EXECUTIVE SUMMARY

This Background Review Document (BRD) reviews available data and information regarding the validation status of the Isolated Chicken Eye (ICE)[1] 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 objective of this BRD is to describe the current validation status of the ICE 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The information summarized in this BRD is based on publications obtained from the peer-reviewed literature, as well as unpublished information submitted to the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) in response to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* notices requesting high quality *in vivo* rabbit eye test and *in vitro* ocular irritation data for ICE, the Isolated Rabbit Eye (IRE), the Hen's Egg Test – Chorioallantoic Membrane (HET-CAM), and the Bovine Corneal Opacity and Permeability (BCOP) test methods. An online literature search identified three publications that contained relevant ICE test results for an evaluation of test method accuracy[2] and reliability[3]. Submitted unpublished ICE data and detailed *in vivo* data for two additional studies allowed for an evaluation of test method accuracy[2] and reliability[3] for a total of five studies.

Other published and unpublished ICE test method studies are reviewed in **Section 9.0** (Other Scientific Reports and Reviews).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

The ICE test method is an organotypic model that provides short-term maintenance of normal physiological and biochemical function of the chicken eye in an isolated system. In this test method, damage by the test substance is assessed by determination of corneal swelling, opacity, and fluorescein retention. Each measurement is either converted into a quantitative score used to calculate an overall Irritation Index, or assigned a qualitative categorization that is used to assign an *in vitro* irritancy classification. Either of these outcomes can then be used to predict the *in vivo* ocular irritation potential of a test substance. A histopathological assessment can also be included on a case-by-case basis to discriminate

---

[1] In order to maintain consistency among the isolated eye methods, ICE is used throughout the BRD as opposed to CEET (Chicken Enucleated Eye Test), which is used by the test method developer.

[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.

borderline cases (i.e., substances that produce results that preclude assignment to a single category).

The ICE test method has not yet been considered by U.S. Federal agencies for regulatory use where submission of testing data is required. However, some companies have found the ICE test method useful 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 *in vitro* test methods that are capable of detecting false negative corrosives and severe irritants.

The ICE test method protocols used in the various studies considered in this BRD are similar, but not identical. The essential principles of the test method include the enucleation of eyes from chickens obtained from a slaughterhouse, mounting in a specially-designed apparatus and testing for damage that may have occurred during the isolation process, treating the eyes with a test substance, collecting corneal thickness, opacity and permeability data, and evaluating the data in relation to a prediction model. The primary difference among these protocols was the number of treated eyes per test substance. Acceptable ranges for negative control responses, historical data used to establish these ranges, and procedures to determine the optimum quantity of test substance to be applied have not been published.

A total of 175 substances in five studies can be used to evaluate ICE test method accuracy, 85 of which were proprietary compounds, consisting largely of products or formulations. The ICE test method has been used to test a variety of chemical and product classes. The chemical classes tested included, but were not limited to, alcohols, acids, hydrocarbons, surfactants, inorganic chemicals, acyl halides, alkalis, solvents, esters, heterocyclics, ketones, onium compounds, and organophosphates. The proprietary compounds tested included, but were not limited to, detergents, pesticides, silicone powder, ink, toilet cleaners, and thermal paper coatings.

Some of the published *in vivo* rabbit eye test data on the substances used to evaluate the accuracy of ICE for detecting ocular corrosives and severe irritants was limited to average score data or the reported irritancy classification. 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, some of the test substances for which there was only limited *in vivo* data could not be used for evaluating test method accuracy and reliability.

Three of the studies received contained original study records. 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 ability of the ICE 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 ICE was assessed separately for each *in vitro-in vivo* comparative study. In the second approach, the accuracy of ICE 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 the ICE test method performance was comparable among the three hazard classification systems. The overall accuracy of the ICE test method ranged from 83% to 87%, depending on the classification system used. Sensitivity and specificity ranged from 50% to 59% and 92% to 94%, respectively. The false positive rate ranged from 6% to 8%, while the false negative rate ranged from 41% to 50%.

According to the accuracy analysis, the chemical class with the highest false positive rate in all three classification systems was alcohols, with false positive rates ranging from 27% to 50%. The chemical class with the next highest false positive rate in all three classification systems was esters, with false positive rates ranging from 11% to 13%. No other chemical classes were consistently overpredicted by all three systems, although for most of the chemical classes tested, the number of substances in each was too few to resolve any definitive overprediction trends by the ICE test method. Alcohols were also consistently underpredicted, with false negative rates ranging from 33% to 50%. Other underpredicted chemical classes were amines/amidines (33% to 50%; GHS and EPA systems only), carboxylic acids (17% to 43%), heterocyclics (33% to 40%), inorganics (50%; EU system only), onium compounds (33% to 40%) and polyethers (100%; EU system only).

