NICEATM Draft Technical Summary

Reduction of the Number of Animals Required for Ocular Irritancy Testing: A Brief Review of the Literature

Introduction

In the interest of reducing the number of animals required for regulatory safety testing, regulatory authorities have revised testing procedures to reduce the minimum number of animals required for ocular and dermal corrosivity and irritancy testing by 50-83%. This reduction was accomplished by changing the requirement for the routine use of six animals to sequential testing of 1-3 animals. The revised procedures now allow for testing to stop and for a substance to be classified as an ocular/dermal corrosive or severe ocular irritant when the corresponding injury occurs in any one animal during the sequential testing.

With the objective of further reducing the number of animals used for ocular corrosion/irritancy testing, the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) and the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) are currently reviewing the validation status of *in vitro* methods proposed for identifying ocular corrosives and severe ocular irritants. In the course of this review, questions arose regarding the reproducibility of the *in vivo* ocular test. The decision by regulatory authorities to reduce the minimum number of animals from six to three (or less) was based on the fact that the animal test was considered sufficiently reproducible such that using fewer animals would not significantly alter the accuracy of the test method for hazard classification and labeling purposes. This draft technical summary briefly reviews the relevant available scientific literature on this topic.

History and Background

The original protocol of Draize et al. (1944) served as the basis for initial regulatory requirements for eye irritation testing that mandated the use of at least six rabbits. In

1981, the U.S. Environmental Protection Agency (EPA) published a report entitled " Eye Irritation Testing: An Assessment of Methods and Guidelines for Testing Materials for Eye Irritancy" (EPA, 1981). This report concluded that initial testing for eye irritation with three animals normally would be sufficient to identify substances that are non-irritating or maximally irritating. However the report noted that testing with additional animals might be necessary to reliably characterize substances of intermediate degrees of irritancy. In 1981, the Organization for Economic Co-operation and Development (OECD) published Test Guideline (TG) 405, which proposed the use of as few as three animals, but also included the provision that additional animals might be necessary in order to clarify equivocal responses.

Several analyses were subsequently published that assessed the consequences of reducing the number of rabbits per test from six to as few as two animals (Guillot et al. 1981; DeSousa et al. 1984; Solti and Freeman 1988; Talsma et al. 1988; Springer et al. 1993; Dalbey et al. 1993; Berdasco et al. 1996). With the exception of Dalbey et al. 1993, each study concluded that reducing the number of rabbits from six to three would not have an unacceptable reduction in the accuracy of ocular irritancy classification/categorization. These analyses were performed using maximum average Draize scores (MAS), internal irritancy classification schemes, and/or regulatory classification schemes as endpoints for comparison. Several of these studies (DeSousa et al. 1984; Talsma et al. 1988; Dalbey et al. 1993) confirmed that correlations between three-animal and six-animal classifications were the highest among substances classified on the extreme ends of the irritancy range (i.e., non-irritants and severe irritants), and that the majority of variability was seen among substances classified in the middle range of irritation. However, Dalbey et al. 1993 was the only study that concluded that this effect justified the continued routine use of six animals. EPA (EPA 1998), the European Union (EU 2004), and the OECD (in revised TG 405) now recommend the use of a maximum of three animals, although additional animals may be tested under certain circumstances (e.g., to confirm weak or moderate responses). The different evaluations on the numbers of rabbits appropriate per study are summarized chronologically in the following sections.

Due to the lack of individual rabbit data in each of these reports, it was not possible to calculate the impact of reduced animal numbers on hazard classification according to EPA (1996), EU (2001), or the United Nations (UN) Globally Harmonized System (GHS) of Classification and Labeling of Chemicals (UN 2003) classification systems. This shortcoming emphasizes the importance of reporting individual animal scores in publications.

Guillot et al. (1981) ***as discussed in EPA (1981)***

"Guillot et al (1981) compared the mean ocular irritation ratings in two groups of three rabbits each with those obtained using six rabbits. Classification differences due to test group size resulted for 25 of the 56 substances tested. In only two cases, however, was a test substance classified as non-irritating based on results in one group of three rabbits while testing with three additional rabbits and with six rabbits resulted in a rating of "slightly irritating". Thus, the data showed that the use of three animals in a preliminary test was adequate in differentiating a positive from a negative response for roughly 96% (54/56) of a wide variety of substances."

