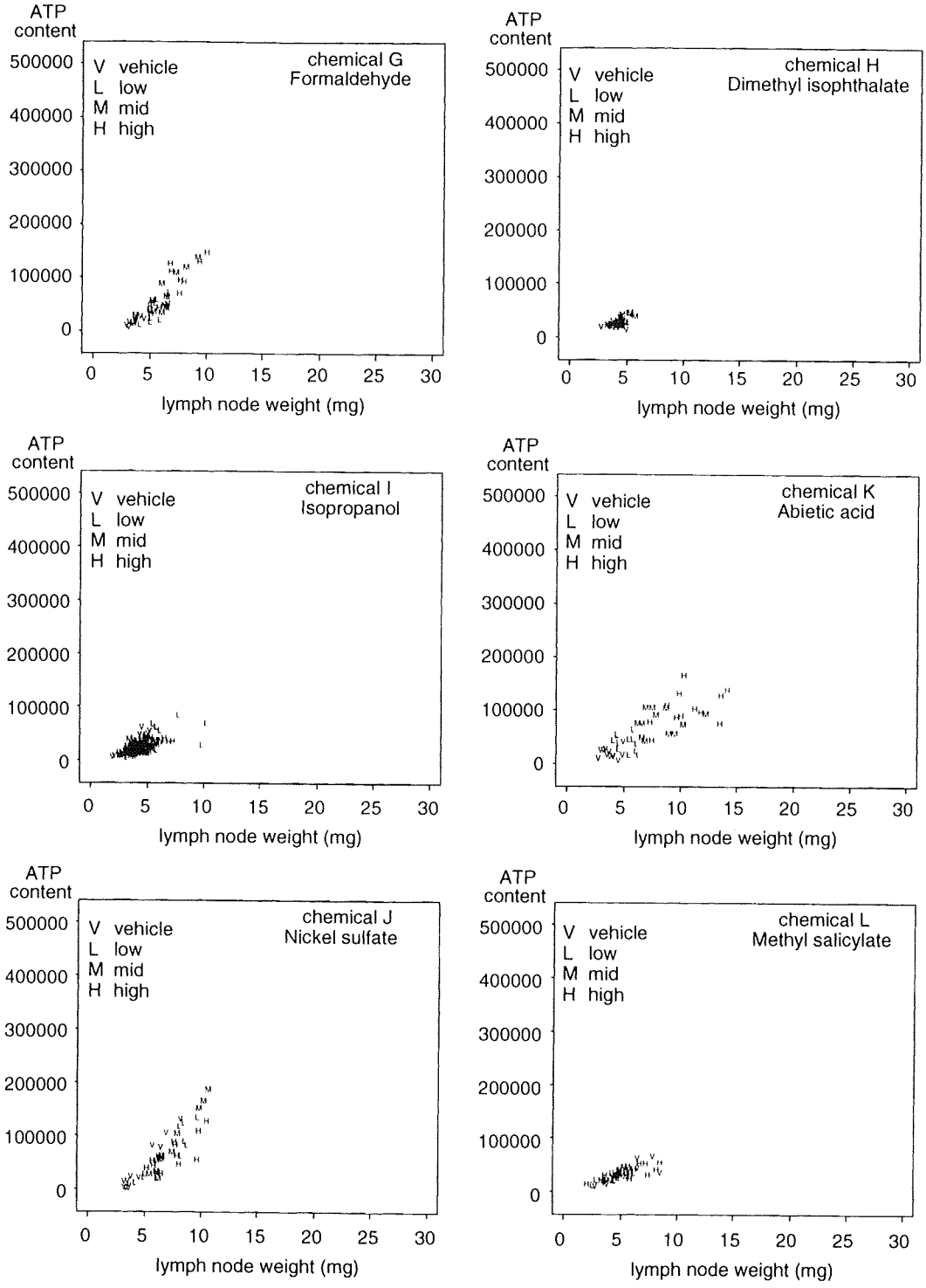

Fig. 2 (*continued*).

response relationships of the chemical were very similar between laboratories and the variation in the SI value for each dose was small, as mentioned above.

3.8. EC3 and measures of relevance

To avoid the problem of multiple counts of the same chemicals from different laboratories, the calculations of EC3 and sensitivity, specificity, accuracy, positive predictivity, and negative predictivity of LLNA-DA were based on the weighted averages of the SI values.

Tables 6(a) and 6(b) show the EC3 results and its classification for LLNA-DA based on the weighted averages for both the studies and the reported EC3 and its classification based on the reported values for LLNA.

The sensitivity, specificity, accuracy, positive predictivity, and negative predictivity of LLNA-DA with regard to the chemicals in the first study, as against those of GPMT/BT and LLNA are shown in Table 7.

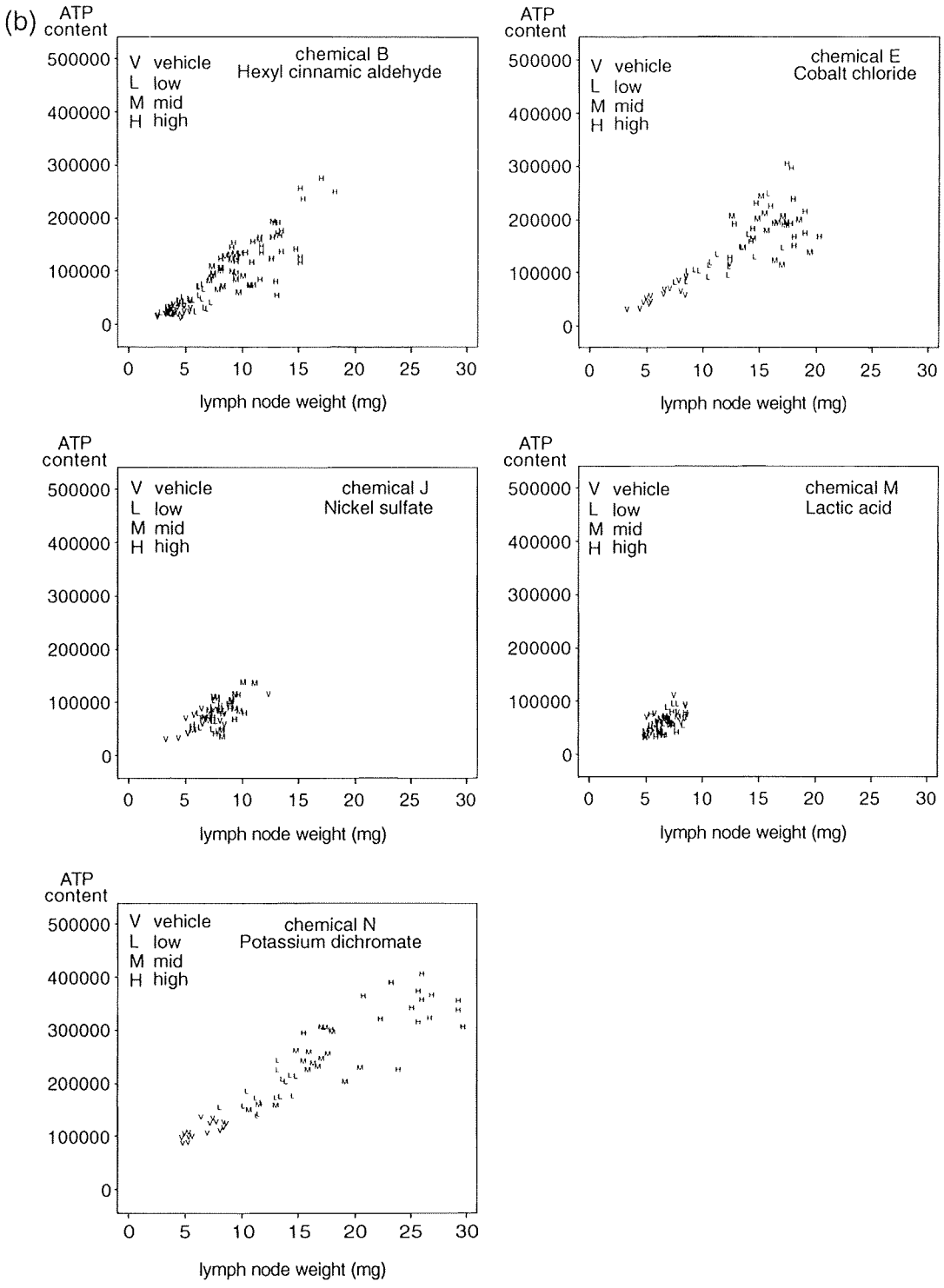

Fig. 2 (*continued*).

To enable comparison of the measurements of LLNA against those of GPMT/BT when the same chemicals were selected, these values calculated only on the basis of the referenced data are shown in Table 7.

The sensitivity, specificity, accuracy, positive predictivity, and negative predictivity values of LLNA-DA against those of GPMT/BT were similar to those of LLNA against those of GPMT/BT. Chemical C (3-aminophenol) was negative for LLNA-DA and positive for LLNA, and chemical J (nickel sulfate) was positive for LLNA-DA and negative for LLNA.

4. Discussion

Researchers have provided considerable evidence for the reliability of LLNA; however, limited evidence is available for the reliability of LLNA-DA. Since the methods involved in LLNA-DA and LLNA are essentially identical, the results of our study provide adequate evidence in support of LLNA-DA as an alternative assay method to LLNA.

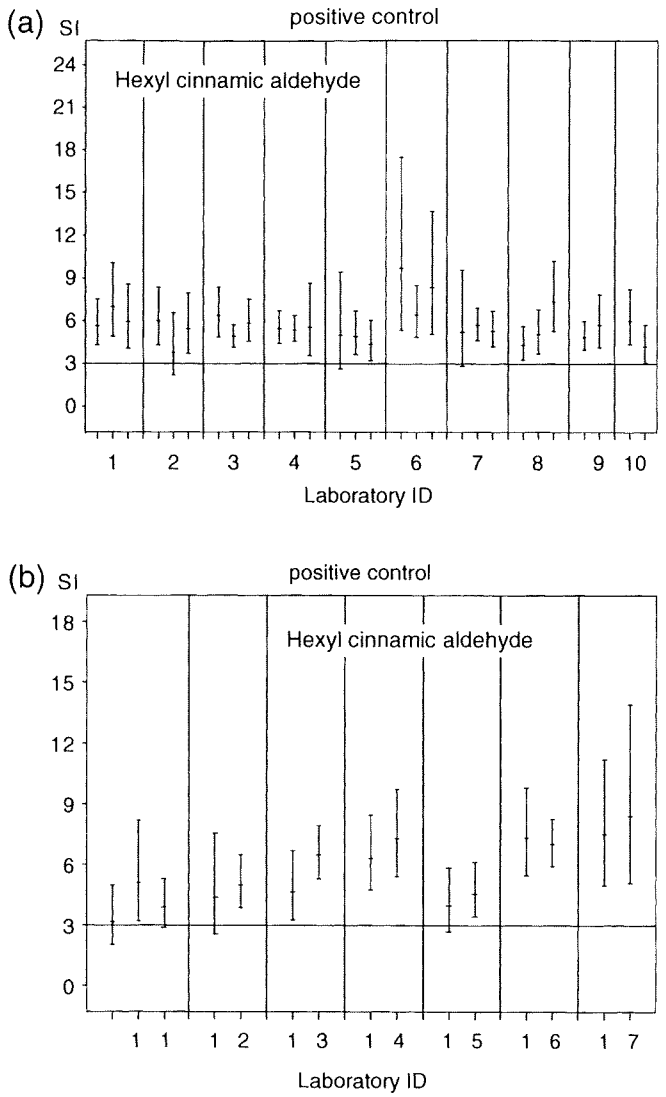

Fig. 3. (a). SI values with 95% confidence intervals obtained for the positive control (25% hexyl cinnamic aldehyde) groups in all the laboratories in the first study. (b). SI values with 95% confidence intervals obtained for the positive control (25% hexyl cinnamic aldehyde) groups in all the laboratories in the second study.

Although several interlaboratory studies on LLNA have been reported, they did not necessarily employ the same protocol; additionally, these studies were conducted by approximately 5 experimental laboratories (Basketter et al., 1991; Kimber et al., 1991, 1998, 1995; Loveless et al., 1996; Scholes et al., 1992). In contrast, one of the distinguishing features of the series of the present 2 studies is that 17 independent experimental laboratories used the same protocol to test chemicals. The fact that the interlaboratory variations were small for most of the chemicals is considered as a significant finding of this study. In particular, chemical B (hexyl cinnamic aldehyde) was tested by all the 17 laboratories; it was observed that the SI value of the interlaboratory variation was small and that the dose–response relationship was considerably similar. These results indicate that LLNA-DA is a robust technique.

In the first study, 2 of the 12 chemicals—chemicals E (cobalt chloride) and J (nickel sulfate)—demonstrated large interlaboratory variations. We considered that this might be attributed to the use of DMSO as the exclusive vehicle for these 2 chemicals. The fact that these 2 chemicals were the only metallic salts could be another reason for the large variations observed. Therefore, in the second study, it was necessary to examine interlaboratory variations with regard to other metallic salts with DMSO as the vehicle. The results of the second study, which used 5 chemicals including these 2 metallic salts, demonstrated small interlaboratory variations for all the chemicals. The small variation observed for the metallic salts could be due to the following reasons. (1) Prior to the study, the developer advised the toxicologists to carefully apply the DMSO solution onto the ears since it is highly hydrophilic, and the presence of moisture in the ears could lead to considerable variation in the results. (2) During the technical-transfer seminar, the participating toxicologists were well trained in all aspects of the experiment, including the application of DMSO solution onto the ears of the mice. Thus, our present finding is that the large variation for the 2 metallic salts in the first study was caused by inappropriate DMSO application, which in comparison with AOO or ACE has unique physical properties in terms of the difficulty involved in its application to the dorsum of the ears. Therefore, this factor was considered when the metallic salts were assessed with LLNA-DA.

