Published in final edited form as:

Regul Toxicol Pharmacol. 2011 February ; 59(1): 191–196. doi:10.1016/j.yrtph.2010.10.004.

Safety Assessment of Allergic Contact Dermatitis Hazards: An Analysis Supporting Reduced Animal Use for the Murine Local Lymph Node Assay

Joseph K. Haseman^a, Judy Strickland^b, David Allen^b, Eleni Salicru^b, Michael Paris^b, Raymond R. Tice^c, and William S. Stokes^{d,*}

^a J. K. Haseman Consulting, Raleigh, North Carolina 27614, USA

^b Integrated Laboratory Systems, Inc., Research Triangle Park, North Carolina 27709, USA

^c The National Institute of Environmental Health Sciences, The National Toxicology Program, Biomolecular Screening Branch, Research Triangle Park, North Carolina 27709, USA

^d The National Institute of Environmental Health Sciences, The National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, Research Triangle Park, North Carolina 27709, USA

Abstract

The original Organisation for Economic Co-operation and Development Test Guideline 429 (OECD TG 429) for the murine local lymph node assay (LLNA) required five mice/group if mice were processed individually. We used data from 83 LLNA tests (275 treated groups) to determine the impact on the LLNA outcome of reducing the group size from five to four. From DPM measurements, we formed all possible four-mice and five-mice combinations for the treated and control groups. Stimulation index (SI) values from each four-mice combination were compared with those from five-mice combinations, and agreement (both $SI < 3$ or both $SI \geq 3$) determined. Average agreement between group sizes was 97.5% for the 275 treated groups. Compared test-by-test, 90% (75/83) of the tests had 100% agreement; agreement was 83% for the remaining eight tests. Disagreement was due primarily to variability in animal responses and closeness of the SI to three (positive response threshold) rather than to group size reduction. We conclude that using four rather than five mice per group would reduce animal use by 20% without adversely impacting LLNA performance. This analysis supported the recent update to OECD TG 429 allowing a minimum of four mice/group when each mouse is processed individually.

*Corresponding author: 530 Davis Drive, P.O. Box 12233, K2-16, Research Triangle Park, North Carolina 27709, USA, Phone: 1-919-541-7997, Fax: 1-919-541-0947, stokes@niehs.nih.gov.

Supplementary Data

The supplementary data include the mean DPM results for each control and treated dose group of each substance, the original SI values, and the proportion of agreement (for $SI < 3$ and $SI \geq 3$) for all possible dose groups of four and five mice.

Conflict of Interest Statement

The authors declare that there are no conflicts of interest.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

local lymph node assay; skin sensitization; alternative test method; animal reduction; sample size; OECD Test Guideline 429

1. INTRODUCTION

The murine local lymph node assay (LLNA)¹ (Dean et al., 2001; Haneke et al., 2001; ICCVAM, 1999; Sailstad et al., 2001) is an alternative skin sensitization test method that requires fewer animals and less time than currently accepted guinea pig tests (e.g., the guinea pig maximization test and the Buehler test) and represents a significant reduction in animal pain and distress. The LLNA is based on the principle that sensitizing chemicals induce lymphocyte proliferation in the lymph nodes draining the test substance application site. Cell proliferation is determined by analyzing the extent of incorporation of a radioactive marker into newly synthesized DNA. Under appropriate test conditions, proliferation is proportional to the dose applied, and provides a means of obtaining an objective, quantitative measurement of sensitization (EPA, 2003; ICCVAM, 2009; OECD, 2010). The LLNA was the first alternative test method evaluated and recommended by the U.S. Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) for consideration by U.S. regulatory agencies. Since 2002, regulatory authorities internationally have recognized the LLNA as an acceptable alternative to guinea pig tests for most testing situations (Stokes and Schechtman, 2008).

