Short communication

LLNA Variability: An Essential Ingredient for a Comprehensive Assessment of Non-Animal Skin Sensitization Test Methods and Strategies

Sebastian Hoffmann

seh consulting + services, Paderborn, Germany

Summary

The development of non-animal skin sensitization test methods and strategies is quickly progressing. Either individually or in combination, the predictive capacity is usually described in comparison to local lymph node assay (LLNA) results. In this process an important lesson from other endpoints, such as skin or eye irritation, i.e., that the variability of reference test results – here the LLNA – must be accounted for, has not yet been fully acknowledged.

In order to provide assessors as well as method and strategy developers with appropriate estimates, the variability of EC3 values from repeated substance testing in the LLNA was investigated using the publicly available NICEATM (NTP Interagency Center for the Evaluation of Alternative Toxicological Methods) LLNA database. Repeat experiments taking the vehicle into account (76 substances) or combining data over different vehicles (38 substances) were analyzed.

In general, variability was higher when different vehicles were used. In terms of skin sensitization potential, i.e., discriminating sensitizers from non-sensitizers, the false positive rate ranged from 14-20%, while the false negative rate was 4-5%. In terms of skin sensitization potency, the rate to assign a substance to the next higher or next lower potency class was approx.10-15% each. In addition, general estimates for EC3 variability are provided that can be used for modelling purposes.

This analysis stresses the importance of considering the LLNA variability in the assessment of skin sensitization test methods and strategies and provides estimates thereof.

Keywords: LLNA variability, skin sensitization, test method assessment, test strategy assessment

1 Introduction

Along with the advances in the life sciences, new testing and non-testing methods for improved, more efficient and animal-free assessment of toxicological hazards are being developed at an increasing rate. Once such a method or strategy has reached a certain level of standardization, it is often evaluated to demonstrate its predictive performance, usually by comparing it to the currently regulated hazard assessment test method it aims to complement or ultimately to replace. In many cases the regulated approach is an animal experiment, so that a dedicated study directly comparing both methods in parallel is not possible due to ethical concerns. In general, this problem is solved by comparing data from the new method/strategy with existing data from the routine methods for the same set of substances. While data quality and reproducibility aspects are controlled and systematically assessed for the new approach, the same rigor cannot be applied to existing data of the routine method. Disregarding these aspects for the routine test methods inevitably results in an overestimation of its predictive performance, which consequently results in unrealistically high expectations for the predictive capacity of the new test method/testing strategy. In order to at least partially compensate for this, traditionally used animal experiments have been thoroughly investigated by deriving estimates of variability. This has, for example, supported the regulatory acceptance of *in vitro* test methods for the human health effects skin irritation and eye irritation/corrosion (Hoffmann et al., 2005; Adriaens et al., 2014).

Various test methods that address the human health endpoint skin sensitization are being developed and many of these have been evaluated systematically by Reisinger et al. (2014). This development was spurred by European regulatory requirements:

Received May 5, 2015;
Accepted July 10, 2015;
Epub July 13, 2015;
http://dx.doi.org/10.14573/altex.1505051

the marketing ban on cosmetics as well as the REACH regulation on chemicals demand or strongly call for skin sensitization assessment of substances without the use of animals (EU, 2006, 2009). The predictive capacity of individual skin sensitization test methods and testing strategies is primarily assessed by comparison with LLNA data (see e.g., Jaworska et al., 2013; Tsujita-Inoue et al., 2014). Furthermore, there have been attempts to circumvent the sub-optimal comparison with animal data by comparison with human data, which has been fueled by a compilation and categorization of human data proposed by Basketter et al. (2014). These comparisons have considered the reference data of the LLNA (and the human data) in a deterministic manner, i.e., without accounting for the aspect of variability, as for example pointed out by Urbisch et al. (2015). Nevertheless, these efforts have culminated in the acceptance of two OECD test guidelines of individual test methods (OECD, 2015a,b), while guidance on testing strategies is being developed.

With the aim to support assessors as well as method and strategy developers of both non-animal methods and testing assessment strategies for skin sensitization hazard and potency of substances, estimates of LLNA EC3 variability from repeat testing data of 68 substances were derived. Repeat tests that used the same vehicle and repeat tests that used different vehicles were analyzed separately. For this purpose, the LLNA data from the publically available NICEATM (NTP Interagency Center for the Evaluation of Alternative Toxicological Methods) LLNA database was used. The impact of the variability on LLNA potency classes was analyzed.

