Research Article

The Borderline Range of Toxicological Methods: Quantification and Implications for Evaluating Precision

Maria Leontaridou [1,2], *Daniel Urbisch* [2], *Susanne N. Kolle* [2], *Katharina Ott* [2], *Denis S. Mulliner* [3], *Silke Gabbert* [1] *and Robert Landsiedel* [2]

[1]Wageningen University, Environmental Economics and Natural Resources Group, Wageningen, The Netherlands; [2]BASF SE, Experimental Toxicology and Ecology, Ludwigshafen, Germany; [3]BASF SE, Computational Chemistry and Biology, Ludwigshafen, Germany

Summary

Test methods to assess the skin sensitization potential of a substance usually use threshold criteria to dichotomize continuous experimental read-outs into yes/no conclusions. The threshold criteria are prescribed in the respective OECD test guidelines and the conclusion is used for regulatory hazard assessment, i.e., classification and labelling of the substance. We can identify a borderline range (BR) around the classification threshold within which test results are inconclusive due to a test method's biological and technical variability. We quantified BRs in the prediction models of the non-animal test methods DPRA, LuSens and h-CLAT, and of the animal test LLNA, respectively. Depending on the size of the BR, we found that between 6% and 28% of the substances in the sets tested with these methods were considered borderline. When the results of individual non-animal test methods were combined into integrated testing strategies (ITS), borderline test results of individual tests also affected the overall assessment of the skin sensitization potential of the testing strategy. This was analyzed for the 2-out-of-3 ITS: Four out of 40 substances (10%) were considered borderline. Based on our findings we propose expanding the standard binary classification of substances into "positive"/"negative" or "hazardous"/"non-hazardous" by adding a "borderline" or "inconclusive" alert for cases where test results fall within the borderline range.

Keywords: non-animal methods, variability, borderline range, skin sensitization

1 Introduction

Skin sensitizers are substances that can lead to an allergic response following skin contact (UNECE, 2011). An individual may be sensitized upon first contact. Subsequent contact can then provoke allergic contact dermatitis (ACD). It is estimated that ACD affects about 20% of the European and North American population at least once in their lifetime, although there is considerable variation of skin sensitization prevalence between different age-sex groups (Thyssen et al., 2007).

Data on skin sensitization potential have to be provided for all substances produced or manufactured above one ton per year under the European chemicals legislation REACH, and for classification and labelling of substances under the European CLP regulation (ECHA, 2016). The assessment of a substance's skin sensitization potential has been traditionally based on data derived from animal tests, such as the guinea pig based tests described in OECD TG 406 (OECD, 1992) or the murine local lymph node assay (LLNA) described in OECD TG 429 (OECD, 2002, 2010).

However, animal welfare concerns and regulatory developments, e.g., the Cosmetics Regulation (EC, 2009) and the REACH legislation (EC, 2006), have driven efforts to move away from animal to non-animal testing. A number of non-animal test methods have been developed (Mehling et al., 2012; Reisinger et al., 2015), two of which, namely the direct peptide reactivity assay (DPRA) (Gerberick et al., 2004, 2007) and the antioxidant response element - nuclear factor erythroid 2 (ARE-Nrf2) luciferase test methods covered by KeratinoSens™ (Natsch et al., 2011), have been validated by the European Centre for Validation of Alternative Methods (ECVAM; Italy) and are described in the OECD TG 442C and

Received June 27, 2016;
Accepted February 16, 2017;
Epub February 23, 2017;
doi:10.14573/altex.1606271

442D (OECD, 2015a,b). LuSens (Ramirez et al., 2014, 2016) also covers the ARE-Nrf2 luciferase test method and is currently undergoing validation. Another non-animal test method, the human cell line activation test (h-CLAT) (Ashikaga et al., 2010, 2006; Sakaguchi et al., 2006, 2010) has recently been validated by ECVAM and is described in OECD TG 442E (OECD, 2016a).

The sequential structure of molecular and cellular mechanisms causing ACD is represented by the "adverse outcome pathway" (AOP) for skin sensitization, consisting of eleven causally linked steps, four of which were defined to be essential and specific ("key events") (OECD, 2012a,b). The DPRA, the ARE-Nrf2 test methods and the h-CLAT cover the first three key events of the skin sensitization AOP.

For hazard classification purposes, i.e., for assessing skin sensitization potential, continuous data obtained from animal tests or from non-animal test methods are dichotomized into binary "positive"/"negative" information (Van der Schouw et al., 1995; Hoffmann and Hartung, 2005). The prediction models used for the DPRA, LuSens and the h-CLAT are described in OECD TG 442C (OECD, 2015a), Ramirez et al. (2014, 2016), and in the OECD TG 442E (OECD, 2016a), respectively. Based on the threshold for classification, a test method's accuracy, i.e., the percentage of true positive and true negative classifications, can be determined (see, for example, Yerushalmy, 1947; Cooper et al., 1979).

The experimental data obtained from a test method are, however, subject to biological and technical variability. Consequently, repeated testing may result in discordant classification results. This affects the precision of a test method, defined as the ability of a test method to deliver concordant results in repeated applications. The problem of intra- and inter-assay variability of *in vitro* methods has been observed earlier (Hothorn, 2002, 2003). Luechtefeld et al. (2016) pointed to a limited intra-assay reproducibility of skin sensitization potential and potency data.

This paper focuses on the intra-assay variability of test methods for skin sensitization potential assessment. Specifically, we analyze limitations with regard to the reproducibility of results when continuous dose-response data are transformed into "toxic"/"non-toxic" outcomes. Kolle et al. (2013), Hoffmann (2015), Dumont et al. (2016) and Dimitrov et al. (2016) analyzed the intra-assay variability of the LLNA. Kolle et al. (2013) showed that for those substances for which the estimated concentration (EC3) led to a stimulation index (SI) value which was relatively close to the threshold for classification (i.e., $SI = 3$), repeated testing resulted in positive and negative classifications of their skin sensitization potential. Kolle et al. (2013) defined a range around the classification threshold of the LLNA, within which discordant outcomes can be expected, by determining coefficients of variation based on individual animal data. This range is called the "borderline range" (BR) (Kolle et al., 2013) or "grey zone" (Dimitrov et al., 2016). The percentage of substances that fall into the BR of a test method's prediction model reflects how limited a test method's precision is.

Analyses of the BR for non-animal test methods used for skin sensitization potential assessment have not been conducted before. The aim of this paper is, therefore, to examine the impact of technical and biological variability on the precision of selected non-animal test methods for skin sensitization potential assessment. Moreover, we examine how the precision of the non-animal test methods and that of the animal test LLNA is affected by variations of the BR. For this purpose, we suggest an approach to quantify BRs for the non-animal test methods DPRA, LuSens, h-CLAT and for the LLNA, based on results obtained from a large number of experiments.

The approach to quantify the BR, and the decision rules for applying the BR to the prediction models of individual test methods are described in Section 2. Results from quantifying the BR for each individual test method are presented in Section 3.1. Borderline substances (i.e., substances that produced test results within the BR) detected in the experimental sets of individual test methods are shown in Section 3.2. In addition, we suggest a decision rule for applying the BR to a combination of the DPRA, LuSens and the h-CLAT in the 2-out-of-3 ITS. Section 3.3 shows borderline substances for the 2-out-of-3 ITS. Section 4 discusses the implications of considering the BR in non-animal test methods' prediction models, the LLNA, and the 2-out-of-3 ITS, respectively.