Regarding the physical form of overpredicted substances, no solids were overpredicted in any classification system, while liquids showed false positive rates ranging from 7% to 10%. Both solids and liquids were underpredicted, however, showing false negative rates ranging from 46% to 70% for solids and 39% to 44% for liquids.

Changes in the ICE test method performance statistics for substances classified according to the GHS classification system were observed when three discordant classes (alcohols, surfactants, and solids) were excluded from the data set; accuracy increased from 83% (120/144) to 92% (69/75), the false negative rate decreased from 50% (15/30) to 29% (2/7) and the false positive rate decreased from 8% (9/114) to 6% (4/68).

Test substances labeled as pesticides were not overpredicted in any classification system, but showed false negative rates ranging from 40% to 60%. Test substances labeled as surfactants were also not overpredicted, but showed false negative rates ranging from 44% to 57%.

Regarding the pH of underpredicted substances for which such information was available, substances with a pH less than 7.00 showed false negative rates of 27% to 40% (3/11 to 4/10) and substances with a pH greater than 7.0 showed false negative rates of 50% to 57% (3/6 to 4/7). However, it is noted that pH information was available for only a portion of the 27 to 32 severe irritant substances (i.e., Category 1, Category I, or R41) for each classification system in the database.

Finally, with respect to the GHS classification system only, as evidenced by an analysis of NICEATM-defined GHS Category 1 sub-groupings, the eight underpredicted substances were more likely to be classified *in vivo* based on persistent lesions (false negative rate of 60% [3/5]), rather than on severe lesions (false negative rate of 28% [5/18]).

A quantitative assessment of intralaboratory data from one study (Prinsen 2000), using scores for each endpoint (i.e., corneal thickness/swelling, corneal opacity, fluorescein retention) and the ICE Irritation Index, indicates the extent of intralaboratory reproducibility of the ICE test method. Four test substances were used in this study. When considering the results of this analysis, note that some test substances had a mean or a standard deviation equal to zero for some endpoints and that scores for corneal opacity and fluorescein retention have a small dynamic range (0 to 4 and 0 to 3, respectively). Corneal thickness measurements within experiments showed %CV values ranging from 0.9 to 6.1 and corneal opacity scores showed %CV values ranging from zero to 86.6 (the highest value was obtained for a nonirritating substance). The %CV values for fluorescein retention were zero for three of the four substances and ranged from zero to 86.6 for the nonirritating substance, although this range is based on only two experiments. Finally, the %CV values for the ICE Irritation Index for the four substances ranged from -86.6 to 41.6, with the same nonirritating substance exhibiting the outlying values (-86.6 and 41.6).

The data from Prinsen (2000) was also used to do a CV analysis on between-experiment values for each endpoint (i.e., corneal thickness/swelling, corneal opacity, fluorescein retention) along with the ICE Irritation Index, for each test substance. When considering the results of this analysis, note that scores for corneal opacity or fluorescein retention have a small dynamic range (0 to 4 and 0 to 3, respectively).

The %CV values for the corneal thickness measurement ranged from 1.8 to 6.3 and those for corneal swelling ranged from 13.9 to 138.7. The %CV values for the corneal opacity score ranged from 8.7 to 95.8. The %CV values for the fluorescein retention score ranged from zero to 141.4. Finally, the %CV values for the ICE Irritation Index ranged from 4.1 to 91.8. Note that for all endpoints considered except corneal thickness, the highest %CV values were obtained for the nonirritating substance.

A qualitative assessment of the data provided for multiple laboratories in one study (Balls et al. 1995) provides an indication of the extent of interlaboratory reproducibility. In an assessment of interlaboratory reproducibility of hazard classification (EPA, EU, or GHS), the four participating laboratories were in 100% agreement in regard to the ocular irritancy classification for 44 to 45 (75% to 76%) of the 59 substances tested *in vitro* in the study, depending on the classification system used. All four laboratories were in 100% agreement on the classification of 60% to 70% of substances classified as corrosives/severe irritants, 85% to 88% of substances classified as nonsevere irritants/nonirritants.

Among the 15 substances classified according to the GHS scheme that exhibited interlaboratory differences in *in vitro* classification, four were classified as alcohols. Two of the 15 substances were classified as cationic surfactants, two were classified as acetates/esters, and two were classified as ketones. Solvents was the product class appearing

most frequently among these substances, with seven of the 15 substances represented. Other product classes represented by multiple substances were chemical intermediates (five substances) and synthetic flavor ingredients (four substances). In regard to physical properties, of the 15 substances with discordant results among the four laboratories, 10 were liquid (seven water soluble) and five were solid (four water insoluble).

