclopentyloxy)ethyl]benzene(cyclopentyl 2-phenylethyl ether)	R38	1	1	1				1	1	1	1
39	methyl stearate	no label	0	0	0	0	0	0	0	0	0	0
40	1-methyl-3-phenyl-1-piperazine	R38	1	1	1				1	1	1	1
41	naphthalene acetic acid	no label	0	0	0	0	0	0	0	0	0	0
42	disodium 2,2'-(1,4-phenylene)bis-(1H-benzimidazole-4,6-disulfonic acid or monosulfonic acid, monosulfonate or disulfonate	no label	0	0	0	0	0	0	0	0	0	0
43	A mixture of isomers: 1-(1,1-dimethylpropyl)-4-ethoxy-cis-cyclohexane; 1-(1,1-dimethylpropyl)-4-ethoxy-trans-cyclohexane	R38	0	1	1	1			1	1	1	1
44	phenylethylalcohol	no label	0	0	0	0	0	0	0	0	0	0
45	(+/-) trans-3,3-dimethyl-5-(2,2,3-trimethyl-cyclopent-3-en-1-yl)-pent-4-en-2-ol	R38	1	1	1				1	1	1	1
46	4-methyl-8-methylenetricyclo[3.3.1.1(3,7)]decan-2-ol	R38	1	1	1				1	1	1	1
47	4-methyl-8-methylenetricyclo[3.3.1.1(3,7)]dec-2-yl acetate	R38	1	1	1				1	1	1	1
48	2-(formylamino)-3-thiophenecarboxylic acid	no label	0	0	0	0	0	0	0	0	0	0
49	isostearic acid monoisopropanolamide	R38	0	0	0	0	1	1	0	1	1	1
50	2-phenylhexanenitrile	no label	0	0	0	0	1	0	0	1	0	0
51	Mixture of isomers: 1-(2-isopropylphenyl)-1-phenylethane 1-(3-isopropylphenyl)-1-phenylethane 1-(4-isopropylphenyl)-1-phenylethane	R38	0	0	0	0	1	1	0	1	1	1
52	propyl (2S)-2-(1,1-dimethylpropoxy)-propanoate	no label	0	0	0	0	1	0	0	1	0	0
53	silane A-1430	no label	0	0	0	0	0	0	0	0	0	0

54	Mixture of isomers: 1-(spiro[4.5]dec-7-en-7-yl)pent-4-en-1-one 1-(spiro[4.5]dec-6-en-7-yl)pent-4-en-1-one	no label	0	1	0	1		1	1	1	1	1
55	terpinyl acetate	R38	0	1	0	1		1	1	1	1	1
56	benzenethiol, 5-(1,1-dimethylethyl)-2-methyl	R38	1	0	1		1		1	1	1	1
57	triethylene glycol	no label	0	0	0	0	0	0	0	0	0	0
58	tri-isobutyl phosphate	R38	1	1	1				1	1	1	1
59	(E,E)-3,7,11-trimethyldodeca-1,4,6,10-tetraen-3-ol	R38	1	1	1				1	1	1	1
60	bis[(1-methylimidazol)-(2-ethyl-hexanoate)], zinc complex	R38	0	1	0	0		0	0	1	0	0

Table 68: Summary of EPISKIN predictions with both endpoints of all laboratories

The resulting predictive capacities for MTT alone and the strategic prediction are shown in Table 69. A substantial increase in sensitivity of 16% points is achieved by a loss of 4% points in specificity. In comparison, EpiDerm resulted for MTT in a specificity of 86.39% and a sensitivity of 57.33%. The AOV-PM resulted in a specificity of 74.75% and a sensitivity 89.33%.

	n	MTT	MTT + PM-total
specificity	99	82.83%	78.79%
sensitivity	75	74.67%	90.67%

Table 69: Specificity and sensitivity summarised over all EPISKIN-laboratories for MTT alone and the combined prediction based on MTT and the total amount of IL1-α released

Finally, the data were further summarised by considering the median classification over the three laboratories for a given chemical. The results with corresponding confidence intervals are presented in Table 70.

	specificity		sensitivity	
		CI		CI
EPISKIN (MTT)	26/33 = 78.8%	61.1% - 91.0%	19/25 = 76.0%	54.9% - 90.6%
EPISKIN (MTT + IL1-α)	26/33 = 78.8%	61.1% - 91.0%	23/25 = 92.0%	74.0% - 99.0%

Table 70: Specificity and sensitivity considering median classification over the EPISKIN-laboratories for MTT alone and the combined prediction based on MTT and the total amount of IL1-α released with 95% confidence intervals

References

1. OECD. OECD Guideline for Testing of Chemicals No. 404: Acute Dermal Irritation/Corrosion. 1-13. 2002. Paris, Organisation for Economic Cooperation and Development.

2. EC. Annex VI of the Directive 67/548/EEC. General Classification and Labelling Requirements for dangerous substances and preparations. *Official Journal of the European Communities* 2001; **L225:** 263-314.

3. UN. Skin Corrosion/Irritation. *UN, Globally harmonized system of classification and labelling of chemicals*. New York and Geneva: UN 2003: 123-135.

4. ECETOC. Skin Irritation and Corrosion: Reference Chemicals Data Bank. Technical Report No. 66, 1-247. 1995. Brussels, European Centre for Ecotoxicology and Toxicology of Chemicals.

5. ISO. Part 4: Measurement methods and results: ISO 5725-1: Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions. In: ISO, ed. *Statistical methods for quality control Vol.2*. Geneva: International Organization for Standardization 1994.

6. Draize JH, Woodard G, Clavery HO. Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membranes. *Journal of Pharmacology and Experimental Therapeutics* 1944; **82:** 377-390.

7. Hoffmann S, Cole T, Hartung T. Skin irritation: prevalence, variability and regulatory classification of existing in vivo data from industrial chemicals. *Regulatory Toxicology and Pharmacology* 2005; **41:** 159-166.

8. Altman DG, Bland JM. Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 1994; **309:** 188.

9. van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. *Investigative Radiology* 1995; **30:** 334-340.

10. Dunnett CW. New tables for multiple comparisons with a control. *Biometrics* 1964; **20:** 482-491.

11. Fentem JH, Archer GEB, Balls M *et al.* The ECVAM international study on in vitro tests for skin corrosivity. 2. Results and evaluation by the management team. *Toxicology in Vitro* 1998; **12:** 483-524.

ANNEX I

Skin Irritation Validation Study Phase I

Introduction

In order to evaluate alternative methods for skin irritation testing, ECVAM currently sponsors a formal validation study of two *in vitro* and one *ex vivo* test system. The aim of this study is to validate *in vitro* skin irritation tests in a formal interlaboratory study, in order to replace the Draize skin irritation test performed on rabbits according to Method B.4 of Annex V to *Directive 67/548/EEC* or OECD TG 404. The primary goal of this validation study is the scientific evaluation of the ability of the *in vitro* tests to reliably discriminate skin irritants (I) from non-irritants (NI), as defined with EU risk phrases (R38; no label) according to the Dangerous Substances Directive, *67/548/EEC*. A secondary goal of this study is to retrospectively analyse the data to assess if the *in vitro* tests reliably discriminate between strong, mild and non-irritants, as defined by the 'Globally Harmonised System (GHS)' for classification and labelling, adopted by the United Nations.

Material and Methods

The two *in vitro* test systems are the EPISKIN and the EpiDerm and the *ex vivo* system is the skin integrity function test (SIFT). The objective of the validation study is to assess the assays' reliability (within and between laboratories) and their relevance (predictive capacity). The validation study is divided into two phases. In the first phase, which is analysed here, twenty blinded chemicals were tested in the lead laboratories of the test systems in three independent runs. This phase allows a preliminary assessment of the within-laboratory reproducibility and the predictive capacity. The lead laboratory for EPISKIN, EpiDerm and SIFT are L'Oréal (France), ZEBET (Germany) and Syngenta (UK).

Both the 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 application of the chemicals, namely trans-epithelial water loss (TEWL) and electrical resistance (ER).

Within-laboratory variability

The within-laboratory variability was analysed with a maximum of four statistical techniques. These range from very rigorous, i.e. aiming to detect optimal reproducibility, to less demanding approaches and they give a complete insight. The EPISKIN- and EpiDerm-data allowed applying a 2-way ANOVA, the most rigorous tool, in which the factors '(experimental) run' and 'chemical', i.e. the blinded chemicals, were modelled. The ANOVA-results regarding the 'run' in terms of the p-value and the relative mean square error are first indicators of the within-laboratory variability. As this model most likely results in significant results due to the large number of chemicals (n = 20), a less rigorous 1-way ANOVA with a Bonferroni post-hoc test comparing the data of the three runs for each single chemical was applied subsequently. Also the data structure of the SIFT allowed this second analysis. For both ANOVA-techniques a significance level of 1% was chosen. In a third step, the correlation according to Bravais-Pearson was calculated for EPISKIN and EpiDerm to compare all three pairs of runs. The SIFT-data did not allow for a meaningful assessment of the correlation. Finally and applicable to all test systems, the predicted classification resulting from the prediction models (PM) were compared between the runs by a simple measure of similarity, i.e. the proportion of identical predictions when comparing all pairs of runs.

Predictive capacity

As the test systems were designed to predict the EU risk phrases, i.e. R38 for skin irritants and no label for non-irritants, the predictions and the respective European classification of the chemicals were combined in 2x2 contingency tables. From these tables the predictive capacity was calculated in terms of sensitivity, specificity, accuracy and positive and negative predictive value (PPV, NPV). ROC-curve analysis was performed to check how shifting of the PM-thresholds of the test systems to discriminate irritants from non-irritants affects the predictive capacity. The sum of sensitivity and specificity was the parameter chosen to assess the ROC, where reproducibility of prediction between the runs was incorporated as a necessary condition. Additionally, the *in vivo* test data, which were used to classify the employed chemicals, were correlated with the endpoints of the new test systems. Therefore, the concept of the dominating median was applied in order to reduce the *in vivo* data to a one-dimensional measure while the loss of information was minimized. Extracting the median for each of the endpoints of the *in vivo* experiment, i.e. erythema and oedema, and choosing the larger one results in the dominating median of a given chemical. In order to maintain the blinding, the data are not shown, but only the correlation coefficients are reported.

The secondary aim, the assessment of the test systems performance in terms of the Globally Harmonised System (GHS) was done in a post-hoc analysis. As no PMs were available, per test method two thresholds were chosen aiming to maximise the accuracy while a high reproducibility between the runs in terms of prediction was included as a condition. In case of ambiguous prediction of a chemical between runs the two identical of the three classifications were chosen. The GHS-classifications of the twenty chemicals were assigned according to their *in vivo* data. Considering the small sample size of Phase 1 and the data-driven nature of the chosen approach, the results will overestimate the test performance.

Results

EpiDerm

ZEBET, the lead laboratory for the EpiDerm assay, submitted the data to ECVAM on 25.05.2004. One operator tested all twenty chemicals in three runs (dates of the runs: 30.04., 05.05., 14.05.). Two chemicals (chemicals code: 33 and 57) were retested once because they did not fulfil the variability criteria of acceptance, i.e. a coefficient of variation (CV) below 30%, in the second run. The data were received on the 14.06.2004 and replaced the respective data in the second run. Despite application problems with chemical 57 in the first run, no further remarks were reported. The results of a 2-way ANOVA are given in Table 1. The respective data are presented in Figure 1.

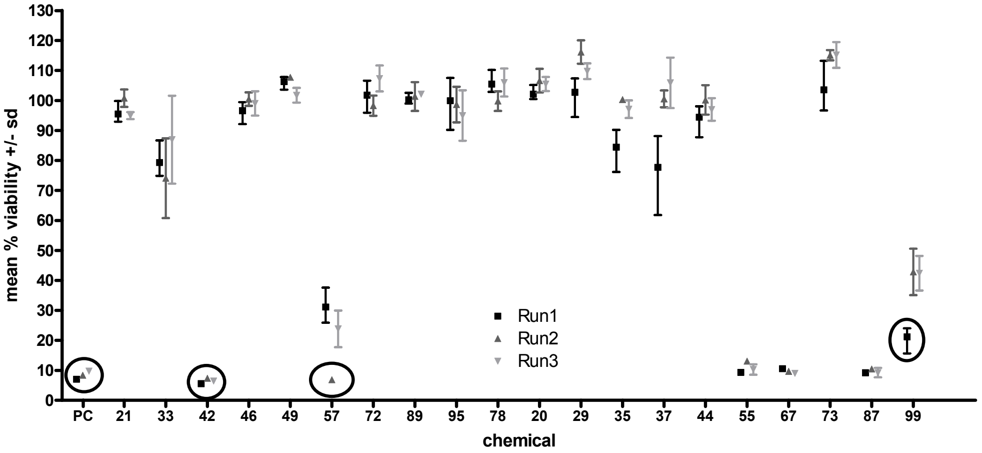


Figure 1: Phase-1 data from ZEBET with EpiDerm (PC: positive control). The encircled chemicals showed significant differences between runs in an one-way ANOVA.