2 Materials and methods

2.1 Test methods

The three non-animal test methods DPRA, LuSens and h-CLAT were developed to address the three key events of the AOP in order to assess a substance's skin sensitization potential. We compared our findings to those of the LLNA, considered the *in vivo* reference test, in order to evaluate the precision of the methods. The number of substances used to quantify the BR was 42 for the DPRA, 26 for LuSens, 13 for h-CLAT and 22 for LLNA. The BR was quantified using results from a large number of experimental runs of each test method. Information about the substances used to determine the BR for each test method, the number of experimental runs conducted and the substance concentrations used is provided in Appendix 1, Tables S1.1-S1.4 in the supplementary file at doi:10.14573/altex.1606271s. Where substance names could not be provided due to data confidentiality substances were numbered consecutively.

The experimental sets to which the BR concept was applied in order to identify borderline substances consisted of 199 substances for the DPRA, 79 for LuSens, 40 for h-CLAT and 22 substances for LLNA, see Bauch et al. (2012) and Urbisch et al. (2015, 2016). The composition of these sets is presented in Appendix 3, Tables S3.1-S3.4 at doi:10.14573/altex.1606271s.

2.1.1 The Local Lymph Node Assay

The Local Lymph Node Assay (LLNA) became the "first choice" animal test for the assessment of skin sensitization potential (Kimber et al., 1994). It is described in OECD TG 429, which was first published in 2002 and updated in 2010 (OECD, 2002, 2010). In the LLNA, the proliferation of lymphocytes in auricular draining lymph nodes induced by test substances is quantified by comparing the mean proliferation in each test group to the mean proliferation in the vehicle treated control group. The ratio of the mean proliferation in each treated group to that in the concurrent vehicle control group, termed the stimulation index (SI), is determined. The classification threshold T of the LLNA is $SI = 3$. If $SI > 3$ a substance is classified a skin sensitizer.

2.1.2 The Direct Peptide Reactivity Assay

The Direct Peptide Reactivity Assay (DPRA) was developed by Gerberick et al. (2004, 2007). The DPRA has been formally validated and the OECD Testing Guideline TG 442C was adopted in 2015 (OECD, 2015a). In the DPRA, depletion of two model peptides containing a cysteine and a lysine residue, respectively, as a reactive nucleophilic center is measured after incubation with a test substance. The classification threshold T of the DPRA is the mean depletion of 6.38% of the two peptides compared to the depletion in the reference controls (OECD, 2015a). If the mean lysine and cysteine peptide depletion is above this threshold, a test substance is considered to be peptide-reactive. According to OECD TG 442C, the DPRA can be used, together with complementary information, to discriminate sensitizers from non-sensitizers. Depending on the regulatory framework, a positive result of the DPRA can serve as standalone information for classifying substances into Category 1 for skin sensitization. However, as emphasized in the ECHA Guidance on Information Requirements and Chemical Safety Assessment Chapter R.4a (ECHA, 2016), the DPRA should not be used in isolation for identifying a skin sensitizer or non-sensitizer.

2.1.3 The ARE-Nrf2 luciferase method

The ARE-Nrf2 luciferase method utilizes the gene induction regulated by the antioxidant response element (ARE) in transgenic human keratinocyte cell lines. OECD TG 442D (OECD, 2015b) was adopted in 2015. The ARE-Nrf2 luciferase method is covered by KeratinoSens™ (Natsch et al., 2011) and LuSens (Ramirez et al., 2014). The LuSens assay was used in this study. In ARE-Nrf2 luciferase methods, the keratinocyte activating potential is determined by measuring luciferase induction after treatment with a test substance relative to concurrent vehicle controls. The classification threshold T for LuSens is $FI = 1.50$, above which a substance is considered to have a keratinocyte activating potential. Similar to the DPRA, LuSens is not considered suitable for classifying substances as skin sensitizers or non-sensitizers if used in isolation (ECHA, 2016).

2.1.4 The human Cell Line Activation Test

The human Cell Line Activation Test (h-CLAT) (Ashikaga et al., 2006, 2010; Sakaguchi et al., 2006, 2010) determines the dendritic cell activating potential by measuring the induction of the expression of the cell surface markers CD54 and CD86 after treatment with a test substance relative to concurrent vehicle controls in immortalized human monocytic leukemia THP-1 cells as surrogate dendritic cells. As indicated in OECD TG 442E (OECD, 2016a), the classification thresholds T for the h-CLAT are CD54 $FI = 1.50$ and CD86 $FI = 2.00$ at relative cell viabilities of at least 50%. As for the DPRA and LuSens, the method only addresses one key event of the skin sensitization AOP and should not be used in isolation to classify skin sensitization potential (ECHA, 2016).

2.1.5 The 2-out-of-3 ITS for characterizing skin sensitization potential

The 2-out-of-3 ITS (Bauch et al., 2012; Urbisch et al., 2015; OECD, 2016b,c; see also Sauer et al., 2016) is an integrated testing strategy for the assessment of skin sensitization potential. According to this approach, 2 out of 3 concordant test results using the DPRA, ARE-NrF2 luciferase method, and the h-CLAT determine the prediction. The ARE-NrF2 luciferase method can be covered by LuSens or KeratinoSens™. The 2-out-of-3 ITS addresses the first three consecutive key events of the AOP for skin sensitization and is a selected case study for integrated approaches to testing and assessment (IATA) (Urbisch et al., 2015). Applying the BR concept to the 2-out-of-3 ITS provides a measure for evaluating the performance of this specific IATA case.

2.2 Approach to quantifying the borderline range (BR)

The first step in assessing a test method's precision limit was to develop an approach to quantify the BR. The BR denotes the area around the classification threshold for which a test method's prediction model may deliver discordant results in repeated applications. For each test method considered, we derived the BR from the pooled standard deviation of a test method's results, SD_p (Eq. 1), pooled across substances i, and concentrations j (i.e., the dose in case of the LLNA). The notation used is explained in Table 1.

Tab. 1: Notation for calculating the pooled standard deviation SD_p of experimental results per substance and concentration (dose in case of the LLNA)

Notation	Explanation
T	Classification threshold in a test method's prediction model
i	Substance (i=1, …, n)
j	Concentration tested per substance i (j=1, …, k_i)
$r_{i,j}$	Number of replicates per substance i and concentration j
l	Replicate per substance i and concentration j (l=1, …, $r_{i,j}$)
$y_{i,j,l}$	Test result of substance i, concentration j and replicate l
$y_{i,j}$	Arithmetic mean of test results for substance i and concentration j

We use the pooled standard deviation SD_p to define the BR around a prediction model's classification threshold T:

$$BR = \{T - SD_p, T + SD_p\}. \qquad (1)$$

Thus, it is assumed that the BR is symmetric around the classification threshold. For a given test method, the SD_p of experimental results retrieved from testing different substances and concentrations is calculated as follows:

$$SD_p = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(r_{i,j}-1)*\sigma_{i,j}^2}{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(r_{i,j}-1)}}, \qquad (2)$$

where $\sigma_{i,j}^2$ is the variance of results for substance i and concentration j. The standard deviation per substance i and concentration j is given by:

$$\sigma_{i,j} = \sqrt{\frac{\sum_{l=1}^{r_{i,j}}(y_{i,j,l}-\bar{y}_{i,j})^2}{(r_{i,j}-1)}}, \qquad (3)$$

which acknowledges that different replicates can be generated for a certain concentration. A numerical example illustrating the approach described above is presented in Appendix 2 at doi:10.14573/altex.1606271s. We consider the part of the distribution of test results that is close to the classification threshold T to be most relevant to determine the BR according to Equation 1. Therefore, we use SD_p values from pre-defined ranges of test results around the threshold. As it cannot be determined *ex ante* how broad or narrow a range should be, we conducted a sensitivity analysis considering different ranges. In this way, we gain insight into the relationship between the size of the BR and the number of borderline substances. The ranges are shown in Table 2. Note that the BR approach suggested in this paper goes beyond that described by Kolle et al. (2013), who calculated the BR only for the LLNA based on individual animal data.