Furthermore, these 2 studies provided substantial historical data with regard to the ATP content for the vehicle control group that used AOO, ACE, or DMSO with LLNA-DA. These data could be referred to by new laboratories that are considering the use of this assay. As in the case of LLNA studies, data from these studies regarding DMSO appear to demonstrate the toxicity of the chemical (Wright et al., 2001). We observed a higher ATP content when DMSO was used as a solvent than when AOO or ACE was used. This tendency may cause the SI values to change depending on the vehicle used in the experiment because a high ATP content in the vehicle control group would lead to relatively low SI values.

These studies also present certain limitations. First, the results are representative of only 14 chemicals. Although it may be recommended that the assay be tested using several chemicals, the

Table 4(a)
The weighted average of the SI values and the variance component, τ^2, in the first study

Chemical	Low-dose group			Middle-dose group			High-dose group		
	SI	95%CI	τ^2	SI	95%CI	τ^2	SI	95% CI	τ^2
A: 2,4-Dinitrochlorobenzene	2.5	(2.0, 3.0)	0.03	3.9	(3.1, 4.8)	0.03	10.3	(8.4, 12.8)	0.04
B: Hexyl cinnamic aldehyde	1.4	(1.2, 1.6)	0.00	2.8	(2.5, 3.3)	0.01	5.1	(4.2, 6.2)	0.03
C: 3-Aminophenol	1.4	(1.2, 1.7)	0.00	2.0	(1.6, 2.4)	0.00	2.2	(1.7, 2.9)	0.02
D: Glutaraldehyde	1.0	(0.7, 1.5)	0.02	2.3	(1.2, 4.6)	0.13	3.6	(2.4, 5.2)	0.03
E: Cobalt chloride	6.1	(2.7, 13.9)	0.11	5.0	(1.9, 13.2)	0.29	7.4	(2.4, 23.3)	0.42
F: Isoeugenol	2.7	(2.2, 3.4)	0.00	3.4	(2.4, 4.7)	0.02	6.7	(5.5, 8.3)	0.00
G: Formaldehyde	1.8	(1.1, 3.1)	0.07	2.7	(2.1, 3.4)	0.00	3.4	(2.5, 4.7)	0.01
H: Dimethyl isophthalate	1.2	(1.0, 1.4)	0.00	1.1	(0.9, 1.3)	0.00	0.9	(0.7, 1.2)	0.02
I: Isopropanol	1.1	(0.8, 1.4)	0.04	0.9	(0.8, 1.0)	0.00	0.9	(0.8, 1.1)	0.03
J: Nickel sulfate	2.7	(1.1, 6.5)	0.24	3.1	(1.4, 6.9)	0.20	3.1	(0.8, 12.1)	0.62
K: Abietic acid	2.1	(1.8, 2.4)	0.00	3.7	(3.1, 4.3)	0.00	5.4	(3.5, 8.3)	0.04
L: Methyl salicylate	0.9	(0.6, 1.3)	0.03	1.1	(0.7, 1.6)	0.04	1.2	(0.8, 1.9)	0.04

The variance component τ^2 represents the interlaboratory variance for the log-transformed SI, which is obtained by decomposing the total variance into the between variance and within variance by performing meta-analysis with a random effect model. Since τ^2 indicates variance, its value is greater than 0, and a higher value indicates greater interlaboratory variation.

Table 4(b)
The weighted averages of the SI values and the variance component,(τ^2, in the second study

Chemical	Low-dose group			Middle-dose group			High-dose group		
	SI	95%CI	τ^2	SI	95%CI	τ^2	SI	95% CI	τ^2
B: Hexyl cinnamic aldehyde	1.7	(1.4, 2.0)	0.00	3.8	(3.1, 4.6)	0.01	5.9	(4.8, 7.2)	0.01
E: Cobalt chloride	2.0	(1.5, 2.6)	0.02	3.0	(2.0, 4.5)	0.07	3.2	(2.1, 4.9)	0.07
J: Nickel sulfate	1.1	(0.7, 1.6)	0.06	1.3	(1.0, 1.6)	0.01	1.2	(0.8, 1.8)	0.07
M: Lactic acid	1.0	(0.8, 1.1)	0.00	0.7	(0.6, 0.9)	0.00	0.8	(0.7, 0.9)	0.00
N: Potassium dichromate	2.3	(1.8, 3.0)	0.02	3.3	(2.2, 4.8)	0.06	5.1	(4.1, 6.3)	0.02

The variance component τ^2 represents the interlaboratory variance for the log-transformed SI, which is obtained by decomposing the total variance into the between variance and within variance by using meta-analysis with a random effect model. Since τ^2 indicates variance, it takes on a value greater than 0, and a larger value indicates greater interlaboratory variation.

chemicals used in the present studies were selected from a wide range of chemicals, and their skin sensitization potentials were determined by the application of the LLNA method.

Further, the precision of the measurements of relevance was low because only 12 chemicals were tested by this assay method; therefore, even a difference in only a single chemical would affect the sensitivity. Since the study demonstrated the strong reliability of the assay, further assessments using other known chemicals should be conducted in other studies. Idehara et al. (in press) report the results of the intralaboratory study.

Another limitation is with regard to the quality of the data. It was extremely difficult to ensure complete compliance with good laboratory practice (GLP) in these studies. However, although the experiments involved in the studies were not conducted in complete accordance with GLP, the format file for data recording of individual experiments was devised at the planning stage of the study, and the data files collected for all the experiments complied with this format. Furthermore, since all the data used for the analyses were based on the database, if required, we can provide the database regarding the ATP content values obtained for the individual animals with the standard protocol that was used here.

Unlike LLNA, LLNA-DA measures the ATP content. It is an extremely simple method for measuring the ATP content during an experiment, and it yields quick results. However, since the ATP content of the LNCs decreases with time, while performing LLNA-DA, it is necessary to comply with the time of operation from lymph node excision to the determination of ATP content. Measuring the LNW as an internal control is recommended. The plot of ATP content against LNW, as in Fig. 2, might aid in roughly checking the compliance.

In conclusion, these 2 studies provide valuable evidence for the reliability of LLNA-DA.

Table 5(a)
Judgment based on SI values greater than 3 for LLNA and the referenced values for LLNA and GPMT/BT in the first study

Chemical	LLNA	GPMT/BT	Laboratory									
			1	2	3	4	5	6	7	8	9	10
A: 2,4-Dinitrochlorobenzene	+	+	+	+	+	+	+	+	+	+	+	+
B: Hexyl cinnamic aldehyde	+	+	+	+	+	+	+	+	+	+	+	+
C: 3-Aminophenol	+	+nonstd	−		−					−		
D: Glutaraldehyde	+		+	+			−					
E: Cobalt chloride	+	+				−		+		+		
F: Isoeugenol	+	+				+	+				+	
G: Formaldehyde	+	+	+	+			−					
H: Dimethyl isophthalate	−	−	−		−				−			
I: Isopropanol	−	−	−	−	−	−	−	−	−	−	−	−
J: Nickel sulfate	−	+				−		+		+		
K: Abietic acid	+	+		+				+	+			
L: Methyl salicylate	−	−			−				−			−

"nonstd" indicates a nonstandard animal.

Table 5(b)
Judgment based on SI values greater than 3 for LLNA and the referenced values for LLNA and GPMT/BT in the second study

Chemical	LLNA	GPMT/BT	Laboratory						
			11	12	13	14	15	16	17
B: Hexyl cinnamic aldehyde	+	+	+	+	+	+	+	+	+
E: Cobalt chloride	+	+	−		−	+			+
J: Nickel sulfate	−	+	−	−		−		−	
M: Lactic acid	−	−	−		−		−	−	
N: Potassium dichromate	+	+	+	+			+		+

Table 6(a)
EC3 and chemical classification in the first study

Chemical	LLNA-DA		LLNA	
	EC3	Classification	EC3	Classification
A: 2,4-Dinitrochlorobenzene	0.06	Extreme	0.04	Extreme
B: Hexyl cinnamic aldehyde	11.1	Weak	8.4	Moderate
C: 3-Aminophenol	–	Negative	3.2	Moderate
D: Glutaraldehyde	0.3	Strong	0.1	Extreme
E: Cobalt chloride	–	(Positive)[a]	<0.5	Strong
F: Isoeugenol	1.9	Moderate	1.8	Moderate
G: Formaldehyde	3.0	Moderate	0.7	Strong
H: Dimethyl isophthalate	–	Negative	–	Negative
I: Isopropanol	–	Negative	–	Negative
J: Nickel sulfate	2.7	Moderate	–	Negative
K: Abiet ic acid	7.9	Moderate	14.7	Weak
L: Methyl salicylate	–	Negative	–	Negative

The EC3 for LLNA-DA is based on the weighted average. The SI values obtained for chemical E (cobalt chloride) with LLNA-DA was greater than 3 for all the doses; however, since the dose–response relationship yielded a v-shaped curve, the EC3 could not be determined.

[a] Although the weighted averages of the SI values were greater than 3 for all the doses, the EC3 and classification were determined because the dose–response relationship exhibited a v-shaped curve.

Table 6(b)
EC3 and chemical classification in the second study

Chemical	LLNA-DA		LLNA	
	EC3	Classification	EC3	Classification
B: Hexyl cinnamic aldehyde	8.1	Moderate	8.4	Moderate
E: Cobalt chloride	3.0	Moderate	<0.5	Strong
J: Nickel sulfate	–	Negative	–	Negative
M: Lactic acid	–	Negative	>25	Negative
N: Potassium dicromate	0.2	Strong	0.1	Strong

Table 7
Sensitivity, specificity, accuracy, positive predictivity, and negative predictivity in the study

	n	Sensitivity	Specificity	Accuracy	Positive predictivity	Negative predictivity
LLNA-DA vs. GPMT/BT	11	87.5% (7/8)	100% (3/3)	90.9% (10/11)	100% (7/7)	75.0% (3/4)
LLNA-DA vs. LLNA	12	87.5% (7/8)	75.0% (3/4)	83.3% (10/12)	88% (7/8)	75.0% (3/4)
LLNA vs. GPMT/BT	11	87.5% (7/8)	100% (3/3)	90.9% (10/11)	100% (7/7)	75.0% (3/4)

For LLNA-DA, the judgment was based on the weighted average of the SI values. For LLNA and GPMT/BT, judgments were based on the referenced data.

Acknowledgment

This study was partially supported by a grant from the Ministry of Health, Labour and Welfare (principal investigator: Yasuo Ohno) and Japanese Society for Alternatives to Animal Experiments.

We would like to express our appreciation to the editors and the reviewers of the original manuscript, who made insightful suggestions that improved the clarity of the presentation.

References

Basketter, D. A., Blaikie, L., Dearman, R. J., Kimber, I., Ryan, C. A., Gerberick, G. F., et al. (2000). Use of the local lymph node assay for the estimation of relative contact allergenic potency. *Contact Dermatitis, 42*, 344–348.

Basketter, D. A., Casati, S., Gerberick, G. F., Griem, P., Philips, B., & Worth, A. (2005). Skin sensitisation. *Alternatives To Laboratory Animals, 33*(Supplement 1), 83–103.

Basketter, D. A., Evans, P., Fielder, R. J., Gerberick, G. F., Dearman, R. J., & Kimber, I. (2002). Local lymph node assay: Validation, conduct and use in practice. *Food and Chemical Toxicology, 40*, 593–598.

Basketter, D. A., Gerberick, G. F., & Kimber, I. (1998). Strategies for identifying false positive responses in predictive skin sensitization tests. *Food and Chemical Toxicology, 36*, 327–333.

Basketter, D. A., Gerberick, G. F., Kimber, I., & Loveless, S. E. (1996). The local lymph node assay: A viable alternative to currently accepted skin sensitization tests. *Food and Chemical Toxicology, 34*, 985–997.

Basketter, D. A., Lea, L. J., Cooper, K. J., Ryan, C. A., Gerberick, G. F., Dearman, R. J., et al. (1999). Identification of metal allergens in the local lymph node assay. *American Journal of Contact Dermatitis, 10*, 207–212.

Basketter, D. A., Lea, L. J., Dickens, A., Briggs, D., Pate, I., Dearman, R. J., et al. (1999). A comparison of statistical approaches to the derivation of EC3 values from local lymph node assay dose responses. *Journal of Applied Toxicology, 19*, 261–266.

Basketter, D. A., & Scholes, E. W. (1992). Comparison of the local lymph node assay with the guinea-pig maximization test for the detection of a range of contact allergens. *Food and Chemical Toxicology, 30*, 65–69.

Basketter, D. A., Scholes, E. W., Kimber, I., Botham, P. A., Hilton, J., Miller, K., et al. (1991). Interlaboratory evaluation of the local lymph node assay with 25 chemicals and comparison with guinea pig test data. *Toxicology Methods, 1*, 30–43.

Dean, J. H., Twerdok, L. E., Tice, R. R., Sailstad, D. M., Hattan, D. G., & Stokes, W. S. (2001). ICCVAM evaluation of the murine local lymph node assay. II. Conclusions and recommendations of an independent scientific peer review panel. *Regulatory Toxicology and Pharmacology, 34*, 258–273.

