EXECUTIVE SUMMARY

This Background Review Document (BRD) reviews available data and information regarding the validation status of the Hen's Egg Test – Chorioallantoic Membrane (HET-CAM) test method for identifying ocular corrosives and severe irritants. The test method was reviewed for its ability to predict ocular corrosives and severe/irreversible effects as defined by the U.S. Environmental Protection Agency (EPA) (EPA 1996), the European Union (EU) (EU 2001), and the United Nations (UN) Globally Harmonized System (GHS) of Classification and Labeling of Chemicals (UN 2003). The objectives of this BRD is to describe the current validation status of the HET-CAM test method, including what is known about its accuracy and reliability, the scope of the substances tested, and the availability of a standardized test method protocol.

The information summarized in this BRD is based on publications obtained from the peer-reviewed literature, as well as unpublished information submitted to the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) in response to two *Federal Register* (*FR*) Notices requesting high quality *in vivo* rabbit eye test and *in vitro* ocular irritation data for HET-CAM, the Isolated Chicken Eye (ICE), the Isolated Rabbit Eye (IRE), and the Bovine Corneal Opacity and Permeability (BCOP) test methods. An online literature search identified 214 publications that contained HET-CAM test method results and protocol information; of these publications, detailed *in vivo* and *in vitro* data were available for 12 studies[1] that allowed for an evaluation of test method accuracy[2] and reliability[3].

Other published and unpublished HET-CAM test method studies are reviewed in **Section 9.0** (Other Scientific Reports and Reviews). This section discusses HET-CAM studies that could not be included in the performance analyses, because of the lack of appropriate study details test method results and/or the lack of appropriate *in vivo* rabbit eye reference data.

The HET-CAM test method uses the chorioallantoic membrane (CAM), which is a vascular fetal membrane composed of the fused chorion and allantois. The method is proposed to provide information on the effects that may occur in the conjunctiva following exposure to a test substance. Published reviews note that chicken-embryo models have long been used as models by embryotoxicologists and virologists. (Parish 1985; Luepke and Kemper 1986). Extending the use of chicken embryos, the HET-CAM test method was proposed by Luepke (1985) and Luepke and Kemper (1986). It was assumed that acute effects induced by a test substance on the small blood vessels and proteins of this soft tissue membrane are similar to

[1] Sufficient information was available for 10 of these publications to assess test method accuracy when compared to the GHS (UN 2003), EPA (1996), and EU (2001) classification systems. For two publications, sufficient information was only available to assess test method accuracy when compared to the EU (2001) classification system.

[2] (a) The closeness of agreement between a test method result and an accepted reference value. (b) The proportion of correct outcomes of a test method. It is a measure of test method performance and one aspect of "relevance". The term is often used interchangeably with "concordance."

[3] 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 intralaboratory repeatability.

effects induced by the same test substance in the eye of a treated rabbit. The CAM has been proposed as a model for a living membrane (such as the conjunctiva) since it comprises a functional vasculature. Additionally, evaluation of coagulation (i.e., protein denaturation) may reflect corneal damage that may be produced by the test substance. The CAM is evaluated for the development of irritant endpoints (hyperemia, hemorrhage, and coagulation). Depending on the method used to collect data on the endpoints (time to development, severity of observed effect) qualitative assessments of the irritation potential of test substances are made.

U.S. Federal regulatory agencies were surveyed to determine whether HET-CAM test method data have been considered for regulatory use where submission of testing data is required. Responses indicated that such data have not been provided to surveyed regulatory agencies.

The HET-CAM test method is currently used by some companies for the identification of ocular corrosives and severe irritants in a tiered testing strategy on a case-by-case basis. In this strategy, positive *in vitro* test results are considered in a weight-of-evidence decision as to whether to classify the substance as an ocular corrosive or severe irritant. Negative results and suspected false positive *in vitro* results proceed to standard *in vivo* testing or to validated *in vitro* test methods that are capable of detecting false negative corrosives and severe irritants.

