Impact of Reducing the Sample Size on the Performance of the LLNA

E Salicru[1], J Haseman[2], M Paris[1], J Strickland[1], D Allen[1], W Stokes[3]
[1]ILS, Inc./NICEATM, RTP, NC; [2]J.K. Haseman Consulting, Raleigh, NC; [3]NICEATM, NIEHS, RTP, NC

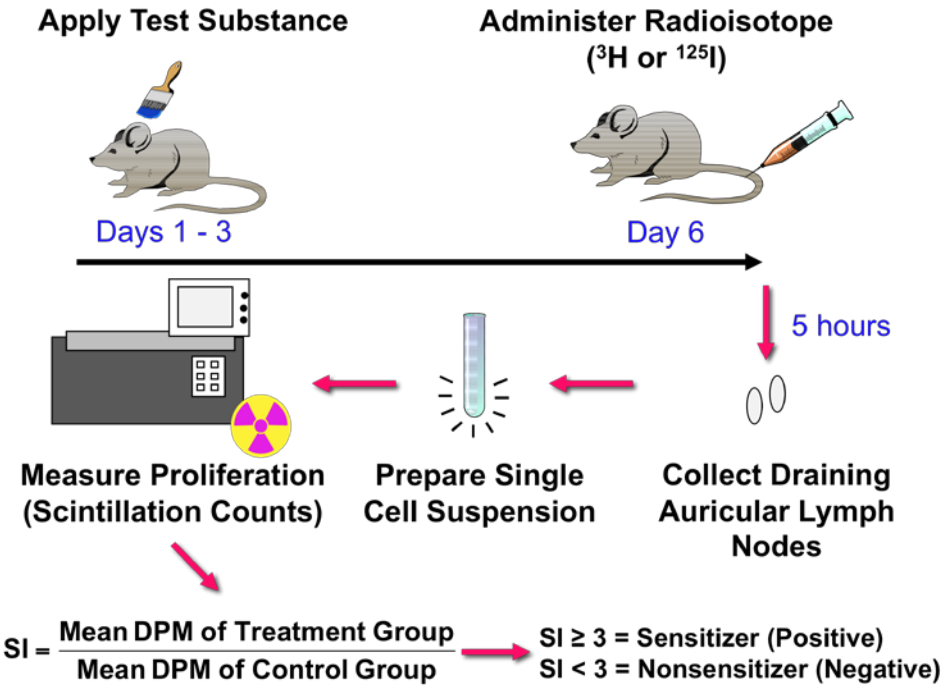
Allergic contact dermatitis (ACD) is an adverse health effect that results in lost workdays and can significantly diminish quality of life. To minimize the occurrence of ACD, regulatory authorities require testing to identify substances that may cause skin sensitization. The local lymph node assay (LLNA) is an alternative test method that virtually eliminates pain and distress associated with testing of substances for skin sensitization potential. OECD Test Guideline 429, which describes the LLNA, includes a requirement of at least four mice per group if the lymph nodes from all mice in the treatment group are pooled and a minimum of five mice per group if the lymph nodes from each mouse are processed separately. To determine if data collected from four individual animals would suffice, NICEATM used data from 83 LLNA tests (275 treated groups) to empirically determine the impact on the LLNA outcome of reducing the number of mice in each group from five to four. The average likelihood of agreement (both stimulation index [SI] < 3 or both SI ≥ 3) between LLNA outcomes with either four or five mice per group was 97.5% for the 275 treated groups. When comparing results on a test-by-test basis, there was complete agreement between outcomes with four or five mice per group for 90% (75/83) of the tests. For the remaining eight tests, there were some differences in classification between five and four mice samples with the overall agreement averaging 83%. Much of the disagreement was due to the closeness of the SI to three, not to the reduction in sample size. The practical impact of reducing the sample size from five to four mice per group on the interpretation of experimental results appears to be minimal and, therefore, using four rather than five mice per group would not impact the overall performance of the LLNA for identifying potential skin sensitizers. NICEATM and ICCVAM have recently submitted a proposal to OECD to recommend that TG 429 be updated to include a requirement of a minimum of four animals per group. ILS supported by NIEHS contract N01-ES-35504.