Dearman, R. J., Hilton, J., Basketter, D. A., & Kimber, I. (1999). Cytokine endpoints for the local lymph node assay: Consideration of interferon-g and interleukin 12. *Journal of Applied Toxicology, 19*, 149–155.

Durand, G., De Burlet, G., Virat, M., & Nauman, B. D. (2003). Use of the local lymph node assay in the evaluation of the sensitizing potential of pharmaceutical process intermediates. *Contact Dermatitis, 49*, 148–154.

Ehling, G., Hecht, M., Heusener, A., Huesler, J., Gamer, A. O., van Loveren, H., et al. (2005). An European inter-laboratory validation of alternative endpoints of the murine local lymph node assay: First round. *Toxicology, 212*, 60–68.

Ehling, G., Hecht, M., Heusener, A., Huesler, J., Gamer, A. O., van Loveren, H., et al. (2005). An European inter-laboratory validation of alternative endpoints of the murine local lymph node assay: 2nd round. *Toxicology, 212*, 69–79.

FDA (2002). *Guidance for industry—immunotoxicology evaluation of investigational new drugs.*

Gerberick, G. F., Ryan, C. A., Kern, P. S., Dearman, R. J., Kimber, I., Patlewicz, G. Y., et al. (2004). A chemical dataset for evaluation of alternative approaches to skin-sensitization testing. *Contact Dermatitis, 50*, 274–288.

Gerberick, G. F., Ryan, C. A., Kimber, I., Dearman, R. J., & Basketter, D. A. (2000). Local lymph node assay: Validation assessment for regulatory purposes. *Toxicology, 11*, 3–18.

Haneke, K. E., Tice, R. R., Carson, B. L., Margolin, B. H., & Stokes, W. S. (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. *Regulatory Toxicology and Pharmacology, 34*, 274–286.

Hastings, K. L. (2001). Pre-clinical methods for detecting the hypersensitivity potential of pharmaceuticals: Regulatory considerations. *Toxicology, 158*, 85–89.

Hatao, M., Hariya, T., Katsumura, Y., & Kato, S. (1995). A modification of the local lymph node assay for contact allergenicity screening: Measurement of interleukin-2 as an alternative to radioisotope-dependent proliferation assay. *Toxicology, 98*, 15–22.

Idehara, K., Yamagishi, G., Yamashita, K., & Ito, M. (2008). Characterization and evaluation of a modified local lymph node assay using ATP content as a non-radio isotopic endpoint. *Journal of Pharmacological and Toxicological Methods, 58*, 1–10 (this issue).

Kimber, I. (2001). The local lymph node assay and potential application to the identification of drug allergens. *Toxicology, 158*, 59–64.

Kimber, I., Hilton, J., Botham, P. A., Basketter, D. A., Scholes, E. W., Miller, K., et al. (1991). The murine local lymph node assay: Results of an inter-laboratory trial. *Toxicology Letters, 55*, 203–213.

Kimber, I., Hilton, J., Dearman, R. J., Gerberick, G. F., Ryan, C. A., Basketter, D. A., et al. (1998). 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, 53*, 563–579.

Kimber, I., Hilton, J., Dearman, R. J., Gerberick, G. F., Ryan, C. A., Basketter, D. A., et al. (1995). An international evaluation of the murine local lymph node assay and comparison of modified procedures. *Toxicology, 103*, 63–73.

Lee, J. K., Park, J. H., Park, S. H., Kim, H. S., & Oh, H. Y. (2002). A nonradioisotopic endpoint for measurement of lymph node cell proliferation in a murine allergic contact dermatitis model, using bromodeoxyuridine immunohistochemistry. *Journal of Pharmacological and Toxicological Methods, 48*, 53–61.

Loveless, S. E., Ladics, G. S., Gerberick, G. F., Ryan, C. A., Basketter, D. A., Scholes, E. W., et al. (1996). Further evaluation of the local lymph node assay in the final phase of an international collaborative trial. *Toxicology, 108*, 141–152.

Normand, S. L. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine, 18*, 321–359.

OECD (1992). Organization for Economic Co-operation and Development—OECD guidelines for testing of chemicals. *No. 406: Skin sensitization.*

OECD (2002). Organization for Economic Co-operation and Development—OECD guidelines for testing of chemicals. *No. 429: Skin sensitization: Local lymph node assay.*

OECD (2005). Organization for Economic Co-operation and Development—OECD series on testing and assessment. *No. 34: Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment.*

Omori, T., Ikarashi, Y., Kanazawa, Y., Idehara, K., Kojima, H., Sozu, T., et al.(2008). Validation studies on an alternative endpoint for the local lymph node assay (LLNA-DA): Importance of study management. *Alternatives to Animal Testing and Experimentation*, 14, Special Issue, 429–432.

Omori, T., & Sozu, T. (2007). Variance of the stimulation index for the local lymph node assay. *Alternatives to Animal Testing and Experimentation, 12*, 321–359.

Sailstad, D. M., Hattan, D., Hill, R. N., & Stokes, W. S. (2001). ICCVAM evaluation of the murine local lymph node assay. I. The ICCVAM review process. *Regulatory Toxicology and Pharmacology, 34*, 249–257.

Scholes, E. W., Basketter, D. A., Sarll, A. E., Kimber, I., Evans, C. D., Miller, K., et al. (1992). The local lymph node assay: Results of a final inter-laboratory validation under field conditions. *Journal of Applied Toxicology, 12*, 217–222.

Takeyoshi, M., Yamasaki, K., Yakabe, Y., Takatsuki, M., & Kimber, I. (2001). Development of non-radio isotopic endpoint of murine local lymph node assay based on 5-bromo-2′-deoxyuridine (BrdU) incorporation. *Toxicology Letters, 119*, 203–208.

Wright, Z. M., Basketter, D. A., Blaikie, L., Cooper, K. J., Warbrick, E. V., Dearman, R. J., et al. (2001). Vehicle effects on skin sensitizing potency of four chemicals: Assessment using the local lymph node assay. *International Journal of Cosmetic Science, 23*, 75–83.

Yamashita, K., Idehara, K., Fukuda, N., Yamagishi, G., & Kawada, N. (2005). Development of a modified local lymph node assay using ATP measurement as an endpoint. *Alternatives to Animal Testing and Experimentation, 11*, 136–144.

Mutagenesis vol. 23 no. 3 pp. 171–182, 2008
Advance Access Publication 2 April 2008
doi:10.1093/mutage/gen015

REVIEW

Statistical issues in the use of the comet assay

David P. Lovell* and Takashi Omori[1]

Department of Biostatistics, Postgraduate Medical School, University of Surrey, Daphne Jackson Road, Manor Park, Guildford, Surrey GU2 7WG, UK and [1]Department of Biostatistics, Kyoto University School of Public Health, Onogawa 16-2, Tsukuba, Japan

The comet or single-cell gel electrophoresis assay is now widely used in regulatory, mechanistic and biomonitoring studies using a range of *in vitro* and *in vivo* systems. Each of these has issues associated with the experimental design which determine to a large extent the statistical analyses than can be used. A key concept is that the experimental unit is the smallest 'amount' of experimental material that can be randomly assigned to a treatment: the animal for *in vivo* studies and the culture for *in vitro* studies. Biomonitoring studies, being observational rather than experimental, are vulnerable to confounding and biases. Critical factors in any statistical analysis include the identification of suitable end points, the choice of measure to represent the distribution of the comet end point in a sample of cells, estimates of variability between experimental units and the identification of the size of effects that could be considered biologically important. Power and sample size calculations can be used in conjunction with this information to identify optimum experimental sizes and provide help in combining the results of statistical analyses with other information to aid interpretation. Interpretation based upon the size of effects and their confidence intervals is preferred to that based solely upon statistical significance tests. Statistical issues associated with the design and subsequent analyses of current validation studies for the comet assay include the identification of acceptable levels of intra- and inter-laboratory repeatability and reproducibility and criteria for dichotomizing results into positive or negative.

Introduction

The comet or single-cell gel electrophoresis assay is now widely used as a quick, sensitive and cheap method for measuring DNA strand breaks in eukaryotic cells for the investigation of genetic damage associated with exposures to potentially genotoxic agents. The method has evolved over the last 20 years since first described by Östling and Johanson (1) and is now used in regulatory, mechanistic and biomonitoring studies in a range of species in *in vitro* and *in vivo* systems (2,3). The assay has the advantage that it can be carried out in non-proliferating cells and single cells from different tissues can be evaluated in large numbers. Guidelines based upon the Organisation for Economic Co-operation and Development (OECD) genotoxicity guidelines are being developed for the alkaline version of the *in vivo* comet assay [single cell gel (pH > 13)] and a validation study of the *in vivo* comet assay is currently being planned by the Mammalian Mutagenesis Study Group (MMS)/Japanese Center for the Validation of Alternative Methods (JaCVAM) (4). *In vitro* methods are being developed with the aim of future validation.

An overview of the uses of the comet assay is given by Collins (3) and the recent Comet Workshop (papers in this issue) showed the range of applications from traditional genotoxicity *in vivo* and *in vitro* studies, through mechanistic studies to its use in ecotoxicology such as for aquatic toxicology. Other examples of its use include investigations on industrial chemicals, pharmaceuticals, biocides, agrochemicals and food chemicals such as additives. It is also used as a biomarker in cancer and nutrition studies (papers in this issue).

A number of protocols have been developed for use in different types of investigations, for instance, for the neutral and the alkaline versions. Guidelines and recommendations for the conduct of studies have been published (5,6). A number of modifications of the comet assay have been developed, for instance, to measure cross-links by determining the reduction of induced DNA migration (7) and for investigating base excision repair and nuclear excision repair (8).

Continuing development of the assay means that a range of statistical methods may be used. Many of them are likely to be interchangeable and, although giving numerically different results, are likely to lead to qualitatively similar conclusions. Where different statistical methods produce different conclusions, this is often an indication that a careful inspection of the data is needed. It is, though, unlikely that a single statistical method will meet every requirement (9).

This paper not only addresses generic issues associated with the comet assay but also discusses specific issues associated with the protocols being developed for *in vivo* and *in vitro* studies and which will be used in validation studies. The statistical input into the development of experimental design is emphasized. Many of the experimental design and statistical analysis issues associated with comet assay are common to many other genotoxicity studies and, in general, to other experimental systems. Comments made here may apply to other assays. It is stressed that the size of an effect [and some indication of the confidence interval [CI] associated with it] is more important than determining statistical significance by itself.

Image analysis and end points

The comet has a complex form which after visualization can be simplified to a set of multivariate data representing the shape. Image analysis methods are capable of collecting a large amount of information on the image and various proprietary automated systems exist and public domain programmes have been developed for image analysis (10). Automated scanners can measure many different components relating to the shape of the comet.

*To whom correspondence should be addressed. Tel: +44 1483 688609; Fax: +44 1483 688501; Email: d.lovell@surrey.ac.uk

Debate continues over whether manual or automatic scoring is best (4). For instance, while manual scoring using an eyepiece micrometre to measure tail length is permitted, image analysis is recommended (11). The key point is that the method should be consistent over a study or combination of studies. The method of scoring is relevant if meta-analysis is planned and where the absolute size of effect is important. It is important to avoid bias in the identification of cells to measure comets (3). Cells with large tails may overlap and thus may not be selected for measurement which could violate the assumptions underlying statistical tests that the cells represent a random sample.

Three measures of DNA migration are commonly used: tail length, tail moment and % of the DNA in tail (% tail DNA). Metrics on tail or head length and moment are measured in arbitrary units and may vary from study to study or from laboratory to laboratory. Tail length is considered unsatisfactory as a measure because the length only increases at relatively low damage levels and is sensitive to the background intensity of the image analysis system which affects the criteria for determining the end of the tail (3). Tail moment, an index taking account both the migration of the genetic material and the relative amount of DNA in the tail, can be calculated a number of ways. The Olive tail moment, for instance, is the product of the tail length and the % tail DNA. The % tail DNA is a measure of the relative fluorescent intensity in the head and tail (3). The % tail DNA values are constrained to a maximum of 100 and a minimum of 0 with no variability at the extremes and maximum variability at intermediate values such as 50%.

The % tail DNA has the advantage that it can be 'standardized' over studies while tail length and moment, although consistent within a study, may not be comparable across studies. There is a increasing emphasis on the use of the % tail DNA as the preferred metric or the primary end point (12) and it was recognized as the most suitable primary end point at the International Workshop on Genotoxicity Test Procedures at San Francisco in 2005 (4). Hartmann *et al.* (11), for instance, in describing the various measures suggest that 'there is much to recommend the use of per cent DNA in tail'. Relative tail intensity (the % tail DNA) was linearly related to DNA damage over a wide range of damage and is related to DNA break frequency. Collins (3) viewed % tail DNA as the most useful measure because it covered a wide range of damage (from 0 to 100%), was independent of the threshold settings of the image analysis program used and gave some 'feel' for what the comet looked like. In contrast, tail moment was not linearly related to dose and did not provide an indication of what the comet looked like. One complication with % tail DNA is that the presence of zero values would complicate statistical analysis. Collins (3), however, suggests that a check of whether cells are in satisfactory condition for the assay is that untreated control cells should have a background level of breaks (i.e. ~10% DNA in tail) and there are suggestions that negative control cells should have between 10 and 20% DNA in tail which would obviate statistical problems.

Other variations on the measures made include 'comet moments' (13) and 'tail inertia'(14). Bowden *et al.* (15) developed a 'tail profile' which identifies more DNA damage than measured by the tail moment. They derived a 'profile plot', a visual representation of a series of comets on a slide, which could identify features in the data that were not otherwise apparent.

