6.0 IRE TEST METHOD ACCURACY

6.1 Accuracy of the IRE Test Method

A critical component of an ICCVAM evaluation of the validation status of a test method is an assessment of the accuracy of the proposed test method when compared to the current reference test method (ICCVAM 2003). This aspect of assay performance is typically evaluated by calculating:

- accuracy (concordance): the proportion of correct outcomes (positive and negative) of a test method
- sensitivity: the proportion of all positive substances that are classified as positive
- specificity: the proportion of all negative substances that are classified as negative
- positive predictivity: the proportion of correct positive responses among substances testing positive
- negative predictivity: the proportion of correct negative responses among substances testing negative
- false positive rate: the proportion of all negative substances that are falsely identified as positive
- false negative rate: the proportion of all positive substances that are falsely identified as negative.

The ability of the IRE test method to correctly identify ocular corrosives and severe irritants, as defined by the GHS, EPA, and EU[1], was evaluated separately for each *in vitro-in vivo* comparative study (i.e., publication) reviewed in **Sections 4.0** and **5.0**. The three ocular hazard classification systems considered during this analysis use different classification systems and decision criteria to identify ocular corrosives and severe irritants based on *in vivo* rabbit eye test results (see **Section 4.0**). All three regulatory classification systems are based on individual animal data in terms of the magnitude of the response and, for the EPA (1996) and GHS (UN 2003), on the extent to which induced ocular lesions fail to reverse by day 21. Thus, to evaluate the accuracy of the IRE test method for identifying ocular corrosives and severe irritants, individual rabbit data collected at different observation times was needed for each substance. However, these data were not consistently available in the reports considered, which limited the number of test results that could be used for assessment of test method accuracy. For example, the CEC (1991) collaborative study provided a EU ocular irritancy classification (i.e., R41, R36, nonirritant [EU 2001]) for the 21 substances tested but did not provide individual *in vivo* rabbit eye data, which precludes an accuracy analysis based on the GHS and EPA classification systems. Furthermore, most of the *in vivo* classifications used for the analyses presented in this section are based on the results of a single study. Unless otherwise indicated, variability in the *in vivo* classification is unknown.

[1] For the purposes of this analysis, an ocular corrosive or severe irritant was defined as a substance that would be classified as Category 1 according to the GHS classification system (UN 2003), Category I according to the EPA classification system (EPA 1996), or as R41 according to the EU classification system (EU 2001) (see **Section 1.0**).

In addition, the accuracy assessments conducted were based on IRE test data that were evaluated by investigators using different endpoints for evaluation and different decision criteria to classify the irritancy potential of test substances. As discussed in **Section 2.2.12**, some IRE studies were conducted using the Draize scoring system to evaluate corneal opacity with or without area of opacity evaluated. Some studies also included other endpoints, such as fluorescein retention or penetration and epithelial integrity (based on slit-lamp observations and/or histology). Furthermore, not all studies evaluated or reported data for all of the time points typically measured in IRE (i.e., 0.5, 1, 2, 3, 4 hours). For example, the CEC (1991) collaborative study reported corneal opacity, corneal swelling, and fluorescein retention at one and four hours. In the Balls et al. (1995) validation study, corneal opacity and corneal swelling were the only endpoints reported and were evaluated at one and four hours. In contrast, Gettings et al. (1996; REET 1) only reported the mean extent of corneal swelling across time (1 to 4 hours). Guerriero et al. (2004) reported maximal corneal opacity (opacity x area), maximal corneal swelling, fluorescein penetration (intensity x area) and assessment of epithelial integrity (i.e., 0.5, 1, 2, 3, and 4 hours). In this study, the decision criteria (Prediction Model) for identification of a severe irritant were based on exceeding cut-off values for any of these parameters (maximum corneal opacity ≥ 3; maximum corneal swelling ≥ 25%; maximum corneal fluorescein penetration ≥ 4; any pitting, stippling, mottling, sloughing, or ulceration of epithelium) (see **Appendix A)**. IRE data from each of these four studies were converted into an irritancy classification using the decision criteria outlined in Guerriero et al. (2004), since these were the only decision criteria that specifically were designed to detect severe ocular irritants (see **Appendix A)**.

