Skin Irritation Validation Study

Phase I: Interim Analysis

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.

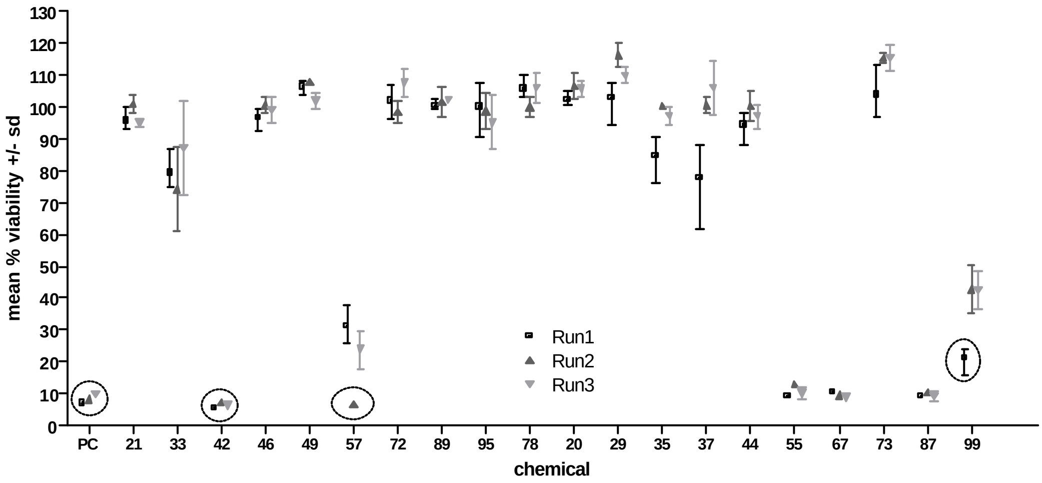


Figure 1: Phase-1 data from ZEBET with EpiDerm (PC: positive control). The encircled chemicals showed significant differences between runs in an one-way ANOVA.

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	-	-	-
S	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	S
PM	non-irritant	10	4	14
	irritant	1	5	6
	S	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			
		Non-irritant	Mild irritant	Irritant	S
GHS-PM	Non-irritant	9	4	0	13
	Mild irritant	0	1	0	1
	Irritant	0	2	4	6
S		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.

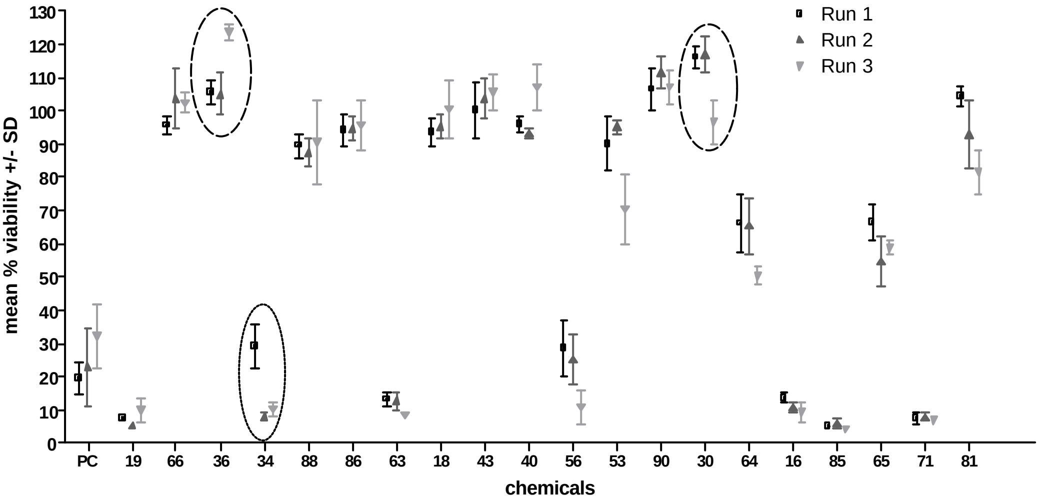


Figure 2: Phase-1 data from L'Oreal with EPISKIN (PC: positive control). The encircled chemicals showed significant differences between runs in an one-way ANOVA.

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	-	-	-
S	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	S
PM	non-irritant	10	3	13
	irritant	1	6	7
	S	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			S
		Non-irritant	Mild irritant	Irritant	
GHS-PM	Non-irritant	9	2	0	11
	Mild irritant	0	2	0	2
	Irritant	0	3	4	7
	S	9	7	4	20