The current U.S. and original international test guidelines for the LLNA, U.S. Environmental Protection Agency (EPA) Health Effects Test Guidelines on Skin Sensitization OPPTS 870.2600 (EPA, 2003) and Organisation for Economic Co-operation and Development (OECD) Test Guideline (TG) 429 for Skin Sensitisation: LLNA (OECD, 2002) are based on similar LLNA protocols. Both guidelines include a comparative assessment of lymph node cell proliferation in treated and control groups of mice by measuring the incorporation of ³H-thymidine or ¹²⁵I-iododeoxyuridine (measured as disintegrations per minute [DPM]) into the DNA of draining auricular lymph nodes. The stimulation index (SI) is the ratio of the mean treated group DPM to the vehicle control group DPM. If the SI ≥ 3 the substance is classified as a skin sensitizer. If the SI < 3 the substance is classified as a nonsensitizer.

The current EPA and the original OECD test guidelines have specific differences however. EPA OPPTS 870.2600 requires at least five mice per group and the collection of individual animal data so that interanimal variability can be assessed. The original version of OECD TG 429 allowed for as few as four mice per dose group when the lymph nodes of the mice in each dose group were pooled. When individual animal data were collected, consistent with the EPA test guideline, OECD TG 429 required at least five mice per dose group. Because many international animal care and use regulations require that the minimum number of mice necessary be used for testing, many laboratories opted to collect pooled data from only four mice.

¹Abbreviations used: DPM, disintegrations per minute; EPA, U.S. Environmental Protection Agency; ICCVAM, Interagency Coordinating Committee on the Validation of Alternative Methods; LLNA, murine local lymph node assay; NICEATM, National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods; OECD, Organisation for Economic Co-operation and Development; OPPTS 870.2600, U.S. Environmental Protection Agency Office of Prevention, Pesticides and Toxic Substances Health Effects Test Guidelines on Skin Sensitization 870.2600; SI, stimulation index; TG, test guideline; TG 429, OECD TG 429 - Skin Sensitisation: Local Lymph Node Assay

Recently, OECD TG 429 was updated (OECD, 2010) to allow for a minimum of four mice per dose group whether the lymph nodes are processed individually for each mouse or whether the lymph nodes of the mice in each dose group are pooled. This update to OECD TG 429 was based on the analysis contained herein and conducted by the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) in support of an ICCVAM evaluation to determine if the LLNA would continue to support the same level of public health protection if the number of animals in each LLNA dose group is reduced from five to four.

1.1 OBJECTIVES

Collecting lymph nodes from individual mice has several advantages over pooling lymph nodes. Interanimal variability can be assessed, which allows for a statistical comparison of differences between test substance and vehicle control groups, along with an opportunity to identify outlier responses using statistical tests such as Dixon's test (Dixon and Massey, 1983). Identifying outlier responses may prevent false results for substances that produce responses near threshold values. Substances that normally would induce an SI value just above or below 3 might be incorrectly classified due to a low or high outlier value, respectively, if the outlier is not identified and excluded.

The purpose of this analysis was to determine whether the requirement of five mice per dose group for individual animal data collection in the original OECD TG 429 and EPA 870.2600 protocols could be reduced to four mice per dose group without adversely affecting the relevance of the LLNA. Because the "true" underlying sensitizer status for individual chemicals may not be known, this investigation focuses primarily on the degree of agreement between outcomes with groups of four or five mice rather than on which observed outcome was the "correct" one.

Although the SI value is the primary determinant of the LLNA outcome, a statistical test might be used in addition to the SI decision criterion. In fact, the EPA test guideline includes a requirement that investigators also submit LLNA data for statistical comparisons of the mean DPM values for treated and vehicle control groups (EPA, 2003). For this reason, we also used a Student's *t*-test based on log-transformed DPM data to compare each dosed group with its concurrent vehicle control. We compared the frequency of LLNA outcomes with either four- or five-mice group sizes using $SI \geq 3$ or statistical significance to classify substances as positive.

2. METHODS

This retrospective evaluation used individual animal data from LLNA tests submitted to NICEATM. These data were submitted by six laboratories that used inbred CBA mice, the strain recommended in LLNA test guidelines by OECD (OECD, 2002) and EPA (EPA, 2003). The 78 substances tested include individual chemicals and proprietary formulations from 83 LLNA tests. There were two tests for formaldehyde and five tests for hexyl cinnamic aldehyde. Each test consisted of three or four dose groups and a vehicle control group.