A limitation of the available IRE data is that the numbers of endpoints used by various investigators differed. In the studies by Guerriero et al (2004), four different ocular endpoints were used. Comparatively, data from the other studies (e.g., Balls et al. 1995) were conducted with between one and three endpoints. In order to make use of all the available data, an "Expanded Data Set" was developed and evaluated. In this data set, any substance evaluated by any of the studies that would be classified as an *in vitro* severe irritant based on a positive result using any of the four ocular endpoints was identified as a corrosive or severe irritant and included into the database (CEC 1991; Balls et al. 1995; Gettings et al. 1996). Substances in these other studies that were not identified as ocular corrosives or severe irritants could not be used, since a positive result in any of the omitted endpoints might have resulted in a severe irritant classification. For example, in Gettings et al. (1996), only corneal swelling was measured. Substances that produced corneal swelling ≥ 25% were classified as severe irritants and were included in the "Expanded Data Set." However, 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 those endpoints been evaluated.

In addition to the analysis for the "Expanded Data Set", an analysis based on a "Pooled Data Set" was conducted. Both performance analyses were included to increase the number of test substances evaluated given the limitations of each data set. The Expanded Data Set was used to identify ocular corrosives or severe irritants based on a positive response in any of the four ocular endpoints used by Guerriero et al. (2004); the decision criteria for the IRE test method used in the BRD performance analyses. However, there is bias associated with this data set, because negative responses could not be included in studies where there were less than four ocular endpoints evaluated (e.g., Guerriero et al. 2004), because any omitted endpoint could have resulted in a positive response had it been tested and, therefore, only positive outcomes that met the decision criteria for any single endpoint could be included. A Pooled Data Set was included in the analysis that used all available data from the four studies and included negative responses. However, this data set is also limited in that it includes data with positive or negative outcomes from studies in which less than four ocular parameters were evaluated.

Using the classification systems discussed in **Section 5.0**, the *in vitro* irritancy potential of each substance was determined using data supplied in the published report or submitted in response to the *FR* Notice request for data (**Section 5.2**). For the "per study" accuracy analysis, two different types of analyses were used. In the first analysis, the IRE ocular irritancy potential of each substance in each report was determined (**Appendix C**). When the same substance was evaluated in multiple laboratories (see Balls et al. 1995 in **Appendix C**), the IRE ocular irritancy potential for each test was determined. Subsequently, based on the majority of ocular irritancy classification calls, an overall IRE ocular irritancy classification was assigned (e.g., if two tests classified a substance as a nonsevere irritant and three tests classified a substance as a severe irritant; the overall *in vitro* irritancy classification for the substance would be severe irritant). When there was an even number of different irritancy classifications for substances (e.g., two tests classified a substance as a nonsevere irritant and two tests classified a substance as a severe irritant), the more severe irritancy classification was used for the overall classification for the substance (severe irritant, in this case). Once the ocular irritancy potential classification was determined for each substance in a report, the ability of the IRE test method to identify ocular corrosives and severe irritants, as defined by the three different regulatory classification systems (EPA 1996, EU 2001, UN 2003), was determined for each report (**Appendix D**).

In the second analysis used in the "per study" evaluation, each irritancy classification obtained on the same substance tested in multiple laboratories was used separately to assess test method accuracy (i.e., results were not combined across multiple laboratories to develop an overall IRE ocular irritancy classification). The ability of the IRE test method to identify ocular corrosives and severe irritants, as defined by the three different classification systems, was then determined for reports where multiple results were available for the substances tested. This approach was applied to the CEC (1991) and the Balls et al. (1995) studies, the only reports that included multiple laboratory study data.

6.1.1 GHS Classification System: IRE Test Method Accuracy

Accuracy analyses using the GHS regulatory classification system (UN 2003)[2] were conducted on data obtained from three reports (Balls et al. 1995; Gettings et al. 1996; Guerriero et al. 2004). To the extent possible, severe ocular irritants were identified from the *in vitro* data in these reports using the Guerriero et al. (2004) IRE test method scoring system

[2] For the purpose of this accuracy analysis, *in vivo* rabbit study results were used to identify GHS (UN 2003) Category 1 irritants (i.e., severe irritants); substances classified as GHS Category 2A and 2B irritants were identified as nonsevere irritants.