Source of variation	degrees of freedom	sum of squares	Mean Square	Relative mean square	F-value	p-value
Run	2	664.1	332.1	0.020	12.8	8.5E-06
Chemical	20	318835.1	15941.8	0.972	616.0	0
Interaction	40	4251.7	106.3	0.006	4.1	7.5E-10
Residuals	126	3260.8	25.9	-	-	-
Σ	188	327011.7	16406.0	-	-	-

Table 1: 2-way ANOVA for the EpiDerm data from ZEBET (Phase 1)

Although the three model parameters ‘run’, ‘chemical’ and ‘interaction’ are highly significant, the chemicals account for more than 97% of the variation in terms of the relative mean square. The differences between runs were low (relative mean square: 2%) indicating a good reproducibility.

Calculating an ANOVA for each of the twenty chemicals and the positive control resulted in significant p-values smaller than 1% for chemicals 42, 57, 99 and the positive control. Thus, the major part of chemicals was well reproducible between the runs. Focusing on the significant results of the four chemicals (encircled data in figure 1) revealed that chemical 42 and the PC had low variability within each run so that a minor viability increase of 1%-2% caused significance. For chemical 57, the run, which had to be repeated, gave a significantly lower viability. Considering chemical 99, the first run resulted in a significantly lower viability.

Taking the mean values per run into account and correlating these with each other resulted in a value of 0.973 when comparing the first with the second run, in 0.980 when comparing the first and the third run and in 0.990 when comparing the second and the third run.

Applying the PM, i.e. classifying the chemicals by the threshold of 50% as either irritants (<50%) or non-irritants, resulted in identical classifications between the runs, i.e. a similarity of 100%.

The predictive capacity in terms of sensitivity, specificity, accuracy, PPV and NPV of the EpiDerm in the lead-laboratory together with the respective 2x2-contingency table is shown in table 2. The accuracy of 75% indicates a promising overall performance of the test method. All misclassifications were chemicals with borderline *in vivo* scores, i.e. around the classification threshold of the European system of 2.

EpiDerm		European classification		
		no label	R38	Σ
PM	non-irritant	10	4	14
	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%

Table 2: 2x2-contigency table and predictive capacity for EpiDerm in Phase 1

A preliminary ROC-analysis revealed that all thresholds between 43% and 74% of viability would result in the maximum sum of sensitivity and specificity, i.e. 146.46%. Thus the SOP-threshold of 50% is chosen in a way that is reproducible and optimises the test performance.

Correlating the viability means of each run with the dominating median *in vivo* scores of the chemicals demonstrated a strong negative correlation (Bravais-Pearson) throughout (first run: -0.719; second run: -0.700; third run: -0.673).

In terms of the GHS, the performance of the test systems was derived from a contingency table (Table 3). The data driven threshold were chosen as 90% and 50%, i.e. chemicals with viability above 90% were classified as GHS-non-irritants, chemicals with viability between 50% and 90% as GHS-mild-irritants and chemicals with viability below 50% as GHS-irritants. With these thresholds, the reproducibility was reduced to a similarity of prediction of 93%.

EpiDerm		GHS-classification			Σ
		Non-irritant	Mild irritant	Irritant	
GHS-PM	Non-irritant	9	4	0	13
	Mild irritant	0	1	0	1
	Irritant	0	2	4	6
	Σ	9	7	4	20

Table 3: 3x3-contigency table according to the GHS for EpiDerm in Phase 1

Despite an accuracy of 70%, this data analysis indicates that EpiDerm is not capable to distinguish the three GHS-classes as the mild-irritants are assigned to all PM-classes.

EPISKIN

L'Oreal, the lead laboratory for the EPISKIN assay, submitted the data to ECVAM on 25.05.2004. One operator tested all twenty chemicals in six runs between 29.03.2004 and 17.05.2004 with ten chemicals per run. No remarks were reported.
The results of a 2-way ANOVA are given in table 4. The respective data are presented in Figure 3.

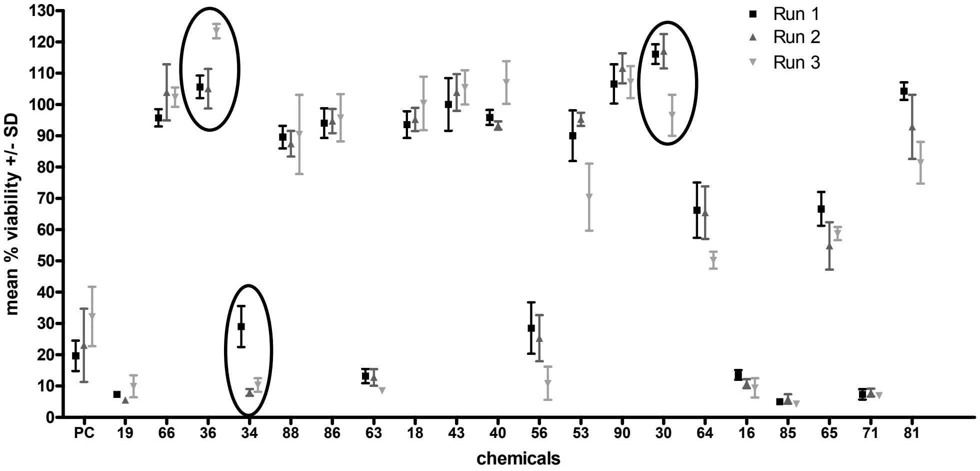


Figure 2: Phase-1 data from L'Oreal with EPISKIN (PC: positive control). The encircled chemicals showed significant differences between runs in an one-way ANOVA.

Source of variation	degrees of freedom	sum of squares	Mean Square	Relative mean square	F-value	p-value
Run	2	316.3	158.1	0.010	5.0	0.0081
Chemical	20	318384.4	15919.2	0.979	503.2	0
Interaction	40	5967.7	149.2	0.009	4.2	0
Residuals	126	3986.2	31.6	-	-	-
Σ	188	328654.6	16258.2	-	-	-

Table 4: 2-way ANOVA for the EPISKIN data from L'Oreal (Phase 1)

Besides the highly significant parameters 'chemical' and 'interaction', the parameter to assess reproducibility, 'run', is borderline significant with a relative mean square of 1%, indicating a good reproducibility.
Calculating an ANOVA for each of the twenty chemicals and the positive control resulted in significant p-values smaller than 1% for chemicals 36, 34 and 30. Thus, the major part of chemicals was well reproducible between the runs.

Focusing on the significant results of the three chemicals (encircled data in Figure 2), the Bonferroni post-test revealed for chemical 30 a significant lower viability in the third run, for chemical 34 a significant lower viability in the first run and for chemical 36 a significant higher viability in the third run.

Taking the mean values per run into account and correlating these with each other with the coefficient of correlation according to Bravais-Pearson resulted in a value of 0.989 when comparing the first with the second run, in 0.965 when comparing the first and the third run and in 0.970 when comparing the second and the third run.

Applying the PM, i.e. classifying the chemicals by the threshold of 50% as either irritants (<50%) or non-irritants, resulted in the identical classifications between the runs, i.e. a similarity of 100%.

The predictive capacity of the EpiDerm in the lead-laboratory together with the 2x2-contingency table is given in table 5. The accuracy of 80% indicates a promising overall performance of the test method. Again, all misclassifications were chemicals with borderline *in vivo* scores.

EPISKIN		European classification		
		no label	R38	Σ
PM	non-irritant	10	3	13
	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%

Table 5: 2x2-contigency table and predictive capacity for EPISKIN in Phase 1

The ROC-analysis, as indicated above, revealed that all thresholds between 30% and 50% of viability would result in an almost maximum sum of sensitivity and specificity, i.e. 157.58. The maximum sum of 159.60 would have been achieved with a threshold between 67% and 70% viability. As this threshold interval is small and entirely data-driven, the SOP-threshold of 50% is chosen in an optimal way.

Correlating the viability means of the runs with the dominating median *in vivo* scores of the chemicals, demonstrated a strong negative correlation (Bravais-Pearson) for each run (first run: -0.761; second run: -0.801; third run: -0.796).

In terms of the GHS, the performance of the test systems was derived from a contingency table (Table 6). The data driven threshold were chosen as 70% and 30%, i.e. chemicals with viability above 70% were classified as GHS-non-irritants, chemicals with viability between 30% and 70% as GHS-mild-irritants and chemicals with viability below 30% as GHS-irritants. With these thresholds, the reproducibility of the predictions between runs was maintained at 100%.

EPISKIN		GHS-classification			
		Non-irritant	Mild irritant	Irritant	Σ
GHS-PM	Non-irritant	9	2	0	11
	Mild irritant	0	2	0	2
	Irritant	0	3	4	7
	Σ	9	7	4	20

Table 6: 3x3-contigency table according to the GHS for EPISKIN in Phase 1

Despite an accuracy of 75%, this data analysis indicates that EPISKIN is not capable to distinguish the three GHS-classes as the mild-irritants are assigned to all PM-classes.

SIFT

Syngenta, the lead laboratory of the SIFT, submitted the data to ECVAM on 04.06.2004. One operator tested all twenty chemicals three times in a total of 17 experiments between the 16.03.2004 and the 04.05.2004. Two to four chemicals were tested per experiment. Several remarks were reported: for eleven of the total of 300 cells cell damage was observed; one chemical stained the cells; dissolving and dry skin was reported once each.

Although the SOP of the SIFT is lacking a formal procedure to deal with aberrant data, the Grubbs-test for outliers was applied with a significance level of 1%. Eight of the eleven damaged cells were identified as outliers. Additionally, three further outliers were detected. Nevertheless, the aberrant data are a minor issue, as only one of the outliers has an effect on the result of the PM.

Removing the outliers and analysing each of the chemicals with a 1-way ANOVA and a post-hoc Bonferroni (significance level of 1%) resulted for TEWL in no significant result and for ER in one significant result (chemical 15). Thus a good reproducibility is indicated. However, for several chemicals, e.g. 59 and 94 with TEWL or 91 with ER, the variability of the measurement within the runs prevented additional significant results.

Applying the PM, i.e. a TEWL-threshold of 10 and an ER-threshold of 4, and comparing the classifications between the runs per chemical, resulted for TEWL in ambiguous classifications between runs for the three chemicals 59, 60 and 61. For chemical 60, the aberrant run has a TEWL of 10.026 close to the threshold. For chemical 61, an outlier resulted in an aberrant run. Considering ER, five chemicals had ambiguous classifications (13, 15, 54, 61 and 70).

The predictive capacity of the SIFT in the lead-laboratory together with the 2x2-contingency table is given in table 7. The accuracy of 45% indicates a discouraging overall performance of the test method.

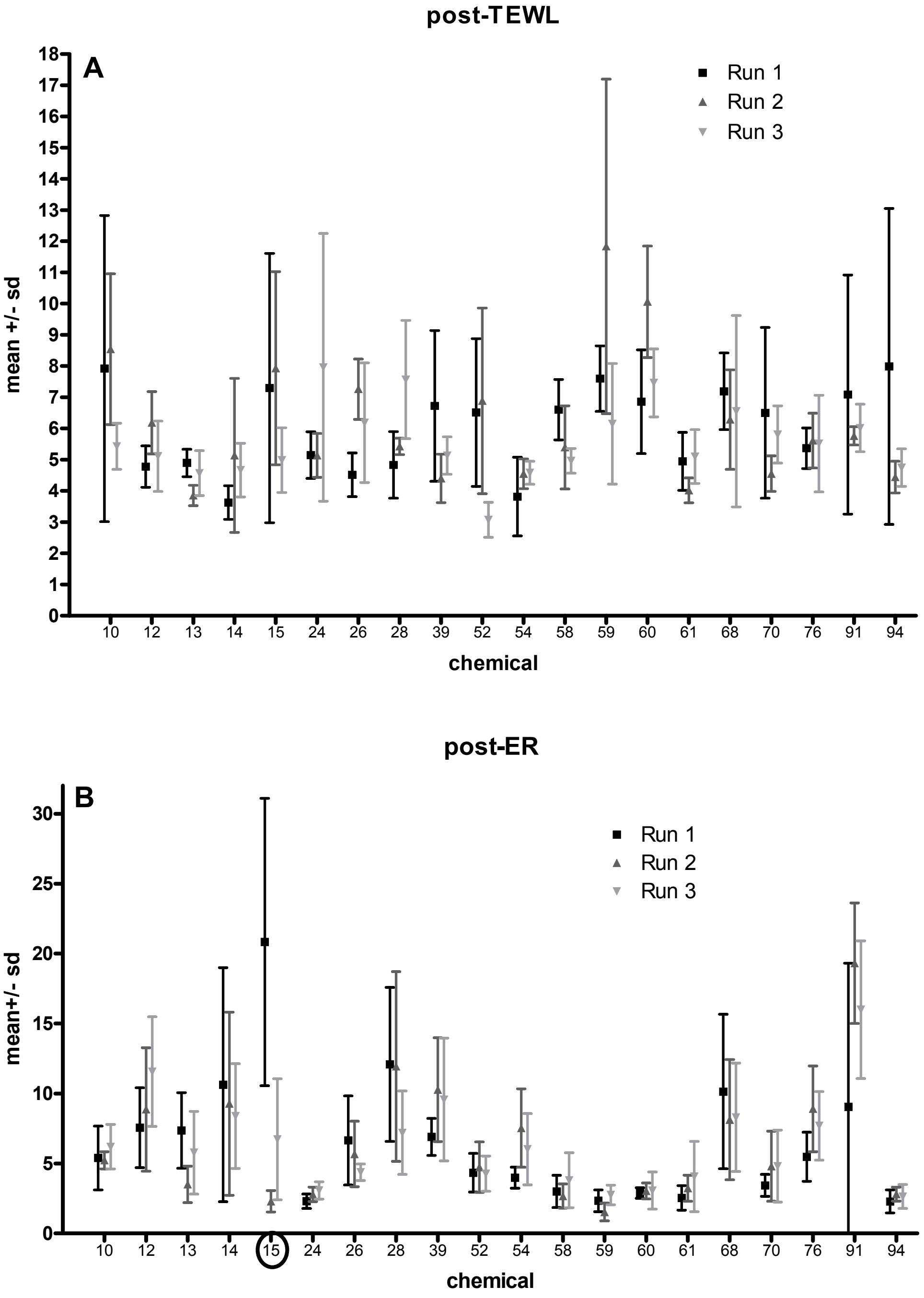


Figure 3: Phase-1 data from Syngenta with SIFT. A: post-TEWL. B: post-ER. The encircled chemical showed significant differences between runs in an one-way ANOVA.