An alternative scoring system is to classify DNA migration data using a five-category classification scheme (0, no damage to 4, almost all the DNA in the tail). The system is manual and relies on a subjective assessment based upon comparisons with standard images. Collins (3) illustrates the five classes (0–4). Each grade is equivalent to ~20% band on the % tail DNA score. The scores for a sample of 100 comets from a slide can be combined to provide an overall score for the slide on an arbitrary score from 0 to 400. This score shows close agreement with scores based upon % tail DNA (3).

Comet cells can also be categorized as responder or non-responder cells based upon the degree of damage and the proportion of responder cells on a slide then used as the measure of damage. Altman and Royston (16), however, point to the costs of dichotomizing continuous variables and that dichotomizing at the median is comparable with losing a third of the data.

Comparisons of results with different comet end points may be useful. Lee and Steiner, in the context of environmental studies, suggest the use of both tail moment and % tail DNA data in the analysis (17). Other data collected in comet assay are measures of cell toxicity, damage or viability and information on the percentage of 'hedgehogs' (cells with a small or non-existent head and large, diffuse tails). These data are not usually included in the formal statistical analysis of the comet measures but are important for an assessment of the quality of the study.

Experimental unit

The concept of the experimental unit is fundamental to the statistical analysis of designed experiments. Misspecification of the experimental unit can lead to serious misinterpretation of the statistical analysis. The US National Institute of Standards and Technology defines the experimental unit as 'the entity to which a specific treatment combination is applied', the US Food and Drug Administration as 'the standard subject to which a treatment is applied and a measurement is made'. More precisely, it is the smallest amount of experimental material that can be randomly assigned to a treatment.

Both the animal (6) and the culture in *in vitro* studies (11) have been clearly identified as the experimental unit (Figure 1). In some protocols, cells may be scored from a number of slides and a summary statistic for the slide may be used in the analysis. It is possible that there may be appreciable variability between slides and this may need to be taken into account in the statistical analysis.

The individual cell may be the smallest unit which can be measured but cells from the same animal or culture are all assigned to the same treatment and repeated measures taken from the same experimental unit are likely to be autocorrelated or more similar to one another than two cells each taken from different samples. The degree of similarity is measured by the intra-class correlation (ICC). The ICC compares the within-group variance with the between-group variance and is calculated as $\sigma_w^2/\sigma_w^2 + \sigma_b^2$ from the estimates derived from the analysis of variance (ANOVA) table.

There is a need with *in vitro* designs to ensure that there is adequate replication with the culture having a similar 'role' as the experimental unit to the animal in the *in vivo* test (Figure 2). There may be differences between cultures, between sub-cultures within a culture and between cells within a subculture. Any analysis needs to take into account these different levels of variability otherwise 'hidden' levels of variability can distort the estimates of variability and lead to errors in interpretation.

An *in vitro* design where each of a series of subcultures receives a different treatment and the cells within the subculture are treated as the experimental unit in the analysis may lead to

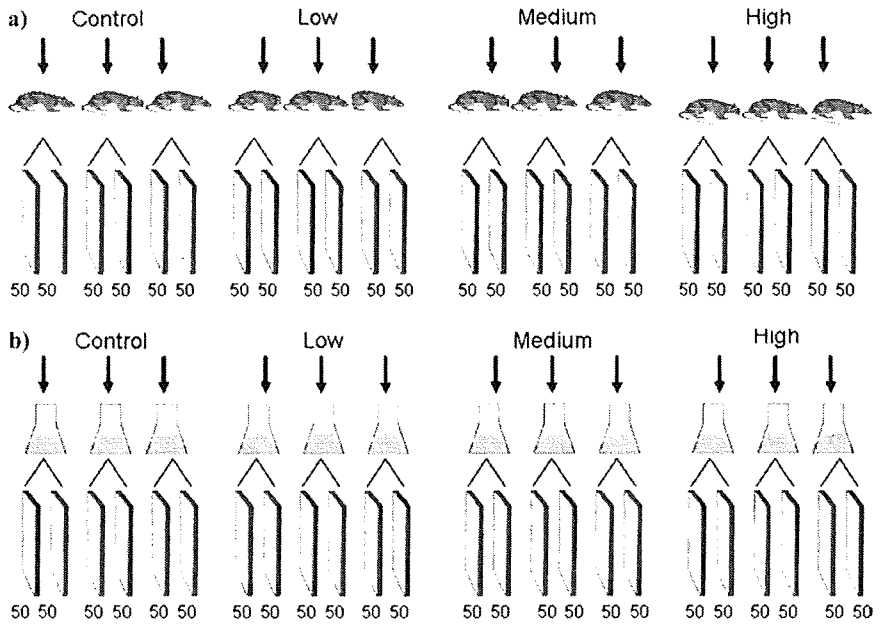

Fig. 1. Hierarchical/nested designs. Schematic diagram of hierarchical/nested designs of (**a**) *in vivo* comet assay and (**b**) *in vitro* comet assay showing experimental design with replicate animals or culture in each of four treatment groups, replicate slides from each experimental unit and 50 cells scored per slide (9). Wiklund and Agurell (27) recommended designs with 50 cells from three slides per culture and either four or five animals per group or two or three cultures per treatment group in *in vitro* studies.

significant but artifactual results such as apparent non-dose-related effects (Figure 2c). (A similar *ex vivo* design with single subcultures per treatment has the same problem.) These small but significant differences between subcultures probably represent their underlying variability rather than a true treatment effect. The more cells that are measured the more likely a significant difference will be detected. An experiment with replicate subcultures will provide a valid estimate of subculture variability and a valid, if low power, test of the treatments for that 'specific' culture (Figure 2b). For power calculations, replication of cultures is needed for an estimate of the variability across cultures.

Common to all, experimental studies should be the standard design features of randomization, replication and blocking together aimed at reducing biases and managing uncontrollable variables. Examples include randomization of the position of slides on platforms, the use of electrophoretic runs as blocks and the blind scoring of cells.

Within sample distributions

Many distributions of comet measures (e.g. % tail DNA) within a sample (intra-sample) from a culture or animal are not normally distributed but rather may be asymmetric, skewed, bi- or multi-modal, a mixture of different distributions or just idiosyncratic especially if an administered compound has caused some DNA damage. Some of the end points may include many small or zero values plus some extreme values. Finding any 'best' measure may be difficult if the distribution is not simple. In the case of normally distributed data, the 'central value' can be described by the mean and the spread by the 'standard deviation' (SD) but description of the distribution may be difficult if a number of parameters are needed to explain the distribution. In these cases, it is unlikely that a single statistical distribution will be able to describe the distribution and any single measure will capture only part of the information in the sample data and may be unrepresentative of any actual value.

Various 'statistics' have been suggested to represent the sample. For instance, the 90th and 95th percentiles have been suggested because they 'capture' the upper tail of the distribution. (However, for a precise estimate, this may require large numbers of cells—more than the 50 cells per slide often measured.) Values of the mean, median and 75th percentile are usually highly correlated.

Duez *et al.* (18) noted that the heterogeneity of the distribution curves of comet measures made the use of standard parametric and non-parametric methods difficult because assumptions underlying them were violated. They suggested using either the median or 75% percentile of the sample in subsequent analyses. They concluded that a trend analysis on medians of the samples was satisfactory. They also noted that non-parametric tests such as the Kruskal–Wallis and Mann–Whitney tests were oversensitive in detecting small differences between replicate samples and were not suitable for use in detecting genotoxic effects.

Data can be transformed to try to make the distribution of the data conform to normality. The logarithmic transformation has the convenient property that back transformation (taking antilogs) to biologically meaningful values is easier than with other transformations. The problem of the logarithm of zero values can be overcome by the addition of small positive values (such as 0.001) to the data. It is important to appreciate, though, that while a transformation may correct one violation, say normality, it can result in another, such as heterogeneity of variances. The logarithmic transformation is a special case from the family of Box–Cox power transformations. The optimum value for the power term in the transformation can be

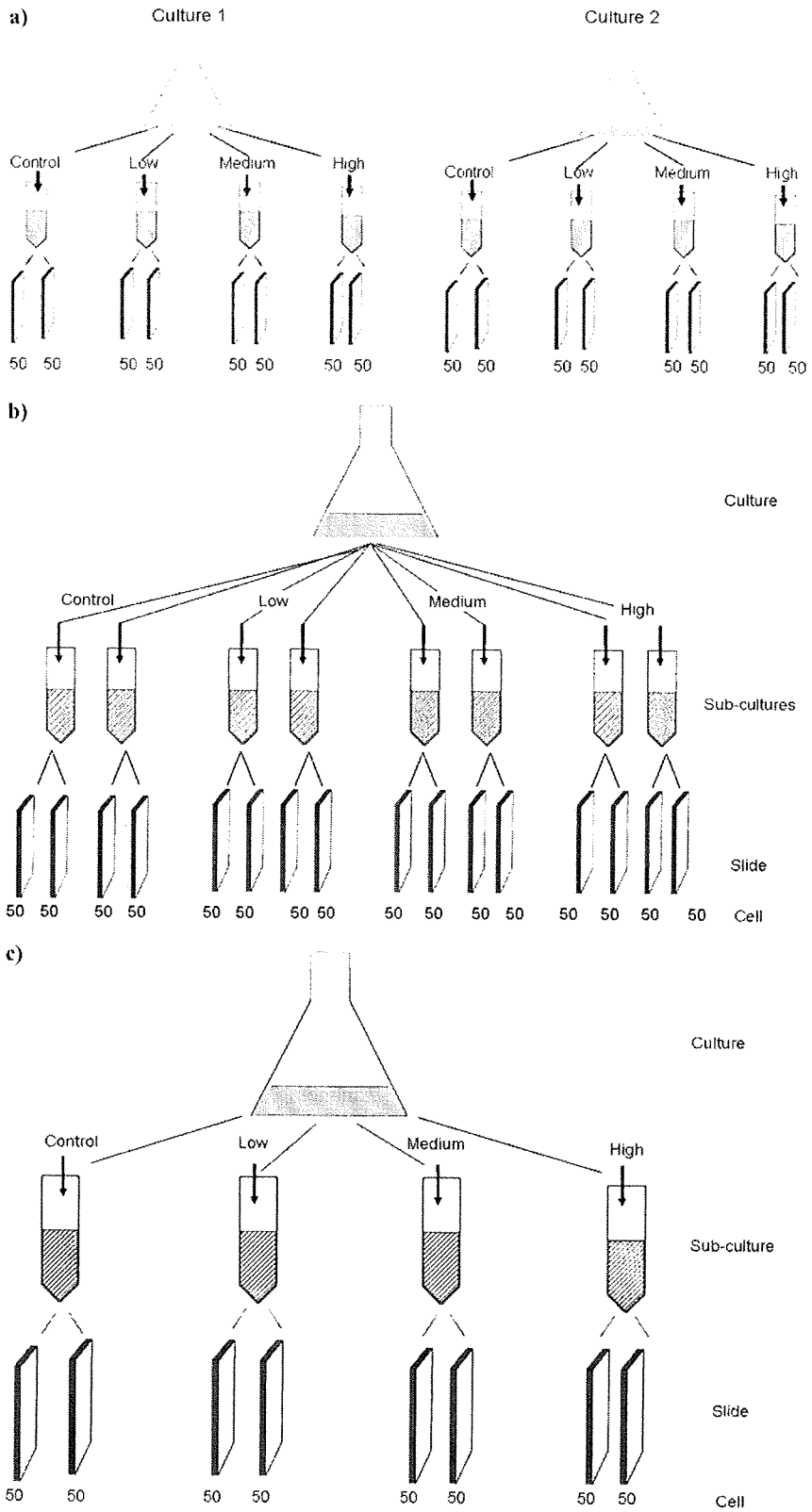

Fig. 2. *In vitro* comet assay designs. (**a**) Replicate cultures (cultures treated as blocks) providing a measure of inter-culture variability. (**b**) Replicate subculture providing estimates of variability for comparisons within a specific culture. (**c**) Design with no replication of subcultures. This is vulnerable to inter-subculture variability if the cells are treated as the experimental unit. The effect of any variability in subcultures is confounded with any treatment effects leading potentially to artifactual results. The effect is magnified as the number of cells scored increases.

derived from an analysis of the data sets using a range of values for the power term (19).

In the case of % tail DNA which is measured on a scale from 0 to 100%, the data may be suitable for transformation by a logistic or arcsin transformation. Collins *et al.* (20), for instance, suggested the use of either an angular or arcsin transformation or the use of generalized linear models with binomial error distribution and a logit link function.

A number of probability distributions have been proposed for modelling the distributions. These include the Weibull, exponential, logistic, normal, log-normal and log-logistic distributions (21). Others investigated include the Poisson, beta, gamma, Erlang and Weibull (22). Debon *et al.* (23) suggest the use of the sum of two Gaussian (normal) curves.

Ejchart and Sadlej-Sosnowska (22) found that the Weibull distribution was the best fit to data from an *in vitro* dose–response study. The Weibull is widely used in other fields, such as to model the failure of mechanical components, and can be characterized by two parameters relating to the shape (α) and scale (β) of the distribution which both increased with increasing dose. Ejchart and Sadlej-Sosnowska argued that changes in the values of these parameters would, therefore, be evidence for genotoxicity and suggested using simulations to derive CIs for the estimates (22).