In case of the DPRA, the BR was quantified using results from repeatedly testing 42 substances with different substance concentrations. The tests were conducted in a GLP-certified laboratory of BASF SE, yielding 446 individual results including the positive control (see Appendix 1, Tab. S1.1 at doi:10.14573/altex.1606271s). The cysteine and lysine depletion per substance and concentration were randomly paired. For each pair, we determined the mean peptide depletion rate (MPD) per substance and concentration. The ranges of test results considered for calculating the SD_p are presented in Table 2. For each SD_p, the corresponding BR was determined according to Equation 3.

The BR in the LuSens prediction model was calculated using test results from repeatedly testing 26 substances, including the positive and negative control, yielding 2206 individual results covering different concentrations per substance (see Appendix 1, Tab. S1.2 at doi:10.14573/altex.1606271s). Again, experiments were conducted in a GLP-certified laboratory of BASF SE (using the Multimode Reader TriStar2 luminometer from Berthold Technologies, Germany), using a luciferase $FI = 1.50$ as classification threshold T. Based on the available dataset the SD_p was calculated for defined ranges of test results (Tab. 2). Only test substance concentrations with at least 70% relative viability were included in the analysis. The BR corresponding to each SD_p was determined according to Equation 3.

The BR around the classification threshold of the h-CLAT was calculated using test results from 13 substances tested during routine (in house) test applications, yielding 528 individual measurements covering different substances and concentrations

Tab. 2: Number of substances, individual results, and range of test results used for determining the pooled standard deviation SD_p around the classification threshold T in the prediction model of the test methods

Test method	Number of substances tested (*n*)	*T*	Range of test results around *T* used for calculating SD_p	Number of individual results from testing *n* substances (with k_i concentrations per substance)
DPRA	42	Mean lysine and cysteine depletion (%); *T* = 6.38%	MPD ≤ 20% MPD ≤ 13% 3.38% ≤ MPD ≤ 9.38%	238 210 76
LuSens	26	*FI* = 1.5	*FI* ≤ 5 *FI* ≤ 3	508 491
h-CLAT	13	CD54 *FI* = 2	*FI* CD54 ≤3 *FI* CD54 ≤ 5	513 473
		CD86 *FI* =1.5	*FI* CD86 ≤3 *FI* CD86 ≤ 5	474 403
LLNA	22	*SI* = 3	0≤ *SI* ≤ 6 1 ≤ *SI* ≤ 5 2 ≤ *SI* ≤ 4 2.5 ≤ *SI* ≤ 3.5	381 270 96 39

T, classification threshold; *SI*, stimulation index; *FI*, fold induction; CD54/CD86, cell surface marker expression; MPD, mean peptide depletion; see Appendix 1 at doi:10.14573/altex.1606271s for list of substances in the substances sets used for quantifying the borderline range (BR).

Tab. 3: Decision rule for concluding on the overall test result of LuSens from two consecutive concentrations in a run

	Concentration x	Concentration (x+1)	Overall test result
Non-animal test method results	N	N	N
	P	P	P
	B	B	B
	N	B	N
	B	P	B

N, negative test result, indicating that a substance has no keratinocyte activating potential; P, positive test results, indicating that a substance has keratinocyte activating potential; B, substances with test results within the borderline range (BR)

(see Appendix 1, Tab. S1.3 at doi:10.14573/altex.1606271s). The SD_p was quantified for defined ranges of fold inductions (*FI*) of CD54 and CD86 expressions documented in Table 2, and for substance concentrations with at least 50% relative viability. The BR corresponding to each SD_p was calculated according to Equation 3.

The BR of the LLNA was quantified using test results obtained from the 22 performance standard (PS) substances (ICCVAM, 2009) that were repeatedly tested according to GLP, yielding 479 test results for substances tested at different concentrations (see Appendix 1, Tab. S1.4 at doi:10.14573/altex.1606271s). Like for the non-animal test methods, the SD_p was determined for different data ranges (Tab. 2).

2.3 Decision rules for identifying borderline substances in experimental sets tested with individual non-animal methods

The BR, determined according to the approach described in Section 2.2, can be applied to experimental sets of substances tested with non-animal methods. The aim is to detect those substances for which results fall within the BR and, hence, for which a clear-cut classification is not possible with sufficient confidence.

Depending on the prediction model of the individual non-animal methods, the application of the BR differed. In case of the DPRA, substances were defined as borderline if the mean depletion rate was within the BR (see also Tab. 6 in Section 3.1).

As described in Ramirez et al. (2014), the prediction model of LuSens requires that two consecutive concentrations per run reveal results above the classification threshold in order to classify the test substance as positive. For LuSens, therefore, we first established decision rules for determining the result across all concentrations considered in a run. As illustrated in Table 3, for a given BR around the classification threshold of the LuSens prediction model, the outcome of a run was concluded to be positive if all results were above the upper margin of the BR. If the first concentration (denoted x in Tab. 3) gave a negative result and the next concentration (x+1) was either borderline or negative, it was concluded that the overall test outcome of the run was negative. If LuSens revealed a borderline result for a certain concentration x and the next concentration (x+1) was borderline or positive, the substance was considered borderline.

Tab. 4: Decision rules for LuSens and h-CLAT to conclude on the overall test result from repeated runs[a]

Combinations of dichotomized results from repeated runs	Overall conclusion
N, N	N
P, P	P
B, B	B
B, P, P	P
B, N, N	N
B, B, N	B
B, B, P	B
N, P, B	B
N, N, B, B	B
P, P, B, B	B
N, N, P, B	B
P, P, N, B	P
N, P, B, B	B

[a] Combinations do not imply a defined order of results;
N, negative test result, i.e., a substance does not have keratinocyte (LuSens) or dendritic cell (h-CLAT) activating potential;
P, positive test result, i.e., a substance has keratinocyte (LuSens) or dendritic cell (h-CLAT) activating potential;
B, test result falls within the borderline range (BR) determined for either LuSens or h-CLAT.

In case of the h-CLAT, at least one of the test results of either the CD54 expression or the CD86 expression from at least one of the runs in an experiment has to fall into the BR for the experimental result to be borderline. Hence, the conclusion on the overall result of the experiment (positive, negative) is based on results from just one concentration.

Second, we established a decision rule that allows concluding on the overall test result across runs. This was necessary because the testing protocols for LuSens and the h-CLAT require conducting two or more runs in order to classify a substance according to the results. In case of LuSens, a complete experiment consists of two independent runs (each of which covers different concentrations, see Ramirez et al. (2014)). If the results from two runs are discordant, a third run has to be conducted and the conclusion on a substance's skin sensitization potential is based on the majority outcome. Similarly, a test substance is tested in the h-CLAT in two independent runs. If the results are discordant, another run has to be performed (OECD, 2016a). Acknowledging that dichotomized test results can be positive, negative or borderline, adopting a final conclusion on a substance's sensitization potential may take up to four runs. The corresponding decision rules are shown in Table 4.

2.4 Decision rules for identifying borderline substances tested with the 2-out-of-3 ITS

The BR of the prediction models of individual non-animal test methods changes the possible outcomes of each method to negative, positive, or borderline. Since test results of borderline substances can (by definition) not unambiguously be denoted as positive or negative, these results cannot be compared with results from a reference animal test in order to conclude whether the test result is FP (i.e., erroneously classified as positive) or FN (i.e., erroneously classified as negative).