The results stress the importance of accounting for LLNA variability in the assessment of skin sensitization test methods/strategies and provide estimates thereof.

2 Material and methods

The publicly available NICEATM LLNA data compilation[1], version of December 23, 2013, was the sole data source. In total, it reports results of 1060 experiments using 35 different vehicles for 677 different substances and formulations, specifying the vehicle used to apply the substance and the EC3 value in %, i.e., the estimated concentration that induces a three-fold stimulation index as compared to the respective vehicle. Experiments not inducing three-fold stimulation were considered non-sensitizers and were reported as "NC", i.e., non-classified, while for experiments with data with insufficient dose-response for the calculation of EC3 (i.e., nonmonotonic) the EC3 value was reported as "IDR", i.e, insufficient dose-response. The data were considered to be sufficiently curated for the purpose of this evaluation and were not verified against the primary sources. After exclusion of "IDR" experiments, LLNA experiments, for which the same CAS number and name or synonyms were listed, were identified (a total of 454 experiments for 72 substances/mixtures) and respective EC3 values were grouped for analysis, once for substances with repeat experiments using the same vehicle ("same-vehicle" approach) and once for substances with repeat experiments using different vehicles ("different-vehicle" approach).

For both approaches median EC3 values – a location measurement that is robust against aberrant values and that also could be derived for substances with both EC3 and "NC" results – were calculated from all repeat experiments of a given substance. This median was used to assign each substance to one of five potency classes: extreme: median < 0.1%; strong: 0.1% ≤ median < 1.0%; moderate: 1.0% ≤ median < 10.0%; weak: 10.0% ≤ median ≤ 100%; non-sensitizer: NC (ECETOC, 2003). In case of two repeat experiments with one EC3 value and one "NC" result, the substance was conservatively assigned to the class that corresponded to the EC3 value.

Substances with medians in the same potency class were grouped for variability analysis. The impact of variability of EC3 of repeat experiments on potency class assignment was analyzed by determining per group the proportion of all individual EC3 data that would result in a different potency class than the median EC3.

In addition, variability was described for each substance by the standard deviation (SD) of log-transformed (base 10) EC3 of the repeat experiments. Substances that were non-sensitizing in at least one repeat test were excluded. In addition, substances with only two repeat experiments were excluded, as a sample size of at least three repeats was considered sufficient for an acceptably precise SD estimation. From this set of SD, estimated for both the "same-vehicle" and the "different-vehicle" approach, a general estimate of SD variability was derived.

3 Results and discussion

For the "same-vehicle" approach, filtering of the database resulted in 53 different substances with a total of 76 substance-vehicle combinations (Fig. 1A). For example, repeat experiments were available for seven different vehicles for 1,4-dihydroquinone. In total, 356 LLNA EC3 values were used for calculations. Data were distributed over all five potency classes. The "extreme" class with 12 substances and 49 experiments was the least populated (Tab. 1A). Applying the "different-vehicle" approach resulted in 38 substances and a total of 333 experiments for analysis (Tab. 1B). LLNA potency classes were unequally populated, both in regard to the amount of substances and the amount of experiments. For both approaches, Table 1 summarizes the proportion of more and less severely classified individual experiments for each potency class. For the "same-vehicle" analysis, on average in 9.3% of the cases a less severe classification was observed, while a more severe classification was present in 15.2% of the cases. For the analysis that combined experiments using different vehicles, the misclassification rates were 14.1% and 15.3, respectively. The majority of misclassification was one class above or below the median class. Reducing the potency classes to dichotomous hazard classes

