EXECUTIVE SUMMARY

This Background Review Document (BRD) reviews available data and information regarding the validation status of the Isolated Rabbit Eye[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 Labelling of Chemicals (UN 2003). The objective of this BRD is to describe the current validation status of the IRE 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The information summarized in this BRD is based on publications obtained from the peer-reviewed literature, as well as unpublished information submitted to the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) in response to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 data and *in vitro* ocular irritation data for the IRE, Isolated Chicken Eye (ICE), Bovine Corneal Opacity and Permeability (BCOP), and the Hen's Egg Test – Chorioallantoic Membrane (HET-CAM) test methods. An online literature search identified 14 publications that contained IRE test method results and protocol information; of these publications, detailed *in vivo* data were obtained for four studies. Submitted IRE and detailed *in vivo* data for these four studies allowed for an evaluation of test method accuracy[2] and reliability[3].

Other published and unpublished IRE test method studies are reviewed in **Section 9.0** (Other Scientific Reports and Reviews). This section discusses IRE 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 IRE test method was developed by Burton et al. (1981) and proposed as a preliminary *in vitro* screen for the assessment of severe eye irritants. This organotypic test method is also referred to as the Rabbit Enucleated Eye Test (REET) (e.g., Guerriero et al. 2004). The principal advantage of the IRE test is that it eliminates the use of live animals for ocular irritancy testing and thus the pain and suffering potentially associated with the *in vivo* Draize rabbit eye test. Another advantage of the IRE test method is that it typically uses eyes isolated from euthanized rabbits used for other research purposes or from animals sacrificed commercially as a food source. In the IRE, liquid or solid substances are placed directly on the corneal surface of isolated rabbit eyes, which are held and maintained in a temperature-

[1] Exposure of the isolated rabbit eye to irritants can produce corneal opacity, corneal swelling, an increase in permeability to sodium fluorescein dye, and observable effects on the corneal epithelium. These endpoints can be quantified and used to evaluate the potential eye irritation of substances.

[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.

controlled chamber. After a 10-second exposure, followed by rinsing, the treated eye may be evaluated for corneal opacity, corneal swelling, fluorescein penetration, and effects on the corneal epithelium at various times over a four-hour observation period. Various decision criteria based on use of one to four ocular endpoints have been employed to identify ocular irritants of varying levels of severity depending on the nature of the test substances used (e.g., surfactant-based or chemically diverse) (Burton et al. 1981; Chamberlain et al. 1997; Cooper et al. 2001; Jones et al. 2001; Gettings et al. 1996). However, Guerriero et al. (2004) provided decision criteria (prediction model) using all four of these endpoints to specifically identify chemically diverse test substances as ocular corrosives or severe irritants.

The IRE 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 IRE test method useful for the identification of ocular corrosives and severe irritants in a tiered testing strategy on a case-by-case basis. 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 IRE test method protocols used in the various studies considered in this BRD are similar, but not identical. The essential principles of the test method protocol include procurement and enucleation of the eyes, a preliminary assessment of the integrity of the isolated eye (especially the corneal surface), equilibration of the eyes in a physiological environment, application of the test substance, incubation, temporal quantization of corneal damage using various endpoints (i.e., corneal opacity score, corneal swelling calculation, fluorescein penetration score, and evaluation of epithelial integrity), evaluation of data in relation to a prediction model, and assignment of an irritancy level based on graded responses (e.g., nonirritant, mild, moderate, or severe irritant) or as all or none responses (e.g., either a nonsevere irritant or a corrosive/severe irritant). However, given the various uses and applications of the IRE test method by different investigators and laboratories, and the evolution of the test method over time, a number of laboratory-specific differences have been noted regarding the conduct of the test method. Variations in the publicly available IRE protocols include evaluation of one to four endpoints, different prediction models or *in vitro* classification systems, and differences in the number of controls, among other methodological variations.

Some of the published *in vivo* rabbit eye test data on the substances used to evaluate the accuracy of IRE for detecting ocular corrosives and severe irritants was limited to average score data or a 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 were necessary to calculate the appropriate EPA (1996), EU (2001), and GHS (UN 2003) ocular irritancy hazard classifications. 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 as described in this BRD.

Only a few of the reports provided original *in vitro* test result data. However, summary *in vitro* data were 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.