Tice *et al.* (6) suggested the use of the H statistic as a measure of the migration patterns among cells within a sample. H, the coefficient of dispersion, is calculated as the variance/mean and is sometimes referred to as Fisher's coefficient of dispersion or the variance–mean ratio. In the case of quantitative measures, H is equivalent to the coefficient of variation (CV) times the SD. The larger the CV, the larger the value of H. One problem with H is that it can be susceptible to one or a small number of outliers. The coefficient of distribution is also used to see if data are distributed according to a Poisson distribution where $H = 1$. Values of $H > 1$ suggest over-dispersion of the data.

The fit of the data to a distribution has often been tested using a goodness of fit statistic such as the Kolmogorov–Smirnov test. However, problems can arise in the interpretation of a goodness of fit test as the null hypothesis that the data fit a normal distribution is likely to be rejected when the sample size is large simply because real data are unlikely to be a perfectly distributed. Duez *et al.* (18) found that another test for normality, the Shapiro–Wilks test, was very sensitive for detecting non-normality of untransformed and log-transformed tail length and tail moment when applied to samples of 100 cells. A similar issue arises with tests of the assumptions of equal variances. Care is, therefore, needed to avoid 'trawling' for a best distribution. The finding that data 'conform' to a particular distribution does not mean that this distribution is the 'correct' one. There should be some biological basis for the choice of a distribution.

In practice, concerns about the assumptions underlying the ANOVA methodology are, in decreasing order of concern: independence, equal (homogeneous) variances and normality (24). Of these, independence is by far the most important, while the ANOVA is robust enough to conduct when the within-group variances differ by a factor of two (some even say five) and where normality is a minor violation (24).

Complications arising from the complex distribution of comet end points may be in part avoided because of the implications of the central limit theorem. While the original distribution of a set of data may not be normally distributed, the distribution of the means of random samples from the distribution will be approximately normal, particularly as the sample size increases. Median values for a slide may, therefore, represent data which are amenable to standard statistical analyses.

Suggested approaches for statistical analysis

Currently, there is no consensus on standard statistical methods for the analysis of comet data (6). Ideally, if a series of related studies are planned then a standard method of statistical analysis should be used for the analysis and interpretation of the data. Different laboratories using different statistical methods may result in different interpretations of results particularly if the criteria for a positive result are based upon statistical significance using each laboratory's favoured method.

In practice, however, comet assay data should be basically straightforward to analyse with the exception that the measures of damage to the cells in a sample have a complex distribution especially if there has been an effect of the chemical. As discussed above, there is probably no simple statistical/mathematical distribution that would explain the observed distributions and this makes statistical analysis using the individual cell scores difficult. On the other hand, analyses concentrating on a single measure from each animal (the experimental unit) may provide robust results which, with care, can be interpreted satisfactorily. Duez *et al.* (18) has listed some of the standard methods available.

Statistical analysis can be carried out by using parametric approaches such as ANOVA techniques which reduce, in the special case of two groups, to one of a range of *t*-tests based upon the degree of variability in the two groups. ANOVA methods, part of the wider general linear model (GLM) approach, can be used to further explore the difference within a group of means by specific contrasts. Some contrasts have clearly defined hypotheses such as tests for linear and quadratic trends in a dose–response experiment. More sophisticated designs and analyses move from a traditional hypothesis testing approach into a modelling methodology where estimates are derived for various model components and attempts made to identify the best fitting and hopefully the most predictive model.

Non-parametric methods shadow the simpler parametric tests: the Mann–Whitney, the *t*-test; the Kruskal–Wallis, the one-way ANOVA and the Jonckheere–Terpstra trend test, the linear dose–response trend test. Non-parametric tests are slightly less powerful than their parametric equivalents but give potentially a more accurate Type I error rate when the assumptions underlying parametric tests are violated. (A Type I error is the risk of rejecting the null hypothesis in a statistical test when, in fact, it is true.) Importantly, the non-parametric tests may be distribution free but are not assumption free, so are probably as vulnerable, if not more so, to differences in the distributions between the groups. Non-parametric tests aim to ensure that the correct Type I error rates are maintained but are less suitable for more complex designs, estimation and model fitting. The distribution of the comet end point can create complications in finding an appropriate transformation of the data and the assumptions underlying parametric analyses may be challenged even if not violated. Small sample sizes (e.g. 4 or 5 units per group) also mean that comparisons using non-parametric tests may have low power even when there are quite large treatment effects.

Qualitative data (present/absent) can be analysed by chi-square and Fisher exact tests of 2 × 2 tables and chi-square

tests of difference between groups and trends which mirror ANOVA approaches but with appreciably less power. The choice of experimental unit for inclusion in the analysis is, however, a critical concern for the appropriate analysis and interpretation of such tests.

One-sided tests are directional and slightly increase the power of a statistical analysis. There is an argument that the statistical test should be two sided if results where either a decrease or an increase in DNA migration could be envisaged.

The ANOVA, especially the one-way ANOVA is an 'omnibus' test of an overall difference between means. A more 'targeted' hypothesis is of a linear or other specified dose response which also has more power. A linear effect may be found even when there is no significant difference between means as a consequence of the increased power of the specific hypothesis test compared with the general hypothesis. This should be borne in mind when the 'rule' that no further testing should be carried out between groups if the overall hypothesis test of a difference between means is not significant in the ANOVA is applied. Another argument against formalized analyses is the complications that arise if an experiment is only declared satisfactory if there is a significant difference between the positive and negative control groups but which does not take into account the complications that can arise from small sample sizes and unequal variances as a result of variable responses in the positive control group.

Multiple comparison approaches are sometimes used to address concerns that when a large number of comparisons (e.g. between pairs of treatments) are made, there is a risk of Type 1 errors (declaring results significant when they are not.) A common method often used in the analysis of comet assay (and other toxicological) data is Dunnett's test (25). This is a specialized multiple comparison test that allows a comparison of a single control group with all other groups. This test was specifically designed to adjust the error rate when multiple comparisons are made between a number of new treatments and the standard treatment group with the objective of avoiding wrongly replacing a satisfactory standard treatment with a new treatment which just happened to perform better by chance in a single particular study. Dunnett's test aims to keep the experiment-wise (or family-wise in contrast to the individual error rate) error rate at 0.05 which means that on average only 1 in 20 experiments will reach a false conclusion. The implication is that testing is done at a more conservative α value so in effect lowering the power of the design but without taking any account of any other structure in the design. A multiple comparison procedure in effect 'dampens' down the number of significant results reported. There are a number of different multiple comparison methods available, each addressing a different aspect of the comparisons across a range of treatments and with different properties. The Bonferroni correction, for instance, is a highly conservative approach which carries out hypothesis testing at the α/n level where n is the number of multiple comparisons being made. The use of multiple comparison methods, however, remains controversial, with some statisticians arguing against their indiscriminate use (26).

Hierarchical design, random effect models and generalized linear modelling

The comet assay is a hierarchical or nested design with animals (in the *in vivo* design) and cultures (in the *in vitro* design) within doses, a number of slides from each animal or culture and a number of cells measured per slide (Figure 1). The statistical models underlying these designs go under various names (hierarchical linear models, multilevel models, mixed-effects models, random-effects models, random coefficient regression models and covariance components models). The models make use of information on the various levels of variability in the design but are quite complex, need sophisticated software and can be complicated to interpret and explain. Their advantage is that they are able to provide estimates for the variability at each level in the design and make use of information at the cell level so increasing the power of the study somewhat. However, if there is appreciable between animal or culture variability, the extra power available may be small. There is also the added difficulty that the variability between cells within the same animal may have a complex distribution which may be difficult to include in the model.

Wiklund and Agurell (27) and Verde *et al.* (21) provide examples of more sophisticated statistical analyses where attempts have been made to model both the variability between samples and between cells within a sample based upon GLM approaches. The GLM is a generalization of the ordinary least squares approach (used in the ANOVA, analysis of covariance and multivariate ANOVA) and is a special case of the generalized linear model (GLZ). The generalized linear model is a unified method used to extend the GLM approach to incorporate responses other than those based upon the normal distribution. Nelder and Wedderburn (28) developed the concept of the GLZ which placed all the commonly used models, binomial, logit, probit and normal in a unified framework. Generalized linear modelling uses a link function which can be considered equivalent to the transformations applied in traditional analyses and provides the 'link' between the linear part of the model and the random part of the model.

The GLZ can be further generalized. Generalized linear mixed models (GLMM) are an extension of the GLZ with random effects and is also called a generalized linear mixed-effects model. Verde *et al.* (21) used it in their modelling approach to the analysis of comet data. Generalized estimating equations (GEE) are another extension of GLZ involving algorithmic adjustments used to model longitudinal or clustered data and to estimate regression coefficients. GEE use a 'working' correlation matrix as an approximation of the true within subject/unit correlation for each unit (29).

Wiklund and Agurell (27) provide concise recommendations regarding the design and statistical analysis of comet assay studies. They used simulations to identify the optimum number of cultures or animals, slides per culture or animal and cells per slide based upon data derived from studies performed in house. They investigated the performance of a number of standard statistical methods on a range of scenarios of *in vitro* and *in vivo* study results. The non-parametric tests investigated (the Kruskal–Wallis and Jonckheere–Terpstra trend test) were generally less efficient than the corresponding parametric tests. The use of parametric linear trend tests was recommended as they generally performed better that the corresponding overall tests for treatment differences especially when the dose–response pattern was monotonic. They noted that the 90th percentile is not affected by extreme outliers but focuses on the upper part of distribution. They recommended, however, using the mean of the log-transformed tail moment data and the 90th percentile of the log-transformed tail length as the end point in the analysis. They cautioned strongly against the use of the untransformed mean tail moment. They recommended designs

with 50 cells from three slides per culture and either four or five animals per group or two or three cultures per treatment group in *in vitro* studies. They suggested analysis based upon a GLM/ANOVA approach which included factors for treatment groups, experimental conditions such as electrophoresis runs and cultures or animals. Similar simulation approaches could be applied to the use of % tail DNA as the end point.

Verde *et al.* modelled tail moment data using GLMM (20). They recommend choosing from a set of distributions that give the best fit to the data. The approach is based upon the use of survival models and the distributions derived from the family of accelerated life models to develop a two-level hierarchical model of within and between individual tail moment measures. Treatment effects were assessed by the construction of 'probability over damage' graph which visualized the degree of damage produced by the treatment by plotting the probability of the damage to a cell being greater than a certain value.

A number of statistical software packages such as SPSS, Genstat and SAS as well as R (a public domain open source statistical analysis software language) have procedures for carrying out analyses using some of these models. For instance, the SAS procedures PROC GLM and MIXED can be used for the analysis of comet data. PROC GLM is the SAS procedure for analysing data using a GLM approach. PROC GLM is able to handle repeated measures by including various postulated correlation structures in the analysis. SAS PROC MIXED was developed for the analysis of designs where there is a mix of both random and fixed effects. It is based upon approaches developed by Wolfinger (30) and is more powerful but more complex to use than PROC GLM. The general linear mixed model in PROC MIXED directly models the covariance structure so dealing with problems that might arise from inefficient analyses and incorrect conclusions being drawn from ignoring the problems associated with the correlations between repeated measures.

Control groups

Two comparisons using control groups are relevant: comparisons between the concurrent positive and negative control groups and comparisons of the concurrent controls with historical control information. An important ethical issue relates to the purpose of the positive controls in *in vivo* studies and how big a group size is needed. As discussed earlier, formal statistical tests may fail to show statistical significance if the variability of response of the positive control group is large and sample sizes are small.

It is often an expectation, such as meeting a requirement for entry to a validation study, that a laboratory can show evidence of a successful record of carrying out the assay by providing historical control data. The compilation of such data sets can be developed as part of a formal quality control (QC) process using the range of statistical methods available (31). Assessment against these criteria could be useful both for the laboratory and the regulator. The development of Bayesian approaches to make the optimum use of historical control data is one potential development.

Hauschke *et al.* (32) argue for an approach which relates the classification of a result as positive or negative to the size of the response in the positive control group. This involves determining the maximum safe dose by incorporating a biologically meaningful threshold value (f) which is fraction of the difference between the positive control and vehicle control responses

Dose–response modelling

The number of doses to include in a comet assay remains an important consideration with the need to ensure non-linear dose–response curves (or downturns) can be detected and positive responses at multiple dose levels reinforcing the biological relevance of results (4). Specific contrasts based upon linear, quadratic and more complex dose–response relationships can be formally tested using the ANOVA approach. Experimental designs to investigate dose–response relationships should include an adequate number of doses over the region of interest and adequate replication and reproduction.

In the context of dose–response modelling, it is important to note that the identification of a dose as a no-observable effect level (NOEL) using a statistical test does not mean that a threshold exists or that effects do not occur below this level. The NOEL classifies a result into an effect/no-effect dichotomy. This may be wrongly interpreted as implying that the response is either non-linear or thresholded. The NOEL detectable in an experiment is a function of the statistical test applied to the data. The larger the experiment carried out, the smaller the difference between a negative control and a treated group that it is capable of detecting as statistically significant. In contrast a small, poorly designed study with appreciable variability is liable to fail to detect effects and thus provide estimates of NOEL above the level where effects should be detected. This is a well-known limitation of the NOEL methodology (33).