[1] http://1.usa.gov/1IdmCmw

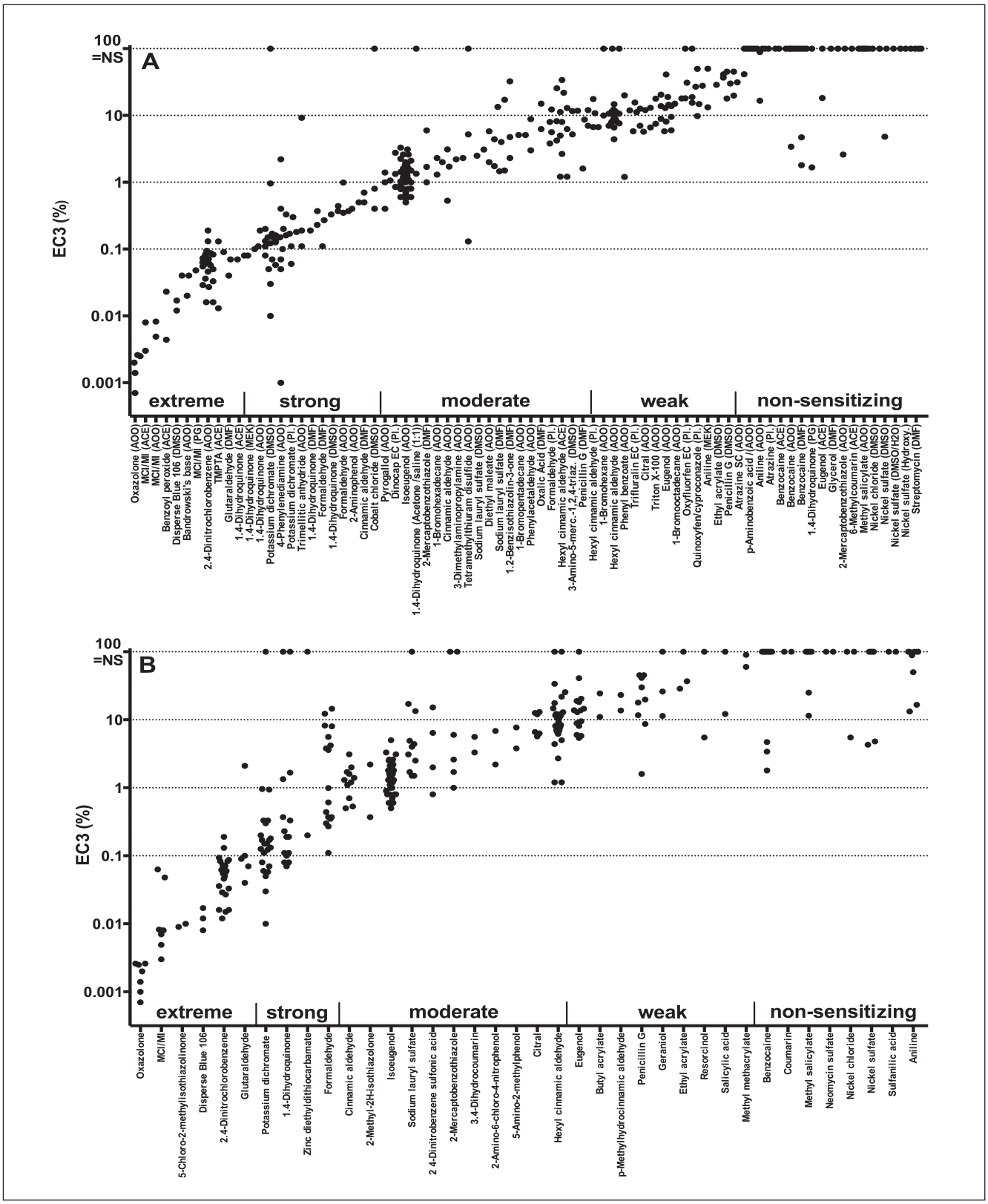

Fig. 1: LLNA EC3 (%) of substances with repeat experiments using A) the same vehicle (76 substance-vehicle combinations; NS: non-sensitizer) and B) different vehicles (38 substances)

AOO: acetone:olive oil (4:1 by volume); ACE: acetone; DMSO: dimethyl sulfoxide; PG: propylene glycol; DMF: dimethylformamide; MEK: methyl ethyl ketone; Pl.: Pluronic® L92; Hydroxy.: hydroxypropyl cellulose in methanol

Tab. 1: Variability of categorizations of repeat testing

using A) the same vehicle based on LLNA EC3 for 76 substances and B) different vehicles for 38 substances grouped by potency (median classification of repeats used for potency class assignment; NS: non-sensitiser; subs.: substances)