Of the 83 test results, 50 tests yielded positive results (i.e., maximum $SI \geq 3$) and 33 tests yielded negative results (i.e., all $SI < 3$). Among the 277 dose groups and the 67 control groups, the number of mice per group ranged from two to nine (Table 1). Two dose groups, one with two mice and one with three, contained too few mice for the comparison of LLNA outcomes and were excluded from $SI \geq 3$ criterion analyses. LLNA test results were evaluated on a dose-by-dose basis as well as on a test-by-test basis, recognizing that the dose

groups within a test used a common vehicle control group. Also, in certain laboratories, a common vehicle control group was used for multiple chemicals.

For each LLNA test that used five mice per dose group, SI values were calculated for all possible four-mice combinations in both the treated and vehicle control groups (25 possible combinations per test). The SI value of each of these combinations was compared with the SI value determined from all five mice. The proportion of outcomes with four mice that agreed with the outcome based on five mice was determined. The outcomes agreed if (1) both protocols produced $SI < 3$ or (2) both produced $SI \geq 3$.

For each LLNA test that had more than five mice per group, a similar procedure was applied. In these cases, however, it was necessary to form all possible four-mice and five-mice combinations from the full dataset. This resulted in significantly more possible combinations of samples (e.g., 8100 possible combinations for tests with six animals per dose group compared to 25 possible combinations for tests with five mice per dose group).

For those tests with more than five mice per dose group, we examined the relative impact of animal-to-animal variability and sample size reduction on the disagreement in study outcome. That is, we compared the disagreement related to reducing the sample size from five to four mice per dose group to the disagreement that would result from simply taking a second sample of five mice per dose group.

In addition to the $SI \geq 3$ criterion, formal statistical testing was also considered. All data were log-transformed prior to statistical analyses to normalize the frequency distribution. A Student's *t*-test was used to compare each dose group with its concurrent vehicle control, and statistically significant differences ($p < 0.05$) between treated and vehicle control groups were regarded as positive test results (i.e., sensitizers). All other results ($p > 0.05$) were regarded as negative (i.e., nonsensitizers). Power calculations based on a two-sided Student's *t*-test were also conducted using a Web-based statistics program (DanielSoper.com Statistics Calculators version 2.0 [<http://www.danielsoper.com/statcalc/calc49.aspx>]) to determine the impact of reducing the sample size from five to four mice per group.

3. RESULTS

3.1 Use of the SI to Identify Sensitizers

Table 2 shows the frequency of the various SI values among the 275 dose groups, together with the average agreement between LLNA outcomes with four or five mice per group. Only 12% (34/275) of the dose groups had less than 100% agreement between four- or five-mice outcomes. Disagreement was limited to those SI values from 2.1 to 4.7, but some dose groups in this range produced 100% agreement (see Table 2 and Supplementary Tables 1–6). Note also that, as expected, the degree of disagreement was greatest at SI values closer to 3 (Table 2). The overall average agreement between outcomes with four or five mice was 97.5%. The individual dose group results are summarized in Supplementary Tables 1–6.

Table 2 shows the greatest disagreement in outcome for SI values approaching 3. To determine whether the closeness of the SI to 3, rather than the reduction in sample size, was primarily responsible for this disagreement, we assessed the agreement between SI values for two studies each having five mice per group, and two studies, one with four mice per group and the other with five mice per group. Table 3 illustrates these calculations for two dose groups with SI values close to 3: one with $SI = 2.8$ (based on six mice) and the other with $SI = 3.2$ (based on six mice). In both of these cases, some five-mice studies resulted in an $SI \geq 3$ and some resulted in an $SI < 3$. The same was true for four-mice studies. For the dose group with $SI = 2.8$, reducing the sample size from five to four mice resulted in 44.6%

disagreement in SI values, where one was ≥ 3 and one was < 3 (see Table 3). By comparison, evaluation of two studies with five animals per group (i.e., not reducing the sample size) resulted in 40.1% disagreement. The dosed group that had $SI = 3.2$ yielded a similar result. These results illustrate that when the SI was close to 3, the disagreement in SI value produced by reducing the sample size from five to four was primarily the result of the variability in responses among mice, coupled with the closeness of the SI to 3.