SIFT		European classification		Σ
		no label	R38	
PM	non-irritant	7	7	14
	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%

Table 7: 2x2-contigency table and predictive capacity for SIFT in Phase 1

For the SIFT, the ROC-approach was not applied. Presenting the mean-values over all runs for the twenty chemicals arranged by endpoint and European classification together with the endpoint-specific thresholds, clearly showed that the performance of the SIFT was not threshold dependent (Figure 4). Moving the thresholds did not substantially improve the assay performance. Correlating the mean-values of both endpoints with the *in vivo* rabbit data resulted in a coefficient of correlation according to Bravais-Pearson of –0.06 for TEWL and of 0.40 for ER (*in vivo* data not shown to maintain blinding). For TEWL there is almost no correlation, where for ER the correlation is opposed to the SOP-threshold, according to which the irritation potential and the ER should be negatively correlated.

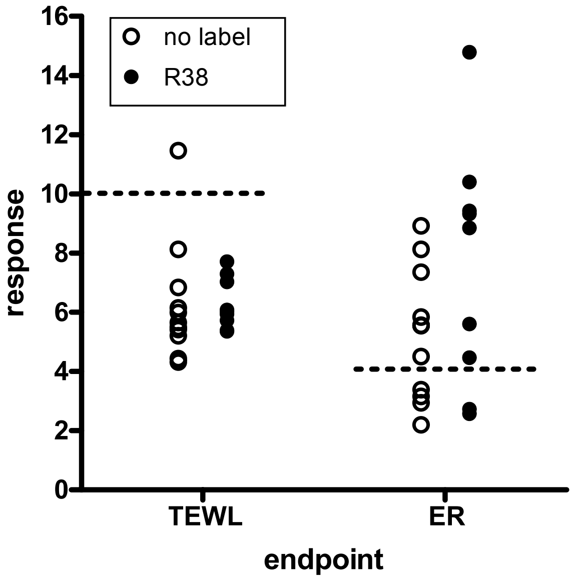


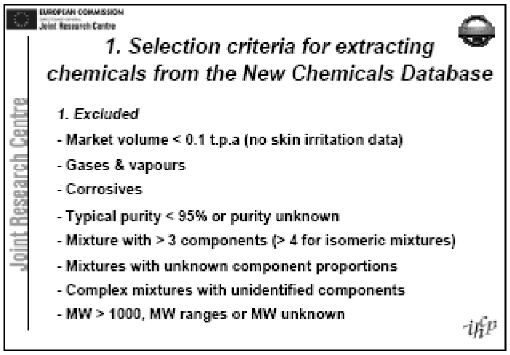
Figure 4: Mean values over the runs per endpoint of SIFT in Phase 1 arranged by European classification and with the endpoint-specific threshold as dotted lines

Conclusion

Based on the good within-laboratory reproducibility and on the acceptable predictive capacity of EpiDerm and EPISKIN, bearing the borderline *in vivo* data of the misclassifications in mind, it is recommended to assess these two test systems in the planned second part of this validation study, i.e. Phase 2. The poor predictive capacity of the SIFT suggests that this assay needs further development and that it should not proceed to Phase 2.

The post-hoc analysis of the EpiDerm and the EPISKIN showed that the two test systems were designed to meet the needs of the European classification of skin irritation. GHS-mild-irritant chemicals cannot be discriminated from the other two GHS-classes. Nevertheless, it is foreseen to conduct a similar post-hoc analysis with the larger data set, which will be generated in Phase 2, in order to confirm the findings of Phase 1.

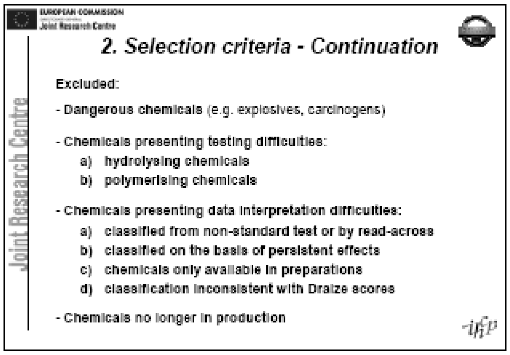
Annex II: Chemical selection criteria


EUROPEAN COMMISSION
Joint Research Centre
Joint Research Centre
ifp

1. Selection criteria for extracting chemicals from the New Chemicals Database

1. Excluded

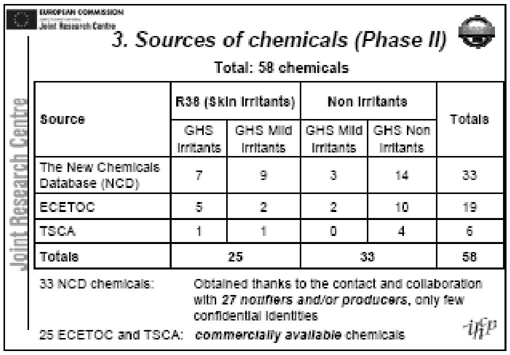
- Market volume < 0.1 t.p.a (no skin irritation data)
- Gases & vapours
- Corrosives
- Typical purity < 95% or purity unknown
- Mixture with > 3 components (> 4 for isomeric mixtures)
- Mixtures with unknown component proportions
- Complex mixtures with unidentified components
- MW > 1000, MW ranges or MW unknown


EUROPEAN COMMISSION
Joint Research Centre
Joint Research Centre
ifp

2. Selection criteria - Continuation

Excluded:

- Dangerous chemicals (e.g. explosives, carcinogens)
- Chemicals presenting testing difficulties:
  a) hydrolysing chemicals
  b) polymerising chemicals
- Chemicals presenting data interpretation difficulties:
  a) classified from non-standard test or by read-across
  b) classified on the basis of persistent effects
  c) chemicals only available in preparations
  d) classification inconsistent with Draize scores
- Chemicals no longer in production


EUROPEAN COMMISSION
Joint Research Centre
Joint Research Centre
ifp

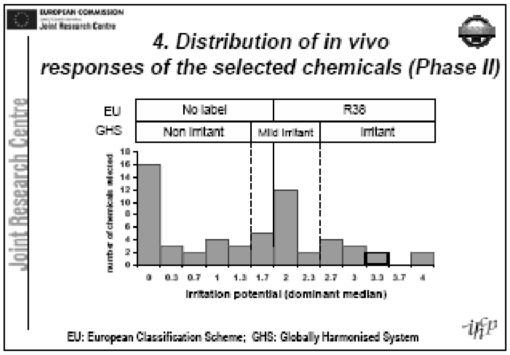
3. Sources of chemicals (Phase II)

Total: 58 chemicals

Source	R38 (Skin Irritants)		Non Irritants		Totals
	GHS Irritants	GHS Mild Irritants	GHS Mild Irritants	GHS Non Irritants	
The New Chemicals Database (NCD)	7	9	3	14	33
ECETOC	5	2	2	10	19
TSCA	1	1	0	4	6
Totals	25		33		58

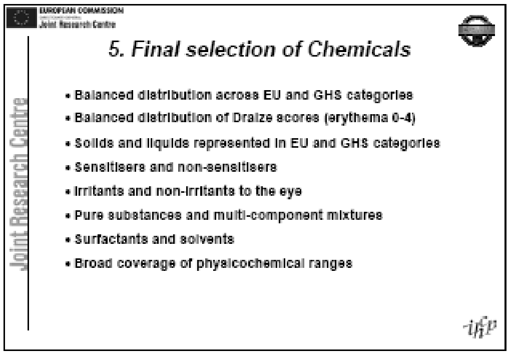
33 NCD chemicals: Obtained thanks to the contact and collaboration with *27 notifiers and/or producers*, only few confidential identities

25 ECETOC and TSCA: *commercially available* chemicals



4. Distribution of in vivo responses of the selected chemicals (Phase II)

EU: European Classification Scheme; GHS: Globally Harmonised System


EUROPEAN COMMISSION
Joint Research Centre
Joint Research Centre
ifp

5. Final selection of Chemicals

- Balanced distribution across EU and GHS categories
- Balanced distribution of Draize scores (erythema 0-4)
- Solids and liquids represented in EU and GHS categories
- Sensitisers and non-sensitisers
- Irritants and non-irritants to the eye
- Pure substances and multi-component mixtures
- Surfactants and solvents
- Broad coverage of physicochemical ranges

ANNEX III: List of chemicals (including codes, classification, purity and physical state)

	EpiDerm				EPISKIN							Purity [%]			
no.	ZEBET code		IIVS code	BASF code	L'Oréal code		Unilever code	Sanofi code	chemical identification	CAS number	class				phys. state
	Ph.I	Ph.II	Phase II	Phase II	Ph.I	Ph.II	Phase II	Phase II				typical	lower limit	Upper limit	
1	42	163	103	759	85	403	135	740	2-chloromethyl-3,5-dimethyl-4-methoxypyridine hydrochloride	86604-75-3	R38 I	98.5	98	100	S
2	-	578	113	527	-	355	569	493	1-bromo-4-chlorobutane	6940-78-9	no label NI	98 *			L
3	-	808	933	179	-	288	280	688	1-bromohexane	111-25-1	R38 I	>98.5 *			L
4	-	187	789	213	-	367	505	565	1-decanol	112-30-1	R38 I	98.8 *			L
5	-	969	190	745	-	746	467	161	3-chloro-4-fluoronitrobenzene	350-30-1	no label NI	98 *			
6	-	630	315	671	-	300	948	342	3-diethylaminopropionitrile	5351-04-2	no label NI	99.8 *			L
7	-	455	867	225	-	149	952	605	3-mercaptohexanol	51755-83-0	no label NI	98.5	97	100	L
8	-	254	110	586	-	567	929	183	4-methylthio-benzaldehyde	3446-89-7	no label NI	98.2 *			L
9	72	207	115	546	43	125	985	526	2,6-dimethyl-4-nitrobenzeneamine	16947-63-0	no label NI	99.5	99	100	S
10	-	936	656	258	-	773	143	496	allyl heptanoate	142-19-8	no label MI	98.1 *			L
11	-	541	160	686	-	572	427	814	allyl phenoxyacetate	7493-74-5	no label NI	100 *			L
12	46	747	249	347	81	636	375	949	2-ethylhexyl 4-aminobenzoate	26218-04-2	no label NI	99	98.5	100	S
13	33	864	915	989	19	818	299	974	1-[4-(2-dimethylaminoethoxy)phenyl] -2-phenylbutan-1-one	68047-07-4	R38 MI	99	95	100	S
15	-	735	337	900	-	204	134	691	a-terpineol	98-55-5	R38 I	98.4 *			L
16	73	975	262	487	18	137	267	518	capryl-isostearate	209802-43-7	no label NI	99	95	100	L
17	87	966	239	622	63	779	261	902	2-methyl-3-[(1,7,7-trimethylbicyclo[2.2.1] hept-2-yl)oxy]-1-propanol, bornyl isomer	128119-70-0	no label MI	96	93	100	L
18	-	501	334	633	-	782	813	973	butyl methacrylate	97-88-1	R38 I	>99 *			L
19	-	384	755	169	-	768	543	147	2,5-dimethyl-4-oxo-4,5-dihydrofuran-3-yl acetate	4166-20-5	no label NI	98	97.5	100	L
20	-	462	431	295	-	836	844	368	cyclamen aldehyde	103-95-7	R38 I	> 98 *			L
21	-	570	480	805	-	682	608	616	A mixture of: 5-exo-decylbicyclo[2.2.1]hept-2-ene; 5-endo-decylbicyclo[2.2.1]hept-2-ene	22094-85-5	no label MI	99.6	99.5	100	L
22	-	319	714	343	-	346	822	371	diethyl phthalate	84-66-2	no label NI	99.7 *			L