described in **Section 6.1**. For example, two ocular parameters included in the recommended protocol, fluorescein penetration and assessment of epithelial integrity, were not assessed in the IRE studies by Balls et al. (1995) and Gettings et al. (1996). The GHS classification assigned to each test substance is shown in **Appendix D**. The performance characteristics (i.e., accuracy, sensitivity, specificity, positive predictivity, negative predictivity, false positive rate, and false negative rate) were determined for each of the three studies based on the available *in vivo* reference data for the substances tested in these studies (**Table 6-1**). Of the three studies, Balls et al. (1995) provided IRE data for substances tested in multiple laboratories; the first set of accuracy calculations for these studies in **Table 6-1** (n = 59) represents the results obtained using the consensus call for each test substance, while the second set of accuracy calculations for each study represents the results obtained when each independent test result from each laboratory was considered separately (n = 236).

6.1.1.1 *Balls et al. (1995)*

Based on the available *in vivo* rabbit eye data, 54 of the 59 test substances could be assigned a classification using the GHS system (UN 2003) (**Table 6-1**). The remaining five substances had inadequate *in vivo* data for assigning a classification according to the GHS system (UN 2003). Based on these 54 substances, the IRE test method had an accuracy of 54% (29/54), a sensitivity of 68% (15/22), a specificity of 44% (14/32), a false positive rate of 56% (18/32), and a false negative rate of 32% (7/22) (**Table 6-1**).

For Balls et al. (1995), using the second approach in which the result of each IRE test is considered separately and GHS classification was possible (n = 216/236), the IRE test method has an accuracy of 60% (130/216), a sensitivity of 72% (63/88), a specificity of 52% (67/128), a false positive rate of 48% (61/128) and a false negative rate of 28% (25/88) for identifying ocular corrosives and severe irritants according to the GHS system (UN 2003) (**Table 6-1**).

6.1.1.2 *Gettings et al. (1996)*

In this study, based on the provided *in vivo* rabbit eye test data, 24 of the 25 test substances could be classified according to the GHS system (UN 2003). Using these data, the IRE test method has an accuracy of 67% (16/24), a sensitivity of 63% (10/16), a specificity of 75% (6/8), a false positive rate of 25% (2/8), and a false negative rate of 38% (6/16) (**Table 6-1**).

6.1.1.3 *Guerriero et al. (2004)*

Based on the available *in vivo* rabbit eye data, 38 of 44 substances could be classified according to the GHS system (UN 2003). Five excluded substances (including two glycols) were classified in the report as severe irritants based on *in vitro* data only (e.g., pH > 11 or < 2) and could not be used for this analysis. In addition, *in vivo* data was not provided for the sixth excluded substance. For the 38 substances, the IRE test method has an accuracy of 79% (30/38), a sensitivity of 100% (11/11), a specificity of 70% (19/27), a false positive rate of 30% (8/27), and a false negative rate of 0% (0/11) (**Table 6-1**).

Table 6-1 Evaluation of the Performance of the IRE Test Method In Predicting Ocular Corrosives and Severe Irritants Compared to the *In Vivo* Rabbit Eye Test Method, as Defined by the GHS Classification System, by Study

Data Source	Anal.[1]	N[2]	Accuracy		Sensitivity		Specificity		Positive Predictivity		Negative Predictivity		False Positive Rate		False Negative Rate	
			%	No.[3]	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.
Balls et al. 1995[4]	A	54/59	54	29/54	68	15/22	44	14/32	45	15/23	67	14/21	56	18/32	32	7/22
Balls et al. 1995[5]	A	216/236	60	130/216	72	63/88	52	67/128	51	63/124	73	67/92	48	61/128	25	28/88
Gettings et al. 1996	B	24/25	67	16/24	63	10/16	75	6/8	83	10/12	50	6/12	25	2/8	38	6/16
Guerriero et al. 2004	C	38/44	79	30/38	100	11/11	70	19/27	58	11/19	100	19/19	30	8/27	0	0/11
Expanded Data Set[6]	-	76/91	68	52/76	100	33/33	44	19/43	58	33/57	100	19/19	56	24/43	0	0/33
Pooled Data Set[7]	-	107/149	65	70/107	70	33/47	62	37/60	59	33/56	73	37/51	38	23/60	30	14/47

A = 1 and 4 hour corneal opacity and swelling.