When the data for each of the 83 LLNA tests were examined on a test-by-test basis, reducing the sample size from five to four resulted in 100% agreement for 90% (75/83) of the tests. For the remaining eight tests, some differences in classification were likely in five- and four-mice studies (see Table 4), with average agreement of 83%. As previously discussed, much of the disagreement was due to the closeness of the SI to 3, not to the reduction in sample size.

Interestingly, hexyl cinnamic aldehyde, a known human skin sensitizer also positive in guinea pig tests (ICCVAM, 1999), yielded a maximum $SI < 3$ at the highest dose of 10% and would therefore not be classified as a sensitizer at this concentration. Four-mice studies yielded a positive response ($SI \geq 3$) more often than the five-mice studies (38% versus 28%) (Table 4). Potassium dichromate was also a known sensitizer in human and guinea pig tests (ICCVAM, 1999), but the categorization of the other five substances with maximum $SI \geq 3$ as “true” sensitizers was uncertain. If we assume that the $SI \geq 3$ criterion correctly classifies all six substances with maximum $SI \geq 3$ (i.e., all six substances are “true” sensitizers), then there was a small loss in power when the sample size was reduced from five to four mice. However, this difference in power was small; and, for all six chemicals, the likelihood was still quite high (77%–96%) that the compound would be identified as a sensitizer using a study of four mice. Note that there was no difference in power between studies of four and five mice for 90% of the studies. Thus, not only does the reduction in sample size from five to four have little impact on reliability using the $SI \geq 3$ criterion, but it also appears to have little impact on the accuracy of classification.

3.2 Use of Statistical Significance to Identify Sensitizers

To evaluate the possible use of formal statistical tests to supplement (or replace) the $SI \geq 3$ criterion in the definition of potential skin sensitizers, Student's *t*-test was used to compare DPM for treated and vehicle control mice (Table 5). This statistical test classified far more chemicals as sensitizers than did the $SI \geq 3$ criterion (132 dose groups statistically different from vehicle controls vs. 99 dose groups with $SI \geq 3$). Statistical significance ($p < 0.05$) was achieved for many dose groups that produced SI values well below 3.

The calculations in Table 5 are based on responses observed in 277 different dose groups. More generally, consider the 17 studies carried out at BASF – The Chemical Company (Supplementary Table 2) as an example. The mean vehicle control DPM response and the corresponding standard deviation (SD), on a log scale, can serve as the baseline for a power calculation for detecting 2-, 2.5-, 3-, and 3.5-times increases in response. The SD of the log-transformed data was assumed to be the same in the dosed and vehicle control groups, an assumption consistent with the data from multiple laboratories obtained to date. Delta was the standardized difference to be detected and was the key input variable into the power calculation program. The power calculations in Table 6 are based on a two-sided Student's *t*-test and assume an underlying normal distribution for the log-transformed data.

These results show that if the underlying variability among vehicle control mice corresponds to that seen in an average BASF study, then in 76.8% of the four-mice tests an underlying SI value of 2.5 would be identified as statistically significant ($p < 0.05$). The likelihood is increased to 87.9% for five-mice tests. This power calculation was consistent with the

empirical results summarized in Table 5 and showed that using statistical analyses, even with four-mice test data, would have an excellent chance of detecting a substance that produced an SI response of 2.5, whereas using the $SI \geq 3$ criterion would not. Whether or not such relatively low SI effects should be considered a result of skin sensitization is a matter of scientific judgment and will be discussed later. These results are illustrated in Figure 1 showing power curves generated for four- and five-mice groups for various SI values based on the vehicle control group data from BASF. The curves were constructed by specifying different values of delta, which could reflect different SI values, different underlying variabilities, or a combination of the two. By showing the variation in power, the likelihood of identifying a sensitizing effect as statistically significant was demonstrated and revealed that reduction in sample size from five to four mice only reduced the power of the analysis by up to 13%.

4. DISCUSSION

Our results demonstrate that a reduction in sample size from five to four mice has virtually no impact on LLNA results that are at least one unit away from the threshold for a positive or negative response (i.e., $SI > 4$ or $SI < 2$). Our analyses provided additional information on (1) how frequently such outcomes are seen in practice, (2) the range of SI values for which there was an impact on study outcome, and (3) the impact of the closeness of the SI to 3 and the underlying variability in animal response on the agreement/disagreement in study outcome. Table 2 did not assess the accuracy of the LLNA with different sample sizes but rather depicted the ability of four- or five-mice studies to produce the same outcome using the $SI \geq 3$ criterion.