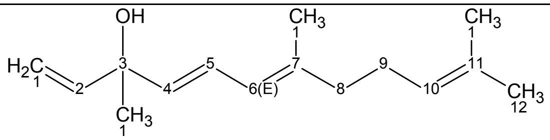
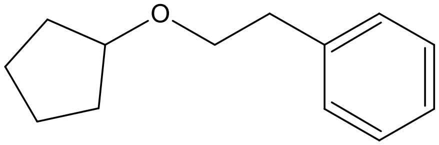
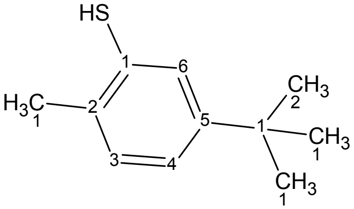
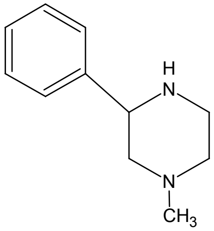
23	-	215	363	780	-	658	673	330	di-n-propyl disulphide	629-19-6	R38 I	99.2 *			L
24	-	593	222	547	-	799	920	404	di-propylene glycol	25265-71-8	no label NI	99 *			L
25	-	353	488	913	-	699	875	446	dipropylene glycol monobutyl ether	29911-28-2	no label NI	> 99 *			L
26	-	628	430	639	-	189	159	549	3,4-dimethyl-1H-pyrazole	2820-37-3	no label NI	99	96	100	S
27	-	503	109	184	-	716	833	170	2-isopropyl-2-isobutyl-1,3-dimethoxypropane	129228-21-3	R38 I	97.6	94.8	100	L
28	95	581	713	697	30	266	447	889	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	no label NI	97.7	97	100	S
29	67	613	148	562	16	495	732	642	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	R38 MI	96	94	98	L
30	20	750	519	287	90	341	855	558	Mixture of: diethyl cis-1,4-cyclohexane dicarboxylate; diethyl trans-1,4-cyclohexanedicarboxylate	0072903-27-6	no label NI	99	96	100	L
31	-	595	659	848	-	255	236	154	A 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 (mix).	R38 MI	99.6	98	100	L
32	21	762	528	162	40	139	827	662	2S-(2-furyl)-5R-hydroxy-4R-(1R,2-dihydroxy)ethyl-6S-hydroxymethyl-1,3-dioxane	7089-59-0	no label NI	99.5	99	100	S
33	-	282	600	188	-	201	416	398	heptyl butyrate	5870-93-9	no label MI	> 95 *			L
34	-	276	232	719	-	815	684	726	hexyl salicylate	6259-76-3	R38 - MI	> 98 *			L
35	49	537	906	466	36	752	908	233	cyclohexadecanone	2550-52-9	no label NI	99.2	99	100	S
36	-	977	706	953	-	893	877	701	isopropanol	67-63-0	no label NI	100 *			L
37	99	338	972	425	71	144	866	598	[2-(cyclopentyloxy) ethyl]benzene (cyclopentyl 2-phenylethyl ether)	not allocated	R38 I	98	94.9	99.3	L
39	-	794	723	722	-	379	326	817	methyl stearate	112-61-8	no label NI	99 *			S
40	57	849	291	988	56	168	707	696	1-methyl-3-phenyl-1-piperazine	5271-27-2	R38 I	99	95	100	S
41	-	477	876	133	-	538	241	934	naphthalene acetic acid	86-87-3	no label NI	96 *			S
42	89	890	409	859	66	298	824	323	disodium 2,2'-(1,4-phenylene)bis-(1H-benzimidazole-4,6-disulfonic acid or monosulfonic acid, monosulfonate or disulfonate	180898-37-7	no label NI	97.1	96.1	98	S

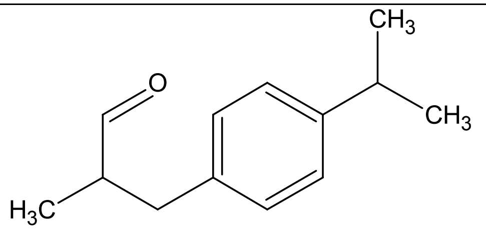
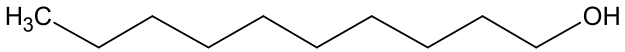
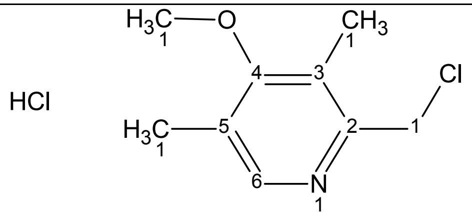
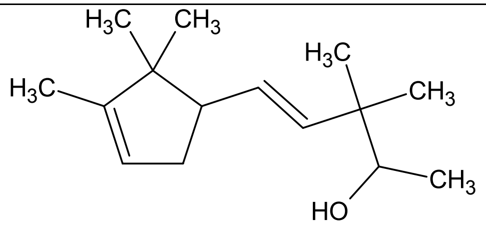
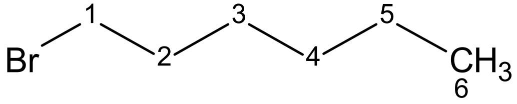
43	-	345	471	542	-	399	756	865	A mixture of isomers: 1-(1,1-dimethylpropyl)-4-ethoxy-cis-cyclohexane;1-(1,1-dimethylpropyl)-4-ethoxy-trans-cyclohexane	181258-87-7 (cis), 181258-89-9 (trans)	R38 MI	99	95	99.9	L
44	-	535	152	927	-	274	583	606	phenylethylalcohol	60-12-8	no label NI	99.6 *			L
45	-	971	676	821	-	269	680	959	(+/-) trans-3,3-dimethyl-5-(2,2,3-trimethy l-cyclopent-3-en-1-yl)-pent-4-en-2-ol	107898-54-4	R38 I	98.4	98	99	L
46	-	981	885	252	-	308	914	270	4-methyl-8-methylenetricyclo [3.3.1.1(3,7)]decan-2-ol	122760-84-3	R38 MI	99.4	95	100	S
47	-	521	278	385	-	359	997	238	4-methyl-8-methylenetricyclo [3.3.1.1(3,7)]dec-2-yl acetate	122760-85-4	R38 MI	98.1	96	100	L
48	-	202	788	694	-	724	122	708	2-(formylamino)-3-thiophenecarboxylic acid	43028-69-9	no label NI	96	94	98	S
49	78	894	838	185	88	485	718	151	isostearic acid monoisopropanolamide	152848-22-1	R38 MI	95	90	100	L
50		585	271	965		961	111	580	2-phenylhexanenitrile	3508-98-3	no label MI	99.5	95	99.9	L
51	29	947	872	126	65	568	121	223	Mixture of isomers: 1-(2-isopropylphenyl)-1-phenylethane 1-(3-isopropylphenyl)-1-phenylethane 1-(4-isopropylphenyl)-1-phenylethane	52783-21-8 (mix.)	R38 MI	96	95	99	L
52	35	648	445	235	53	883	164	366	propyl (2S)-2-(1,1-dimethylpropoxy)-propanoate	0319002-92-1	no label NI	99.6	98	100	L
53	-	459	158	832	-	197	123	856	silane A-1430	2530-87-2	no label NI	99.7 *			L
54	-	903	208	807	-	982	828	638	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	no label NI	96	90	99	L
55	-	637	968	286	-	800	625	400	terpinyl acetate	80-26-2	R38 MI	>= 95 *			L
56	-	120	141	476	-	461	321	247	benzenethiol, 5-(1,1-dimethylethyl)-2-methyl *(NB: CAS name from company)*	7340-90-1	R38 I	94	92.5	96.5	L
57	-	576	124	200	-	964	992	539	triethylene glycol	112-27-6	no label NI	99.8 *			L
58	-	797	917	749	-	743	655	456	tri-isobutyl phosphate	126-71-6	R38 MI	99.7 *			L
59	55	106	481	588	34	830	874	439	(E,E)-3,7,11-trimethyldodeca-1,4,6,10-tetraen-3-ol	125474-34-2	R38 I	95	91.8	97	L
60	-	666	356	132	-	880	119	820	bis[(1-methylimidazol)-(2-ethyl-hexanoate)], zinc complex	not allocated	R38 MI	99.6	99.1	99.7	L
#	37	-	-	-	64	-	-	-	A mixture of: (E)-oxacyclohexadec-12-en-2-one; (E)-oxacyclohexadec-13-en-2-one; a) (Z)-oxacyclohexadec-12-en-2-one b) (Z)-oxacyclohexadec-13-en-2-one	not allocated	no label NI	96	92	99	L
#	44	-	-	-	86	-	-	-	diisononyl cyclohexane-1,2-dicarboxylate	166412-78-8	R38 MI	99	90	100	L

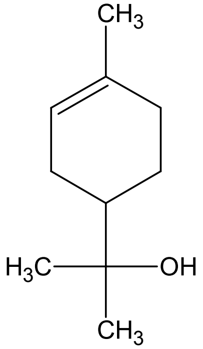
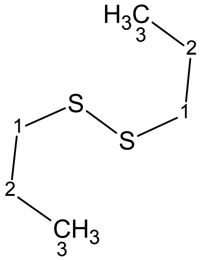
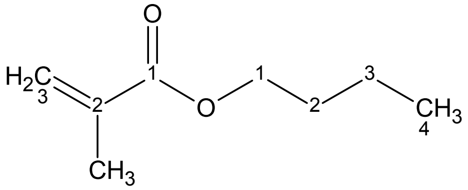
14		226	804	869		294	739	309	confidential-1	not disclosed			
38		514	955	523		717	520	529	confidential-2	not disclosed			
# Phase I only												* information from ECETOC or TSCA	

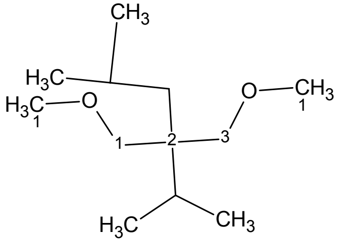
Annex IV: Structures of chemicals

R38 - GHS irritants

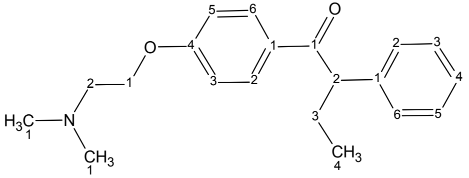
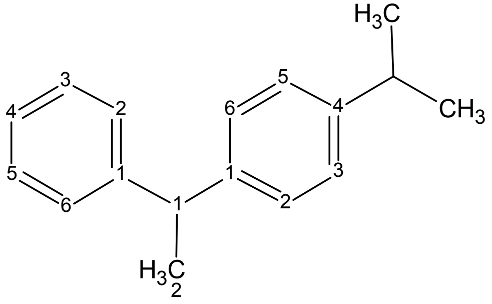
chemical number	*IUPAC Name (NCD confidential; Annex I, ELINCS, EINECS published)*	Molecular Structure	*Source*
59	(E,E)-3,7,11-trimethyldodeca-1,4,6,10-tetraen-3-ol *(disclosure by company agreement)*		NCD
37	[2-(cyclopentyloxy)ethyl]benzene (cyclopentyl 2-phenylethyl ether) *(disclosure by company agreement)*		NCD
56	benzenethiol, 5-(1,1-dimethylethyl)-2-methyl *(NB: CAS name from company)*		NCD
40	1-methyl-3-phenyl-1-piperazine *(disclosure by company agreement)*		NCD