DeSousa et al. (1984)

These authors examined the statistical consequences of reducing the number of rabbits per test to five, four, three, or two animals. Data included in the analysis was obtained from three separate laboratories; one laboratory tested 55 chemicals, another tested 11 chemicals, and another tested one chemical, for a total of 67 substances. No substances were tested more than once. The substances spanned a wide range of chemical classes and the full range of irritancy potential (based on Draize scores). In their analysis, the authors used the maximum average Draize score (obtained at 1 hour or at 1, 2, 3, 7, 14, and 21 days post treatment) from in-house six-rabbit tests. From the 67 six-rabbit test results, scores for all possible subsample combinations of two to five rabbits were calculated. The original maximum average six-rabbit score was subtracted from each corresponding subsample score, and the difference plotted versus the original six-rabbit score to provide a measure of variability of the subsample scores. From these plots, subsample prediction intervals were calculated for the six-rabbit scores.

The authors found that prediction intervals for two, three, and four animals were comparable, although decreasing the sample size did increase the prediction interval. The authors also noted that the greatest discordance from the six-animal test occurred in the middle of the Draize scale (i.e., extremes at the high or low end of Draize scores produced very little difference in prediction intervals). An additional analysis was performed based on an in-house classification system (Texaco Single-Digit Toxicity Classification – SDTC) that classified the ocular irritancy of test substances according to their maximum average Draize score into five categories (**Table 1**).

Table 1. Texaco Single-Digit Toxicity Classification for Eye Irritation

Single-Digit Toxicity Classification (SDTC)	Explanation	Draize rabbit scores
0	Minimally irritating	0-15
1	Slightly irritating	>15-25
2	Moderately irritating	>25-50
3	Severely irritating	>50-80
4	Extremely irritating	>80-110

From DeSousa et al. (1984)

Overall, for the three-rabbit subsamples, there was 93% (1241/1340) agreement with the classification obtained when the maximum average score for the six-rabbit test was used. When the analysis was limited to the 10 substances classified as severely irritating (those with maximum average Draize scores from >50 to 80) and extremely irritating (those with maximum average Draize scores from >80 to 110), there was 84% (168/200) agreement. However, the authors recognized that the overall analysis could be biased by the limited number of substances inducing mid-ranged Draize scores, where the greatest variability would be expected to occur. Based on the results of their analysis, the authors concluded that a three-animal test system would be acceptable for most ocular irritant classification systems used in the petrochemical industry, with additional three-animal testing being performed in the case of tests with high variation in results.

Solti and Freeman (1988)

This poster presentation describes a subsample analysis of petrochemicals classified based on six-animal tests. Based on a scatterplot of eye irritation scores, it appears that the full range of irritancy responses were represented. From each of the original test groups (n = 6 animals) data for three animals was randomly selected for comparison to the original classification. A correlation coefficient of 0.975 was calculated from the resulting analysis of 30 studies when comparing the irritation scores among three animals versus six animals. Both the U.S. Occupation Safety and Heath Administration (OSHA) and the European Economic Community (EEC, now referred to as the European Union - EU) labeling schemes were used to evaluate the studies. Because the OSHA regulations did not define criteria for a three-animal test, the six-animal rules were extrapolated down to three animals (i.e., positive response in ≥ 2/6 animals = 1/3 animals). Accordingly, using a random subpopulation of three animals instead of six animals for labeling would have resulted in different OSHA labeling in 10% (3/30) of the studies (one overclassified [3% false positive rate], two underclassified [7% false negative rate]), while labeling would have been different in 7% (2/30) of the studies (both overclassified [7% false positive rate; 0% false negative rate]). The authors concluded that using a reduced number of animals in safety evaluations of petrochemicals would not greatly impact on hazard labeling decisions.

Talsma et al. (1988)

These authors performed a subsample analysis of 155 chemical and petroleum products that covered the full range of irritancy responses, in which they also evaluated the ability of irritation scores derived from two-, three-, four-, or five-animal tests to predict a six-animal Draize score. Similar to the approach used by DeSousa et al. (1984), the authors used in their analysis the maximum average Draize score (obtained at 1, 2, 3, 7, 14, and 21 days post treatment) from in-house six-rabbit tests. Also, similar to DeSousa et al. (1984), the authors applied scores from each subsample to an in house (Amoco) four-category classification system based on specific ranges of Draize scores (**Table 2**). For comparison, they also classified each substance according to the five-category SDTC system of DeSousa et al. (1984) (see **Table 1**).

Table 2. Amoco Eye Irritation Classification System

Rating	Explanation	Draize rabbit scores
0	Minimally irritating	0-15
1	Mildly irritating	16-30
2	Irritating	31-50
3	Extremely irritating/eye damage	>50-80

Modified from Talsma et al. (1988)

Talsma et al. found that, for a three-rabbit test, the correlation of randomly selected subset scores (i.e., maximum average score) with the six-rabbit Draize score was 0.99 and, overall, the number of correct classifications achieved was 94% (2912/3100) accurate in predicting the six-rabbit classification using the Amoco classification system, and 91% (2813/3100) using the SDTC classification system. When the analysis was limited to the 23 substances classified in the SDTC system as severely irritating (those with maximum average Draize scores from >50 to 80) and extremely irritating (those with maximum average Draize scores from >80 to 110), there was 89% (408/460) agreement.