As the SI approaches 3, different studies may produce different classifications using the $SI \geq 3$ criterion, regardless of sample size, because of the naturally occurring variability among mice. In fact, most of the discordance between four- and five-mice studies shown in Table 2 was not due to the reduction in sample size. For SI values close to 3, we demonstrated that a second study with five mice per group would show almost the same level of disagreement with the first five mice per group study, as would a second study with only four mice per group (Table 3). This important concept was illustrated with two examples from dose groups that yielded SI values close to the decision criterion of $SI \geq 3$ ($SI = 2.8$ and $SI = 3.2$).

Where there is disagreement and $SI < 3$, the four-mice groups may actually be more likely than the five-mice groups to produce an $SI \geq 3$. In the example for $SI = 2.8$ (10% hexyl cinnamic aldehyde), a sample of four mice would have a 38% chance (85/225) of producing an $SI \geq 3$, compared with only a 28% (10/36) chance for a sample of five mice. In that sense, four mice could be regarded as having greater power than five mice for these data. In fact, as the number of mice increases, the likelihood of classifying $SI = 2.8$ as a sensitizer using the $SI \geq 3$ criterion approaches zero.

If SI responses less than 3 are considered indicative of sensitization, then a formal statistical test should be used instead of the $SI \geq 3$ criterion, which will likely not detect them. Although not detected by the $SI \geq 3$ criterion, the $SI = 2.8$ yielded by the maximum dose of a 10% dose of hexyl cinnamic aldehyde is highly significant ($p < 0.01$) by Student's *t*-test. Not only is it highly significant statistically, but also the same compound at this dose produced SI values of 2.2, 4.1, 4.2, and 6.6 in four other studies. Higher doses produced even higher SI values (see Supplementary Tables 1 and 2). The sensitizer LLNA outcomes for hexyl cinnamic aldehyde are consistent with outcomes of guinea pig tests and experience in humans (Basketter et al., 1999).

The greater power of a statistical test to detect an increase in the SI (relative to the use of the $SI \geq 3$ criterion) must be balanced against the fact that certain SI effects below 3 may reflect an irritant response rather than a true sensitizing effect (Basketter et al., 1998). Thus, identifying such effects as “significant” might contribute more to a higher false positive rate than to a lower false negative rate. Use of the $SI \geq 3$ criterion implicitly assumes that avoiding such false positives is an important study objective, even if it increases the false negative rate. However, using the $SI \geq 3$ criterion, the LLNA is already a sensitive test for identifying skin sensitizers, even if some substances are incorrectly identified as sensitizers (Basketter et al., 1999b).

In 75 out of 83 tests conducted on 78 substances with four or five mice per group, 100% agreement between outcomes was observed. These results were similar to those of Basketter et al. (2009) who used a smaller published dataset (i.e., 17 substances) from their own laboratories to determine whether the outcomes of pooled tests with four mice were different from the outcomes with five mice. For 16 (94%) of the 17 substances (15 positive and 2 negative) evaluated, the pooled four-mice LLNA tests produced results identical to those from five mice. The discordant substance, streptomycin sulfate, produced one marginally positive result ($SI = 3.2$) in one laboratory and one clearly negative result ($SI = 1.3$) in another laboratory in the four-animal assay but produced negative results in the five-animal assay at three laboratories ($SI = 1.2, 1.3, \text{ and } 1.9$) (Kimber et al., 1998).

The current analysis yielded eight tests (for eight substances) for which four- and five-mice studies were not in complete agreement with respect to $SI \geq 3$ and $SI < 3$ (see Table 4). Assuming that the $SI \geq 3$ criterion correctly classified the six substances with $SI \geq 3$ as sensitizers, there was a small loss in power when the sample size was reduced from five mice to four mice. However, for all six substances, the likelihood was still high (77%–96%) that the compound would be identified as a sensitizer ($SI \geq 3$) using four mice. Thus, not only does the reduction in sample size from five to four mice have little impact on reliability using the $SI \geq 3$ criterion, but it also appears to have little impact on the accuracy of classification.