The skin sensitization potential, however, is assessed by a combination of the results of non-animal test methods addressing different steps of the AOP (Jaworska, 2016; Kleinstreuer et al., 2016; Strickland et al., 2016). One of the simplest, yet successful, ways to do this, is the 2-out-of-3 ITS (Bauch et al., 2012; Urbisch et al., 2015). The 2-out-of-3 ITS uses dichotomized results of individual non-animal test methods (i.e., positive or negative). If a borderline/ambiguous outcome of an individual test method is considered in the 2-out-of-3 ITS, the overall conclusion on the skin sensitization potential of a test substance may also be borderline/ambiguous (or negative or positive). The 2-out-of-3 ITS assigns equal weight to each test method. The order of results of the individual methods does not matter. Consequently, one test method yielding a borderline/ambiguous result will not change the overall result of the 2-out-of-3 ITS if the other two methods provide concordant – negative or positive

Tab. 5: Decision rules to conclude on the overall result using the 2-out-of-3 ITS when considering borderline substances in individual non-animal testing methods

Dichotomized result from non-animal test methods A/B/C[a]	Overall conclusion
N, N, N	N
N, N, B	N
N, B, B	B
N, B, P	B
P, P, P	P
P, P, B	P
B, B, B	B

[a] A/B/C does not imply a sequential order of testing in the 2-out-of-3 ITS;
N, negative test result, i.e., substance does not have a peptide reactivity (DPRA), keratinocyte activating (LuSens) or dendritic cell activating (h-CLAT) potential;
P, positive test result, i.e., substance has peptide reactivity (DPRA), keratinocyte activating (LuSens) or dendritic cell activating (h-CLAT) potential;
B, test result falls within the borderline range (BR) of the DPRA, LuSens or the h-CLAT prediction model.

Tab. 6: Ranges of test results considered for quantifying the SD$_p$, SD$_p$ and the BR in the prediction models of the non-animal test methods DPRA, LuSens and h-CLAT, and of the animal test LLNA[a]

Test method	Range of test results around *T* considered for calculating the SD$_p$	Pooled standard deviation (SD$_p$)	Borderline range (BR)

DPRA	MPD ≤ 20%
	MPD ≤ 13%
	3.38% ≤ MPD ≤ 9.38%
		5.03%
		3.49%
		1.52%
			MPD = {1.35%, 11.41%}
			MPD = {2.89%, 9.87%}
			MPD = {4.86%, 7.9%}

LuSens	*FI* ≤ 5
	FI ≤ 3
		0.244
		0.229
			FI = {1.26, 1.74}
			FI = {1.27, 1.73}

h-CLAT	CD54 *FI* ≤3
	CD86 *FI* ≤3
		0.190
		0.260
			CD54 *FI* = {1.81, 2.19}
			CD86 *FI* = {1.24, 1.76}

	CD54 *FI* ≤ 5
	CD86 *FI* ≤ 5
		0.255
		0.301
			CD54 *FI* = {1.74, 2.26}
			CD86 *FI* = {1.2, 1.81}

LLNA	0≤ *SI* ≤ 6
	1 ≤ *SI* ≤ 5
	2 ≤ *SI* ≤ 4
	2.5 ≤ *SI* ≤ 3.5
		0.709
		0.639
		0.498
		0.353
			SI = {2.20, 3.71}
			SI = {2.36, 3.64}
			SI = {2.5, 3.5}
			SI = {2.65, 3.53}

SI, stimulation index; *FI*, fold induction; CD54/CD86, cell surface marker expression; MPD, mean peptide depletion rate
[a] See Appendix 1, Tables S1.1-S1.4 at doi:10.14573/altex.1606271s for list of substances included in the sets for calculating the BR.
Source: own calculations

– results. But, if test results of two non-animal test methods fall into the BR of their prediction models, the overall outcome is considered borderline. Likewise, the overall conclusion on the result of the 2-out-of-3 ITS is borderline if the three methods yielded positive, negative and borderline/ambiguous results, respectively. Table 5 lists the overall outcome of the 2-out-of-3 ITS depending on the results of the prediction models of the individual non-animal test methods.

3 Results

3.1 Quantification of the borderline range (BR) for the DPRA, LuSens, h-CLAT and LLNA

Table 6 shows for each test method the ranges of test results used for calculating the SD_p, the corresponding SD_p and the retrieved BR values of the test methods' prediction models.

If a substance is tested with any of the test methods shown in Table 2, and if the result falls within the BR of its prediction model, a clear-cut conclusion about the substance's response in this test method is not possible with sufficient confidence. If, for instance, the BR: *SI* = {2.89%, 9.87%} is selected for the DPRA prediction model and a substance reveals a mean peptide depletion within this range, the result can neither be concluded to be negative nor to be positive. Instead, such test result would have to be qualified as "borderline" because the result is likely to vary in repeated runs.

3.2 Identification of borderline substances in experimental sets tested with the DPRA, LuSens, h-CLAT and LLNA

Substances for which test results fell within the BRs of the test methods' prediction models are listed in Table 7. Obviously, an increase in the BR referring to a certain prediction model caused the number of borderline results to increase. Depending on the size of the BR, we found the number of borderline substances to be between 20 and 57 (of 199) in case of the DPRA. Of the 79 substances tested with LuSens, 4 and 5 were considered borderline. Regarding the h-CLAT, the number of borderline substances varied between 8 and 10 (of 40), and in case of the LLNA, the number of substances considered borderline varied between 5 and 7 (of 22). A detailed list of all substances considered borderline under different BRs is presented in Appendix 3 at doi:10.14573/altex.1606271s.

Table 8 presents a list of substances considered borderline for each BR listed in Table 7. With regard to the largest BR considered for the DPRA (i.e., mean depletion between 1.35% and 11.41%), 11 of the 57 substances considered borderline were positive in the LLNA. Regarding the smallest BR considered, 9 of the 20 substances in the set tested with the DPRA revealed negative and 11 positive test results in the LLNA. Of these, one substance (salicylic acid) was also considered borderline in the LLNA (see Tab. 8). As illustrated in Figure 1, most substances considered borderline were non-sensitizers in the LLNA.

Tab. 7: Number and percentage of borderline substances in the experimental sets tested with the DPRA, LuSens, h-CLAT and LLNA[a]

Test method	Number of substances in the set	Borderline range (BR)	Number (percentage) of borderline substances	Borderline substances tested positive in the LLNA	Borderline substances tested negative in the LLNA

DPRA	199	MPD = {1.35, 11.41}
		MPD = {2.89, 9.87}
		MPD = {4.86, 7.9}
			57 (28%)
			35 (17%)
			20 (10%)
				11
				10
				8
					46
					25
					12

LuSens	79	*FI* = {1.26, 1.74}
		FI = {1.27, 1.73}
			6 (7%)
			5 (6%)
				5
				4
					1
					1

h-CLAT	40	CD54 *FI* = {1.81, 2.19}
		CD86 *FI* = {1.24, 1.76}
			8 (20%)	8	0
					0

		CD54 *FI* = {1.74, 2.26}
		CD86 *FI* = {1.2, 1.81}
			10 (25%)	10	0
					0

LLNA	22	*SI* = {2.20, 3.71}
		SI = {2.36, 3.64}
		SI = {2.5, 3.5}
		SI = {2.65, 3.53}
			7 (27%)
			7 (27%)
			6 (23%
			6 (23%)
				3
				3
				2
				2
					4
					4
					4
					4

BR, borderline range; *SI*, stimulation index; *FI*, fold induction; CD54/CD86, cell surface marker expression; MPD, mean peptide depletion rate

[a] See Appendix 3, Tables S3.1-S3.4 at doi:10.14573/altex.1606271s for list of substances in the experimental sets to which the BR approach was applied.