A total of 149 substances and formulations obtained from four studies that included a variety of chemical and product classes have been tested in the IRE test method. The chemical classes with the greatest amount of *in vitro* IRE data are alcohols, amines, carboxylic acids, esters, formulations, heterocyclic compounds, ketones, onium compounds (e.g., ammonium nitrate), and sulfur compounds. The formulations tested include hair shampoos, soaps, personal care cleansers, and detergents. Most common product classes tested in the IRE assay are active pharmaceutical ingredients, chemical/synthetic intermediates, cleaners, raw materials, soaps and detergents, solvents, and surfactants.

The existing database of substances tested using the four ocular endpoints needed to identify a severe irritant (corneal opacity, corneal swelling, fluorescein penetration, and epithelial integrity) was limited to the Guerriero et al. (2004) study. Because this was a small dataset (n=38), substances in the CEC (1991), Balls et al. (1995), and Gettings et al. (1996) studies that had been identified as ocular corrosives/severe irritants using appropriate decision criteria for identification of a severe irritant (i.e., a corneal opacity score greater than or equal to 3, or a corneal swelling equal to or greater than a 25%) were considered together with the test results obtained by Guerriero et al. (2004). This database is referred to as the "Expanded Data Set." This database has limitations, however, since negative (i.e., true or false negative) outcomes are not considered in those studies using fewer than four ocular endpoints.

Substances that were identified as ocular corrosives/severe irritants based on *in vitro* results by any single endpoint were, therefore, included as part of the expanded data set. Substances in the CEC (1991), Balls et al. (1995), and Gettings et al. (1996) studies that were identified as nonsevere irritants or nonirritants, based on *in vitro* results, were not included in the expanded data set. These substances were not included because an evaluation that included any of the omitted endpoints might have resulted in a severe irritant classification. For example, a substance that did not produce ≥ 25% corneal swelling might have produced a corneal opacity score, fluorescein penetration score, or damage to the epithelium that would have classified it as a severe irritant had these endpoints been evaluated.

A pooled data set consisting of substances from all available studies within a regulatory classification system was also analyzed. For example, using the GHS classification system, data from the Balls et al. (1995), Gettings et al. (1996), and Guerriero et al. (2004) studies were pooled for this analysis. While this pooled data set included all available data within a classification system, it was also limited by variability in the number of ocular endpoints. With the exception of the Guerriero et al. (2004) data in which four endpoints were used, the number of endpoints ranged from one (i.e., corneal swelling) to three (i.e., corneal opacity, corneal swelling, and fluorescein retention) in the other studies. Having less than four ocular endpoints could potentially reduce the likelihood of a positive response using the BRD all-or-none decision criteria.

The accuracy evaluation of the IRE test method was limited to the substances evaluated in four *in vitro-in vivo* comparative studies. The ability of the IRE 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 IRE was assessed separately for each *in vitro-in vivo* comparative study using the decision criteria (prediction model) of Guerriero et al. (2004), where possible, to identify corrosives/severe irritants. In the second approach, the accuracy of IRE was assessed after pooling data across *in vitro-in vivo* comparative studies that used similar protocols, same method of data collection, and the decision criteria of Guerriero et al. (2004). 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 IRE test method performance was comparable among the three hazard classification systems. The overall accuracy of the IRE test method obtained by pooling all studies ranged from 64% to 69%, depending on the classification system used. Sensitivity and specificity ranged from 69% to 76% and 60% to 65%, respectively. The false positive rate ranged from 35% to 40%, while the false negative rate ranged from 24% to 30%. When the analysis is restricted to Guerriero et al. (2004) in which the four ocular endpoints were used in the decision process, an accuracy of 79%, a sensitivity of 100%, and a specificity of 70% were obtained across all classification systems. In this analysis, the false positive rate was 30% and the false negative rate was 0% across all classification systems.

For the expanded data set and using the GHS ocular hazard classification system, the accuracy was 68% (52/76), the false positive rate was 56% (24/43), and the false negative rate was 0% (0/33). The expanded data set used for this evaluation include the 38 substances evaluated by Guerriero et al. (2004) and an additional 38 substances tested by Balls et al. (1995) and Gettings et al. (1996) and classified by IRE as severe irritants, 22 of which were also severe irritants *in vivo* and 16 of which were nonsevere irritants or nonirritants *in vivo*. The performance of the expanded data set is potentially confounded by the exclusion of substances with true negative outcomes (matching *in vivo* and *in vitro* nonsevere or nonirritant classifications), which would affect both specificity and the false negative rate.