INTRODUCTION

LLNA Test Method

- The murine local lymph node assay (LLNA) was the first alternative test method evaluated and recommended by the U.S. Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) (ICCVAM 1999; Dean et al. 2001; Haneke et al. 2001; Sailstad et al. 2001).
- U.S. and International regulatory authorities recognize the LLNA as an acceptable alternative to guinea pig tests for most skin sensitization testing situations. The Organisation for Economic Co-operation and Development (OECD) has published Test Guideline (TG) 429 for Skin Sensitisation: LLNA (OECD 2002).
- Compared to currently accepted guinea pig tests (e.g., guinea pig maximization test and Buehler test) the LLNA:
  - Requires fewer animals
  - Requires less time
  - Avoids potential animal pain and distress

LLNA Methodology and Calculation of Results



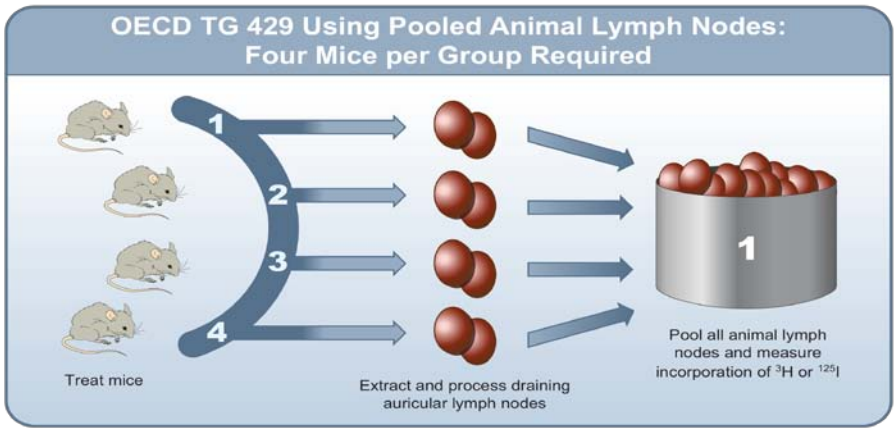
Purpose of Analysis

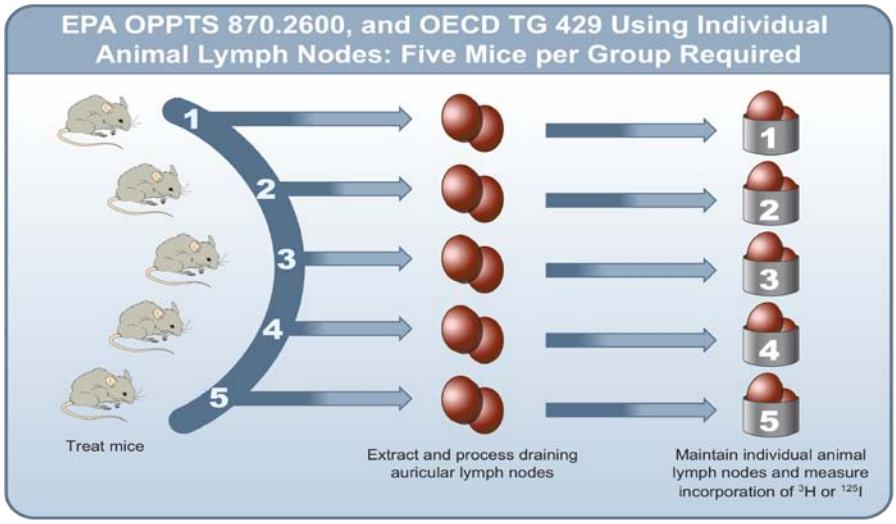
- To determine whether the five mice per group required for individual animal data collection could be reduced to four mice per group without adversely affecting the accuracy of the LLNA.

- To evaluate the usefulness of a statistical test used in addition to (or possibly even in place of) the SI decision criterion although the primary determinant of the LLNA outcome is the magnitude of the SI value.

LLNA Testing Guidelines

- Current U.S. and international LLNA testing guidelines use similar LLNA protocols (based on the ICCVAM-recommended LLNA test method protocol [ICCVAM 2009]).





Advantages of Collecting Individual Animal Data

- Allows for an assessment of inter-animal variability.
- Allows for a statistical comparison of differences between test substance and vehicle control groups.
- Allows for identification of outlier responses by using statistical tests such as Dixon's test (Dixon and Massey 1983).
  - Identifying outlier responses may avoid false negative; or false positive results for substances that produce responses near SI of 3. Substances that normally would induce an SI just above or below 3 might be incorrectly classified due to a low outlier value, because the resulting mean SI may be moved above or below 3 if the outlier is not identified and excluded.