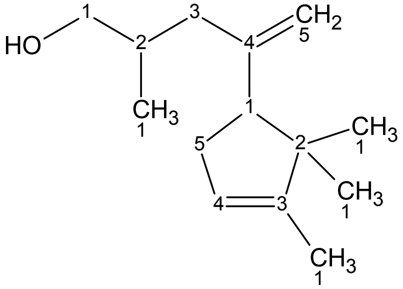
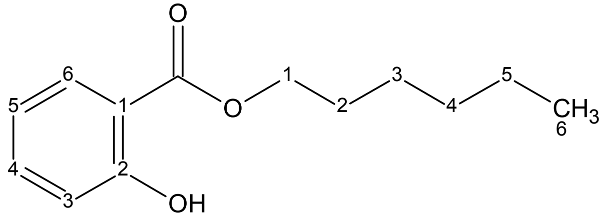
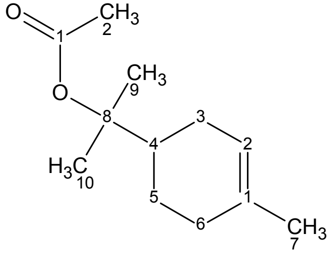
20	cyclamen aldehyde **EINECS Name** 3-p-cumenyl-2-methylpropionaldehyde		ECETOC
4	1-decanol **EINECS Name** decan-1-ol		ECETOC
1	2-chloromethyl-3,5-dimethyl-4-methoxypyridine hydrochloride (disclosure by company agreement)	 H$_3$C 1 O CH$_3$ 1 4 3 Cl HCl H$_3$C 1 5 2 1 6 N 1	NCD
45	(+/-) *trans*-3,3-dimethyl-5-(2,2,3-trimethyl-cyclopent-3-en-1-yl)-pent-4-en-2-ol (Annex I published name)	 H$_3$C CH$_3$ H$_3$C H$_3$C CH$_3$ CH$_3$ HO	NCD
3	1-bromohexane (EINECS name)		ECETOC

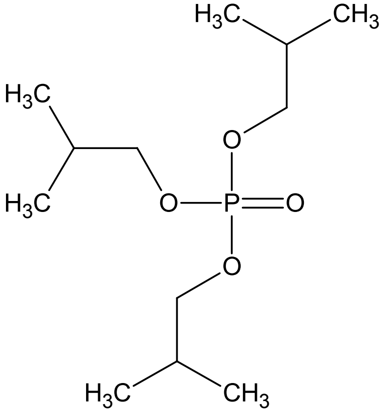
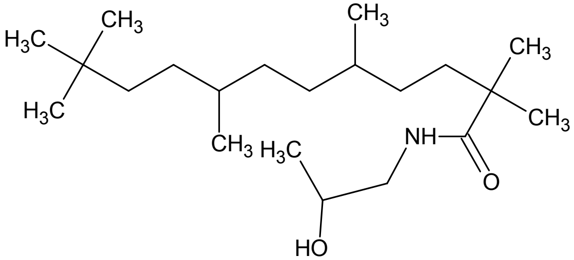
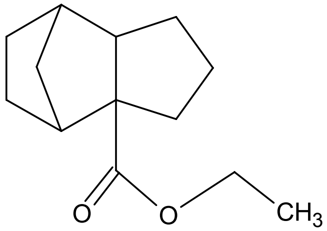
15	a-terpineol EINECS Name p-menth-1-en-8-ol		ECETOC
23	di-n-propyl disulphide EINECS Name dipropyl disulphide		ECETOC
18	butyl methacrylate (EINECS name) n-butyl methacrylate (Annex I published name)	 O H2C 3 2 1 O 1 2 3 CH3 4 CH3	TSCA

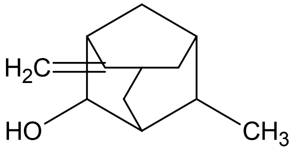
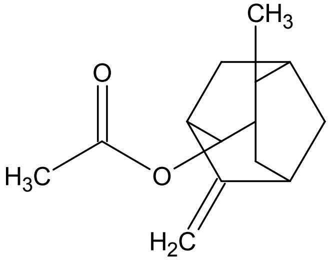
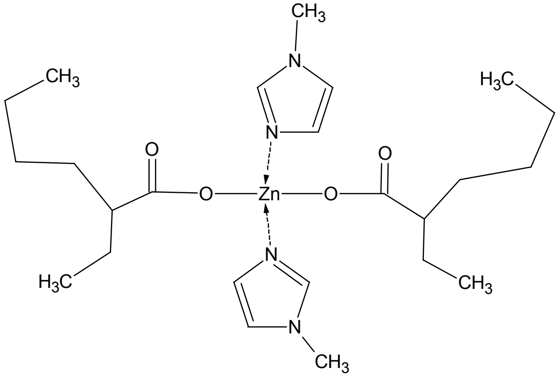
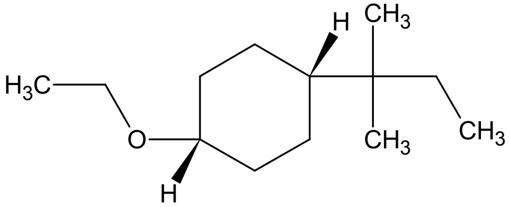
27	2-isopropyl-2-isobutyl-1,3-dimethoxypropane (disclosure by company agreement)		NCD

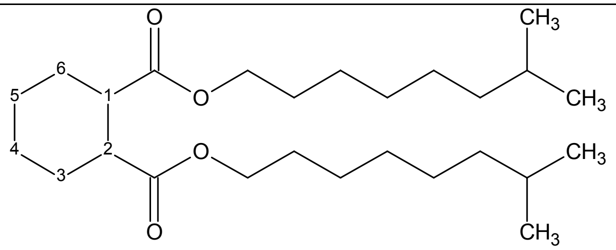
R38 – GHS mild irritants

13	1-[4-(2-dimethylaminoethoxy)phenyl]-2-phenylbutan-1-one *(disclosure by company agreement)*		NCD
51	Mixture of isomers: 1-(2-isopropylphenyl)-1-phenylethane 1-(3-isopropylphenyl)-1-phenylethane 1-(4-isopropylphenyl)-1-phenylethane *(structure shown)* *(disclosure by company agreement)*		NCD

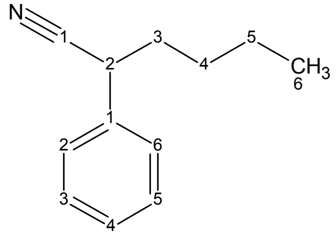
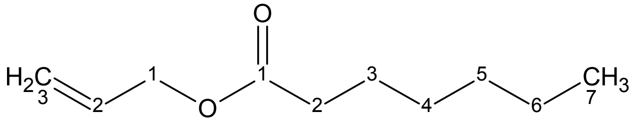
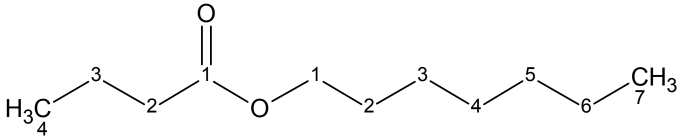
29	Mixture of isomers: 2-methyl-4-(2',2',3'-trimethyl-3'-cyclopenten-1'-yl)-4-penten-1-ol 56% (1'R,2R) & 40%(1'R,2S) isomer *(disclosure by company agreement)*		NCD
34	hexyl salicylate *(EINECS name)*		ECETOC
55	terpinyl acetate **<u>EINECS Name</u>** **p-menth-1-en-8-yl acetate**		ECETOC

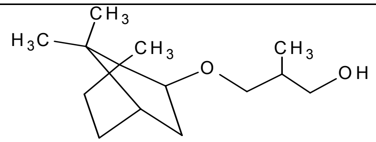
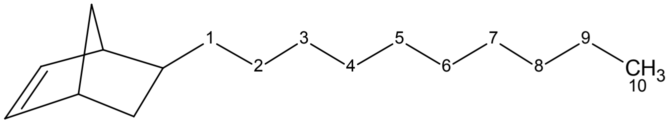
58	tri-isobutyl phosphate *(EINECS name)*		TSCA
49	isostearic acid monoisopropanolamide (disclosure by company agreement)		NCD
31	A mixture of isomers: ethyl *exo*-tricyclo[5.2.1.02,6]decane-*endo*-2-carboxylate; ethyl *endo*-tricyclo[5.2.1.02,6]decane-*exo*-2-carboxylate (Annex I published name) Ratio of isomer 1: 35-60 Ratio of isomer 2: 40-65		NCD

46	4-methyl-8-methylenetricyclo[3.3.1.13,7]decan-2-ol (Annex I published name)		NCD
47	4-methyl-8-methylenetricyclo[3.3.1.13,7]dec-2-yl acetate (Annex I published name)		NCD
60	bis[(1-methylimidazol)-(2-ethyl-hexanoate)], zinc complex (Annex I published name)	 CH$_3$ N N CH$_3$ H$_3$C O O O Zn O N N H$_3$C CH$_3$ CH$_3$	NCD
43	A mixture of isomers: 1-(1,1-dimethylpropyl)-4-ethoxy-cis-cyclohexane; 1-(1,1-dimethylpropyl)-4-ethoxy-trans-cyclohexane (disclosure by company agreement)		NCD

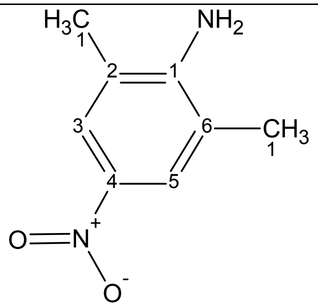
Phase I only	diisononyl cyclohexane-1,2-dicarboxylate		NCD

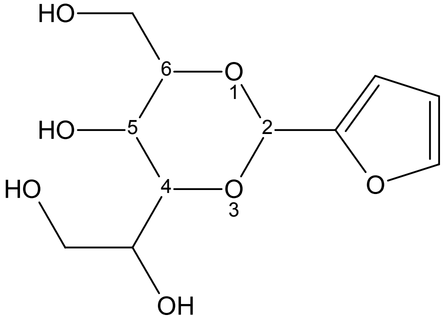
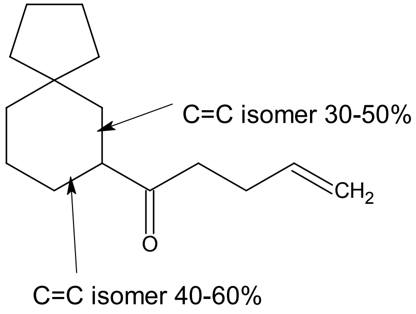
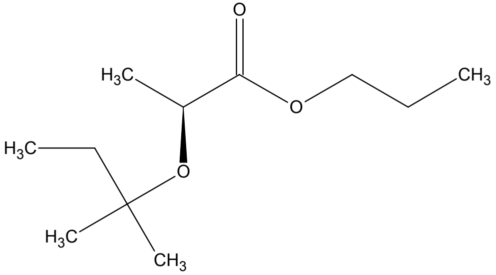
Non-R38 – GHS mild irritants

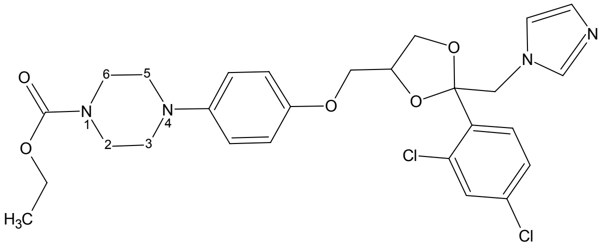
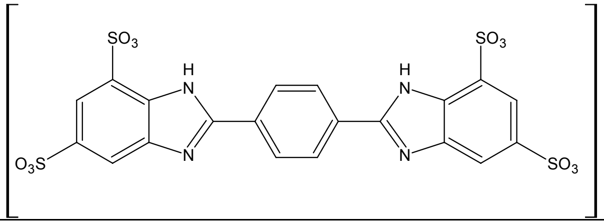
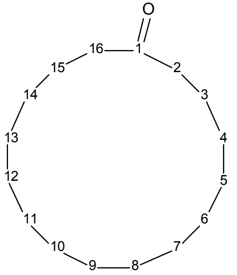
50	2-phenylhexanenitrile *(Annex I name)*		NCD
10	allyl heptanoate *(EINECS name)*		ECETOC
33	heptyl butyrate *(EINECS name)*		ECETOC

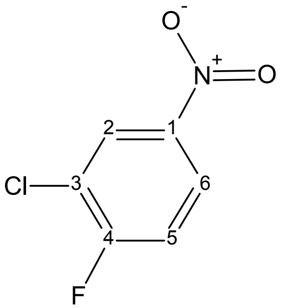
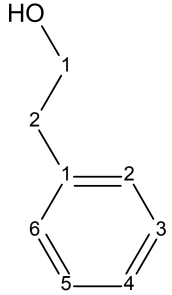
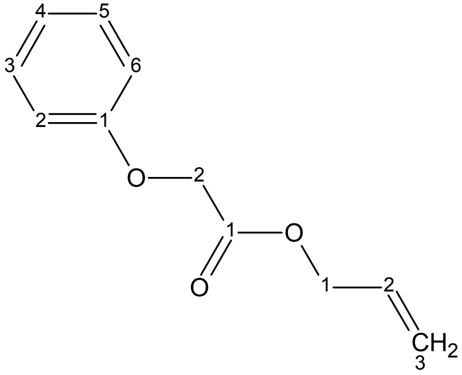
17	2-methyl-3-[(1,7,7-trimethylbicyclo[2.2.1]hept-2-yl)oxy]-1-propanol, bornyl isomer (disclosure by company agreement)		NCD
21	A mixture of: 5-exo-decylbicyclo[2.2.1]hept-2-ene; 5-endo-decylbicyclo[2.2.1]hept-2-ene (disclosure by company agreement)		NCD