Mean endpoint values (i.e., fluorescein retention, corneal opacity, corneal swelling) and the ICE Irritation Index for each substance were provided for each of the four laboratories participating in the study. To provide a quantitative assessment of interlaboratory variability, individual laboratory ICE test results were used to calculate a mean, standard deviation, and the %CV for corneal opacity, fluorescein retention, corneal swelling, and the Irritation Index for each substance tested. Mean and median %CV values for all 59 substances were calculated to provide an assessment of overall variability. Traditionally, mean/median %CV values of less than 35% have been considered satisfactory for biologically-based test methods (Fentem et al. 1998). For ICE, a wide range of %CV values for individual substances is evident for all endpoints. The mean/median %CV values were 39%/36% (ranging from 0 to 159%) for fluorescein retention, 47/37% for corneal opacity (ranging from 0 to 159%), 77%/75% for corneal swelling (ranging from 31 to159%), and 35%/32% (ranging from 10 to 98%) for the Irritation Index. When only severe irritants (GHS Category 1, based on *in vivo* data) are considered, the %CV values are lower for all endpoints, with corneal swelling (mean of 72%, median of 69%) the sole endpoint with a mean/median %CV value greater than 35%. Of the four liquid substances with a CV < 35% for corneal swelling (2,2-dimethylbutanoic acid, 2,6-dichlorobenzoyl chloride, benzalkonium chloride 5%, and cetylpyridinium bromide 10%), two were water insoluble. No solid substances had a CV < 35% for corneal swelling. It is noteworthy that some of the corneal swelling values reported in the data are greater than 80% and therefore above the reported historical maximum range of 60-80%. However, different depth measuring devices may have been used by the participating laboratories to determine corneal thickness, which, unless normalized, would have contributed to the increased variability and/or the excessive values calculated for this evaluation (Prinsen M, personal communication).

Common physicochemical characteristics do not appear among the substances showing the most variable responses (defined as CV >70% for any of the endpoints). Of the 37 substances with significant variability in at least one endpoint, 18 are solids (of a total of 19 solids, 12 of which are water soluble) and 19 are liquids (of a total of 40 liquids, 14 of which are water soluble). However, some chemical classes appear to predominate among the 37 substances with CV values greater than 70%, including seven surfactants (of 12 tested), five heterocyclic compounds (of six tested), four acetate/esters (of six tested), and four acids (of six tested). Therefore, the majority of substances tested from these chemical classes exhibited increased interlaboratory variability.

Balls et al. (1995) also determined the interlaboratory correlation between ICE test method endpoint data generated by each laboratory for all substances tested, as well as for subsets of test substances (water-soluble, water-insoluble, surfactants, solids, solutions, and liquids). Interlaboratory correlation coefficients generally spanned a range of 0.6 to 0.9 depending on the specific subsets of substances being evaluated. However, the range of correlation

coefficients for some endpoints was larger (e.g., correlation coefficients for ICE-Mean Swelling ranged from 0.210 to 0.757 when testing substances that are insoluble in water).

Review of the mean *in vitro* data from this study indicates that wide ranges of corneal swelling values were recorded for the five insoluble test substances that were classified as ocular corrosives/severe irritants. For all five substances, the same laboratory produced the highest values, with mean corneal swelling percentages ranging from 1.5 to 6 times greater than the next highest mean corneal swelling value for the same substance tested by the other three laboratories. In addition, of the 14 remaining ocular corrosives/severe irritants (soluble and surfactant combined), a considerably higher value was reported for corneal swelling by the same laboratory for 12 substances. This trend was also apparent for nonsevere irritants/nonirritants.

Although the interlaboratory variability for fluorescein retention or corneal opacity was not as pronounced for the insoluble ocular corrosives/severe irritants, and could not be associated with a single laboratory, the ranges of correlation coefficients for these endpoints are also relatively high. Therefore, the apparently large interlaboratory variability noted among these substances cannot be attributed to a single laboratory or to a single endpoint.

At least one eye is traditionally included in each ICE study as a negative/vehicle control (isotonic saline). Individual eye data that could be used to perform a CV analysis on between-experiment values for each of the test method endpoints (i.e., corneal thickness/swelling, corneal opacity, fluorescein retention) along with the ICE Irritation Index for each test substance were obtained from negative control eyes. This analysis revealed that responses in the negative control eye remain relatively consistent.

Concurrent positive control substances have not been employed in the ICE test method, and therefore, an evaluation of historical positive control data is not possible.

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