The HET-CAM test method protocols used in the various studies considered in this BRD are similar, but not identical. Examples of some of the test method components that differed among the HET-CAM protocols used to generate data include:

- relative humidity during egg incubation ranged from 52.5 to 62.5%,
- volume or quantity of the test substance applied to the CAM (when reported) was either 0.1 or 0.3 mL for liquids and 0.3 g for solids,
- number of replicate eggs per test substance ranged from 3 to 6, and
- some studies included concurrent positive control substances, while others did not.

In addition to the various test method protocol permutations in the published literature, there were several HET-CAM analysis methods utilized to assess acute eye irritation. The analysis methods that are described in the literature include: Irritation Score (defined as IS(A) and IS(B)), Q-Score, S-Score, mtc value, and the IS and ITC method. All of these analysis methods are reviewed and evaluated in the BRD. Furthermore, the data available allowed for additional assessments based on the concentration tested *in vivo* and *in vitro*.

A total of 260 substances and formulations were evaluated in the studies. A variety of chemical and product classes have been tested in the HET-CAM assay. The chemical classes with the greatest number of substances tested are alcohols, carboxylic acids, and organic salts. For some of the test substances that were identified as formulations, components of the formulation and the relative concentrations of the components were available. The most common product classes tested are solvent, shampoo, surfactants, and cosmetics.

Some of the published *in vivo* rabbit eye test data on the substances used to evaluate the accuracy of HET-CAM for detecting ocular corrosives and severe irritants was limited to average score data or the reported irritancy classification based on a laboratory specific classification scheme. However, detailed *in vivo* data, consisting of cornea, iris and conjunctiva scores for each animal at 24, 48, and 72 hours and/or assessment of the presence or absence of lesions at 7, 14, and 21 days was necessary to calculate the appropriate EPA (1996), EU (2001), and GHS (UN 2003) ocular irritancy hazard classification. Thus, a portion of the test substances for which there was only limited *in vivo* data could not be used for evaluating test method accuracy and reliability as described in this BRD.

None of the studies provided original test result data. However, summary *in vitro* data was available for all of the test substances evaluated such that they could be assigned *in vitro* irritancy classifications for comparison to the available *in vivo* reference data.

The accuracy evaluation of the HET-CAM test method was limited to the substances evaluated in 10 to 12 *in vitro-in vivo* comparative studies. The ability of the HET-CAM test method to correctly identify ocular corrosives and severe irritants, as defined by the EPA (1996), the EU (2001), and the GHS (UN 2003) was evaluated using two approaches. In the first approach, the accuracy of HET-CAM was assessed separately for each *in vitro-in vivo* comparative study. In the second approach, the accuracy of HET-CAM was assessed after pooling data across *in vitro-in vivo* comparative studies that used the same method of data collection and analysis. While there were some differences in results among the three hazard classification systems evaluated (i.e., EPA [EPA 1996], EU [EU 2001], and GHS [UN 2003]), the accuracy analysis revealed that HET-CAM test method performance was comparable among the three hazard classification systems (see **Table ES-1**).

Table ES-1 Ranges of Performance Statistics for Evaluated Analysis Methods for GHS, EPA, and EU Classification Systems

Analysis Methods	Accuracy	Sensitivity	Specificity	False Positive Rates	False Negative Rates
IS(A)-10	48-50%	24-25%	100%	0%	75-76%
IS(A)-100	85%	100%	83%	17%	0%
IS(B)-10	65-68%	68-70%	64-67%	33-36%	30-32%
IS(B)-100	51-57%	87-93%	40-47%	52-59%	6-13%
Q-Score	61-64%	100%	43-46%	54-57%	0%
S-Score	44-50%	36-44%	60-67%	33-40%	56-64%

Abbreviations: EPA = U.S. Environmental Protection Agency, EU = European Union, GHS = Globally Harmonized System.
A single value indicates the same percentage results for all three hazard classification systems.

Most of the substances evaluated by the IS(A)-10 and IS(A)-100 analysis methods were formulations. For the IS(A)-10 analysis method, which evaluated mostly surfactant-based formulations, the false negative rates ranged from 75% to 76%, while the false positive rate

was 0% for all classification systems. Comparatively, the IS(A)-100 analysis method, which evaluated primarily oil-water formulations, had a higher false positive rate than false negative rate.