Using the expanded data set, the chemical classes that were overpredicted (i.e., were false positives) in the IRE test method according the GHS classification system were ketones (67%, [4/6]), esters (67%, [4/6]), and alcohols (60%, [6/10]). Among the 10 surfactants tested, the false positive rate was 67% (2/3) and the false negative rate was 0% (0/7). The seven cationic surfactants included in this group had a false positive rate of 100% (1/1) and a false negative rate of 0% (0/6). Twelve surfactant-based formulations had a false positive rate of 100% (2/2) with no false negative outcomes (0/10).

For the pooled data set (when results were compared to the GHS ocular hazard classification system) the accuracy was 65% (70/107), the false negative rate was 30% (14/47), and the false positive rate was 38% (23/60). The pooled data set used for this evaluation includes 38 substances evaluated by Guerriero et al. (2004), 54 substances tested by Balls et al. (1995), and 24 tested by Gettings et al. (1996). For nine substances tested in common, consensus regulatory calls were used for comparison of *in vitro* and *in vivo* data.

In order to further evaluate discordant responses of the IRE test method relative to the *in vivo* hazard classification, several accuracy subanalyses were performed using both the expanded

data set and the pooled data set. These included specific classes of chemicals with sufficiently robust numbers of substances ($n \geq 5$), as well as certain properties of interest considered relevant to ocular toxicity testing (e.g., pesticides, surfactants, pH, physical form). Because the international community will soon adopt the GHS classification system for hazard labeling (UN 2003), and considering that there were only modest differences in overall IRE test method accuracy among the three regulatory classification systems (i.e., EPA, EU, GHS), these sub-analyses are focused only on the GHS classification system, using the expanded data set.

Using the expanded data set, with regard to physical form of the substances overpredicted by the IRE test method, liquids had a higher overprediction rate (83%, [19/23]) than solids (25%, [5/20]). The highest false positive rate, based on pH was 33% (2/6) for substances with pH > 7.

No substances in the expanded data set were underpredicted (i.e., were false negatives) by the IRE test method. Thus, an analysis of underprediction based on chemical class, physical form, pH, or NICEATM GHS Category I subclassification was not possible.

Using the pooled data set, the chemical classes that were overpredicted (i.e., were false positives) in the IRE test method according to the GHS classification system were ketones (67%, [4/6]), alcohols (55%, [6/11]), and amines (50%, [3/6]). Among the 13 surfactants tested, 40% (2/5) were overpredicted and 12% (1/8) were underpredicted. Of 25 surfactant-based formulations, 25% (2/8) were overpredicted and 38% (6/16) were underpredicted.

Using the pooled data set, with regard to physical form of the substances overpredicted by the IRE test method, liquids had a higher overprediction rate (49%, (18/37) than solids (22%, [5/23]). The highest false positive rate, based on pH was 33% (2/6) for substances with pH > 7.

In the pooled data set, the highest underprediction rate (i.e., were false negatives) was for carboxylic acids (67%, [4/6]) and organic compounds (50%, [3/6]). The underprediction rate for liquids and solids were similar at 29% (8/28) and 32% (6/19), respectively. The underprediction rate for surfactants was 12% (1/8) and for surfactant-based formulations was 38% (6/16). Underprediction rates of 25-37% (1/4 to 7/19) were obtained for Category 1 subgroups 1 to 4.

In the original draft IRE BRD (NICEATM 2004), no data was provided for the assessment of intralaboratory repeatability and reproducibility. Therefore, an analysis of intralaboratory reliability still could not be conducted.

The original IRE test method reliability analysis included an evaluation of interlaboratory reproducibility using both qualitative and quantitative approaches. While the quantitative analysis was unaffected by the reclassification of the ocular irritancy of some test substances, the qualitative analysis (correct classification as an ocular corrosive/severe irritant or as a non-corrosive/non-severe irritant) of the individual laboratory test results obtained for the

EC/HO validation study (Balls et al., 1995) and for the CEC (1991) collaborative study was affected.

Overall, in the Balls et al. (1995) study, the number of substances with 100% agreement among the four participating laboratories was 59 to 63% (35 to 37/59). The number of substances with 75% agreement among laboratories was 22 to 25% (13 to 15/59). The number of substances with 50% agreement among laboratories was 15% (9/59).

Overall, in the CEC (1991) study, the number of substances with 100% agreement among the three participating laboratories was 81% (17/21). The number of substances with 67% agreement among laboratories was 14% (3/21), while the number of substances with 33% agreement was 5% (1/21).

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