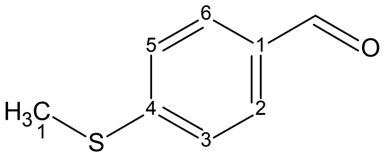
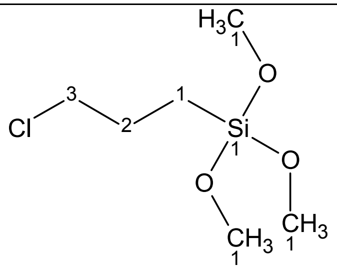
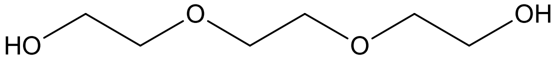
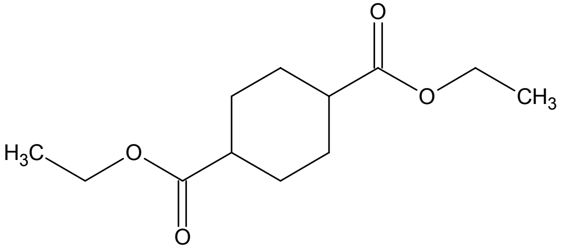
Non-R38 – GHS non irritants

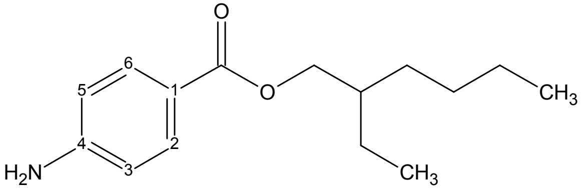
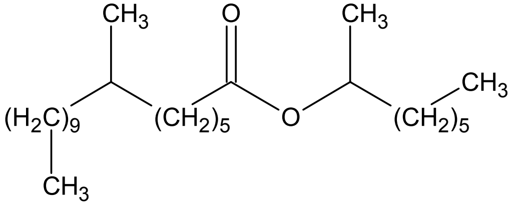
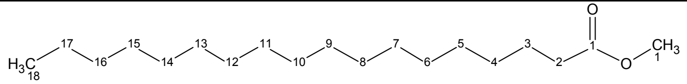
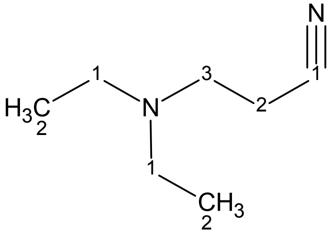
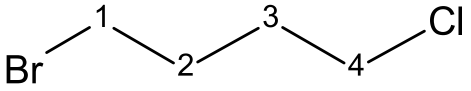
9	2,6-dimethyl-4-nitrobenzeneamine *(disclosure by company agreement)*	 H3C 1 NH2 2 1 3 6 CH3 1 4 5 O N + O -	NCD

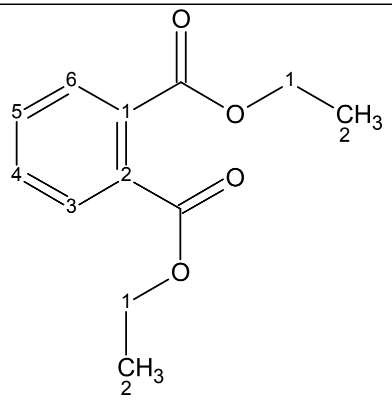
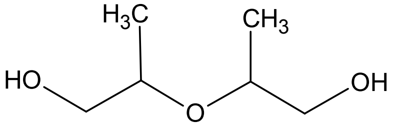
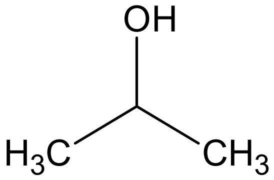
32	2S-(2-furyl)-5R-hydroxy-4R-(1R,2-dihydroxy)ethyl-6S-hydroxymethyl-1,3-dioxane *(disclosure by company agreement)*	 HO 6 O 1 HO 5 2 HO 4 O 3 O OH	NCD
54	Mixture of isomers: 1-(spiro[4.5]dec-7-en-7-yl)pent-4-en-1-one (40-60%) 1-(spiro[4.5]dec-6-en-7-yl)pent-4-en-1-one (30-50%) *(disclosure by company agreement)*	 C=C isomer 30-50% CH₂ O C=C isomer 40-60%	NCD
52	propyl (2S)-2-(1,1-dimethylpropoxy)- propanoate *(disclosure by company agreement)*		NCD

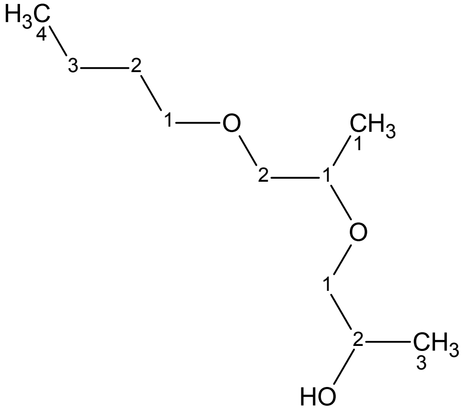
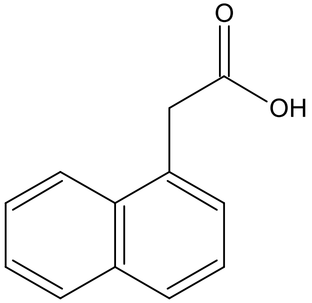
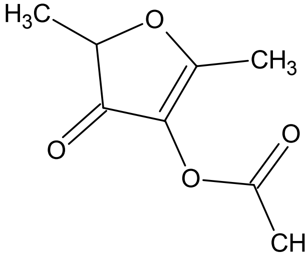
28	ethyl cis-4-[4-[[2-(2,4-dichlorophenyl)-2-(1H-imidazol-1-ylmethyl)-1,3-dioxolan-4-yl]-methoxy]phenyl]piperazine-1-carboxylate *(disclosure by company agreement)*		NCD
42	disodium 2,2'-(1,4-phenylene)bis-(1H-benzimidazole-4,6-disul fonic acid or monosulfonic acid, monosulfonate or disulfonate *(disclosure by company agreement)*	 4^- $2H^+$ $2Na^+$	NCD
35	Cyclohexadecanone *(disclosure by company agreement)*		NCD

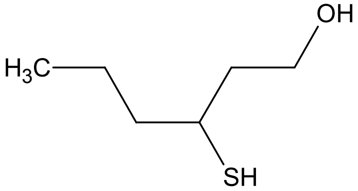
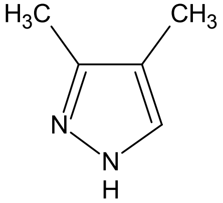
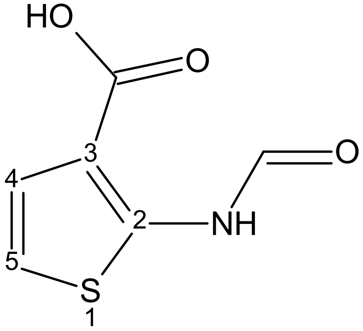
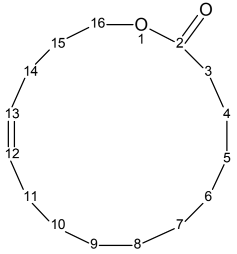
5	3-chloro-4-fluoronitrobenzene **EINECS Name** 2-chloro-1-fluoro-4-nitrobenzene		ECETOC
44	phenylethylalcohol **EINECS Name** 2-phenylethanol		ECETOC
11	allyl phenoxyacetate *(EINECS Name)*		ECETOC

8	4-(methylthio)benzaldehyde *(EINECS Name)*		ECETOC
53	Silane A-1430 **EINECS Name** 3-chloropropyltrimethoxysilane	 H3C O Cl 3 2 1 Si O O CH3 CH3 1 1 1 1 1	TSCA
57	triethylene glycol **EINECS Name** 2,2'-(ethylenedioxy)diethanol		TSCA
30	A mixture of: diethyl cis-1,4-cyclohexanedicarboxylate; diethyl trans-1,4-cyclohexanedicarboxylate (disclosure by company agreement)		NCD

12	2-ethylhexyl 4-aminobenzoate (Annex I published name)		NCD
16	capryl-isostearate (disclosure by company agreement)	 CH3 O CH3 CH3 (H2C)9 (CH2)5 O (CH2)5 CH3	NCD
39	methyl stearate (EINECS name)	 O H3C 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 O CH3 1	ECETOC
6	3-diethylaminopropionitrile (EINECS name)		ECETOC
2	1-bromo-4-chlorobutane (EINECS name)		ECETOC

22	diethyl phthalate (EINECS name)		ECETOC
24	di-propylene glycol EINECS Name oxydipropanol		ECETOC
36	isopropanol EINECS Name propan-2-ol		ECETOC

25	dipropylene glycol monobutyl ether EINECS Name 1-(2-butoxy-1-methylethoxy)propan-2-ol		ECETOC
41	naphthalene acetic acid EINECS Name 1-naphthylacetic acid		TSCA
19	2,5-dimethyl-4-oxo-4,5-dihydrofuran-3-yl acetate (disclosure by company agreement)		NCD

7	3-mercaptohexanol (disclosure by company agreement)		NCD
26	3,4-dimethyl-1H-pyrazole (disclosure by company agreement)		NCD
48	2-(formylamino)-3-thiophenecarboxylic acid (disclosure by company agreement)		NCD
Phase I only	A mixture of: (E)-oxacyclohexadec-12-en-2-one; (E)-oxacyclohexadec-13-en-2-one; a) (Z)-oxacyclohexadec-12-en-2-one and b) (Z)-oxacyclohexadec-13-en-2-one (disclosure by company agreement)		NCD

Annex V: In vivo data of chemcials

no.	source	Classifications		Dominant median	dominant endpoint	no. of experiments	no. of rabbits	Rabbit erythema scores (average of scores after 24, 48, 72 h)															Erythema median/ mean	Rabbit oedema scores (average of scores after 24, 48, 72 h)															Oedema median/ mean	remark
		EU	GHS					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
1	NCD	R38	I	2.7	B																																			
2	ECETOC	no label	NI	0.0	B	1	3	0	0	0													0	0	0	0													0	
3	ECETOC	R38	I	2.7	E	1	3	2.7	2	2.7													2.7	0	2.7	2													2	
4	ECETOC	R38	I	2.3	E	1	4	2.3	2.3	2.3	1.7												2.3	2	1	1	1												1	single scores of 0.5, 1.5, 2.5
5	ECETOC	no label	NI	1.0	E	1	6	1	1	1	1.7	1.7	1										1	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	0	0	0													0	
7	NCD	no label	NI	0.0	B																																			
8	ECETOC	no label	NI	1.0	E	1	3	1	1.3	0.3													1	0	0	0													0	
9	NCD	no label	NI	0.3	E																																			
10	ECETOC	no label	MI	1.7	E	1	4	1.3	2	1.7	2												1.7	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	0.7	0.3	0.3												0.3	0.3	0	0	0.3												0	single scores of 0.5
12	NCD	no label	NI	0.7	E																																			
13	NCD	R38	MI	2.0	E																																			
14																																								
15	ECETOC	R38	I	2.7	O	3	11	1.7	2	2.3	2	2.7	2	1.7	2	2	1.7	2					2	2	2.3	3	3	3	2.7	1.7	2.7	2	0.7	3					2.7	
16	NCD	no label	NI	1.0	E																																			
17	NCD	no label	MI	1.7	E																																			
18	TSCA	R38	I	3.0	E	1	6	3	3	2.7	3	3	3										3	2.7	3.7	2.3	2.7	2.7	4										2.7	
19	NCD	no label	NI	0.0	B																																			
20	ECETOC	R38	I	2.3	O	4	15	2.7	2	2	2	2	2	2	2	2	2	2.7	2	2	2	2	2	3	3	2.7	2.3	2.7	2	1.7	2.7	2.7	2.3	3	1.3	1	2	1.3	2.3	
21	NCD	no label	MI	1.7	E																																			
22	ECETOC	no label	NI	0.0	E	2	7	0.7	0	0	0	0	0	0									0	0	0	0	0	0	0	0									0	
23	ECETOC	R38	I	3.0	E	1	3	1.7	3	3													3	0	0	0													0	
24	ECETOC	no label	NI	0.0	E	2	7	1	0	0	0	0	0	0									0	0	0	0	0	0	0	0									0	
25	TSCA	no label	NI	0.0	E	1	6	0.3	0	0	0	0	0										0	0	0	0	0	0	0										0	
26	NCD	no label	NI	0.0	B																																			
27	NCD	R38	I	4.0	E																																			
28	NCD	no label	NI	0.0	B																																			
29	NCD	R38	MI	2.0	B																																			
30	NCD	no label	NI	1.3	E																																			
31	NCD	R38	MI	2.0	O																																			
32	NCD	no label	NI	0.0	B																																			
33	ECETOC	no label	MI	1.7	E	1	4	1.7	2	0.7	2												1.7	0	0.3	0	0.7												0.3	single scores of 0.5, 1.5
34	ECETOC	R38	MI	2.0	B	4	15	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1.3	2	0.7	2	2	2	2.7	1.7	2	2.7	1.3	2	1	1	2	
35	NCD	no label	NI	0.0	B																																			
36	ECETOC	no label	NI	0.3	E	1	3	1.7	0.3	0.3													0.3	0	0	0													0	
37	NCD	R38	I	3.0	E																																			
38																																								
39	ECETOC	no label	NI	1.0	E	1	3	1	2.3	1													1	0	2	0													0	
40	NCD	R38	I	3.3	E																																			
41	TSCA	no label	NI	0.0	B	1	6	0	0	0	0	0	0										0	0	0	0	0	0	0										0	no scores after 48h; two application sites per rabbit (scores were averaged)
42	NCD	no label	NI	0.0	B																																			