METHODS OF EVALUATION

- Retrospective evaluation based on individual animal data from LLNA tests submitted to the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM).
- Six laboratories – all used CBA mice [as recommended in LLNA TGs - OECD 2002; EPA 2003].
- 83 individual LLNA tests representing 78 substances (individual chemicals and proprietary formulations) with three or four dose groups and a vehicle control group per test.
  - 50 tests yielded positive results (i.e., maximum SI ≥ 3).
  - 33 tests yielded negative results (i.e., all SI < 3).
- The number of individual mice per group ranged from two to nine among the 277 treated groups and the 67 control groups (**Table 1**). Two treated groups, one with two mice and one with three mice, contained too few mice for the comparison of LLNA outcomes and were excluded from SI ≥ 3 criterion analyses.
- LLNA test results were evaluated both on a dose-by-dose basis and a test-by-test basis (for SI ≥ 3 or SI < 3).

Table 1 Number of Animals per Dose Group 277 Treated Dose Groups and 67 Control Groups[1]

Sample Size[2]	Number of Control Groups	Number of Treated Groups	Number of Tests[3]
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	277	83

[1] The total number of control groups is less than the total number of 83 tests because some control groups were used for multiple tests.

[2] Number of mice.

[3] Maximum number of animals in any dose group for that test.

- For each LLNA test that used 5 mice/group, all possible DPM responses were randomly sampled to create 4 mice/group samples for both the vehicle control and treated groups (25 possible combinations per test). The SI value of each combination 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.

- Agreement could occur in two ways:
  - Both approaches could produce SI < 3, or
  - Both could produce SI ≥ 3.
- For each LLNA test that had more than 5 mice/group, a similar procedure was applied, but in these cases, it was necessary to form all possible five mice and four mice combinations from the full dataset (8100 possible combinations for tests with six animals compared to 25 possible combinations for tests with 5 mice/group).
- For tests with more than five mice, the relative impact of animal-to-animal variability and sample size reduction on study outcome was examined.
  - The disagreement related to reducing the sample size from five to four mice per group was compared to the disagreement that would occur by simply taking a second sample of five mice per group.
- In addition to the SI ≥ 3 criterion, statistical testing was also conducted.
  - All data were log transformed prior to statistical analyses.
  - A Student's *t*-test was used to compare each dose group with its concurrent vehicle control:
    - Statistically significant differences ($p<0.05$) were regarded as positive test results (i.e., sensitizers).
    - When $p>0.05$, results were regarded as negative (i.e., nonsensitizers).
  - Power calculations, based on a two-sided Student's *t*-test, were 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.

Use of the SI to Identify Sensitizers

- When evaluating the data on the basis of dose groups, 88% (241/275) of the treated groups had 100% agreement between five and four mice outcomes; 12% (34/275) of the treated groups had less than 100% agreement.
  - Disagreement was limited to those SI values from 2.1 to 4.7, but some treated groups in this range produced 100% agreement (see **Table 2, Figure 1**).
  - Disagreement increased as the SI approached 3 (**Table 2, Figure 1**). The overall average agreement between outcomes with four or five mice was 97.5%.

Table 2 Stimulation Index Frequency and Agreement of Five Mice and Four Mice Sample Sizes for Local Lymph Node Assay Outcome[1] for 275 Treated Groups

SI	Frequency of SI (number of dose groups)	Agreement[2] Between Study Outcomes (%)
< 2.1	154	100.0
2.1 – 2.5	16	90.1
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
> 4.7	70	100.0
Total	**275**	**97.5**

Abbreviations: SI = stimulation index.

[1] Proportion of samples with SI ≥ 3 plus proportion of samples with SI < 3.

[2] Average agreement between study outcomes based on 5 mice/group and those based on 4 mice/group.

Figure 1 Agreement[1] of Five Mice and Four Mice Sample Sizes for Local Lymph Node Assay Outcome[2] for 275 Dose Groups

Abbreviations: SI = stimulation index.

[1] Average agreement between study outcomes based on 5 mice/group and those based on 4 mice/group.

[2] Proportion of samples with SI ≥ 3 plus proportion of samples with SI < 3.