When Student's *t*-test was used instead of the $SI \geq 3$ criterion, reducing the sample size from five mice to four mice decreased the power slightly (Figure 1). However, for $SI \geq 3$, the power differences between studies of four mice and five mice were minimal. As shown in our analysis, a statistical test based on four mice per group will generally identify more sensitizers than using the $SI \geq 3$ criterion based on five mice per group (Tables 3, 4, and 6; Figure 1). Therefore, even if a formal statistical test is used in addition to (or rather than) the $SI \geq 3$ criterion, reducing the sample size from five to four mice per group appears to have little practical impact on the interpretation of experimental results.

5. CONCLUSIONS

We conclude that, when using the $SI \geq 3$ criterion, which is the primary method for classifying the outcomes of LLNA tests, the reduction in sample size from five to four mice has essentially no impact on the observed LLNA test outcome except for substances that produce maximum SI values close to 3 (i.e., >2 and <4). For these substances, the outcomes may differ, but any such differences reflect primarily the inherent variability among mice and the closeness of the SI value to 3 rather than the impact of reducing sample size. Empirical examination of data from 83 LLNA studies confirms that a reduction in sample size from five to four mice per group is unlikely to affect the overall accuracy of study results using the $SI \geq 3$ criterion.

ICCVAM recently reduced the minimum number of animals from five to four mice per dose group in an updated ICCVAM-recommended LLNA protocol (ICCVAM, 2009) based on the analyses reported here. Importantly, the updated protocol still calls for the collection of lymph nodes from individual mice rather than the pooling of lymph nodes for a treatment group. The collection of individual animal data permits the assessment of interanimal variability, identification of outlier responses, and statistical comparison of the treated and vehicle control group responses.

Based on the updated ICCVAM-recommended LLNA protocol and the analyses reported herein, on July 22, 2010, the OECD adopted revisions to TG 429 that reduced the minimum number of animals per dose group from five to four when lymph nodes from individual mice are processed separately (OECD, 2010). This international adoption of the four-animal group size will now facilitate the collection of individual animal data in those countries that require that the minimum number of animals be used for testing, and provide for a 20% reduction in the number of animals required for each LLNA test where individual animal data are collected. Regulatory agencies and other data end users can then take full advantage of the additional information provided by individual animal data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences. ILS staff are supported by NIEHS contract N01-ES 35504. The views expressed in this manuscript do not necessarily represent the official position of any US Federal agency.

We acknowledge the following companies and organizations and their representatives that submitted LLNA data to NICEATM for the evaluation of new applications and modifications of the LLNA: BASF – The Chemical Company, USA (Mr. Charles Hastings); The German Institute for Occupational Safety and Health, Germany (Ms. Heidi Ott); Dow AgroSciences, USA (Dr. Michael Woolhiser); DuPont, USA (Dr. Fredrick O'Neal); European Crop Protection Association, Belgium (Dr. Philip Botham); and European Federation for Cosmetic Ingredients, Belgium (Dr. Peter Ungeheuer).

References

- Basketter DA, et al. The impact of LLNA group size on the identification and potency classification of skin sensitizers: a review of published data. *Cutan Ocul Toxicol* 2009;28:19–22. [PubMed: 19514920]
- Basketter DA, et al. Strategies for identifying false positive responses in predictive skin sensitization tests. *Food and Chemical Toxicology* 1998;36:327–333. [PubMed: 9651050]
- Basketter DA, et al. Threshold for classification as a skin sensitizer in the local lymph node assay: A statistical evaluation. *Food and Chemical Toxicology* 1999b;37:1167–1174. [PubMed: 10654593]
- Dean JH, et al. ICCVAM evaluation of the murine local lymph node assay: II. Conclusions and recommendations of an independent scientific peer review panel. *Regulatory Toxicology and Pharmacology* 2001;34:258–273. [PubMed: 11754530]
- Dixon, WJ.; Massey, FJ. *Introduction to Statistical Analysis*. McGraw Hill; New York, NY: 1983.
- EPA, U. S.. *Book Health Effects Test Guidelines: OPPTS 870.2600 - Skin Sensitization*. U.S. Environmental Protection Agency; Washington, DC: 2003. *Health Effects Test Guidelines: OPPTS 870.2600 - Skin Sensitization*. Editor, (Ed.)^Eds.)
- Haneke KE, et al. ICCVAM evaluation of the murine local lymph node assay: III. Data analyses completed by the national toxicology program interagency center for the evaluation of alternative

toxicological methods. *Regulatory Toxicology and Pharmacology* 2001;34:274–286. [PubMed: 11754531]