B = Mean corneal swelling at 0.5, 1, 2, 3 and 4 hours and a numerical irritation scale ranging from 0-4 based on the swelling measurements.

C = Maximum corneal opacity, mean corneal swelling, maximum fluorescein uptake and evaluation of epithelial integrity (1, 2, 3, 4 hours).

[1] Anal. = Analytical method used to transform the sample data into IRE classification.

[2] N = Number of substances included in this analysis/number of substances in the study.

[3] The data on which the percentage calculation is based.

[4] Performance calculated using the overall *in vitro* classification based on the majority and/or most severe classification among the four laboratories.

[5] Performance calculated using each individual *in vitro* classification from each of the four laboratories.

[6] Expanded Data Set includes substances classified as corrosives/severe irritants based on *in vitro* results by any single endpoint.

[7] Pooled Data Set includes data from Balls et al. (1995), Gettings et al. (1996), Guerriero et al. (2004). Consensus calls were used for substances tested more than once.

6.1.1.4 *Expanded Data Set*

Based on the available *in vivo* rabbit eye data and using the Expanded Data Set described in **Section 6.1,** 76 of 91 could be classified according to the GHS system (UN 2003). For the 76 substances classified, the IRE test method has an accuracy of 68% (52/76), a sensitivity of 100% (33/33), a specificity of 44% (19/43), a false positive rate of 56% (24/43), and a false negative rate of 0% (0/33).

6.1.1.5 *Pooled Data Set*

An additional analysis using pooled data from the Balls et al. (1995), Gettings et al. (1996), and Guerriero et al. (2004) studies is shown in **Table 6-1**. In this pooled data set, consensus calls were used for *in vitro* results on nine substances that were tested in more than one laboratory. In this pooled data set, an accuracy of 65% (70/107), a sensitivity of 70% (33/47), a specificity of 62% (37/60), a false positive rate of 38% (23/60), and a false negative rate of 30% (14/47) was obtained.

6.1.1.6 *Discordant Results According to the GHS Classification System*

In order to evaluate discordant responses of the IRE test method relative to the *in vivo* hazard classification, several subanalyses were performed. The subgroup analyses were conducted for both the Expanded Data Set (n = 76) and the Pooled Data Set (n = 107). These analyses included specific classes of chemicals with sufficiently robust numbers of substances (n ≥ 5), as well as certain properties of interest considered relevant to ocular toxicity testing (e.g., surfactants, pH, physical form).

6.1.1.7 Expanded Data Set

As shown in **Table 6-2,** various subgroups of test substances impacted the performance of the IRE test method in the Expanded Data Set. For example, when substances were divided according to chemical class and there were at least five test substances included, false positive rates were greatest for alcohols (60% [6/10]), amines (60% [3/5]), esters (67% [4/6]), heterocycles (50% [4/8]), and ketones (67% [4/6]). There were no false negatives observed for any chemical class.

When physical properties were considered, liquids had a false positive rate of (83% [19/23]) and solids had a false positive rate of (25% [5/20]).

Of 10 surfactants that were assigned a GHS classification (UN 2003), 67% (2/3) were overpredicted. Three nonionic surfactants produced a false positive rate of 50% (1/2) and a 100% false negative rate (1/1). There were no anionic surfactants identified. Of 12 surfactant-based formulations tested, a 100% (2/2) false positive response was produced and none produced false negative responses (0/10).

Table 6-2 False Positive and False Negative Rates of the IRE Test Method, by Chemical Class and Properties of Interest, for the GHS[1] Classification System (Analysis Based on the Expanded Data Set)