- **Table 3** provides two examples, each with six animals per group, that indicate that inter-animal variability, rather than reduction in sample size, was responsible for the disagreements in outcome at SI values close to 3.
  - For the treated group with SI = 2.8, reducing the sample size from five to four mice per group resulted in 44.6% disagreement in SI values, where one was ≥ 3 and one < 3 (see **Table 3**).
  - By comparison, simply taking a second study of five animals (i.e., not reducing the sample size) resulted in 40.1% disagreement. A similar result is noted for the treated group having SI = 3.2.

Table 3 Dose Group Examples that Show the Effect of Sample Size on the Agreement of Local Lymph Node Assay Outcome[1] for Stimulation Index Values Close to the SI ≥ 3 Decision Criterion

Agreement of LLNA Outcome	Two Studies (5 mice/group)	Two Studies (one with 4 mice/group and one with 5 mice/group)
SI = 2.8 (10% hexyl cinnamic aldehyde)		
Agreement (SI ≥ 3)	7.7% (10/36 x 10/36)	10.5% (10/36 x 85/225)
Agreement (SI < 3)	52.2% (26/36 x 26/36)	44.9% (26/36 x 140/225)
Total Agreement	59.9%	55.4%
Disagreement (one SI ≥ 3; one SI < 3)	40.1%	44.6%
SI = 3.2 (1% dipropylene triamine)		
Agreement (SI ≥ 3)	56.2% (27/36 x 27/36)	50.7% (27/36 x 152/225)
Agreement (SI < 3)	6.2% (9/36 x 9/36)	8.1% (9/36 x 73/225)
Total Agreement	62.4%	58.8%
Disagreement (one SI ≥ 3; one SI < 3)	37.6%	41.2%

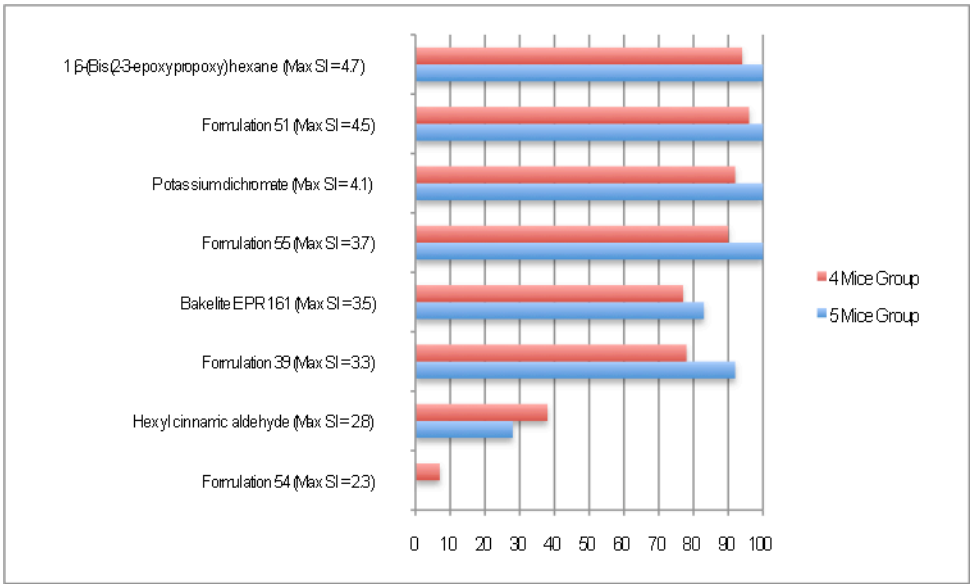
Abbreviations: LLNA = murine local lymph node assay; SI = stimulation index.

[1] Proportion of samples with SI ≥ 3 plus proportion of samples with SI < 3. Numbers in parentheses show the calculation of the agreement percentages.