43	NCD	R38	MI	2.0	B							
44	ECETOC	no label	NI	1.0	E	2	7	1.3 1.3 2 1 1 0.7 0.3	1	0 1 1 0 0.7 0 0	0	
45	NCD	R38	I	2.7	E							
46	NCD	R38	MI	2.0	B							
47	NCD	R38	MI	2.0	B							
48	NCD	no label	NI	0.0	B							
49	NCD	R38	MI	2.0	E							
50	NCD	no label	MI	1.7	E							
51	NCD	R38	MI	2.0	E							
52	NCD	no label	NI	0.7	E							
53	TSCA	no label	NI	0.0	B	1	6	0 0 0 0 0 0	0	0 0 0 0 0 0	0	
54	NCD	no label	NI	1.3	E							
55	ECETOC	R38	MI	2.0	B	3	11	1.7 2 2 2 2 2 2 2 1.7 2 1.3	2	1 2 2 3 2 2.3 2 2 1 0.7 0.3	2	
56	NCD	R38	I	3.3	O							
57	TSCA	no label	NI	0.0	B	1	6	0 0 0 0 0 0	0	0 0 0 0 0 0	0	
58	TSCA	R38	MI	2.0	E	1	6	2 2.3 1.7 2 2.3 2	2	1 1.3 0.7 1.3 1 0	1	two application sites per rabbit (scores were averaged)
59	NCD	R38	I	4.0	E							
60	NCD	R38	MI	2.0	E							
61#	NCD	no label	MI									
62#	NCD	R38	MI									

#Chemicals tested in Phase I only.

Annex VI

Analysis of within-assay variability of Phase I

A criterion on variability is usually applied to identify any excessive variation. Increased variability might be an indicator of experimental problems, so that the result obtained should be questioned. In this regard, the controls (positive and negative) are most crucial, as a rejection of a control usually would demand the rejection of the whole experiment. Such a variability criterion should be based on historical empirical data and should be applied independently from the PM, respectively its cut-off.

In order to find a suitable and generally applicable measure of variability for EpiDerm and EPISKIN, the range, the SD and the CV were plotted against the mean viability of samples/controls. Linear regressions were fitted for each variability measure in order to highlight their dependence on the response, i.e. cell viability.

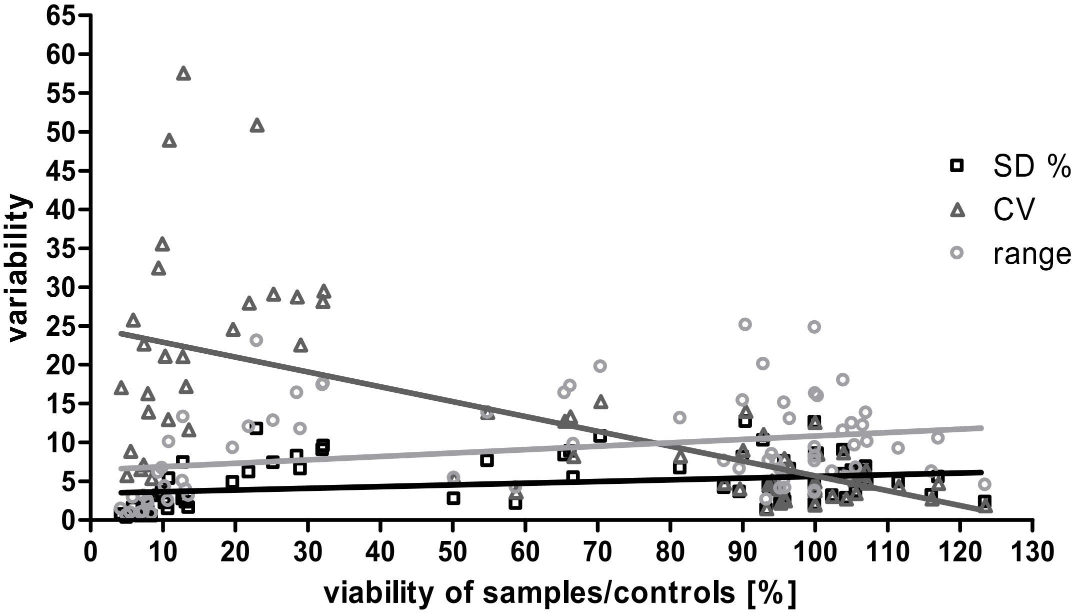


Figure 1: EPISKIN data from Phase I

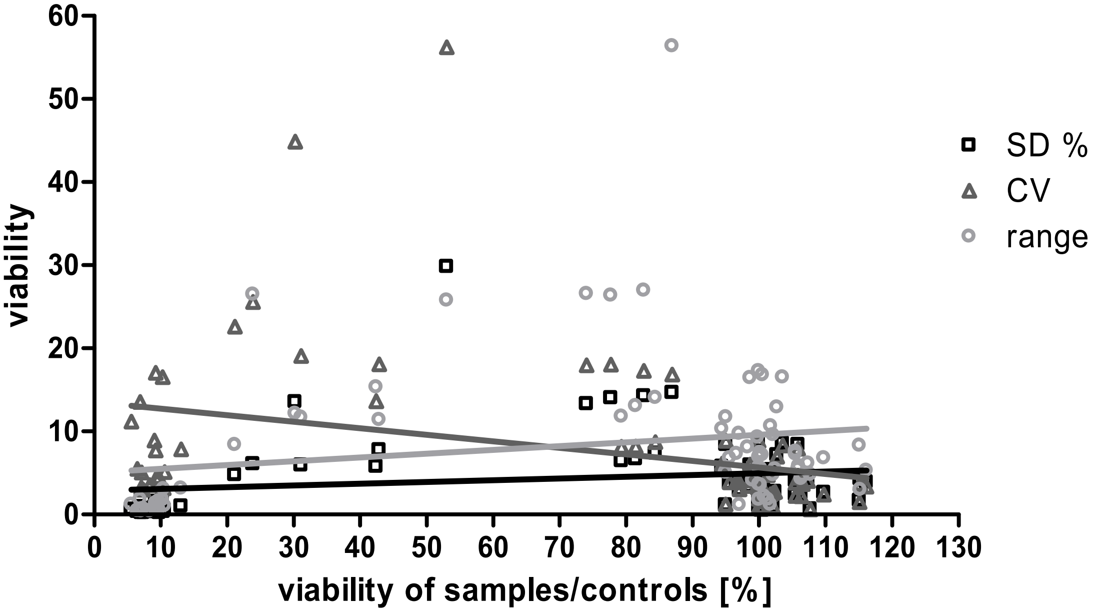


Figure 2: EpiDerm-data from Phase I

Both figures look similar: The CV is decreasing and the range and the SD are increasing with increasing viability. For both test the CV linear regression decreases stronger than the other regression lines increase. For both datasets, the SD is the most stable measure of variability over the whole range, i.e. it has the smallest absolute slope value. The CV seems less suitable as it might result in rejection of data for lower viabilities due to the bias in this measure for small viability values.

Additionally the data of Phase A were plotted with the range covered by the replicates (Figures 3 and 4). In the figures the samples, which have the largest range in responses, are highlighted in red (EPISKIN: 5 samples/controls; EpiDerm: 6 samples/controls).

The CV-criterion of 30% would reject 6 EPISKIN-controls/samples of which only two belong to the 5 highlighted chemicals and 2 EpiDerm-controls/samples, which belong to the six highlighted chemicals. In contrast, the highlighted chemicals have the overall largest SDs and of course ranges, so that a cut-off-value can easily be chosen. E.g. a common SD-cut-off of 11.0 would reject all six EpiDerm chemicals and three EPISKIN-chemical (2 in the first set and one, marked by an arrow in Fig. 4, in the second set.)

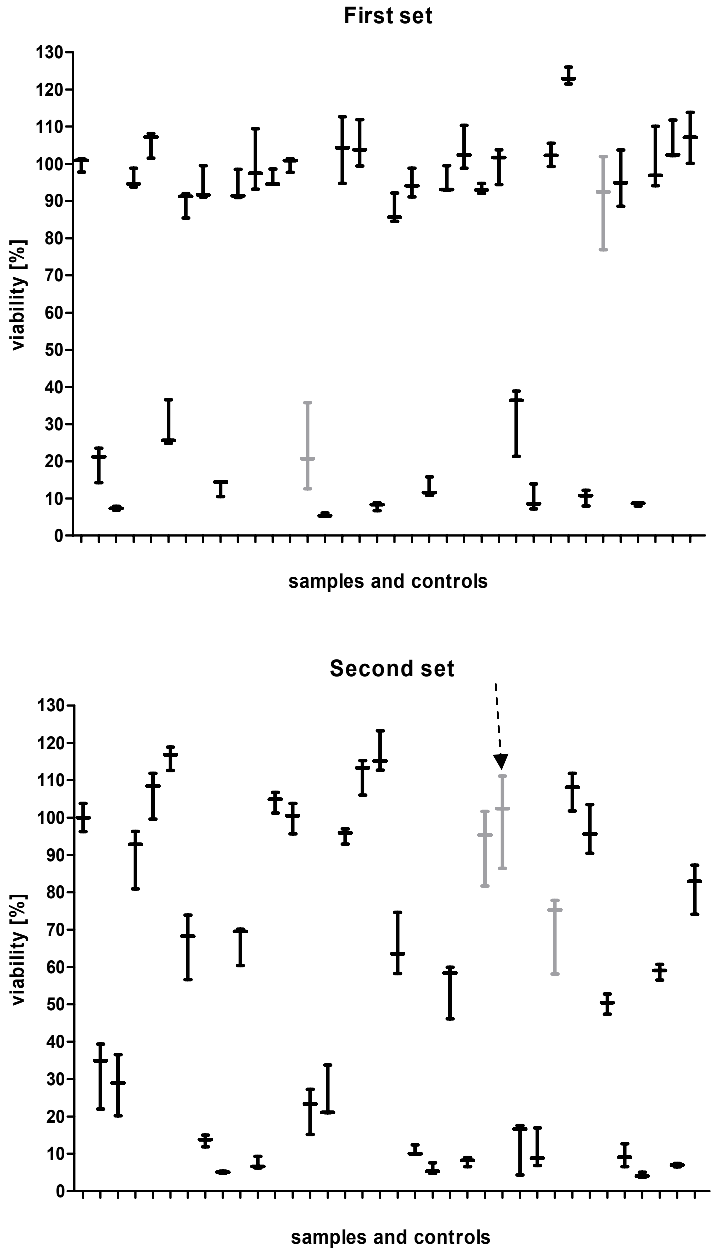


Figure 3: EPISKIN-data of Phase A. For each sample/control the range, i.e. the distance between the smallest and the highest value, and the median are presented. The most variable samples/controls are highlighted in grey.