Design of experiment approaches

The comet assay continues to develop and now exists in a number of forms. Further extension of the methodology makes the assay a good candidate for systematic development using a design of experiment (DOE) methodology. This approach finds multiple factors (and interactions between them) that affect results appreciably and identifies the levels of these factors which optimize results while minimizing the number of experiments that need to be run and the material used (34,35). Such an approach is ideal for areas such as of protocol development where there are a number of factors that may affect results. DOE methodology builds on the work of R. A. Fisher on factorial designs in the 1920s in which Fisher demonstrated that DOE approaches (systematic and simultaneous variation of experimental conditions) is both economically and scientifically more efficient than the traditional one factor at a time approach. The DOE approach is now widely used in industrial settings such as the manufacturing and chemical industries to identify optimal conditions for processes to operate under. DOE methodology could, for instance, be an efficient approach to identifying optimum allocation or use of resources in *in vitro* studies where there are a number of different factors where conditions could be varied affecting cell preparation, assay running and visualization. Developments of high-throughput methods capable of investigating large numbers of samples for applications such as REACH screening are an area where the use of DOE methodology could be extremely effective. The use of DOE methods to optimize new or modify comet protocols such as the development of standard

protocols for validation studies could also be a productive approach.

Power and sample sizes

A key concept in the design of a study is a determination of the number of experimental units needed. A range of software packages, web-based resources, books and formulae are available for estimating sample sizes for a given power and vice versa. Power is defined as the probability of detecting an effect of a specified size if it is present and is related to the Type II or beta error associated with hypothesis testing. Most formulations represent very simple situations: comparison of two groups for differences in means or proportions. More complex hypotheses such as tests for specific dose–response relationships are more difficult as the power depends very much on the specific hypothesis being tested. Statistical packages such as nQuery Advisor have options for sample sizing for more complex designs and hypotheses. An alternative approach to the more complex problem is simulation and modelling of the design (36).

In the case of quantitative end points, four things are needed to determine sample sizes: the significance level the hypothesis will be tested at, the chosen power (conventionally 80 or 90%), the size of effect considered biologically important and some measure of the variability of the experimental units (e.g. the between-unit SD). Note that the sample size is for the number of experimental units. If the power for a specific sample size is required, then the sample size is entered instead. Ignoring strata in the design can lead to serious misinterpretation.

For qualitative end points, besides the alpha and beta levels, the control and treated proportions are needed to obtain sample sizes. The sample sizes needed are likely to be appreciably larger with qualitative compared with quantitative end points because of the lower information content of qualitative data.

The background level is important in determining the size of effect that can be detected by a design. This level affects how easy it is to detect absolute as opposed to relative changes. For instance, with a low background level, a small absolute difference may equate to a large-fold change while with a high background level a large absolute difference will equate to a smaller fold change. This becomes important, for instance, if the negative control group was to have little or no variability for a measure like % tail DNA.

The challenge in power and sample size calculation for the comet assay is to identify what size of effect can be considered biologically important and to have appropriate measures of the inter-experimental unit SD. One source of such measures is from previous studies or data from the literature. For example (37), an estimate of the inter-individual SD of % tail DNA of comets from buccal cells from seven healthy, young female non-smokers was 6.1% [taken from day 0; Figure 5 of reference (37)]. As another example, Frenzilli *et al.* (38) reported mean (SD) comet lengths in leukocytes of 16.5 (4.6) for 39 children from Pisa, 26.3 (9.6) for 16 healthy and 16.0 (4.3) in 27 tumour-affected children from Belarus.

It is important to remember than any power/sample size calculation is only an approximation and depends upon the assumptions, particularly of the inter-experimental unit (SD). The simple calculations are also based upon the assumption that a *t*-test will be appropriate for the analysis. If the treated animals are appreciably more variable, then the sample sizes can be underestimates depending upon the nature of the response. Alternatively, in some case where there is an appreciable difference in variability between the groups, a transformation may be appropriate and may result in the power being retained. Power calculations are possible with log-transformed data.

An alternative approach is based upon the work of Cohen (39). He developed the concept of expressing the size of differences based upon effect sizes in SD units calling 0.2 small, 0.5 medium and 0.8 large effects. A simple rule of thumb is that for a two-sided test of two group means at 80% power, that for every halving of the effect size in SD units the sample size in each group increases by ~4 (Table I).

Although the approach is useful and widely applied in circumstances where information of biologically important differences and variability is difficult to obtain, it is not without its critics (40).

Figure 3 shows the implications of reducing sample sizes below $n = 5$. Based upon standard methods, a two-sided test with $n = 5$ has 80 and 90% power to detect difference of 2.02 and 2.35 SD units, respectively, in a two-sample *t*-test. The comparable values for sample sizes of $n = 4$ are difference of

Table I. Table of sample sizes for various effect sizes

Effect size (SD units)	Power	
	80%	90%
0.125	1006	1346
0.25	253	338
0.5	64	86
1.0	17	23
2.0	6	7
4.0	3	3

Calculations derived from nQuery Advisor and based upon sample sizes for comparison between two groups with within-group SD of 1 unit. Two-sided test with $\alpha = 0.05$.

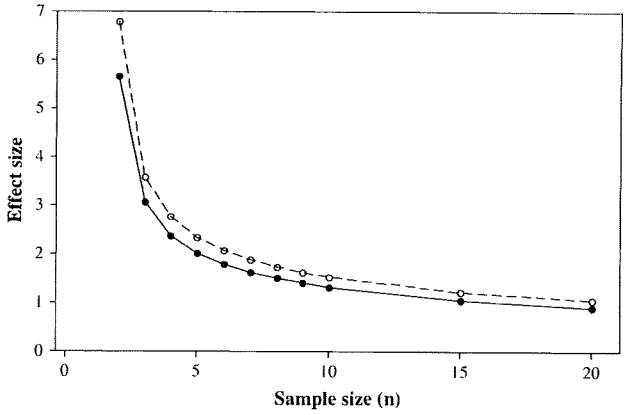

Fig. 3. Size of effect detectable with 80% (closed circles and continuous line) and 90% (open circles and dashed line) for a two-sample comparison in a two-sided test with $\alpha = 0.05$. Effect size measured in SD units. The figure shows the implications of reducing sample sizes below $n = 5$. Based upon standard methods a two-sided test with $n = 5$ has 80 and 90% power to detect difference of 2.02 and 2.35 SD units, respectively, in a two-sample *t*-test. The comparable values for sample sizes of $n = 4$ are difference of 2.38 and 2.77 SD units and the graph shows the appreciable information gain from an extra experimental unit when sample sizes are small.

2.38 and 2.77 SD units and the graph shows the appreciable information gain from an extra experimental unit when sample sizes are small.

Hierarchical designs of the comet assay share features in common with hierarchical or cluster randomized controlled trials. These are common in studies involving interventions in units such as general practices or schools where individuals within the same practice or school show similarities (41). In the case of hierarchical or cluster designs, the sample sizes depend upon both n, the number of clusters, and m, the number of individuals within a cluster. Sample sizes for cluster trials make use of the ICC coefficient. If the ICC is zero, then the observations are basically independent but as the ICC gets bigger the standard errors get larger. When the ICC is large, then large numbers of individuals will be needed in the clusters to obtain satisfactory power (42–44).

A failure to take clustering into account results in the standard errors of the estimates being underestimated and, potentially, leading to false conclusions. Even small ICC values can have big effects on the size of an estimate. An ICC of just 0.05 with cluster sizes (m) of 20 per group can lead to a 30% underestimation of the true precision of an estimate which leads to an increase risks of Type I error (29). The variance of the estimate increases by $(1 + (m - 1)\rho)$ (called the design effect) where m is the number per unit. Table II illustrates the effective sample sizes associated with different ICCs.

Criteria for a positive result

Guidelines explicitly refer to the identification for regulatory purposes of a positive or negative result to categorize the result as genotoxic or non-genotoxic and that an equivocal result may require further testing. The criteria needed for a definitive result are usually not explicitly defined. In some cases, there is reference to the need for a statistically significant result. However, statistical significance is not a measure of the size effect alone but depends upon a number of other factors especially the size of the experiment and the variability of the material. Different laboratories may also use their own statistical methods to define a positive.

The criteria for a positive result should, therefore, be defined. Given the limitations of basing this solely on statistical significance, the criteria should be related to the size of the difference, for instance, in the mean % tail DNA between a negative control and treated group. This approach has the advantage that the study can be explicitly designed to have sufficient power to detect differences large enough to be considered biologically important.

It should always be remembered that dichotomization of results into genotoxic or non-genotoxic leads to a loss of information with the consequence that some weak mutagens will be called negative and disagreements will occur when different criteria are used by different laboratories.

Table II. Effective sample sizes associated with different ICCs

No. of cells per unit (m)	No. of experimental units (n)	Intra-class correlation coefficient (ρ)	Design effect	Effective sample size
100	5	0.05	5.95	84.0
100	5	0.1	10.90	45.9
100	5	0.25	25.75	19.4
100	5	0.5	50.50	9.9

The effective sample size is = mn/DE. The design effect (DE) is $1 + (m - 1)\rho$ and is the ratio of the variance of the measure when the nested or clustered nature of the data is accounted for to the variation when it is not. It is also called the Variance Inflation Factor because it provides an estimate of how much an estimate based upon ignoring the clustering needs to be increased to allow for the clustering.

Replication and repeat experiments

Repeating experiments with adequate replication within them is usually considered good experimental practice. The terminology is not always standardized but a repeat experiment can be considered a separate experiment while replication occurs within an experiment. It is not always clear just how independent repeat experiments are or whether they should be based upon the same or a different design. A further consideration is whether the conditions for a repeat experiment should be decided on before or after the first study. Replicate samples should, in general, be prepared. Replicates can be considered as either biological and/or technical replicates. Biological replicates are samples taken from different independent experimental units such as subjects, animals or cultures. Technical replicates are repeat samples taken from the same animal or culture. They may be replicates from the same unit or from replicate samples from the same experimental unit. The need for biological replication should take precedence over the need for technical replication.

Pooled samples may sometimes be used. However, samples should not be pooled if information on individual experimental units is important and information is required on within-group variability in *in vivo* studies. Pooling samples may make an experiment technically easier because there are an adequate number of cells for analysis but this advantage is offset by the loss of any measure of inter-sample variability, the loss of statistical power and the potential for outliers with idiosyncratic responses masking the effects seen in the other units.

The OECD guidelines (45) note that equivocal results should be clarified by further testing preferably using a modification of experimental conditions. A follow-up study may fail to confirm initial results because of the 'regression to the mean' effect. This is because a study that generated a follow-up experiment may be at the upper end of possible results and the results of subsequent studies are likely to approach the true but smaller effect.

Validation studies

Validation is the process by which the reliability and accuracy of a procedure are established for a specific purpose (46). Reliability is specifically defined as a measure of the degree to which a test method can be performed reproducibly within and among laboratories over time. It is assessed by calculating intra- and inter-laboratory reproducibility and intra-laboratory repeatability. Accuracy is defined as the closeness of agreement between a test method result and an accepted reference value (46).

In an inter-comparison of the comet assay, there are three different levels of potential variability: between laboratories, between experiments carried out in the same laboratory and within an experiment. (This latter can be broken down further into variability between animals/cultures within the same dose level and between cell within the same animal/culture.) The

criteria for acceptable levels of variability are a scientific rather than a purely statistical issue. These criteria should be defined before the study begins.

The International Standards Organization and the American Society for Testing and Materials have developed guidelines for the investigation of repeatability and reproducibility of inter-laboratory comparisons (47,48). Repeatability is defined as the closeness of agreement under identical conditions in the same laboratory using the same conditions (equivalent to a 'best case') (measured by R and the within-laboratory consistency statistic k) and reproducibility the variability between laboratories using the same methods (equivalent to a 'realistic case') (measured by r and the between-laboratory consistency statistic h). Guidelines for conducting the analysis are provided (including issues such as inclusion or exclusion of potential outlier laboratories).

Qualitative agreement between laboratories should be expected for potent positive control chemicals. Determining acceptable levels of variability in a quantitative measure is different from whether a particular individual experiment is significant or not. The criteria need to be defined for assessing how much variability in results using reference chemicals is acceptable.

An alternative definition of accuracy is the proportion of correct outcomes of a test method (46). In validation studies, dichotomization (genotoxic/non-genotoxic) allows the calculation of diagnostic statistics such as sensitivity and specificity. Large sample sizes (of chemicals) are needed for precise estimates or small CIs. These CIs are usually derived by standard methods based upon the binomial distribution. The choice of the cut-off point for dichotomization can be investigated using receiver operator curves but any choice will be a trade-off because some misclassifications will occur so that sensitivity and specificity estimates will be less than one. The prevalence of the classes and criteria for how good the agreement or concordance needs to be to claim that the method is validated need to be predetermined. Scientific judgement is required on how big a sample of chemicals is needed for adequate precision (i.e. width of CIs) of the diagnostic statistics.

Observational and biomonitoring studies

The comet assay is used in biomonitoring and molecular epidemiological studies. Observational studies differ from experimental studies, in that there is, in effect, no choice of who is allocated to the control or exposed group. Individuals are observed unlike experimental studies where animals or cultures are randomly assigned to the treatments. In randomized trials, the design explicitly tries to ensure that the observed effects are not a consequence of some differences in the baseline characteristics of the groups. However, in non-randomized human biomonitoring studies such as case–control and cohort studies, the groups being compared are likely to differ with respect to a large number of potentially uncontrolled confounding factors such as sex, age, weight, diet, cigarette smoking, alcohol consumption, lifestyle and genetic polymorphisms. These factors may be unequally distributed between the control or reference group and the exposed groups. Bias in the selection of the groups is also a major risk for such studies.

Observational studies produce methodological problems but can, if carefully done, generate important results not easily otherwise obtained. Considerable care is needed, however, with such studies to ensure that the problems of confounding and bias do not influence results. Standard statistical methods such as ANOVA make assumptions about randomization which are unlikely to hold for observational studies.