Classification Between Five and Four Mice Studies

- 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, there were differences in classification between five and four mice studies (see **Figure 2**), with the overall agreement averaging 83%.
- Hexyl cinnamic aldehyde (HCA), which yielded a maximum SI < 3 at the highest dose of 10% (for this test), is a sensitizer in guinea pig tests and/or human experience (ICCVAM 1999). For HCA, the four mice study had a higher likelihood (38%) than the five mice study (28%) for detecting this effect (**Figure 2**).
- Potassium dichromate is also a known sensitizer (ICCVAM, 1999), but the categorization of the other five substances with maximum SI ≥ 3 as "true" sensitizers was uncertain. Assuming that the SI≥ 3 criterion classifies all six substances with maximum SI ≥ 3 correctly as "true" sensitizers:
  - There was a small loss in power with a reduced sample size (i.e., five to four mice).
  - The difference in power was small, and the likelihood was still high (77% - 96%) that the six chemicals would be identified as sensitizers using a sample of four mice.

Figure 2 Likelihood of SI ≥ 3 for Local Lymph Node Assay Tests with Less Than Complete Agreement of Five Mice and Four Mice Studies

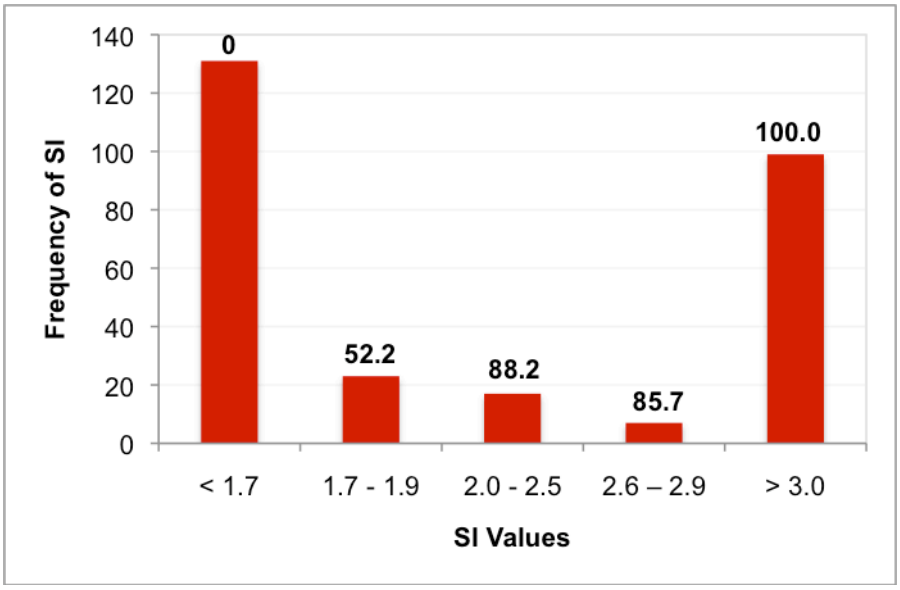


Abbreviations: EPR = epoxy resin; Max = maximum; SI = stimulation index.

Use of Statistical Significance to Identify Sensitizers

- **Figure 3** shows the distribution of statistically significant SI values ($p<0.05$ by Student's test). The statistical test identified more sensitizers than did the SI ≥ 3 criterion (132 treated groups statistically different from vehicle controls vs. 99 treated groups with SI ≥ 3). Statistical significance is evident for many treated groups with SI values below 3.
- The calculations in **Figure 3** are based on specific patterns of responses observed in 277 different treated groups.
  - To account for the interanimal variability in responses, a power calculation was performed for the largest database from a single laboratory, BASF – The Chemical Company (**Table 4**).
    - 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-fold increases in response.
    - The SD (of the log-transformed data) was assumed to be the same in the treated 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.

Figure 3 Distribution of Statistically Significant ($p<0.05$) Stimulation Index Values for 277[1] Dose Groups



Abbreviations: SI = stimulation index.
[1] Includes one dose group of two mice and one dose group of three mice. 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.

Power Calculations

- 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 was increased to 87.9% for five mice tests.
- This power calculation 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 ≥ 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.

Table 4 Post-hoc Power Calculations[1] Based on the BASF Vehicle Control Data

	SI Value			
	3.5	**3.0**	**2.5**	**2.0**
Assumed control response (DPM)[2]	552.3	552.3	552.3	552.3
Log (control response)	6.314	6.314	6.314	6.314
Treated group response (DPM)[3]	1933.0	1656.9	1380.8	1104.6
Log (treated group response)	7.567	7.413	7.230	7.007
Difference (log scale)[4]	1.253	1.099	0.916	0.693
Assumed SD (log scale)	0.4077	0.4077	0.4077	0.4077
Delta[5] = 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.