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	76	56	24/43	0	0/33
***Chemical Class*[6]**					
Alcohol	11	60	6/10	0	0/1
Amide	5	0	0/3	0	0/2
Amine	9	60	3/5	0	0/4
Carboxylic acid	5	67	2/3	0	0/2
Ester	6	67	4/6	-	0/0
Ether	8	40	2/5	0	0/3
Formulation	12	100	2/2	0	0/10
Heterocycle	16	50	4/8	0	0/8
Ketone	6	67	4/6	-	0/0
Onium compound	9	33	1/3	0	0/6
Sulfur compound	7	20	1/5	0	0/2
Properties of Interest					
Liquid/Solution	43	83	19/23	0	0/20
Solids	33	25	5/20	0	0/13
Surfactants[7] - Total	10	67	2/3	0	0/7
-nonionic	3	50	1/2	0	0/1
-anionic	-	-	-	-	-
-cationic	7	100	1/1	0	0/6
Surfactant-based formulations	12	100	2/2	0	0/10
pH - Total[8]	27	24	4/17	0	0/10
-acidic (pH < 7.0)	18	20	2/10	0	0/8
-basic (pH > 7.0)	7	33	2/6	0	0/1
-neutral (pH = 7.0)	2	0	0/1	0	0/1
Category 1 Subgroup[9] - Total	25[11]	-	-	0	0/25
- 4 (CO=4 at any time)	8	-	-	0	0/8
- 3 (severity/persistence)	3	-	-	0	0/3
- 2 (severity)	2	-	-	0	0/2
- 2-4 combined[10]	13	-	-	0	0/13
- 1 (persistence)	12	-	-	0	0/12

[1] GHS = Globally Harmonized System (UN 2003).

[2] N = Number of substances.

[3] False Positive Rate = the proportion of all negative substances that are falsely identified as positive *in vitro*.

[4] False Negative Rate = the proportion of all positive substances that are falsely identified as negative *in vitro.*

[5] Data used to calculate the percentage.

[6] Chemical classes included in this table are represented by at least five substances tested in the BCOP test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh)

[7] Combines single chemicals labeled as surfactants along with surfactant-containing formulations.

[8] Total number of GHS Category 1 substances for which pH information was obtained.

[9] NICEATM-defined subgroups assigned based on the lesions that drove classification of a GHS Category 1 substance. 1: based on lesions that are persistent; 2: based on lesions that are severe (not including Corneal Opacity [CO] = 4); 3: based on lesions that are severe (not including CO = 4) and persistent; 4: CO = 4 at any time.

[10] Subcategories 2 to 4 combined to allow for a direct comparison of GHS Category 1 substances classified *in vivo* based on some lesion severity component and those classified based on persistent lesions alone.

[11] The number of substances evaluated in the Category 1 subgroup analysis may be less than the number of *in vivo* Category 1 subsstances evaluated, since some substances could not be classified into the subgroups used in the evaluation.

Overall, the false positive rate of 27 substances with pH information that assigned a classification according to the GHS system (UN 2003) was 24% (4/17) with no false negative responses (0% [0/10]). Of the eighteen acidic substances (pH < 7.0) tested, the false positive rate was 20% (2/10) with no false negatives produced (0% [0/8]). Of the seven basic substances (pH > 7.0) evaluated, a higher false positive rate of 33% (2/6) was observed with no false negative outcomes (0% [0/1]). The two neutral substances (pH = 7.0) did not produce any false positive or false negative responses.

Finally, for 25 substances that were assigned a GHS classification (UN 2003), there were no incorrect *in vitro* classifications (false positive or false negative) based on whether the ocular lesions were based on either severity (n = 13) or persistence (n = 12).

6.1.1.8 Pooled Data Set

As shown in **Table 6-3,** various subgroups of test substances impacted the performance of the IRE test method in the Pooled Data Set. For example, when substances were divided according to chemical class and there were at least 5 test substances included, false positive rates were greatest for alcohols (55% [6/11]), amines (50% [3/6]), and ketones (67% [4/6]). The false negative rates were greatest for carboxylic acids (67% [4/6]) and organic compounds (50% [3/6]).

When physical properties were considered, liquids had higher false positive rate (49% [18/37]) when compared to solids (22% [5/23]). Liquids had a 29% (8/28) false negative rate compared to a 32% (6/19) false negative rate for solids.

Of 13 surfactants that were assigned a GHS classification (UN 2003), 40% (2/5) were overpredicted and 12% (1/8) were underpredicted. Four nonionic surfactants produced a false positive rate of 33% (1/3) with no false positive responses (0% [0/1]). Of two anionic surfactants identified, no false positives were produced (0% [0/1]), but there was one false negative outcome (100% [1/1]). Seven cationic surfactants were available with one false positive (100% [1/1]) and no false negative outcomes (0% [0/6]). Of 24 surfactant-based formulations, 25% (2/8) were overpredicted and 38% (6/16) were underpredicted.