If confounding cannot be avoided, identifying the causal relationships involving the factors becomes more difficult. There are several more or less complex statistical methods for dealing with confounding including techniques such as matching, stratification and regression. Modelling approaches such as multiple regression, logistic regression and Cox' s proportional hazard modelling are often used. It is important that the report of the study shows if and how adjustments for confounding have been done. However, Műllner *et al.* (49) reported that there is often inadequate reporting in papers of the statistical methods used to adjust for confounding factors. Recommendations for reporting such approaches are given by Campbell (29).

Collins (3) pointed to the need in human epidemiological studies to carry out power calculation to establish group sizes needed and that pilot studies might be needed to estimate intra- and inter-individual variability in the end point under investigation.

Analyses reporting correlations between variables are often reported although it is widely appreciated that an association identified by a significant correlation does not imply causation. In some cases, correlations are reported when a regression analysis would be more appropriate as there are clearly dependent and independent variables. Large sample sizes can result in small but statistically significant correlations but if many end points are measured, large numbers of correlations can be calculated with a serious risk of Type 1 errors: with 10 end points there are 45 possible correlations. Similar multiple comparison problem can arise when subgroup analyses are carried out particularly *post hoc* analyses. Consequently, considerable care should be exercised in the interpretation of significant correlation coefficients.

A large number of factors can affect the quality of biological samples before their analysis. Guidelines on sample collection and processing of samples should be followed to prevent these factors introducing systematic biases into data derived from the analyses of the samples (50).

Recommendations

There is nothing especially unusual about the statistical issues associated with the comet assay. Lovell *et al.* (9) provided a set of recommendations for the statistical analysis of comet data. Experimental design is the critical factor for a successful study. There is unlikely to be a single correct statistical analysis for all designs but there are a number of potentially wrong analyses. If the results of using different statistical approaches produce qualitatively different interpretations, then it is sensible to investigate the data set and identify the cause of the differences.

Identifying the experimental unit is crucial. The experimental unit is the unit to which treatments are randomized. In an *in vivo* study, this is the animal while in *in vitro* studies it is the culture. Statistical analyses which treat the cell rather than the animal or culture as the experimental unit can produce incorrect results with a risk of overestimating the statistical significance of a result.

A clearly defined end point for the comet should be used. The % tail DNA is a suitable end point for analysis and has the advantage of a defined scale from 0 to 100% which is comparable across studies. Statistical analysis of other end

points e.g. tail moment and tail length is possible although they are less directly comparable across studies. Data sets where interpretation of the statistical analyses would differ appreciably between different end points should be investigated to identify the cause of the divergence. Transformation such as the use of logarithms of the end points may be appropriate and, in the case of the % tail DNA, a logistic transformation may be appropriate.

Graphical presentation of results is useful but should not supersede formal statistical methods. Histograms should be presented of the individual data with 'bins' representing the number of cells falling in particular ranges. Care should be taken in comparisons of the shifts across histograms because the measures within the same sample will show autocorrelations so that apparent visual evidence for treatment effects must be critically examined.

Summary statistics for the distribution of cells within a sample can be used for statistical analysis. These could be the mean, the log mean and various percentiles such as the median, 75th and 90th percentile. The more cells measured per unit the more accurate the estimate of these statistics. However, sample sizes of 50 cells per slide are probably satisfactory as the central limit theorem begins to apply when the number of cells is >30. If appreciable variability exists between duplicate slides, then increasing the number of slides would be sensible. A summary statistic like the median value for the slide may be a suitable metric for the statistical analysis.

Statistical analysis could be carried out on the multiple end points collected in the comet assay to see if there are alternative combinations of the measures that could be used in the analysis of a study. Multivariate analysis (MVA) methods could be applied to the data collected on individual comet shape to see whether this could provide extra information for use in the interpretation of results. Similarly, MVA could be used to investigate cells from a sample to see if there is an optimal representative end point for the sample. Estimates of the ICC coefficients could be calculated to help in designing studies with the optimum number of observations at each level in the design.

Statistical tests to identify suitable distributions of the data have limited use because a significant lack of fit may be more a consequence of the sample size than the degree of departure from a distribution.

There should be increased emphasis on the estimate of the size of an effect and its CI rather than solely concentrating on the statistical significance level (*P*-value) determined in a specific experiment using a particular statistical test. The criteria for a positive effect should be based upon size of effect produced that would be considered biologically important. This should form the basis for power and sample size calculations for study designs. A retrospective analysis of data could be carried out to explore potential improvements to statistical analyses and to provide estimates for sample size/power calculations.

A suitable background incidence of % tail DNA should be identified for the negative control group in a study. The implications for the power of the study design of using different expected levels of % tail DNA in negative control samples should be explored.

Historical control data may help in the interpretation of results but should not preclude the need for concurrent control data to be collected in the study. The use of QC statistics should be considered for the monitoring and assessment of historical control data.

A range of parametric tests such as *t*-tests and ANOVA and their non-parametric equivalents are appropriate for analysing simple experiments. Tests of dose-related effects such as linear trends can be used and will have higher statistical power. Care should be taken using tests of proportions such as chi-square and Fisher exact tests because these tests assume independence of the data and can seriously overestimate significance levels if the cell is wrongly considered the experimental unit.

Statistical analyses based upon GLMs/ANOVA methodology provide a general approach to the analysis. The continuing development of more sophisticated methods making use of the hierarchical structure such as random effect modelling (e.g. GEE) may also be suitable approaches. Methods for the effective reporting and interpretation of these more sophisticated analyses will need to be developed. DOE approaches should be considered in the context of developing new or modified protocols for the comet assay.

Biomonitoring studies are observational rather than experimental studies. They are thus vulnerable to the problems of bias and confounding. Especial care is needed in the analysis and interpretation of such studies to avoid drawing incorrect conclusions.

In validation studies, acceptable levels of intra- and inter-laboratory variability should be defined to help assess the reliability of an assay. An adequate number of chemicals are needed to get precise estimates of the accuracy of the method.

Considerable care is needed in the design of *in vitro* studies to ensure that there is adequate replication of cultures. A failure to take into account hidden variability can result in the overestimation of effects such as the identification of artifactual non-dose-related effects as a consequence of differences being detected between subcultures rather than treatments.

Acknowledgements

Conflict of interest statement: None declared.

References

1. Östling, O. and Johanson, K. J. (1984) Microelectrophoretic study of radiation-induced DNA damage in individual mammalian cells. *Biochem. Biophys. Res. Commun.*, **123**, 291–298.
2. Brendler-Schwaab, S., Hartmann, A., Pfuhler, S. and Speit, G. (2005) The *in vivo* comet assay: use and status in genotoxicity testing. *Mutagenesis*, **20**, 245–254.
3. Collins, A. R. (2004) The comet assay for DNA damage and repair: principles, applications, and limitations. *Mol. Biotechnol.*, **26**, 249–261.
4. Burlinson, B., Tice, R. R., Speit, G. *et al.* (2006) Fourth International Workgroup on Genotoxicity Testing: result of the in vivo comet assay workgroup. *Mutat. Res.*, **627**, 31–35.
5. Singh, N. P., McCoy, M. T., Tice, R. R. and Schneider, E. L. (1988) A simple technique for quantitation of low levels of DNA damage in individual cells. *Exp. Cell Res.*, **75**, 184–191.
6. Tice, R. R., Agurell, E., Anderson, D. *et al.* (2000) Single cell gel/comet assay: guidelines for in vitro and in vivo genetic toxicology testing. *Environ. Mol. Mutagen.*, **35**, 206–221.
7. Merk, O. and Speit, G. (1999) Detection of crosslinks with the comet assay in relationship to genotoxicity and cytotoxicity. *Environ. Mol. Mutagen.*, **33**, 167–172.
8. Langie, S. A. S., Knaapen, A. M., Brauers, K. J. J., van Berlo, D., van Schooten, F.-J. and Godschalk, R. W. L. (2006) Development and validation of a modified comet assay to phenotypically assess nucleotide excision repair. *Mutagenesis*, **21**, 153–158.
9. Lovell, D. P., Thomas, G. and Dubow, R. (1999) Issues related to the experimental design and subsequent statistical analysis of in vivo and in vitro comet studies. *Teratog. Carcinog. Mutagen.*, **19**, 109–119.
10. Helma, C. and Uhl, M. (2000) A public domain image-analysis program for the single-cell-gel-electrophoresis (comet) assay. *Mutat. Res.*, **466**, 9–15.

11. Hartmann, A., Agurell, E., Beevers, C. *et al.* (2003) Recommendations for conducting the in vivo alkaline Comet assay. 4th International Comet Assay Workshop. *Mutagenesis*, **18**, 45–51.
12. Kumaravel, T. S. and Jha, A. N. (2006) Reliable comet assay measurements for detecting DNA damage induced by ionising radiation and chemicals. *Mutat. Res.*, **605**, 7–16.
13. Kent, C. R. H., Eady, J. J., Ross, G. M. and Steel, G. G. (1995) The comet moment as a measure of DNA damage in the Comet assay. *Int. J. Radiat.*, **67**, 660–665.
14. Hellman, B. H., Vaghef, H. and Bostorm, B. (1995) The concepts of tail moment and tail inertia in the single cell gel electrophoresis assay. *Mutat. Res.*, **336**, 123–131.
15. Bowden, R. D., Buckwalter, M. R., McBride, J. F., Johnson, D. A., Murray, B. K. and O'Neill, K. L. (2003) Tail profile: a more accurate system for analyzing DNA damage using the Comet assay. *Mutat. Res.*, **537**, 1–9.
16. Altman, D. G. and Royston, P. (2006) The cost of dichotomising continuous variables. *Br. Med. J.*, **332**, 1080.
17. Lee, R. F. and Steiner, S. (2003) Use of the single cell gel electrophoresis/comet assay for detecting DNA damage in aquatic (marine and freshwater) animals. *Mutat. Res.*, **544**, 43–64.
18. Duez, P., Dehon, G., Kumps, A. and Dubois, J. (2003) Statistics of the comet assay: a key to discriminate between genotoxic effects. *Mutagenesis*, **18**, 159–166.
19. Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *J. R. Stat. Soc. B*, **26**, 211–246.
20. Collins, A., Dusinská, M., Franklin, M. *et al.* (1997) Comet assay in human biomonitoring studies: reliability, validation, and applications. *Environ. Mol. Mutagen.*, **30**, 139–146.
21. Verde, P. E., Geracitano, L. A., Amado, L. L., Rosa, C. E., Bianchini, A. and Monserrat, J. M. (2006) Application of public-domain statistical analysis software for evaluation and comparison of comet assay data. *Mutat. Res.*, **604**, 71–82.
22. Ejchart, A. and Sadlej-Sosnowska, N. (2003) Statistical evaluation and comparison of comet assay results. *Mutat. Res.*, **534**, 85–92.
23. Debon, G., Bogaerts, P., Duez, P., Catoire, L. and Dubois, J. (2004) Curve fitting of combined comet intensity profiles: a new global concept to quantify DNA damage by the comet assay. *Chemomet. Intell. Lab. Syst.*, **73**, 235–243.
24. van Belle, G. (2002) *Statistical Rules of Thumb*. Wiley Series in Probability and Statistics, Hoboken, NJ.
25. Dunnett, C. W. (1955) A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.*, **50**, 1096–1121.
26. Finney, D. J. (1995) Thoughts suggested by a recent paper: questions on non-parametric analysis of quantitative data. *J. Toxicol. Sci.*, **20**, 165–170.
27. Wiklund, S. J. and Agurell, E. (2003) Aspects of design and statistical analysis in the Comet assay. *Mutagenesis*, **18**, 167–175.
28. Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Stat. Soc. A*, **135**, 370–384.
29. Campbell, M. J. (2001) *Statistics at Square Two: Understanding Modern Statistical Applications in Medicine*. BMJ Publishing Group, London.
30. Wolfinger, R. D. and Chang, M. (1998) *Comparing the SAS GLM and MIXED Procedures for Repeated Measures*. Cary, NC: SAS Institute Inc.
31. Ryan, T. P. (2000) *Statistical Methods for Quality Improvement*. 2nd edn. John Wiley and Sons, New York, NY.
32. Hauschke, D., Slacik-Erben, R., Hensen, S. and Kaufmann, R. (2005) Biostatistical assessment of mutagenicity studies by including the positive control. *Biom. J.*, **47**, 82–87.
33. Gaylor, D., Ryan, L., Krewski, D. and Zhu, Y. (1998) Procedures for calculating benchmark doses for health risk assessment. *Regul. Toxcol. Pharmacol.*, **28**, 150–164.
34. Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978) *Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building*. John Wiley and Sons, New York, NY.
35. Montgomery, D. C. (1997) *Design and Analysis of Experiments*. John Wiley and Sons Inc., New York, NY.
36. Eng, J. (2004) Sample size estimation: a glimpse beyond simple formulas. *Radiology*, **230**, 606–612.
37. Szeto, Y. T., Benzie, I. F. F., Collins, A. R., Choi, S. W., Cheng, C. Y., Yow, C. M. N. and Tsec, M. M. Y. (2005) A buccal cell model comet assay: development and evaluation for human biomonitoring and nutritional studies. *Mutat. Res.*, **578**, 371–381.
38. Frenzilli, G., Bosco, E., Antonelli, A., Panasiuk, G. and Barale, R. (2001) DNA damage evaluated by alkaline single cell gel electrophoresis (SCGE) in children of Chernobyl, 10 years after the disaster. *Mutat. Res.*, **491**, 139–149.
39. Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Academic Press, New York, NY.
40. Lenth, J. (2007) Statistical power calculations. *Anim. Sci.*, **85**, E24–E29.
41. Bland, J. M. and Kerry, S. M. (1997) Statistics notes. Trials randomised in clusters. *Br. Med. J.*, **315**, 600.
42. Kerry, S. M. and Bland, J. M. (1998) Statistics notes. Analysis of a trial randomised in clusters. *Br. Med. J.*, **316**, 54.
43. Kerry, S. M. and Bland, J. M. (1998) Statistics notes. Sample size in cluster randomisation. *Br. Med. J.*, **316**, 549.
44. Kerry, S. M. and Bland, J. M. (1998) Statistics notes. The intra-cluster correlation coefficient in cluster randomisation. *Br. Med. J.*, **316**, 1455.
45. OECD (1997) *OECD Guidelines for the Testing of Chemicals: Genotoxicity, Revised and New Guidelines, Adopted 1997*. Organization for Economic Cooperation and Development, Paris.
46. ICCVAM (2003) *ICCVAM Guidelines for the Nomination and Submission of New, Revised, and Alternative Test Methods (NIH 03-4508)*. Appendix C: glossary. http://iccvam.niehs.nih.gov/SuppDocs/SubGuidelines/SD_subg034508.pdf.
47. ISO (1986) *Precision of Test Methods—Determination of Repeatability and Reproducibility for a Standard Test by Inter-Laboratory Tests. ISO 5725*. International Standards Organization, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=11832.
48. ASTM (1999) *ASTM E691-99 Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method*. American Society for Testing and Materials.
49. Müllner, M., Matthews, H. and Altman, D. G. (2002) Reporting on statistical methods to adjust for confounding: a cross-sectional survey. *Ann. Intern. Med.*, **136**, 122–126.
50. Holland, N. T., Smith, M. T., Eskenazi, B. and Bastaki, M. (2003) Biological sample collection and processing for molecular epidemiological studies. *Mutat. Res.*, **543**, 217–234.