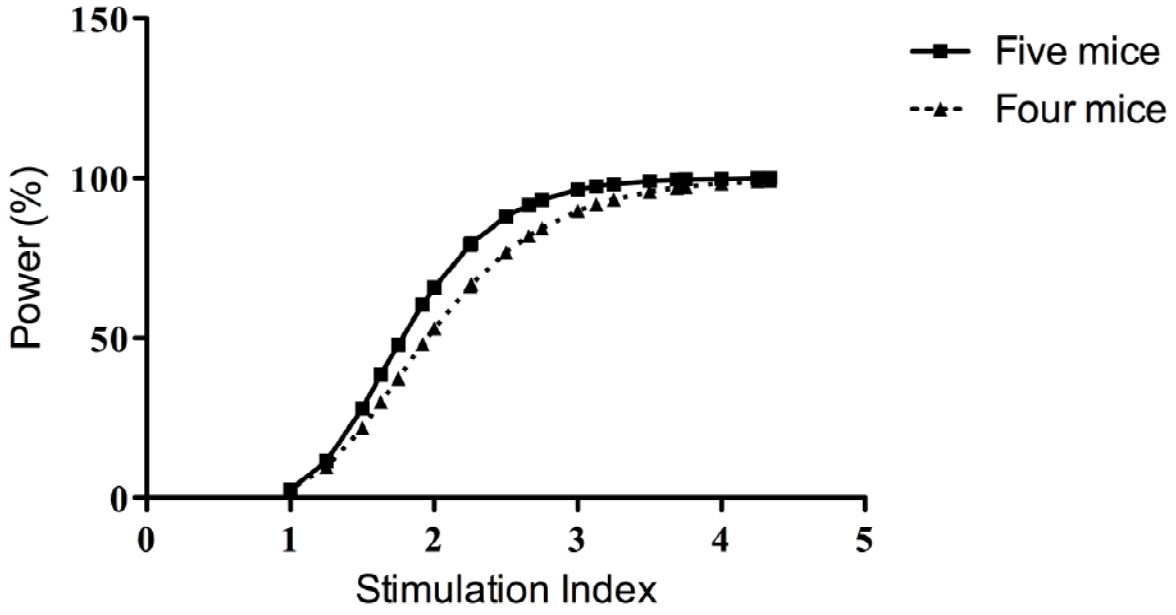
[1] The power calculations are based on a two-sided Student's *t*-test, and assume an underlying normal distribution for the log-transformed data.
[2] Mean of the 17 vehicle control group mean DPMs from BASF.
[3] Mean vehicle control group DPM x SI value.
[4] Difference (on a log scale) between the treated group and vehicle control DPM response.

[5] Delta, the standardized difference, is referred to as “Cohen’s d” in the web-based statistics program (http://www.danielsoper.com/statcalc/calc49.aspx) and was used to perform the power calculations.

Power Curves

- **Figure 4** shows power curves generated by specifying different values of Delta, which could reflect different SI values, different underlying variabilities, or a combination of these two factors. This figure shows that the power of four- vs. five-mouse groups is very similar at the SI ≥ 3 criterion.

Figure 4 Power for Five Mice and Four 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 (triangles) shows the power for 4 mice/group and the solid line (squares) shows the power for 5 mice/group.

- The updated ICCVAM-recommended protocol for the LLNA (ICCVAM 2009) incorporates the results of these analyses and changes the recommended minimum number of animals from five to four mice per group. It still recommends collection and processing of lymph nodes for individual mice rather than pooling the lymph nodes for a treatment group.
  - An updated LLNA test method protocol was transmitted to the Federal regulatory agencies for comment (September 30, 2009. See *Federal Register Notice* FR 74 50212). The agencies will provide comments to NICEATM by March 29, 2010. The Federal Register Notice and comments available at: http://ntp-apps.niehs.nih.gov/iccvampb/searchFR.cfm