Source: own calculations

Tab. 8: Substances with borderline results in the DPRA, LuSens, h-CLAT or LLNA with each substance's sensitization potential and potency class according to the reference test

Test method	Borderline substances[a]	Reference test	Sensitization potential[b,c] according to reference test	Potency class (based on LLNA)[b]
DPRA	Salicylic acid[c]	LLNA	N	Non-sensitizer
	α-Hexyl cinnamic aldehyde[c]	LLNA	P	Weak / Moderate
	Geraniol	LLNA	P	Non-sensitizer
	Benzyl alcohol	LLNA	N	Non-sensitizer
	Tween 80	LLNA	N	Moderate
	3-Dimethylamino propylamine	LLNA	P	Weak
	Cis-6-Nonenal	LLNA	P	Non-sensitizer
	Ethyl vanillin	LLNA	N	Weak
	Undecylenic acid	LLNA	P	Moderate
	2-methoxy-4-methylphenol	LLNA	P	Non-sensitizer
	Ethyl benzoylacetate	LLNA	N	Moderate
	Dihydroeugenol	LLNA	P	Weak
	N,N-Diethyl-m-toluanimde	LLNA	N	Non-sensitizer
	Penicillin G	LLNA	P	Weak
	d,l-Citronellol	LLNA	P	Weak
	Pentachlorophenol	LLNA	P	Weak
	p-tert-Butyl-alpha-ethyl hydrocinnamal (Lilial)	LLNA	P	Weak
	1-Bromobutane	LLNA	N	Non-sensitizer
	Fumaric acid	LLNA	N	Non-sensitizer
	Glucose	LLNA	N	Non-sensitizer
	Sulfanilic acid	LLNA	N	Non-sensitizer
	Isopropyl myristate a	LLNA	N	Weak
	p-Aminobenzoic acid	LLNA	N	Non-sensitizer
	Tartaric acid	LLNA	N	Non-sensitizer
	Zinc sulfate	LLNA	N	Non-sensitizer
	Dioctyl ether	LLNA	N	Non-sensitizer
	2,2-Azobis phenol	LLNA	N	Weak
	Benzaldehyde	LLNA	N	Non-sensitizer
	Farnesal	LLNA	N	Weak
	3-Aminophenol	LLNA	N	Weak
	(+/-) Linalool	LLNA	N	Moderate
	Diethylenetriamine	LLNA	N	Moderate
	Octanoic acid, 4-methyl-2-pentylbutyl ester	LLNA	N	Non-sensitizer
	R(+)-Limonene	LLNA	P	Weak
	Ethylenediamine free base	LLNA	P	Moderate
	Vanillin	LLNA	N	Non-sensitizer
	Cyclamen aldehyde	LLNA	P	Weak
	Benzalkonium chloride	LLNA	N	Non-sensitizer
	Lactic acid	LLNA	N	Non-sensitizer
	Octanenitrile	LLNA	N	Non-sensitizer
	Undec-10-enal	LLNA	N	Moderate

Test method	Borderline substances[a]	Reference test	Sensitization potential[b,c] according to reference test	Potency class (based on LLNA)[b]
	Benzyl benzoate	LLNA	N	Weak
	Methyl 4-hydroxybenzoate (Methylparaben)	LLNA	N	Non-sensitizer
	Butylbenzylphthalate	LLNA	N	Non-sensitizer
	4-Hydroxybenzoic acid	LLNA	N	Non-sensitizer
	Sulfanilamide	LLNA	N	Non-sensitizer
	Cocamidopropyl betaine	LLNA	N	Non-sensitizer
	Benzene,1-methoxy-4-methyl-2-nitro (4-Methyl-2-nitroanisole)	LLNA	N	Non-sensitizer
	Squaric acid diethyl ester	LLNA	N	Strong
	Clofibrate (Ethyl (2-(4-chlorophenoxy)-2-methylpropanoate)	LLNA	N	Non-sensitizer
	α-Amyl cinnamic aldehyde	LLNA	N	Weak
	Streptomycin sulfate	LLNA	N	Non-sensitizer
	α-iso-Methylionone	LLNA	N	Weak
	Carbonic acid, dioctyl ester	LLNA	N	Non-sensitizer
	Hexyl salicylate	LLNA	N	Strong
	Benzyl cinnamate	LLNA	N	Weak
	Benzyl salicylate	LLNA	N	Moderate
LuSens	1-Butanol	LLNA	N	Non-sensitizer
	Benzoyl peroxide	LLNA	P	Weak
	4-Allylanisole	LLNA	P	Weak
	1,2-Dibromo-2,4-dicyanobutane (MDGN, Methyldibromo glutaronitrile)	LLNA	P	Strong
	Imidazolidinyl urea	LLNA	P	Weak
h-CLAT	4-phenylenediamine[c]	LLNA	P	Strong
	Phenyl benzoate[c]	LLNA	P	Weak
	Ethylene diamine[c]	LLNA	P	Moderate
	Aniline	LLNA	P	Weak
	Farnesal	LLNA	P	Weak
	Methyldibromo glutaronitrile[c]	LLNA	P	Strong
	p-Benzoquinone	LLNA	P	Extreme
	Propyl gallate[c]	LLNA	P	Strong
	Citral	LLNA	P	Moderate
	Cobalt chloride	LLNA	P	Strong
LLNA	Salicylic acid[c]	human	N	Non-sensitizer
	Methyl salicylate[c]	human	N	Non-sensitizer
	Chlorobenzene[c]	human	N	Non-sensitizer
	Nickel chloride[c]	human	N	Non-sensitizer
	Phenyl benzoate[c]	human	P	Weak
	Methyl methacrylate[c]	human	P	Weak
	MCI/MI	human	P	Extreme

[a] Substances considered borderline when applying the largest BR considered for each test method (see also Table 6).
[b] Prediction based on (Urbisch et al., 2015); human data were extracted from (Basketter et al., 2014).
[c] N, negative; P, positive

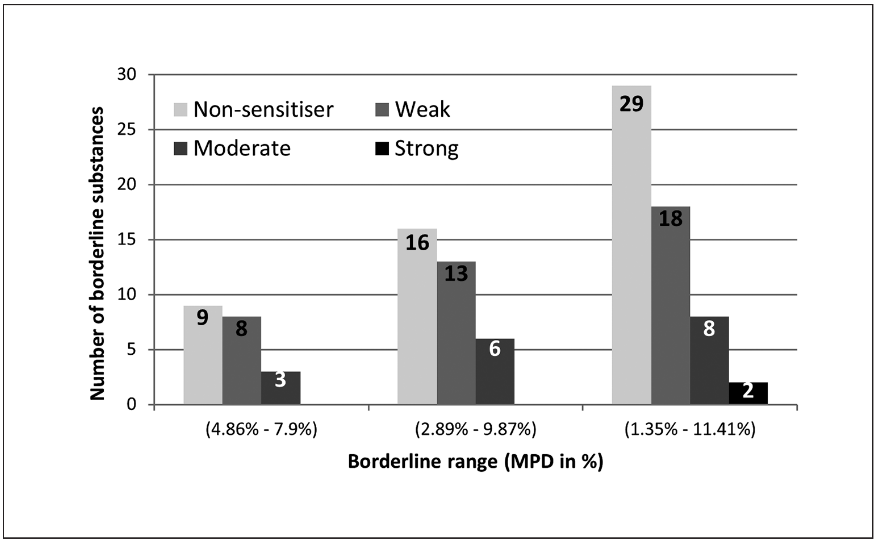

Fig. 1: Number of substances considered borderline and their potency classes (non-sensitizer, weak, moderate, strong) according to LLNA results for different BRs in the DPRA prediction model

X-axis: BRs considered for the DPRA (MPD in %); Y-axis: Number of borderline substances.