ICCVAM. The Results of an Independent Peer Review Evaluation Coordinated by the Interagency Coordinating Committee on the Validation of Alternative Methods and the National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM). National Institute of Environmental Health Sciences; Research Triangle Park, NC: 1999. The Murine Local Lymph Node Assay: A Test Method for Assessing the Allergic Contact Dermatitis Potential of Chemicals/Compounds.

ICCVAM. Recommended Performance Standards: Murine Local Lymph Node Assay. National Institute of Environmental Health Sciences; Research Triangle Park, NC: 2009.

Kimber I, et al. Assessment of the skin sensitization potential of topical medicaments using the local lymph node assay: An interlaboratory evaluation. *Journal of Toxicology and Environmental Health - Part A* 1998;53:563–579. [PubMed: 9561969]

OECD. OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. OECD Publishing; 2002. Test No. 429. Skin Sensitisation: Local Lymph Node Assay.

OECD. OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. OECD Publishing; 2010. Test No. 429. Skin Sensitization: Local Lymph Node Assay.

Sailstad DM, et al. ICCVAM evaluation of the murine local lymph node assay: I. The ICCVAM review process. *Regulatory Toxicology and Pharmacology* 2001;34:249–257. [PubMed: 11754529]

Stokes, WS.; Schechtman, LM. Validation and Regulatory Acceptance of New, Revised, and Alternative Toxicological Methods. Informa Health Care; New York, NY: 2008.

Figure 1. Power for Four-Mice and Five-Mice Samples Based on BASF Vehicle Control Data
Lines show the variation in power (i.e., the likelihood that a sensitizer effect will be identified as statistically significant [$p < 0.05$]). The dashed line (triangle) shows the power for four mice per group and the solid line (squares) shows the power for five mice per group.

Table 1Number of Animals per Dose Group: 277 Treated Dose Groups and 67 Control Groups^a

Sample Size ^b	Number of Control Groups	Number of Treated Groups	Number of Tests ^c
2	0	1	0
3	0	1	0
4	1	8	0
5	37	169	48
6	28	98	35
9	1	0	0
Total	67	27	83

^aThe total number of control groups is less than the total number of 83 tests because some control groups were used for multiple tests.

^bNumber of mice in each group.

^cBased on the maximum number of animals in any dose group for that test.

Table 2

Stimulation Index Frequency and Agreement of Four-Mice and Five-Mice Sample Sizes for Local Lymph Node Assay Outcome^a for 275 Treated Groups

SI	Frequency of SI (number of dose groups)	Agreement ^b Between Study Outcomes (%)
<2.1	154	100.0
2.1–2.5	16	90.1 ^c
2.6	2	85.0
2.7	3	73.3
2.8	2	59.5
3.1	1	56.0
3.2	2	55.5
3.3	4	73.5
3.4	1	88.0
3.5	1	68.0
3.6	1	84.0
3.7	1	90.0
3.8	1	100.0
4.0–4.7	16	97.9 ^d
>4.7	70	100.0
Total	275	97.5

Abbreviations: SI = stimulation index.

^aProportion of studies with SI ≥ 3 plus proportion of studies with SI < 3 .

^bAverage agreement between study outcomes based on five mice per group and those based on four mice per group.

^cThe individual groups had agreement ranging from 81%–100% (11 groups had <100% agreement).

^dThe individual groups had agreement ranging from 92%–100% (5 groups had <100% agreement).