Similar to DeSousa et al. (1984), they also found that the width of the prediction interval for a subsample was inversely proportional to the number of animals evaluated. Talsma et al. (1988) also pointed out that their database was weighted heavily toward minimally irritating substances, which may have affected the outcome of the analysis, given that they too noted that the greatest disagreement occurred among substances classified in the middle range of irritation. The authors concluded that their results indicate that a high level of accuracy can be obtained with reduced numbers of rabbits per test.

Springer et al. (1993)

This report this is part of the published proceedings of the 1991 IRAG Workshop on Updated Eye Irritation Methods. This report detailed an analysis of eye irritation tests on pesticides (n = 48, data submitted to the EPA), cosmetics/consumer products (n = 53; data submitted to the U.S. Food and Drug Administration [FDA] or the U.S. Consumer Products Safety Commission [CPSC]), cleaning products/ingredients (n = 30; data submitted to the FDA or the CPSC), and unspecified chemicals (n = 12; data from

Marzulli and Ruggles 1973). The substances spanned the full range of irritancy potential (based on FHSA and/or EPA classification systems). The only dataset from which substances were tested multiple times was the Marzulli and Ruggles (1973) database. These substances were all borderline compounds with respect to ocular irritancy (i.e., between nonirritating and irritating) which were tested in twelve different laboratories to yield a total of 139 tests.

Unlike DeSousa et al. (1984) and Talsma et al. (1988), this study did not perform a subsample analysis. Data from eye irritation tests from each group of substances were examined to estimate the distribution of positive animal responses for substances classified as irritant or nonirritant. An animal was classified as positive for eye irritancy if any score attained or exceeded the criterion for a positive response for corneal opacity ≥1, conjunctival redness ≥2 or conjunctival chemosis ≥2. There was no attempt to limit the evaluation to substances classified as ocular corrosives or severe irritants. Using data from the six-rabbit tests, probability calculations were performed based on a three-rabbit test (either a one-stage, three-rabbit test, or two-stage approach which sequentially tests up to three rabbits) to determine the likelihood of correctly identifying a substance as irritant or nonirritant. This analysis showed that a high level of accuracy (≥ 94%; actual numbers not available) could be obtained from a sample size of three rabbits in which two positive responses were required to assign an irritant classification (false positive rates ≤ 5% and false negative rates of 1%).

However, applying the EPA classification system (in effect at the time of this evaluation) to a three-animal test, where only one animal is required to assign an irritant classification, resulted in much higher false positive rates (20% to 50%). Based on this evaluation, the authors recommended revising the *in vivo* eye irritation protocol to include testing of only three animals using either a one-stage or two-stage approach.

Dalbey et al. (1993)

Similar to DeSousa et al. (1984) and Talsma et al. (1988), Dalbey et al. (1993) evaluated mean weighted Draize eye scores from subsets of two, three, four, or five rabbits and their predictivity of the score produced by a six-rabbit test. The database consisted of data from 185 six-rabbit eye irritation studies conducted in-house with petroleum-based products. No indication of substances being tested more than once is provided. The substances spanned the full range of irritancy potential (based on Draize scores). The authors used in their analysis average Draize score for cornea, iris, and conjunctiva calculated separately over three days after dosing, or mean weighted Draize scores calculated for cornea, iris, and conjunctiva combined. This study sought to confirm the earlier conclusions of DeSousa et al. (1984) and Talsma et al. (1988), that a three-rabbit test was suitable for classification of eye irritation and thus the classification resulting from each subset was compared to that resulting from a six-animal test. In this comparison, the European Commission (EC, now referred to as the EU] and U.S. Federal Hazardous Substances Act (FHSA) classification systems were considered. In addition, a "workshop classification" (based on recommendations from the 1991 Interagency Regulatory Alternatives Group [IRAG] workshop) was considered. Similar to previous studies, the agreement between subsets and the original six-animal Draize score was directly proportional to the number of animals. Dalbey et al. found that, overall for a three-rabbit test, there was approximately 90% agreement with the Draize scores produced by a six-rabbit test.

With regard to EC classification system, there was 96% (3158/3280) agreement for nonirritants, and 98% (412/420) agreement for irritants. However, it is noteworthy that only 11% (21/185) of the substances considered in this evaluation were classified as irritants. Upon classifying the same data according to the "workshop" classification system, only 41% (76/185) were labeled as nonirritants, along with 42% (77/185) severe irritants, and 17% (32/185) irritants. For the irritant category, there was only 29% (183/640) agreement between the three-animal and six-animal classification. For substances classified as severe irritants, there was 75% (1152/1540) agreement, while there was 100% (1520/1520) agreement for the nonirritants. Using the FHSA classification scheme, there was 88% (1340/1520) agreement for nonirritants, 97% (1488/1540) agreement for severe irritants, but only 55% (351/640) agreement for irritants.