A

LLNA potency class	number		categorization		proportion of categorizations		
	subs.	exps.	more severe	less severe	missed (total)	more severe	less severe
NS	15	64	9	–	14.1%	14.1%	–
weak	15	75	21	6	36.0%	28.0%	8.0%
moderate	20	106	14	16	28.3%	13.2%	15.1%
strong	14	62	10	7	27.4%	16.1%	11.3%
extreme	12	49	–	4	8.2%	–	8.2%
total	76	356	54	33	24.4%	15.2%	9.3%

B

LLNA potency class	number		categorization		proportion of categorizations		
	subs.	exps.	more severe	less severe	missed (total)	more severe	less severe
NS	8	61	12	–	19.7%	19.7%	–
weak	8	38	8	4	31.6%	21.1%	10.5%
moderate	12	128	17	29	36.9%	13.3%	22.7%
strong	4	57	10	14	42.1%	17.2%	24.6%
extreme	6	49	–	4	8.0%	–	8.0%
total	38	333	47	51	29.4%	14.1%	15.3%

of "NS" and "S", i.e., the classes "moderate" to "extreme", resulted in an overprediction proportion (NS as S) of 14.1% and an underprediction (S as NS) proportion of 3.8% (11/292) for the "same-vehicle" approach and in 19.7% and 5.1% (14/272), respectively, for the "different-vehicle" approach.

Repeat experiments also provided the means to generalize LLNA variability. To increase the robustness of this approach, substances with two repeat experiments were excluded. In addition, substances with at least one NC result, which may simply be explained by different test concentration ranges, were disregarded. It needs to be noted that this approach resulted in exclusion of some of the most variable cases potentially resulting in a systematic underestimation of variability. In this regard, it represented a conservative approach and EC3 variability of repeat experiments is likely higher.

For the "same-vehicle" approach 27 substances were considered. SD values ranged from 0.137 to 1.048 with a median SD of 0.252. For the "different-vehicle" approach 11 substances were included. Their SD values ranged from 0.164 to 0.691, while the median was 0.312. Assuming that log-transformed EC3 are approximately normally distributed, the median SD values can, for example, be used to calculate the most likely probability distribution, confidence intervals and probabilities for over- and under-classification for any given EC3. Consider the example that a LLNA test of a substance with unknown sensitization potential resulted in an EC3 point estimation of 20% that would trigger a classification as "weak sensitizer". An approximate 95%-confidence interval (CI) can be calculated in a simple manner by adding and subtracting 2 * SD from the log-transformed median ($\log(20) - 2 * SD = 0.797$; $\log(20) + 2 * SD = 1.805$). Retransformation results in a 95%-CI for the EC3 ranging from 6.27 to 63.83. The likelihood that the substance is a moderate sensitizer, i.e., has an EC < 10%, is 11.6%, while the likelihood that it is a non-sensitizer is as low as 0.3%. Calculating the same example with the "different-vehicle" approach median SD of 0.312 results in a likelihood of 16.7% for "moderate" and of 1.3% for "non-sensitizer".

This relatively simple example demonstrates that the information from repeat LLNA experiments can be used to account for LLNA variability in statistical approaches. First of all, it provides an approach for more appropriate assessment of any new skin sensitization testing method and of testing strategies. Instead of comparing with deterministic EC3 values or classification derived from such values, likelihoods of over- and underclassification can be estimated for each substance to derive more realistic estimates of LLNA predictive capacities for any given substance or set of substances. In this way some of the uncertainty associated with LLNA data can be quantified and accounted for, potentially facilitating discussions about the acceptance of new skin sensitization testing approaches.