Source: Own calculations based on results documented in Table 8

In case of LuSens, doubling the data range for calculating the SD_p increased the BR only marginally. 4 of the 5 substances for which positive test results were within the BRs of LuSens revealed positive but non-borderline results in the LLNA. Of these, one was a non-sensitizer, 3 substances were weak, and one was a moderate sensitizer (see Tab. 8 and Appendix 3 at doi:10.14573/altex.1606271s). Within the BRs of the h-CLAT all substances considered borderline were positive and also sensitizers in the LLNA. Three substances were weak sensitizers. Of these, one substance (phenyl benzoate) was also in the BR of the prediction model of the LLNA. Two substances in the BRs of the h-CLAT were moderate, three strong and one an extreme sensitizer, respectively.

Of the borderline substances in the experimental set of the LLNA, two (i.e., phenyl benzoate, methyl methacrylate) are weak sensitizers and four (i.e., salicylic acid, methyl salicylate, chlorobenzene, nickel chloride) are non-sensitizers (Tab. 4). One substance (MCI/MI) was an extreme sensitizer. Most substances identified as borderline in the LLNA are also discussed in Kolle et al. (2013). For the smallest BR considered, i.e., BR = $\{2.5 \leq SI \leq 3.5\}$, our analysis revealed an equivalent percentage of borderline chemicals (23%) compared to Kolle et al. (2013). Increasing the range around the classification threshold that is used for calculating the SD_p and, in turn, the BR (i.e., BR = $\{2.34 \leq SI \leq 3.63\}$), phenyl benzoate also becomes a borderline substance, causing the percentage of substances falling in the BR of the LLNA to be slightly higher (27%) compared to Kolle et al. (2013) (23%). Note, however, that Kolle et al. (2013) determined the BR by calculating coefficients of variation based on individual animal data instead of pooled animal data.

3.3 Identifying borderline substances in the experimental set tested with the 2-out-of-3 ITS

As shown in Table 9, we found four of 40 (10%) of the substances tested with the 2-out-of-3 ITS to be borderline. This result was robust for all combinations of BRs considered in the prediction models of individual non-animal test methods. A complete list of ITS results, including results for borderline substances, is included in Appendix 3 at doi:10.14573/altex.1606271s. All substances were positive in the LLNA. Of these, one is a weak, one a moderate and two substances are strong sensitizers according to LLNA potency classes. One substance (phenyl benzoate) considered borderline in the 2-out-of-3 ITS was also borderline in the LLNA.

Tab 9: Borderline substances in the experimental set tested with the 2-out-of-3 ITS

Borderline substances	Sensitization potential[a] according to LLNA or human data	Potency class (according to the LLNA)
Phenyl benzoate	P	Weak
Ethylene diamine	P	Moderate
Methyldibromo glutaronitrile	P	Strong
Propyl gallate	P	Strong

[a] Prediction based on (Urbisch et al., 2015).

4 Discussion

4.1 Identification of borderline substances and implications of the BR for assessing substances' skin sensitization potential

The BR defines the area around a prediction model's classification threshold within which repeated testing will more likely show discordant results. That is, within the BR a test method is not precise and conclusions about a borderline substance's skin sensitization potential cannot be adopted with sufficient confidence. This limited precision is caused by technical and biological variability. If a substance yields test results falling into the BR, further testing with other available test methods is required to allow for unanimous discrimination between a positive and a negative test outcome. The probability of a substance with unknown properties generating a borderline result depends on the size of the BR and the distribution of test results. The latter, in turn, depends on the composition of the experimental set of substances. Conclusions regarding the probability of a substance to generate a borderline result are, therefore, only possible for a particular BR and assuming a representative set of substances.

In this study, we quantified the BR for prediction models of three non-animal test methods, their combination in a 2-out-of-3 ITS, and the animal test method LLNA. The BR was derived from the SDp around the individual test methods' classification threshold. We considered different BRs in order to gain insight into the relationship between the size of the BR and the number of substances for which results fall into this range and for which, consequently, a clear-cut conclusion on their skin sensitization potential cannot be adopted with sufficient confidence. Based on the BRs considered and the experimental sets used in our analysis, the percentage of substances considered borderline in the DPRA, LuSens and h-CLAT was between 6% and 28% (Tab. 7). We find that 23%-27% of the performance standard (PS) substances tested with the LLNA fall into its BR. This value is higher than that obtained from the variability assessment in Hoffmann (2015), which may be because Hoffmann (2015) determined the BR from EC3 values and our analysis was based on *SI* values.

For the DPRA, the percentage of substances identified as borderline varied between 10% and 28%. LuSens has a stringent prediction model because two consecutive concentrations in each run, and two or more runs, must show concordant results in order to arrive at a final conclusion about a substance's skin sensitization potential (Ramirez et al., 2014, 2016). Therefore, applying the BR approach to LuSens required two steps to identify borderline substances (i.e., identifying borderline substances within a run and across runs, see Tab. 3 and 4). The stringent prediction model may be a reason why LuSens revealed a relatively small percentage of borderline substances (6% and 7%, depending on the size of the BR, see Tab. 7). The prediction model of the h-CLAT (Ashikaga et al., 2010; Sakaguchi et al., 2010) does not require concordant test results in consecutive concentrations of the same run to conclude on the substance's skin sensitization potential. Furthermore, concordant test results with the expression of cell surface markers CD54 or CD86 from at least two runs within the same experiment are required to conclude on a positive or negative test result (OECD, 2016a). Hence, compared to LuSens, the prediction model of the h-CLAT is less conservative. This may explain the slightly higher percentage of borderline results in the substance set tested with the h-CLAT (20% and 25%, see Tab. 7). All borderline substances in the experimental set of the h-CLAT were sensitizers.

4.2 Precision of the 2-out-of-3 ITS

Following Jowsey et al. (2006), Basketter and Kimber (2009), Reisinger et al. (2015) and ECHA (2016), a single test method cannot be used to predict skin sensitization potential as a stand-alone method. The 2-out-of-3 ITS has been suggested as a suitable approach for the overall assessment of the skin sensitization potential as it is based on the results of three individual test methods (Urbisch et al., 2015). Applying the BR concept to the 2-out-of-3 ITS (Urbisch et al., 2015) revealed four borderline substances in a set of 40 (10%), which is lower than that of the LLNA (27%). The number of borderline substances identified remained constant for all BRs applied to the individual non-animal test methods (see Tab. S3.5 in Appendix 3 at doi:10.14573/altex.1606271s). Our results, therefore, may indicate that the precision of the 2-out-of-3 ITS is higher compared to the LLNA. Again, this result has to be treated with care because the experimental set of the LLNA differed from that of the non-animal test methods used in the 2-out-of-3 ITS. Notwithstanding, the majority rule applied in the 2-out-of-3 ITS reduces the influence of borderline substances on the overall conclusion about a substance's skin sensitization potential for all cases where two of the three methods provide concordant results. This, in turn, increases the overall precision of the 2-out-of-3 ITS compared to the precision of the individual non-animal test methods.

4.3 Implications of the BR approach for evaluating a test method's precision

The share of substances considered borderline in an experimental set depends on the size of the BR, which, in turn, depends on the precision of the experimental method, the specification of the classification threshold, and on the data range around the threshold used for quantifying the BR. We find that the number and the percentage of test results that fall in the BR is higher (lower) the larger (smaller) the BR.