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

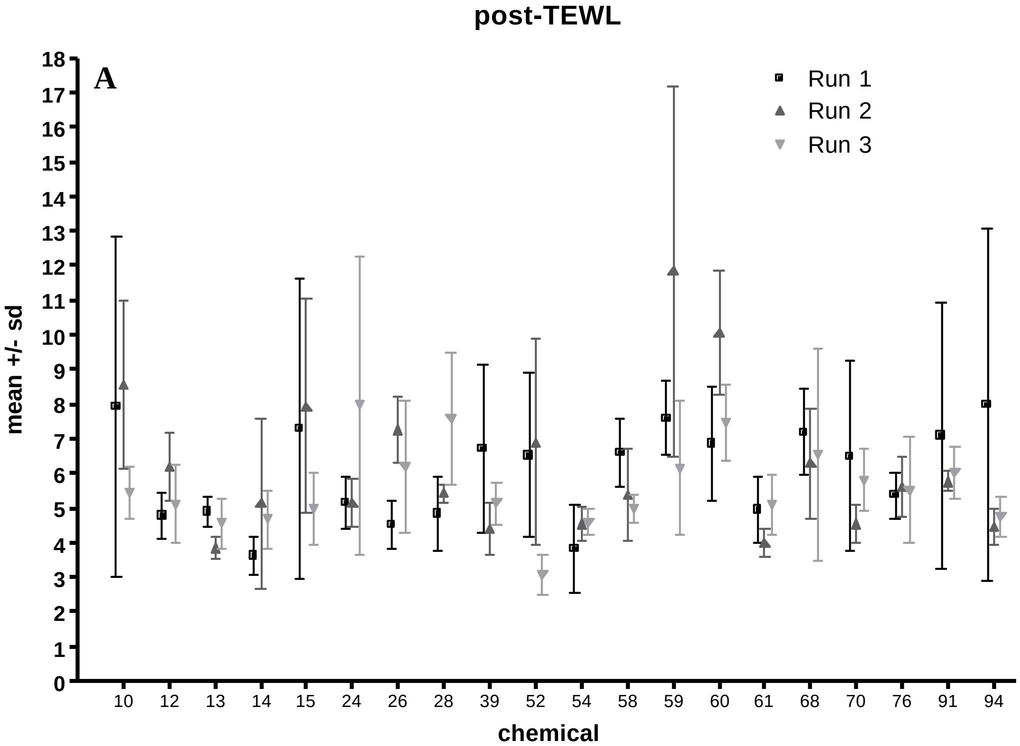
Despite an accuracy of 75%, this data analysis indicates that EPISKIN is not capable to distinguish the three GHS-classes as the mild-irritants are assigned to all PM-classes.

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.



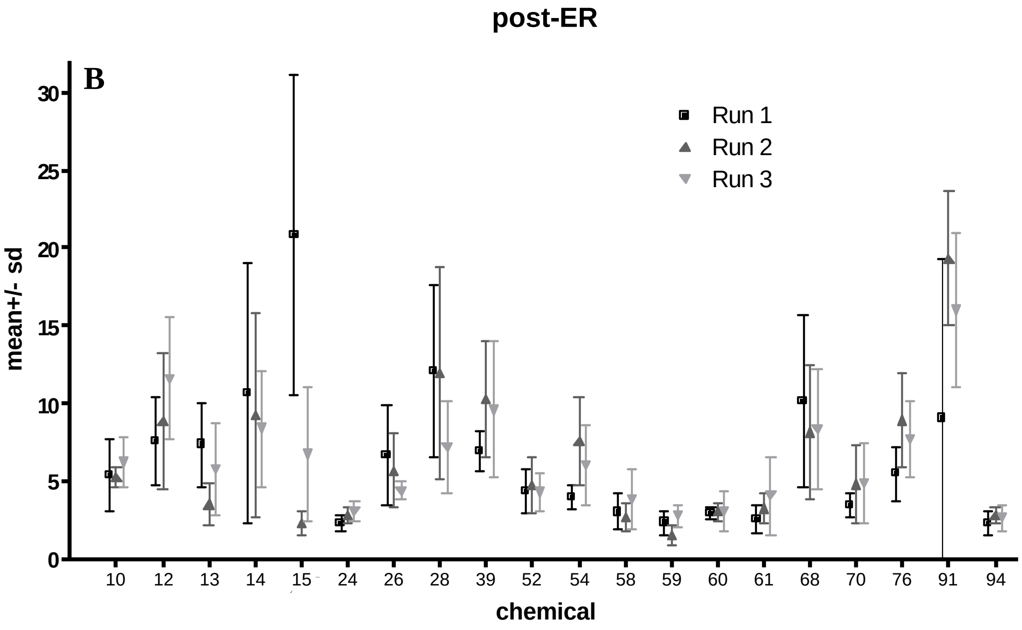


Figure 3: Phase-1 data from Syngenta with SIFT. A: post-TEWL. B: post-ER. The encircled chemical showed significant differences between runs in an one-way ANOVA.

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.

SIFT		European classification		
		no label	R38	S
PM	non-irritant	7	7	14
	irritant	4	2	6
	S	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.

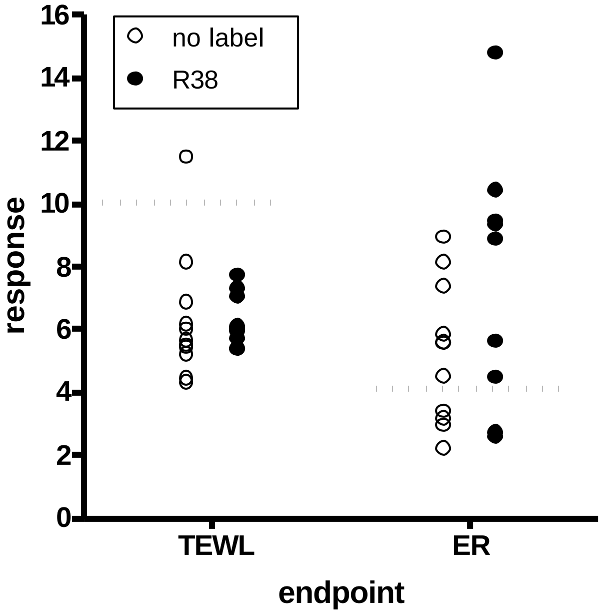


Figure 4: Mean values over the runs per endpoint of SIFT in Phase 1 arranged by European classification and with the endpoint-specific threshold as dotted lines

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.