Received on October 19, 2007; revised on February 6, 2008; accepted on February 18, 2008

An Evaluation of Performance Standards and Non-radioactive Endpoints for the Local Lymph Node Assay

The Report and Recommendations of ECVAM Workshop 65[a]

David Basketter,[1] Amanda Cockshott,[2] Emanuela Corsini,[3] G. Frank Gerberick,[4] Kenji Idehara,[5] Ian Kimber,[6] Henk Van Loveren,[7] Joanna Matheson,[8] Annette Mehling,[9] Takashi Omori,[10] Costanza Rovida,[11] Takashi Sozu,[12] Masahiro Takeyoshi[13] and Silvia Casati[11]

[1]St John's Institute of Dermatology, St Thomas' Hospital, London, UK; [2]Health and Safety Executive, Bootle, UK; [3]Department of Pharmacological Sciences, University of Milan, Milan, Italy; [4]Procter & Gamble Company, Miami Valley Innovation Center, Cincinnati, OH, USA; [5]Daicel Chemical Industries Ltd, Himeji, Japan; [6]Faculty of Life Sciences, The University of Manchester, Manchester, UK; [7]National Institute of Public Health and the Environment, Bilthoven, The Netherlands; [8]Consumer Product Safety Commission, Bethesda, MD, USA; [9]Cognis GmbH, Dusseldorf, Germany; [10]Kyoto University School of Public Health, Kyoto, Japan; [11]ECVAM, IHCP, European Commission Joint Research Centre, Ispra, Italy; [12]Osaka University, Osaka, Japan; [13]CERI, Saitama, Japan

Preface

This is the report of the 65th of a series of workshops organised by the European Centre for the Validation of Alternative Methods (ECVAM).

The main objective of ECVAM, as defined in 1993 by its Scientific Advisory Committee (ESAC), is to promote the scientific and regulatory acceptance of alternative methods which have scientific relevance and which reduce, refine or replace the use of laboratory animals. One of the first priorities set by ECVAM was the implementation of procedures that would enable it to become well-informed about the state-of-the-art of non-animal test development and validation, and of opportunities for the possible incorporation of alternative methods into regulatory procedures. It was decided that this would be best achieved through a programme of ECVAM workshops, each addressing a specific topic, and at which selected groups of independent international experts would review the current status of various types of *in vitro* tests and their potential uses, and make recommendations about the best way forward.

A Workshop on *An Evaluation of Performance Standards and Non-radioactive Endpoints for the Local Lymph Node Assay* was held at ECVAM on 25–27 September 2007, under the chairmanship of David Basketter. The workshop was attended by experts from academia, industry, national organisations, and national and international validation authorities. At present, the local lymph node assay (LLNA) involves the use of radiolabelled thymidine as part of the standard protocol. The aim of the workshop was to review the status of methods which employ non-radioactive endpoints for the LLNA and to consider Performance Standards for their eventual assessment. At the end of the report are listed recommendations that should be considered for progressing toward the validation of relevant and reliable methods.

Key Definitions

To ensure the Performance Standards are applied appropriately, it is necessary to define their domain of applicability. For this purpose, the workshop participants debated in depth what could be considered to represent minor or major modifications to the standard LLNA. The following definitions were agreed:

Minor changes: those that maintain full compliance with OECD Test Guideline (TG) 429 (1), and that incorporate potential changes already foreseen in OECD TG 429. For a change to be considered minor, there is a requirement that the endpoint measured is still one of lymph node cell proliferation.

Address for correspondence: David Basketter, 2 Normans Road, Sharnbrook, Bedfordshire MK44 1PR, UK.
E-mail: *david.basketter@ukonline.co.uk*
Requests for reprints: Silvia Casati, ECVAM, IHCP, Joint Research Centre, European Commission, Via E Fermi, 2749, I-21027 Ispra (VA), Italy.
E-mail: *silvia.casati@jrc.it*

[a]*This document represents the agreed report of the participants as individual scientists.*

Major changes: those that incorporate modifications to the standard LLNA that are broader in scope and of greater substance than those defined as being *minor*. Changes of this type would normally trigger a more thorough validation exercise, but should be considered on a case-by-case basis.

Introduction

The regulatory background

The approach used for the identification of chemicals with a significant degree of skin sensitisation potential is well characterised in the EU, and will soon be within the Globally Harmonised System (GHS; 2). The relevant European legislation includes the Dangerous Substances Directive, *Directive 67/548/EEC* (3), and the Dangerous Preparations Directive, *Directive 1999/45/EC* (4). With the advent of the legislation related to the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) system (5), further emphasis has been placed on the use of the most up-to-date methods, as well as ensuring that decisions are made by using all the available data, and with the minimum of additional animal testing. However, for confirmatory testing, the LLNA is the method of choice within the REACH system.

The tests traditionally used for the identification of chemicals possessing the intrinsic ability to cause skin sensitisation are the guinea-pig maximisation test (GPMT; 6), the Buehler occluded patch test (7) and the LLNA (8). The first two of these use a combination of the induction and elicitation phases in the guinea-pig, with the extent of sensitisation induction being determined as a function of the (erythematous) response to topical challenge. In contrast, the LLNA quantifies the induction response in mice by measuring proliferation in the lymph nodes which drain the site of topical application. The capacity of these methods to identify skin sensitisation hazard has only been formally validated for the LLNA (9–12). However, both within this validation process and via the publication of other datasets, the guinea-pig methods are also recognised to be of sufficient sensitivity and specificity (13–15).

For the purposes of hazard identification, skin sensitisation assays are interpreted in the same manner. In simple terms, if the results in the LLNA are positive (i.e. the stimulation of proliferation in test group lymph nodes is at least 3-times greater than in the concurrent vehicle-only-treated controls), or if at challenge ≥ 30% of the guinea-pigs are positive in a maximisation test, or if ≥ 15% of the guinea-pigs are positive in the Buehler test, then the substance is regarded as a skin sensitiser. The substance can then be classified formally and labelled, according to the EU system, as "R43: May cause sensitisation by skin contact". Thus, labelling can be applied to a chemical substance exclusively on the basis of data from a single animal test. Human experience can only be taken into account if it exists, and even then, it is normally not used to overturn the conclusion from positive animal data (16).

Ultimately, basic hazard identification is not sufficient for protection of human health; it merely represents the first step. Risk assessment and risk management are the processes that deliver human health protection. To permit this, some experts have proposed that, ideally, the relative potencies of skin sensitising chemicals should be determined and considered. The measurement of skin sensitisation potency has been the subject of much discussion in recent years, and expert groups in the EU (17), in European industry (18) and in the World Health Organisation (19) have made closely similar recommendations. Essentially, they all recommended that the optimal strategy is to determine the threshold positive concentration, the EC3 value, in the LLNA. It would not be appropriate to go into any detail of this measurement here, as it has been thoroughly reviewed elsewhere (20). However, what is important, is to appreciate its value for characterising skin sensitisation hazard and facilitating risk assessment (21, 22).

Background to performance standards

Prior to the acceptance of a new test method for regulatory testing, validation studies are conducted to assess its predictive capacity (the ability of the test method to correctly predict or measure the biological effect of interest, also referred to as accuracy) and its reliability (the extent of its intra-laboratory and inter-laboratory reproducibility). The LLNA underwent such a formal assessment before being adopted for use at the regulatory level. However, there might be cases for which a comprehensive validation exercise could be avoided and a simplified procedure applied. General criteria have been established by the validation authorities, and accepted at international level, to identify these cases and provide guidance for their assessment.

The concept of Performance Standards has been introduced as a way to streamline the validation process for test methods that are functionally and structurally similar to existing and adequately-validated test methods (23–25). As defined by the OECD (26), the purpose of Performance Standards is to communicate the basis by which new test methods, both proprietary (i.e. copyrighted, trademarked, registered) and non-proprietary, can be determined to have sufficient accuracy and reliability for specific testing purposes. These Performance

Standards, based on validated and accepted test methods, can be used to evaluate the accuracy and reliability of other analogous test methods (also referred to as "me-too" tests) that are based on the same or similar scientific principles and that measure or predict the same biological or toxic effects.

Performance Standards should be provided by the Management Team of a validation study, and, as appropriate, used in the TGs issued for new test methods. The three main elements of Performance Standards are:

a) The essential structural, functional, and procedural elements of a validated test method that should be included in the protocol of a proposed mechanistically and functionally similar test method. These components include the unique characteristics of the test method, critical procedural details and quality control measures. Adherence to the essential test method components will help to ensure that a proposed test method is based on the same concepts as the corresponding validated test method.

b) A list of recommended reference chemicals that are used to assess the accuracy and reliability of a proposed mechanistically and functionally similar test method. These chemicals are a representative subset of those used to demonstrate the reliability and the accuracy of the validated method.

c) Accuracy and reliability values, which represent the comparable performance requisites that should be achieved by the proposed test method when evaluated by using the list of reference chemicals.

So far, Performance Standards have been developed only for *in vitro* methods, i.e. for *in vitro* skin corrosion testing (27, 28) and *in vitro* skin irritation testing (29). In both cases, they should be used to evaluate the performance of human skin models which are similar to those that have already been validated.

Background to the LLNA

The LLNA identifies chemicals that have skin sensitising potential (8; 30–32). The assay measures sensitising activity as a function of proliferative responses induced in auricular lymph nodes following the repeated topical exposure of mice to several concentrations of the test chemical. In the standard LLNA, the proliferation of draining lymph node cells (LNCs) is measured by using the incorporation of ^{3}H-thymidine and subsequent β scintillation counting. For this purpose, mice are injected intravenously (via the tail vein) with a source of ^{3}H-thymidine, five days after the initiation of exposure to the test chemical.

This approach to measuring the proliferative activity of LNCs was based on studies in which the sensitivity and specificity of various read-outs for lymph node activation and lymphocyte turnover were compared. Of the endpoints considered, the incorporation of radiolabelled thymidine was found to provide the most robust and most reliable correlation with skin sensitising potential.

Although the standard LLNA, which incorporates this approach for the determination of LNC activation and proliferation, has provided a useful and reliable method for identifying skin sensitising chemicals, it is acknowledged that it would be beneficial to have available a version of the LLNA that does not require the use of radioisotopes. For this reason, there has been interest in exploring other relevant read-outs for the assay, including alternative strategies for the measurement of LNC turnover.

Among the approaches that have been explored are:

a) the direct measurement of changes in draining lymph weight and/or cellularity (33, 34);

b) the measurement of other endpoints, such as induced changes in the concentration of ATP, that can serve as surrogates of altered lymph node cellularity;

c) the measurement of induced changes in the relative number of lymphocyte phenotypes found in draining lymph nodes, e.g. alterations in B lymphocytes ($B220^{+}$) number or in the representation of discrete T-lymphocyte sub-sets (CD62L/CD44; 35–37);

d) the characterisation of the elaboration by LNCs of cytokines, such as interleukin-2 (IL-2), a T-lymphocyte growth factor (38–40);

e) the use of non-radioactive methods for the determination of cell turnover in draining lymph nodes, e.g. the use of BrdU (41–43);

f) the use of radioisotypes other than ^{3}H-thymidine, such ^{125}I-uridine (44); and

g) the use of *in vitro*, rather than *in vivo*, radiolabelling of LNCs (45).

However, the challenge is to ensure that such alternative approaches, employing novel read-outs, have sensitivity, specificity, and overall accuracy and reliability, comparable to the standard LLNA. It is for this purpose, and for verifying the acceptability of the performance characteristics of modified assays, that the proposed Performance Standards described here were developed. However, it must be