The BR in a test method's prediction model defines a range in which conclusions on substances' skin sensitization potential cannot be drawn with sufficient confidence. Hence, for substances for which test results fall into the BR, the test result is inconclusive. Furthermore, our results illustrate that the number of substances for which classifications can be made is smaller the broader the BR. This points to a trade-off between a test method's precision (i.e., after removing results that fall in the BR) and the number of substances in the set for which a test method is able to deliver decisive information. So far, normative criteria, defining how broad the BR should be, have not been established. The evidence provided in our study, therefore, does not allow for comparisons of precision (limitations) between individual test methods. Further research must examine how the specification of a prediction model's classification threshold, in combination with the range and distribution of test results used for calculating the BR, affects the size of the BR.

5 Conclusions

Technical and biological variability of non-animal test methods used for assessing skin sensitization potential, and the animal test LLNA, influence the precision of these methods. It is important to recognize that neither the animal test LLNA, often considered "the gold standard", nor non-animal test methods perfectly predict effects in humans (due to limited accuracy) and do not always yield clear-cut results (due to limited precision). A test method's precision constraint caused by intra-assay variability can be captured by quantifying a BR around the classification threshold of the method's prediction model, which is used to transform continuous experimental data into a dichotomous result, being either "positive" (indicating an effect) or "negative" (indicating no effect). The size of the BR depends on the specification of the classification threshold in a test method's prediction model, but also on the range of test results on both sides of the threshold that is considered appropriate for determining the BR. Test substances for which results fall within the BR of a test method could be assessed as positive or negative upon re-testing; thus the result of the test is ambiguous. Quite obviously, any conclusion drawn from experimental data is constrained by uncertainties and this is often neglected in reporting the results. The BR may offer a simple and pragmatic way to take into account that not every experimental result allows for a definite conclusion. We therefore suggest that a measure of precision, i.e., the percentage of substances falling in the BR, should therefore be reported with every test method. Furthermore, when using prediction models that dichotomize data there should always be three potential outcomes: positive, negative or borderline.

The aim of our paper was to suggest an approach to determining the BR, and to illustrate for individual test methods and an ITS how the BR can be applied. As our results illustrate, there is a trade-off between the size of the BR (i.e., the range for which a test method is considered to be of limited precision) and the number of substances for which a test can assess skins sensitization potential with sufficient confidence. Being beyond the scope of this paper, it is a matter of further research to define normative rules for determining the "optimal BR". This, essentially, requires assessing the (social) gains from increasing a test method's precision against (social) costs of making errors.

While the paper focused on skin sensitization as a proof-of-concept case, the BR approach is a generic method and can be applied to other endpoints, tests, and ITS. Further research should quantify the BR for a broader set of (non-animal) test methods. Moreover, examining the precision of test methods for continuous endpoints deserves further attention in order to provide complementary insights into a test methods' precision regarding potency assessment (Slob, 2016).

Another important issue for further research and discussion is how to deal with borderline test results in a regulatory context. One possible option could be to define borderline outcomes per default as positive results. However, this would imply that the upper part of the BR is factually ignored. Alternatively, one could require additional testing with other (non-animal) methods and would thus advocate for redundant test method options. Decision-theoretic approaches such the Bayesian value-of-information approach introduced in Leontaridou et al. (2016) can help to determine the optimal follow-up test in a systematic and transparent way. Finally, the question how borderline substances impact test methods' predictive performance deserves further attention. Since for borderline substances the overall conclusion on their hazardous potential remains inconclusive, they cannot contribute unambiguously to the evaluation of a test method's accuracy. Ignoring a substance's test result being borderline will thus cause over- or underestimation errors of, for example, a test method's sensitivity or specificity. Exploring the size and direction of these errors for different non-animal test methods and analyzing the influence of the size and composition of experimental sets on the number of borderline substances detected will provide complementary insights into the implications of intra-assay variability for comparative evaluations of test methods' predictive performance.

References

Ashikaga, T., Yoshida, Y., Hirota, M. et al. (2006). Development of an in vitro skin sensitization test using human cell lines: The human cell line activation test (h-CLAT): I. Optimization of the h-CLAT protocol. *Toxicol In Vitro 20*, 767-773. doi:10.1016/j.tiv.2005.10.012

Ashikaga, T., Sakaguchi, H., Sono, S. et al. (2010). A comparative evaluation of in vitro skin sensitization tests: The human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). *Altern Lab Anim 38*, 275-284.

Basketter, D. A. and Kimber, I. (2009). Updating the skin sensitisation in vitro data assessment paradigm in 2009. *J Appl Toxicol 29*, 545-550. doi:10.1002/jat.1443

Basketter, D. A., Alepee, N., Ashikaga, T. et al. (2014). Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis 25*, 11-21. doi:10.1097/der.0000000000000003

Bauch, C., Kolle, S. N., Ramirez, T. et al. (2012). Putting the parts together: Combining in vitro methods to test for skin sensitizing potentials. *Regul Toxicol Pharmacol 63*, 489-504. doi:10.1016/j.yrtph.2012.05.013

EC – European Commission (2006). Regulation (EC) No 1907/2006 of the European Parliament and the Council of of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive

1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC.

EC (2009). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. OJ L342, 59-209. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009R1223&-from=EN

ECHA – European Chemicals Agency (2016). Guidance on information requirements and Chemical Safety Assessment. Chapter R.7a: Endpoint specific guidance. Draft version 5.0, June 2016.

Cooper, J. A., Saracci, R. and Cole, P. (1979). Describing the validity of carcinogen screening tests. *Br J Cancer 39*, 87-89. doi:10.1038/bjc.1979.10

Dimitrov, S., Detroyer, A., Piroird, C. et al. (2016). Accounting for data variability, a key factor in in vivo/in vitro relationships: Application to the skin sensitization potency (in vivo LLNA versus in vitro DPRA) example. *J Appl Toxicol 36*, 1568-1578. doi:10.1002/jat.3318

Dumont, C., Barroso, J., Matys, I. et al. (2016). Analysis of the local lymph node assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches. *Toxicol In Vitro 34*, 220-228. doi:10.1016/j.tiv.2016.04.008

Gerberick, G. F., Vassallo, J. D., Bailey, R. E. et al. (2004). Development of a peptide reactivity assay for screening contact allergens. *Toxicol Sci 81*, 332-334. doi:10.1093/toxsci/kfh213

Gerberick, G. F., Vassallo, J. D., Foertsch, L. M. et al. (2007). Quantification of chemical peptide reactivity for screening contact allergens: A classification tree model approach. *Toxicol Sci 97*, 417-427. doi:10.1093/toxsci/kfm064

Hoffmann, S. and Hartung, T. (2005). Diagnosis: Toxic! – Trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations. *Toxicol Sci 85*, 422-428. doi:10.1093/toxsci/kfi099

Hoffmann, S. (2015). LLNA variability: An essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. *ALTEX 32*, 379-383. doi:10.14573/altex.1505051

Hothorn, L. A. (2002). Selected biostatistical aspects of the validation of in vitro toxicological assays. *Altern Lab Anim 30, Suppl 2*, 93-98.

Hothorn, L. A. (2003). Statistics of interlaboratory in vitro toxicological studies. *Altern Lab Anim 31, Suppl 1*, 43-63.

ICCVAM (2009). Recommended Performance Standards: Murine Local Lymph Node Assay. NIH Publication Number 09-7357. Research Triangle Park, NC: National Institute of Environmental Health Sciences.