However, building on substance-specific investigations of the calibrant hexyl cinnamic aldehyde and the positive control isoeugenol (Dearman et al., 2011; Basketter and Cadby, 2004), for which considerable numbers of repeat tests from a single or multiple laboratories are available, this initial work should be considered primarily as a staring point, as several relevant important aspects have not been or have only preliminarily been addressed or discussed. For example, the impact of individual repeats or of specific substances has not been considered here. Furthermore, it needs to be acknowledged that individual EC3 values are point estimates with varying precision, which are greatly determined by the number, range and spacing of test substance concentrations and by the shape of the obtained dose-response curve. For example, testing of a few low concentrations may result in missing the sensitizing potential of a test substance. Another crucial factor affecting the variability is the choice of the vehicle. Vehicle impact has already been explored in some detail, for example by Jowsey et al. (2008), who reported a tendency toward underestimated potency for aqueous vehicles or propylene glycol, and has been briefly reviewed (Anderson et al., 2011). Our analysis supports the general view that repeat testing with different vehicles leads to more variable EC3 values than repeat testing with the same vehicle. Consequently, appropriateness and choice of vehicle for a given substance are important factors in the assessment of LLNA variability.

In addition to its test method/strategy assessment uses, the LLNA variability assessment may be used in the context of risk assessment. For individual sensitizing substances it may – considered together with other relevant substance-specific information – contribute to conducting a probabilistic skin sensitization risk assessment based on an individual EC3 value.

References

Adriaens, E., Barroso, J., Eskes, C. et al. (2014). Retrospective analysis of the Draize test for serious eye damage/eye irritation: Importance of understanding the in vivo endpoints under UN GHS/EU CLP for the development and evaluation of in vitro test methods. *Arch Toxicol 88*, 701-723. http://dx.doi.org/10.1007/s00204-013-1156-8

Anderson, S. E., Siegel P. D. and Meade, B. J. (2011). The LLNA: A brief review of recent advances and limitations. *J Allergy 2011*, 424203. http://dx.doi.org/10.1155/2011/424203

Basketter, D. A. and Cadby, P. (2004). Reproducible prediction of contact allergenic potency using the local lymph node assay. *Contact Dermatitis 50*, 15-17. http://dx.doi.org/10.1111/j.0105-1873.2004.00278.x

Basketter, D., Alépée, N., Ashikaga, T. et al. (2014). Categorization of chemicals according to their relative human skin sensitising potency. *Dermatitis 25*, 11-21. http://dx.doi.org/10.1097/DER.0000000000000003

Dearman, R. J., Wright, Z. M., Basketter, D. A. et al. (2011). The suitability of hexyl cinnamic aldehyde as a calibrant for the murine local lymph node assay. *Contact Dermatitis* 44, 357-361. http://dx.doi.org/10.1034/j.1600-0536.2001.044006357.x

ECETOC (2003). Contact Sensitisation: Classification According to Potency. *Technical Report No. 87*, Brussels.

EU (2006). Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). *Off J Eur Union L 396*, 1-1355.

EU (2009). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. *Off J Eur Union L 342*, p. 1-59.

Hoffmann, S., Cole, T. and Hartung, T. (2005). Skin irritation: Prevalence, variability, and regulatory classification of existing in vivo data from industrial chemicals. *Regul Toxicol Pharmacol 41*, 159-166. http://dx.doi.org/10.1016/j.yrtph.2004.11.003

Jaworska, J., Dancik, Y., Kern, P. et al. (2013). Bayesian integrated testing strategy to assess skin sensitization potency: From theory to practice. *J Appl Toxicol 33*, 1353-1364. http://dx.doi.org/10.1002/jat.2869

Jowsey, I. R., Clapp, C. J., Safford, B. et al. (2008). The impact of vehicle on the relative potency of skin-sensitizing chemicals in the local lymph node assay. *Cutan Ocul Toxicol 27*, 67-75. http://dx.doi.org/10.1080/15569520801904655

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

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

Reisinger, K., Hoffmann, S., Alépée, N. et al. (2014). Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. *Toxicol In Vitro 29*, 259-270. http://dx.doi.org/10.1016/j.tiv.2014.10.018

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. http://dx.doi.org/10.1016/j.yrtph.2014.12.008

Tsujita-Inoue, K., Hirota, M., Ashikaga, T. et al. (2014). Skin sensitization risk assessment model using artificial neural network analysis of data from multiple in vitro assays. Toxicol In Vitro 28, 626-639. http://dx.doi.org/10.1016/j.tiv.2014.01.003

Conflict of interest

The author declares that he has no conflicts of interest.

Correspondence to

Sebastian Hoffmann
seh consulting + services
Stembergring 15
33106 Paderborn
Germany
Phone: +49 5251 8700566
e-mail: sebastian.hoffmann@seh-cs.com