With regard to physical form of the substances tested by these analysis methods, a majority was tested as liquids/solutions *in vitro* and *in vivo*. Therefore, the false negative and false positive rates for these analysis methods were consistent or the same as to the overall false positive and false negative rates. No solids were evaluated using the IS(A)-10 analysis method, while the false negative and false positive rates were 0% for the IS(A)-100 analysis method. For the GHS classification scheme, the evaluation indicated that substances were more likely to be underpredicted if (a) the *in vivo* lesion was based on persistence of effect and (b) if the *in vitro* test concentration was 100%.

The chemical class of substances that was consistently overpredicted according to the GHS classification system (i.e., were false positives) by the IS(B)-10 and IS(B)-100 analysis methods is alcohols (89% to 90% for the IS(B)-10 analysis method and 79% to 88% for the IS(B)-100 analysis method). Additional chemical classes that were overpredicted by both analysis methods were ethers, organic salts, and heterocyclic compounds. Formulations appeared to have the lowest false positive rates for both analysis methods (0% for IS(B)-10 and 23% to 26% for IS(B)-100). The chemical classes that were underpredicted by both the IS(B)-10 and IS(B)-100 analysis methods were amines. Generally, the false negative and false positive rates for the same chemical class were higher for the IS(B)-100 analysis method when compared to the IS(B)-10 analysis method.

With regard to physical form of the substances overpredicted by the IS(B)-10 analysis method, the false positive and false negative rates were 19% and 37% to 38% (7/18), respectively for liquids and 58% to 65% and 0% to 13% for solids. For the IS(B)-100 analysis method, the false positive and false negative rates were 61% to 65% and 0%, respectively for liquids and 48% to 67% and 8% to 24% for solids. The physical form of many of the tested substances was unknown based on the available information.

Information regarding the pH of test substances was available for a subset of the substances tested (29 to 35 substances). Overall, substances were observed to have a higher false positive rate when (a) tested at a 100% concentration (IS(B)-100) and (b) had a pH greater than 7.0. For the GHS classification scheme, the evaluation indicated that substances were more likely to be underpredicted if (a) the *in vivo* lesion was based on persistence of effect and, (b) if the *in vitro* test concentration was 10%.

The accuracy analysis indicated that alcohols and esters are often overpredicted (43 to 50% and 43% false positive rate, depending on the classification system used) in the Q-score analysis method. The numbers of substances among the remaining chemical classes were too few to resolve any definitive trends in overprediction by the Q-Score analysis method. The false negative rate for all chemical classes with a sufficient number of substances ($n \geq 5$) was 0%.

With regard to physical form of the substances overpredicted by the Q-Score analysis method, 14 to 17 were liquids and none were solids. The ranges of false positive and false negative rates for liquids were 56% to 61% and 0%, respectively. The false positive and false negative rates for solids were 0% for both parameters. There was insufficient information for the other evaluated categories (e.g., surfactant-based formulations) to conduct an analysis.

Due to the limited database for the S-Score analysis method, a chemical class evaluation could only be conducted for carboxylic acids/carboxylic acid salts for the GHS classification system. For this chemical class and classification system, the false negative rate was 75% (3/4) and the false positive rate was 0% (0/1).

With regard to physical form of the substances overpredicted by the S-Score analysis method, 14 to 16 were solids. There were no liquids evaluated with analysis method. The false negative rates for solids ranged from 56%-64% (5/9 to 7/11) and the false positive rates ranged from 33% to 40% (2/6 to 2/5). There was insufficient information for the other evaluated categories (e.g., surfactant-based formulations) to conduct an analysis.