Jaworska, J. (2016). Integrated testing strategies for skin sensitization hazard and potency assessment – State of the art and challenges. *Cosmetics 3*, 16. doi:10.3390/cos metics3020016

Jowsey, I. R., Basketter, D. A., Westmoreland, C. and Kimber, I. (2006). A future approach to measuring relative skin sensitisation potential. *J Appl Toxicol 26*, 341-350. doi:10.1002/jat.1146

Kimber, I., Dearman, R. J., Scholes, E. W. and Basketter, D. A. (1994). The local lymph node assay: Developments and applications. *Toxicology 93*, 13-31. doi:10.1016/0300-483X(94)90193-7

Kleinstreuer, N. C., Sullivan, K., Allen, D. et al. (2016). Adverse outcome pathways: From research to regulation scientific workshop report. *Regul Toxicol Pharmacol 76*, 39-50. doi:10.1016/j.yrtph.2016.01.007

Kolle, S. N., Basketter, D. A., Casati, S. et al., (2013). Performance standards and alternative assays: Practical insights from skin sensitization. *Regul Toxicol Pharmacol 65*, 278-285. doi:10.1016/j.yrtph.2012.12.006

Leontaridou, M., Gabbert S., van Ierland, E. C. et al. (2016). Evaluation of non-animal methods for assessing skin sensitisation hazard: A Bayesian value-of-information analysis. *Altern Lab Anim 44*, 255-269.

Luechtefeld, T., Martens, A., Russo, D. P. et al. (2016). Analysis of publically available skin sensitization data from REACH registrations 2008-2014. *ALTEX 33*, 135-148. doi:10.14573/altex.1510055

Mehling, A., Eriksson, T., Eltze, T., Kolle, S. et al. (2012). Non-animal test methods for predicting skin sensitization potentials. *Arch Toxicol 86*, 1273-1295. doi:10.1007/s00204-012-0867-6

Natsch, A., Bauch, C., Foertsch, L. et al. (2011). The intra- and inter-laboratory reproducibility and predictivity of the KeratinoSens assay to predict skin sensitizers in vitro: Results of a ring-study in five laboratories. *Toxicol In Vitro 25*, 733-744. doi:10.1016/j.tiv.2010.12.014

OECD (1992). Test No. 406: OECD Guideline for testing chemicals. Adopted by the Council on 17th July 1992. Skin Sensitisation OECD Publishing.

OECD (2002). Skin Sensitisation: Local Lymph Node Assay. OECD Guideline for the Testing of Chemicals No. 429, Paris. http://www.oecd.org/env/testguidelines

OECD (2010). Test No. 429: Skin Sensitisation: Local Lymph Node Assay, OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris.

OECD (2012a). The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Protein Part 1: Scientific Evidence. OECD Environment, health and safety publications No.168.

OECD (2012b). The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Part 2: Use of the AOP to Develop Chemical Categories and Integrated Assessment and Testing Approaches OECD Environment, health and safety publications No 168.

OECD (2015a). Test No. 442C: In Chemico Skin Sensitisation. Direct Peptide Reactivity Assay (DPRA), OECD Guidelines for the Testing of Chemicals, Section 4 ed. OECD, Paris.

OECD (2015b). Test No. 442D: In Vitro Skin Sensitisation. In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method, OECD Guidelines for the Testing of Chemicals, Section 4 ed. OECD, Paris.

OECD (2016a). Test No. 442E: In Vitro Skin Sensitisation: Human Cell Line Activation Test (h-CLAT). OECD, Paris.

OECD (2016b). OECD Guidance Document on the Reporting of

Defined Approaches to be Used Within Integrated Approaches to Testing and Assessment ENV/JM/MONO(2016)28, doi:10.1787/9789264274822-en

OECD (2016c). OECD Guidance Document on the Reporting of Defined Approaches and Individual Information Sources to be Used Within Integrated Approaches to Testing and Assessment (IATA) for Skin Sensitization ENV/JM/MONO(2016) 29, doi:10.1787/9789264279285-en

Ramirez, T., Mehling, A., Kolle, S. N. et al. (2014). LuSens: A keratinocyte based ARE reporter gene assay for use in integrated testing strategies for skin sensitization hazard identification. *Toxicol In Vitro 28*, 1482-1497. doi:10.1016/j.tiv.2014.08.002

Ramirez, T., Stein, N., Aumann, A. et al. (2016). Intra- and inter-laboratory reproducibility and accuracy of the LuSens assay: A reporter gene-cell line to detect keratinocyte activation by skin sensitizers. *Toxicol In Vitro 32*, 278-286. doi:10.1016/j.tiv.2016.01.004

Reisinger, K., Hoffmann, S., Alépée, N. et al. (2015). Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. *Toxicol In Vitro 29*, 259-270. doi:10.1016/j.tiv.2014.10.018

Sakaguchi, H., Ashikaga, T., Miyazawa, M. et al. (2006). Development of an in vitro skin sensitization test using human cell lines; human cell line activation test (h-CLAT) II. An inter-laboratory study of the h-CLAT. *Toxicol In Vitro 20*, 767-773. doi:10.1016/j.tiv.2005.10.012

Sakaguchi, H., Ryan, C., Ovigne, J. M. et al. (2010). Predicting skin sensitization potential and inter-laboratory reproducibility of a human cell line activation test (h-CLAT) in the European cosmetics association (COLIPA) ring trials. *Toxicol In Vitro 24*, 1810-1820. doi:10.1016/j.tiv.2010.05.012

Sauer, U. G., Hill, E. H., Curren, R. D. et al. (2016). Local tolerance testing under REACH: Accepted non-animal methods are not on equal footing with animal tests. *Altern Lab Anim 44*, 281-299.

Slob, W. (2016). A general theory of effect size, and its consequences for defining the benchmark response (BMR) for continuous endpoints. *Crit Rev Toxicol 47*, 342-351. doi:10.1080/10408444.2016.1241756

Strickland, J., Zang, Q., Kleinstreuer, N. et al. (2016). Integrated decision strategies for skin sensitization hazard. *J Appl Toxicol 36*, 1150-1162. doi:10.1002/jat.3281

Thyssen, J. P., Linneberg, A., Menné, T. and Johansen, J. D. (2007). The epidemiology of contact allergy in the general population – Prevalence and main findings. *Contact Dermatitis 57*, 287-299. doi:10.1111/j.1600-0536.2007.01220.x

UNECE (2011). Chapter 3.4 Respiratory or skin sensitization, Globally Hormonized Systen of Classification and Labelling of Chemicals (GHS), Fourth revised version, United Nations New York and Geneva.

Urbisch, D., Mehling, A., Guth, K. et al. (2015). Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul Toxicol Pharmacol 71*, 337-351. doi:10.1016/j.yrtph.2014.12.008

Urbisch, D., Becker, M., Honarvar, N. et al. (2016). Assessment of pre-and pro-haptens using nonanimal test methods for skin sensitization. *Chem Res Toxicol 29*, 901-913. doi:10.1021/acs.chemrestox.6b00055

Van der Schouw, Y. T., Verbeek, A. L. and Ruijs, S. H. (1995). Guidelines for the assessment of new diagnostic tests. *Investigative Radiology 30*, 334-340. doi:10.1097/00004424-199506000-00002

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports 62*, 1432-1449. doi:10.2307/4586294

Conflict of interest

The authors declare they have no conflicts of interest.

Acknowledgments

We wish to thank Tzutzuy Fernandez Ramirez (BASF, Ludwigshafen, Germany) for providing the datasets used for the LuSens prediction model. Helpful comments from Ekko C. Van Ierland (WUR, Wageningen, The Netherlands) and two anonymous referees are gratefully acknowledged.

Correspondence to

Robert Landsiedel, PhD
BASF Experimental Toxicology and Ecology
BASF SE, GB/TB – Z470
67056 Ludwigshafen, Germany
e-mail: robert.landsiedel@basf.com