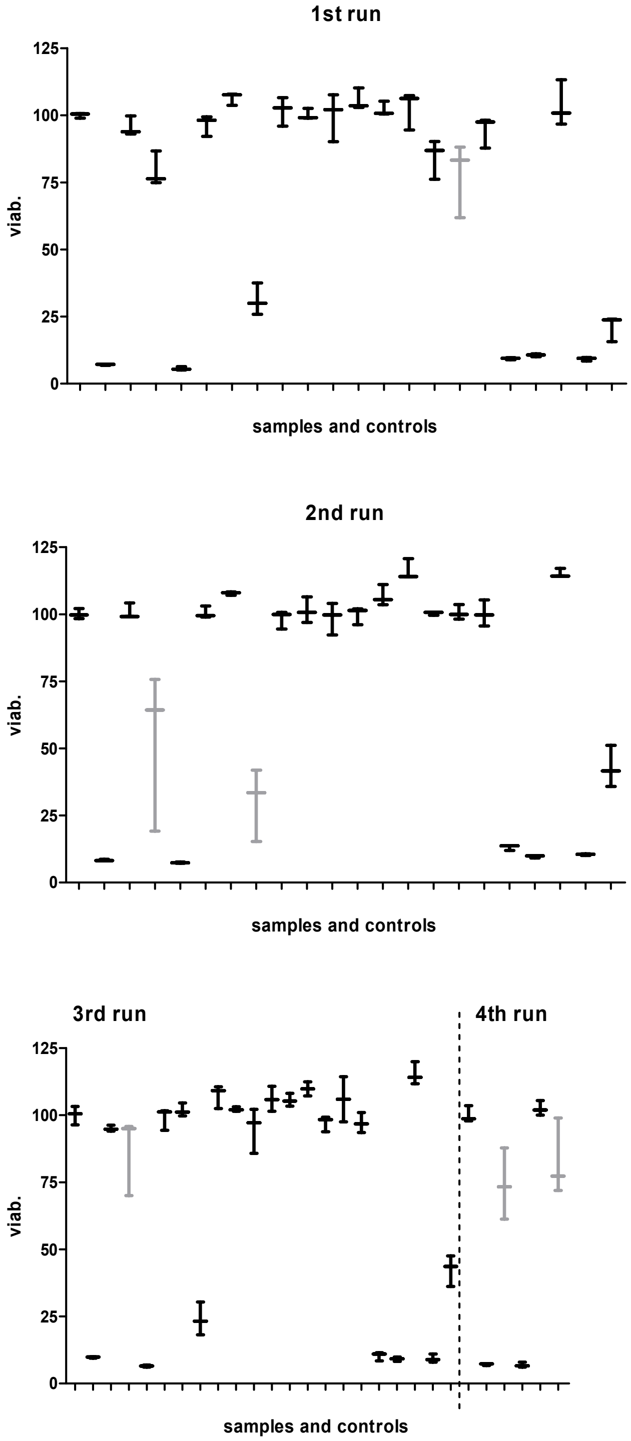


Figure 4: EpiDerm-data of Phase A. For each sample/control the range, i.e. the distance between the smallest and the highest value, and the median are presented. The most variable samples/controls are highlighted in grey.

Annex VII

Interim within-assay variability analysis

Based on the analysis of the Phase I data, the variability criterion allowing the assessment of the data quality of a test compound or included control was modified. The empirical data generated in Phase I showed only small variability of replicates, so that a standard deviation of 11 was defined as the cut-off value. In case this value was exceeded, the testing would have to be repeated.
During the first testing period of Phase II, it became however evident that this criterion cut-off was set too restrictive. In order to be able to adjust it, L'Oréal – the EPIKSIN lead laboratory – provided historical data of negative controls from 58 evaluated batches from the years 2002 to 2004. The respective standard deviations of these data are plotted in Figure 1. Based on this information, the variability criterion cut-off was set to a standard deviation of 18.

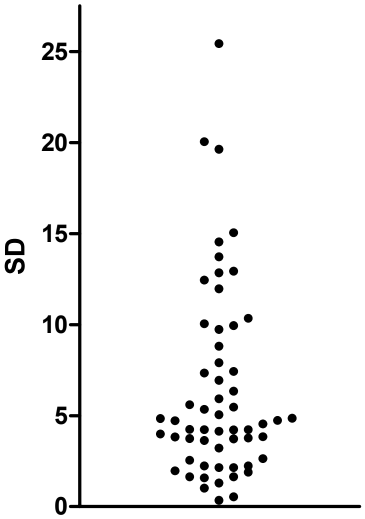


Figure 1: Standard deviations of negative controls of 58 batches tested.

Regarding EpiDerm, it was agreed to have a preliminary analysis of this variability aspect with the MTT-data from the first 30 Phase II chemicals while the coding was maintained. As from each laboratory three runs were already available, the proportion of chemicals with three passing runs depending on four standard deviation cut-off value, i.e. 11, 12, 15 and 18, was established. For each case, we further analysed how many chemicals would require retesting and how many chemicals would fail (Figure 2). Also these data favoured a cut-off value of 18.

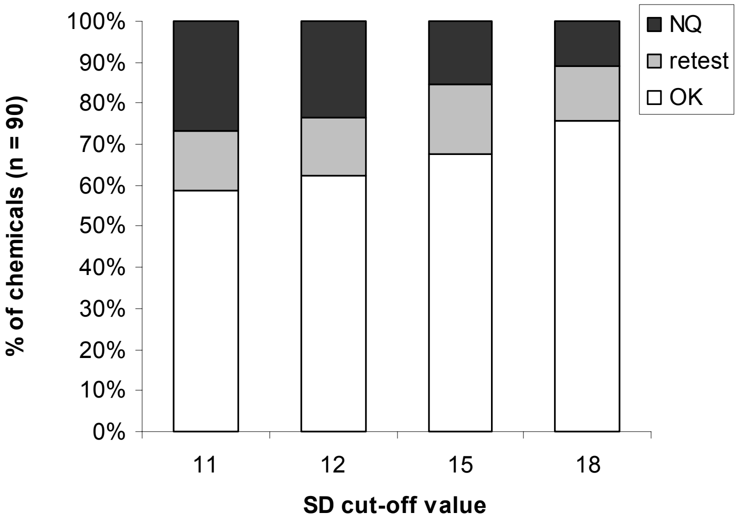


Figure 2: Proportion of qualified (OK), non-qualified (NQ) and to-be-retested chemicals (retest) in the first set of 30 chemicals tested at the three EpiDerm laboratories.

Annex VIII

Predictive values of the test systems

In order to support the relevance assessment in terms of predictive capacity, the negative and positive predictive values were calculated for prevalences of skin irritants, i.e. the proportion of irritant chemicals in a defined population of chemicals, ranging from 0 to 100%. For this exercise a specificity of 88.76% and sensitivity of 60.11% were assumed (see Table 29). Considering the sum of both predictive values, EpiDerm reached the maximum of 1.524 for a prevalence of 40%, i.e. at the intersection point of both curves (Figure 1). For prevalences between 29% and 52%, where the sum is always larger 1.50, the trade-off between both parameters is almost linear. The study prevalence of 26/58 = 43.1%, indicated by the dotted vertical line in Figure 1, fell into this area.

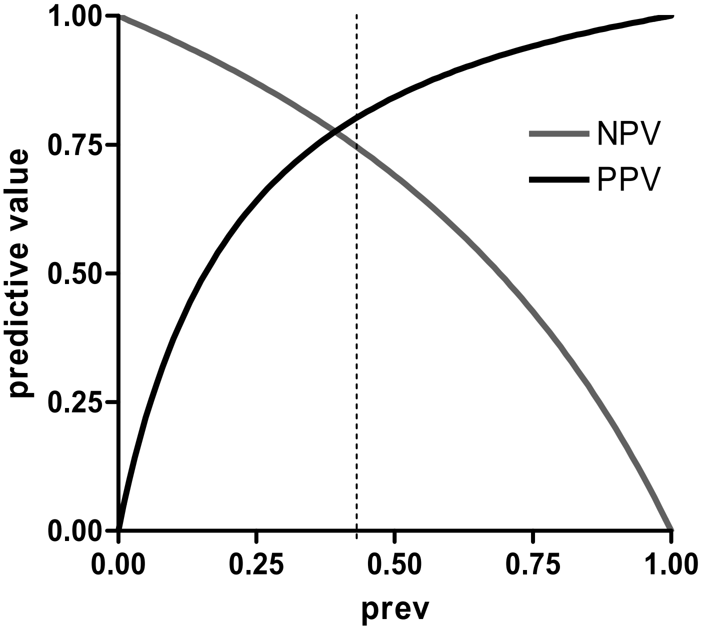


Figure 1: Curves of the negative and positive predicted values over the entire prevalence range for EpiDerm
(The dotted line indicates the study prevalence of 43.1%.)

In the context of the known prevalence of new chemicals[7], which is about 8%, EpiDerm as a stand-alone test would have a negative predicted value (NPV) of 96.2% and a positive predicted value of 29.3%. In other words, in such a scenario only 3.9% of the negative results would be false negative, but over 70% of the positives would be false positives.

For EPISKIN, assuming a specificity of 80.70% and a sensitivity of 77.62% (see Table 54) the sum of both predictive values reached the maximum of 1.584 for a prevalence of 49%, i.e. at the intersection point of both curves. For prevalences between 35% and 63%, where the sum is always larger 1.55, the trade-off between both parameters is almost linear. The study prevalence of 26/58 = 43.1%, indicated by the dotted vertical line in Figure2, fell into this area.

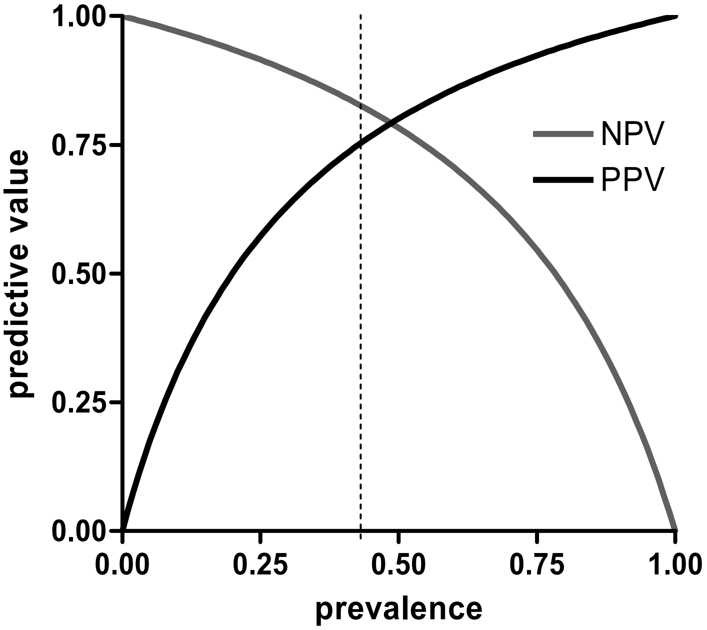


Figure 2: Curves of the negative and positive predicted values over the entire prevalence range for EPISKIN (MTT only)
(The dotted line indicates the study prevalence of 43.1%.)

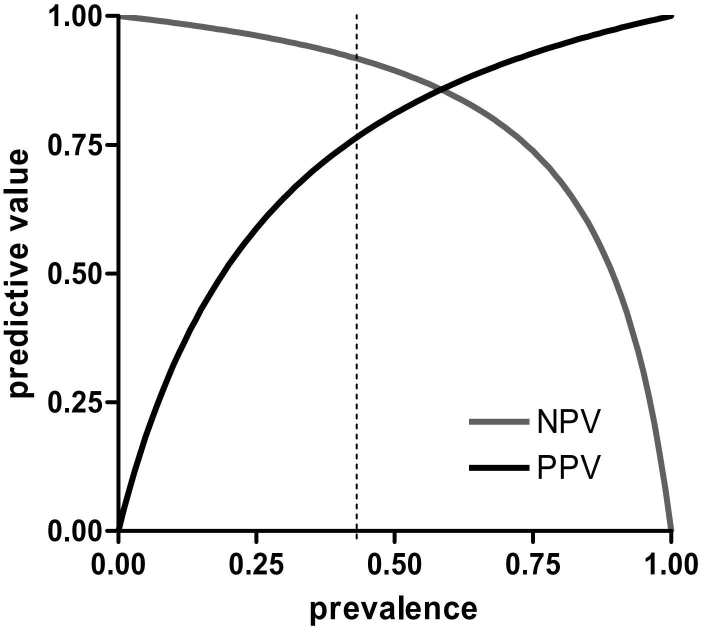


Figure 3: Curves of the negative and positive predicted values over the entire prevalence range for EPISKIN (strategic use of MTT and IL1-α)
(The dotted line indicates the study prevalence of 43.1%.)

In the context of the known prevalence of new chemicals[7], which is about 8%, EPISKIN as a stand-alone test would have a negative predicted value (NPV) of 97.6% and a positive predicted value of 25.9%. In other words, in such a scenario only 2.4% of the negative results would be false negative, but over 70% of the positives would be false positives.

Performing the respective calculations for the predictive capacity of the strategic combinations of both endpoints for EPISKIN with a specificity of 78.79% and a sensitivity of 90.67% (see Table 67), resulted especially in a difference in the NPV (Figure 3). The NPV was increased for low prevalence values as with the increased sensitivity the number of false negative results would be decreased. Indeed, applying this test to the population of new chemicals with a prevalence of 8% would result in a NPV of 99.0% and a PPV of 27.1%.