The analysis of intralaboratory repeatability was evaluated using data from two different publications (Gilleron et al. 1996, 1997) for the IS(B) analysis method. In both studies, the hemorrhage endpoint had a high %CV value (104 to 117). Additionally, the %CV values for the coagulation endpoint were the lowest of the three endpoints evaluated in the HET-CAM test method. However, the actual values were quite disparate between the two studies (e.g., Gilleron et al. 1996 coagulation %CV = 95.69; Gilleron et al. 1997 coagulation %CV = 41.78). The difference in the numbers may be due to several factors including test substances evaluated and differences in the test method protocols used between the two studies. The calculated variability for the endpoints and the overall test method may be exaggerated because of the relatively small values that are obtained from each of the endpoints (5 for hemorrhage, 7 for lysis, and 9 for coagulation). Similar results were obtained from the analysis of intralaboratory reproducibility. The overall irritation score was generally reproducible (%CV values of 53 and 17.5 for the two studies evaluated).

A qualitative assessment of the data provided for multiple laboratories in three to four studies indicates the extent of interlaboratory reproducibility. Given the relatively homogeneous performance of the HET-CAM test method among the three classification systems, the discussions for the individual studies and analysis methods encompasses all three hazard classification systems, unless otherwise indicated. The two to four participating laboratories that used the Q-Score analysis method were in 100% agreement in regard to the ocular irritancy classification for 21 (45%) of the 47 substances analyzed. Comparatively, participating laboratories were in 100% agreement for 12 to 13 (66% to 68%) of the 18 to 19 substances analyzed using the S-Score analysis method, depending on the classification system used. For the IS(B)-10 analysis method, the participating laboratories were in 100% agreement for 84 to 85 (79% to 81%) of 104 to 107 substances evaluated. For the IS(B)-100 analysis method, the participating laboratories in Spielmann et al. (1996) were in 100% agreement for 80 to 81 (82% to 84%) of the 95 to 99 substances evaluated. There was 100% agreement in regard to the GHS ocular irritancy classification for 11 (64% to 69%) of the 16

to 17 substances evaluated in five laboratories using the IS(A) analysis method in Hagino et al. (1999).

The overall reliability statistics, arranged by HET-CAM data analysis method, for the IS(B), IS(B)-10, S-Score and Q-Score are identical to what was discussed previously. For the IS(A) and IS(B)-100 analysis methods, additional data laboratory data was available for a subset of the substances tested for each analysis method. For both of these analysis methods, the addition of the results from additional testing laboratories yielded a concordance pattern consistent with what was observed for Hagino et al. (199) and Spielmann et al, (1996).

A quantitative evaluation of interlaboratory reproducibility was conducted for four studies (CEC 1991; Balls et al. 1995; Spielmann et al. 1996; Hagino et al. 1999) by performing a %CV analysis of *in vitro* scores obtained for substances tested in multiple laboratories. For CEC (1991), two different evaluations were conducted based on the concentration tested *in vitro*. For 14 substances evaluated at 100% concentration, the mean and median %CV values were 31.86 and 33.04, respectively. For 12 substances evaluated at 10% concentration, the mean and median %CV values were 34.6 and 33.1, respectively. For the Balls et al. (1995) study, the average and median %CV values for substances evaluated with the Q-Score were 49.83 and 42.50, respectively. The average and median %CV values for the substances evaluated with the S-Score were 84.42 and 71.90, respectively. For the substances evaluated in Spielmann et al. (1996), the average and median %CV values for substances tested at 10% concentration were 60.17 and 42.65, respectively. For substances tested at 100% concentration in Spielmann et al. (1996), the average and median %CV values were lower: 35.21 and 26.22, respectively. When substances that were tested in three different testing laboratories were removed from the assessment, little change was seen in the mean and median %CV values for both concentrations tested. For Hagino et al. (1999), the average and median %CV for substances classified as GHS Category 1 (UN 2003) were 24.4 and 27.0, respectively. The average and median %CV for substances classified as EPA Category I (EPA [1996]) were 23.86 and 26.0, respectively.

As stated above, this BRD provides a comprehensive summary of the current validation status of the HET-CAM test method, including what is known about its reliability and accuracy, and the scope of the substances tested. Raw and transformed data for the HET-CAM test method will be maintained for future use, so that these performance statistics may be updated as additional information becomes available.