Based on these results, and unlike DeSousa et al. (1984) and Talsma et al. (1988), Dalbey et al. (1993) concluded that the six-rabbit test should continue to be used (at least for the purposes of classifying substances according to the FHSA system), although a three-animal test could be used to screen for nonirritants or the most severe irritants, as these types of substances produced the greatest agreement. They emphasized the finding that the greatest variability was noted among the middle range of irritation, and only the extremes of the scoring scale were most accurate. This observation is consistent with previous evaluations that much of the variability lies within the mid-range

Berdasco et al. (1996)

These authors also performed a subset analysis of ocular irritation tests for 118 substances, by generating scores for five-, four, three, and two-rabbit subsets. The substances included in this analysis included pesticides, antimicrobials, consumer products and industrial chemicals. The substances spanned the full range of irritancy potential. Each substance was assigned an ocular irritancy category based on the EPA (1989) classification system using the six-animal test results, and then according to each subset result. The accuracy of the *in vivo* ocular irritation test using three rabbits instead of six was 96% (113/118) for Category I substances (EPA 1989), with a false negative rate of 10% (5/48) and a false positive rate of 0% (0/48). Based on these results, and similar to Dalbey et al. (1993), the authors concluded that as few as three animals could be used in an initial eye irritation test, with the provision that up to six rabbits might be necessary to clarify equivocal (or disparate) results.

References

Berdasco N, Gilbert K, Lacher J, and Mattsson J. 1996. Low rate of severe injury from dermal and ocular irritation tests and the validity of using fewer animals. Journal of the American College of Toxicology. 15:177-193.

Dalbey W, Rodriguez S, Wilkins K, Cope C. 1993. Reducing the number of rabbits in eye and skin irritancy tests. Journal of the American College of Toxicology 12:347-357.

DeSousa D, Rouse A, Smolon W. 1984. Statistical consequences of reducing the number of rabbits utilized in eye irritation testing: Data on 67 petrochemicals. Toxicology and Applied Pharmacology 76:234-242.

EPA. 1981. Eye Irritation Testing: An Assessment of Methods and Guidelines for Testing Materials for Eye Irritancy. EPA 560/11-82-001. Office of Pesticides and Toxic Substances, EPA, Washington, D.C.

EPA. 1989. Pesticide Assessment Guidelines, Subdivision F, Hazard Evaluation: Human and Domestic Animals, Series 83-4, Primary Eye Irritation. U.S. Government Printing Office, Washington, D.C.

EPA. 1998. Health Effects Test Guideline, OPPTS 870.2400 Acute Eye Irritation. EPA 712-C-98-195. Washington, DC: U.S. Environmental Protection Agency.

EU. 2004. Commission Directive 2004/73/EC of 29 April 2004 adapting to technical progress for the 29th time Council Directive 67/548/EEC on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances. Official Journal of the European Union L152, 1-316.

Guillot JP, Caillard L, Gonnet JF, Clement C. 1981. Chemicals-Ocular and Cutaneous Local Tolerance-“Cosmetic,” A.F.N.O.R. and OECD Protocols. Institut Francais de Rechserches et Essais Biologiques. Centre de Lyon, Les Oncins. b.p. 109, 69210 L’Arbresle.

Marzulli F, Ruggles D. 1973. Rabbit eye irritation test: collaborative study. Journal of the Association of Analytical Chemists 56:905-914.

OECD. 2002. Guideline for testing of chemicals revised guideline 405: Acute Eye Irritation/Corrosion. Available: http://www.oecd.org. [accessed 26 August 2004].

Solti J, Freeman JJ. 1988. Effect of reducing the number of animals in acute toxicity/irritation tests on U.S. and European labeling requirements. The Toxicologist 8:263.

Springer J, Chambers W, Green S, Gupta K, Hill R, Hurley P, Lambert L, Lee C, Lee J, Liu P, Lowther D, Roberts C, Seabaugh V, Wilcox N. 1993. Number of animals for sequential testing. Food and Chemical Toxicology 31:105-109.

Talsma D, Leach C, Hatoum N, Gibbons R, Roger J-C, Garvin P. 1988. Reducing the number of rabbits in the Draize eye irritancy test: A statistical analysis of 155 studies conducted over 6 years. Fundamental and Applied Toxicology 10:146-153.

UN. 2003. Globally Harmonised System of Classification and Labeling of Chemicals (GHS). New York & Geneva: United Nations Publications.