Overall, the false positive rate of 27 substances with pH information that assigned a classification according to the GHS system (UN 2003) was 24% (4/17) with no false negative responses (0% [0/10]). Eighteen acidic substances (pH < 7.0) produced a false positive rate of 20% (2/10) with no false negative outcomes (0% [0/8]). Seven basic substances (pH > 7.0) produced a higher false positive rate of 33% (2/6) with no false positive outcomes (0% [0/1]). Two neutral substances (pH =7.0) did not produce any false positive or false negative responses.

Finally, for 37 substances that were assigned a GHS Category 1 classification (UN 2003), the false negative rate was 32% (12/37). False negative rates were greater for substances classified *in vivo* (according to the GHS classification system) based on persistent lesions (37% [7/19]), rather than severe lesions (28% [5/18]). However, three substances that caused severe lesions *in vivo* (corneal opacity = 4) were false negatives.

Table 6-3 False Positive and False Negative Rates of the IRE Test Method, by Chemical Class and Properties of Interest, for the GHS[1] Classification System (Analysis Based on the Pooled Data Set)

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	107	38	23/60	30	14/47
Chemical Class[6]					
Alcohol	13	55	6/11	50	1/2
Amide	5	0	0/3	0	0/2
Amine	11	50	3/6	20	1/5
Carboxylic acid	12	33	2/6	67	4/6
Ester	10	30	3/10	-	0/0
Ether	9	33	2/6	0	0/3
Formulation	24	25	2/8	38	6/16
Heterocycle	18	44	4/9	11	1/9
Ketone	6	67	4/6	-	0/0
Onium compound	10	33	1/3	0	0/7
Organic	12	17	1/6	50	3/6
Sulfur compound	8	20	1/5	33	1/3
Properties of Interest					
Liquid/Solution	65	49	18/37	29	8/28
Solids	42	22	5/23	32	6/19
Surfactants[7] - Total	13	40	2/5	12	1/8
-nonionic	4	33	1/3	0	0/1
-anionic	2	0	0/1	100	1/1
-cationic	7	100	1/1	0	0/6
Surfactant-based formulations	24	25	2/8	38	6/16
pH - Total[8]	27	24	4/17	0	0/10
-acidic (pH < 7.0)	18	20	2/10	0	0/8
-basic (pH > 7.0)	7	33	2/6	0	0/1
-neutral (pH = 7.0)	2	0	0/1	0	0/1
Category 1 Subgroup[9] - Total	37[11]	-	-	32	12/37
- 4 (CO=4 at any time)	11	-	-	27	3/11
- 3 (severity/persistence)	4	-	-	25	1/4
- 2 (severity)	3	-	-	33	1/3
- 2-4 combined[10]	18	-	-	28	5/18
- 1 (persistence)	19	-	-	37	7/19

[1] GHS = Globally Harmonized System (UN 2003).

[2] N = Number of substances.

[3] False Positive Rate = the proportion of all negative substances that are falsely identified as positive *in vitro*.

[4] False Negative Rate = the proportion of all positive substances that are falsely identified as negative *in vitro*.

[5] Data used to calculate the percentage.

[6] Chemical classes included in this table are represented by at least five substances tested in the BCOP test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh)

[7] Combines single chemicals labeled as surfactants along with surfactant-containing formulations.

[8] Total number of GHS Category 1 substances for which pH information was obtained.

[9] NICEATM-defined subgroups assigned based on the lesions that drove classification of a GHS Category 1 substance. 1: based on lesions that are persistent; 2: based on lesions that are severe (not including Corneal Opacity [CO]=4); 3: based on lesions that are severe (not including CO=4) and persistent; 4: CO = 4 at any time.

[10] Subcategories 2 to 4 combined to allow for a direct comparison of GHS Category 1 substances classified *in vivo* based on some lesion severity component and those classified based on persistent lesions alone.

[11] The number of substances evaluated in the Category 1 subgroup analysis may be less than the number of *in vivo* Category 1 substances evaluated, since some substances could not be classified into the subgroups used in the evaluation.