Conclusions

- Using the SI ≥ 3 criterion, (the primary method for classifying the outcomes of LLNA tests) the reduction in sample size from five to four mice per group would have essentially no impact on the observed LLNA test outcome for strong sensitizers and for obvious nonsensitizers.
  - 88% (241/275) of the treated groups had 100% agreement between five and four mice outcomes; average agreement was 97.5% (**Table 2**).
  - 90% (75/83) of the tests had 100% agreement between five and four mice outcomes.
- For those substances having an SI value close to 3, the outcomes may be different, 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 the sample size (**Table 3**).
- International adoption of the four-animal group size will allow for the collection of individual animal data in those countries that require that the minimum number of animals be used for testing so that regulatory agencies and other data end users can take full advantage of the additional information it provides.
  - NICEATM and ICCVAM have recently submitted a proposal to the OECD to recommend that TG 429 be updated to include a requirement to include a minimum of four animals per group. An OECD Expert Consultation Group has since endorsed the updated TG 429 which will be considered for approval at the 22nd Meeting of the Working Group of National Coordinators of the Test Guidelines Programme, 23-25 March 2010, Paris, France.
- When using a statistical test (e.g., Student's *t*-test) rather than the SI≥ 3 criterion, reducing the sample size from five mice to four mice decreased the power slightly, especially when SI < 3 (**Figure 4**).
- A statistical test based on 4 mice/group will generally identify more sensitizers than using the SI ≥ 3 criterion based on 5 mice/group (**Tables 3, 4; Figure 4**). Therefore, even if a statistical test is used rather than (or in addition to) the SI≥ 3 criterion, the practical impact of reducing the sample size from five to four mice per group on the interpretation of experimental results appears to be minimal.

References

Dean JH, Twerdok LE, Tice RR, Sailstad DM, Hattan DG, Stokes WS. 2001. ICCVAM Evaluation of the Murine Local Lymph Node Assay (LLNA) II: Conclusions and Recommendations of an Independent Scientific Peer Review Panel. *Regul. Toxicol. Pharmacol.* 34, 258-273.

Dixon WJ, Massey FJ. 1983. Introduction to Statistical Analysis, 4th ed. McGraw-Hill, New York.

EPA. 2003. Health Effects Test Guidelines: OPPTS 870.2600 - Skin Sensitization. EPA 712–C–98–197. Washington, DC: U.S. Environmental Protection Agency. Available: http://www.epa.gov/opptsfrs/publications/OPPTS_Harmonized/870_Health_Effects_Test_Guidelines/Series/870-2600.pdf.

Haneke KE, Tice RR, Carson BL, Margolin BH, Stokes WS. 2001. 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. *Regul. Toxicol. Pharmacol.* 34, 274-286.

ICCVAM. 2009. Recommended Performance Standards: Murine Local Lymph Node Assay. NIH Publication No. 09-7357. Research Triangle Park, NC: National Institute of Environmental Health Sciences. Available: http://iccvam.niehs.nih.gov/methods/immunotox/llna_PerfStds.htm

ICCVAM. 1999. The murine local lymph node assay: A test method for assessing the allergic contact dermatitis potential of chemical/compounds. NIH Publication No. 99-4494. Research Triangle Park, NC: National Institute of Environmental Health Sciences. Available: http://iccvam.niehs.nih.gov/methods/immunotox/llna_PeerPanel98.htm

OECD. 2002. Guidelines for the Testing of Chemicals. Test guideline 429. Skin Sensitisation: Local Lymph Node Assay, adopted April 24, 2002. In: OECD Guidelines for Testing of Chemicals. Paris: OECD. Available: http://titania.sourceoecd.org/vl=1176280/cl=42/nw=1/rpsv/ij/oecdjournals/1607310x/v1n4/s30/p1.

Sailstad DM, Hattan DG, Hill RN, Stokes WS. 2001. ICCVAM evaluation of the murine local lymph node assay: I. The ICCVAM review process. *Regul. Toxicol. Pharmacol.* 34, 249-257.

Acknowledgments

This poster was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. ILS staff are supported by NIEHS contract N01-ES 35504. The views expressed on this poster do not necessarily represent the official positions of any Federal agency. Since the poster was written as part of the official duties of the authors, it can be freely copied.

We acknowledge the following companies and organizations that submitted LLNA data to NICEATM to assist in the evaluation of new applications and modifications of the LLNA: BASF – The Chemical Company, U.S.A.; BAuA (The Federal Institute for Occupational Safety and Health), Germany; Dow AgroSciences, U.S.A.; DuPont, U.S.A.; ECPA (European Crop Protection Association), Belgium; and EFfCI (European Federation for Cosmetic Ingredients), Belgium.

NICEATM-ICCVAM Website
http://iccvam.niehs.nih.gov/home.htm