Table 3

Dose Group Examples – Effect of Sample Size on the Agreement of Local Lymph Node Assay Outcome^a for Stimulation Index Values Close to $SI \geq 3$

Agreement of LLNA Outcome	Two Studies (one with four mice per group and one with five mice per group) ^a	
	Two Studies (five mice per group) ^a	
SI = 2.8 (10% hexyl cinnamic aldehyde)		
Agreement ($SI \geq 3$)	7.7% (10/36 × 10/36)	10.5% (10/36 × 85/225)
Agreement ($SI < 3$)	52.2% (26/36 × 26/36)	44.9% (26/36 × 140/225)
Total Agreement ^a	59.9%	55.4%
Disagreement	40.1%	44.6%
SI = 3.2 (1% dipropylene triamine)		
Agreement ($SI \geq 3$)	56.2% (27/36 × 27/36)	50.7% (27/36 × 152/225)
Agreement ($SI < 3$)	6.2% (9/36 × 9/36)	8.1% (9/36 × 73/225)
Total Agreement ^a	62.4%	58.8%
Disagreement	37.6%	41.2%

Abbreviations: LLNA = murine local lymph node assay; SI = stimulation index.

^aProportion of studies with $SI \geq 3$ plus proportion of studies with $SI < 3$. Numbers in parentheses show the calculation of the agreement percentages.

Table 4

Likelihood of $SI \geq 3$ for Murine Local Lymph Node Assay Tests with Less Than Complete Agreement of Four-Mice and Five-Mice Samples

Compound	Maximum SI	Likelihood of $SI \geq 3$ (%)	
		Five-Mice Samples	Four-Mice Samples
Formulation 54	2.3	0 (0/36)	7 (16/225)
Hexyl cinnamic aldehyde	2.8	28 (10/36)	38 (85/225)
Formulation 39	3.3	92 (33/36)	78 (175/225)
Bakelite EPR 161	3.5	83 (30/36)	77 (174/225)
Formulation 55	3.7	100 (36/36)	90 (202/225)
Potassium dichromate	4.1	100 (1/1)	92 (23/25)
Formulation 51	4.5	100 (36/36)	96 (215/225)
1,6-(Bis(2-3-epoxypropoxy) hexane	4.7	100 (36/36)	94 (211/225)

Abbreviations: EPR = epoxy resin; SI = stimulation index.

Table 5Distribution of Statistically Significant ($p < 0.05$) Stimulation Index Values for 277 Dose Groups

SI	Frequency of SI	Percentage of Statistically Significant ($p < 0.05$) SI Values
<1.7	131	0.0
1.7–1.9	23	52.2
2.0–2.5	17	88.2
2.6–2.9	7	85.7
>3.0	99	100.0
Total	277 ^a	47.6

Abbreviation: SI = stimulation index.

^aIncludes one dose group of two mice and one dose group of three mice, which were excluded from the agreement-disagreement analysis since there were only 2 or 3 animals per dose group. Disintegrations per minute for vehicle control and treated groups were compared using Student's *t*-test. Bold values above the columns are percentage of statistically significant ($p < 0.05$) SI values.

Table 6
***Post hoc* Power Calculations^a Based on the BASF Vehicle Control Data**

	SI Value			
	3.5	3	2.5	2
Assumed control response (DPM) ^b	552.3	552.3	552.3	552.3
Log (control response)	6.314	6.314	6.314	6.314
Dose group response (DPM) ^c	1933.0	1656.9	1380.8	1104.6
Log (dose group response)	7.567	7.413	7.230	7.007
Difference (log scale) ^d	1.253	1.099	0.916	0.693
Assumed SD (log scale)	0.4077	0.4077	0.4077	0.4077
Delta ^e = Difference/SD	3.07	2.70	2.25	1.70
Power for five mice	99.0%	96.4%	87.9%	65.8%
Power for four mice	95.7%	89.8%	76.8%	53.0%

Abbreviations: DPM = disintegrations per minute; SD = standard deviation; SI = stimulation index.

^aThe power calculations are based on a two-sided Student's *t*-test, and assume an underlying normal distribution for the log-transformed data.

^bMean of the 17 vehicle control group mean DPMs from BASF.

^cMean vehicle control group DPM × SI value.

^dDifference (on a log scale) between the dose group and vehicle control DPM responses.

^eDelta, the standardized difference, is referred to as "Cohen's d" in the Web-based statistics program (<http://www.danielsoper.com/statcalc/calc49.aspx>) used to perform the power calculations.