6.1.2 EPA Classification System: IRE Test Method Accuracy

Accuracy analyses for ocular corrosives and severe irritancy, as defined by the EPA regulatory classification system[3] were conducted on data obtained from Balls et al. (1995), Gettings et al. (1996), and Guerriero et al. (2004). The EPA classification assigned to each test substance is presented in **Appendix D**. To the extent possible, severe ocular irritants were identified from the *in vitro* data in these reports using the Guerriero et al. (2004) IRE test method scoring system described in **Section 6.1**. The performance characteristics of the three studies are shown in **Table 6-4** and are based on the available *in vivo* reference data for each study. Of the three studies, Balls provided IRE data for substances tested in multiple laboratories; the first set of accuracy calculations for these studies in **Table 6-4** (n = 59) represents the results obtained using the consensus call for each test substance, while the second set of accuracy calculations for each study represents the results obtained when each independent test result from each laboratory was considered separately (n = 236).

6.1.2.1 *Balls et al. (1995)*

Based on the available *in vivo* rabbit eye data, 53 of the 59 substances tested in this study could be assigned an EPA classification (**Table 6-4**) (EPA 1996). The remaining six substances had inadequate *in vivo* data for assigning a classification according to the EPA system (EPA 1996). For the 53 substances that could be evaluated, the IRE test method has an accuracy of 51% (27/53), a sensitivity of 65% (13/20), a specificity of 42% (14/33), a false positive rate of 58% (19/33), and a false negative rate of 35% (7/20) (**Table 6-4**).

For Balls et al. (1995), using the second approach in which the result of each IRE test result is considered separately and test substances could be classified according to the EPA system (EPA 1996) (n = 208/236), the IRE test method has an accuracy of 56% (116/208), a sensitivity of 65% (47/72), a specificity of 51% (69/136), a false positive rate of 49% (67/136), and a false negative rate of 35% (25/72), for identifying ocular corrosives and severe irritants as classified by the EPA (**Table 6-4**).

6.1.2.2 *Gettings et al. (1996)*

Based on the available *in vivo* rabbit eye test data, all 25 test substances could be assigned an EPA ocular hazard classification (EPA 1996). Using these data, the IRE test method has an accuracy of 64% (16/25), a sensitivity of 59% (10/17), a specificity of 75% (6/8), a false positive rate of 25% (2/8), and a false negative rate of 41% (7/17) (**Table 6-4**).

6.1.2.3 *Guerriero et al. (2004)*

Based on the available *in vivo* rabbit eye test data, 38 of the 44 substances could be assigned an EPA hazard classification (EPA 1996) (**Table 6-4**). The remaining six substances had inadequate *in vivo* data for assigning a classification according to the EPA system (EPA 1996). For the 38 substances that could be evaluated, the IRE test method has an accuracy of 79% (30/38), a sensitivity of 100% (11/11), a specificity of 70% (19/27), a false positive rate of 30% (8/27), and a false negative rate of 0% (0/11) (**Table 6-4**).

[3] For the purpose of this accuracy analysis, *in vivo* rabbit study results were used to identify EPA (EPA 1996) Category I irritants (i.e., severe irritants); substances classified as EPA Category II, III, or IV irritants were defined as nonsevere irritants.

Table 6-4 Evaluation of the Performance of the IRE Test Method In Predicting Ocular Corrosives and Severe Irritants Compared to the *In Vivo* Rabbit Eye Test Method, as Defined by the EPA Classification System, by Study

Data Source	Anal.[1]	N[2]	Accuracy		Sensitivity		Specificity		Positive Predictivity		Negative Predictivity		False Positive Rate		False Negative Rate	
			%	**No.[3]**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**
Balls et al. 1995[4]	A	53/59	51	27/53	65	13/20	42	14/33	41	13/32	67	14/21	58	19/33	35	7/20
Balls et al. 1995[5]	A	208/236	56	116/208	65	47/72	51	69/136	41	47/114	73	69/94	49	67/136	35	25/72
Gettings et al. 1996	B	25/25	64	16/25	59	10/17	75	6/8	83	10/12	46	6/13	25	2/8	41	7/17
Guerriero et al. 2004	C	38/44	79	30/38	100	11/11	70	19/27	58	11/19	100	19/19	30	8/27	0	0/11
Expanded Data Set[6]	-	76/91	66	50/76	100	31/31	42	19/45	54	31/57	100	19/19	58	26/45	0	0/31
Pooled Data Set[7]	-	107/149	64	68/107	69	31/45	60	37/62	55	31/56	73	37/51	40	25/62	31	14/45

A = 1 and 4 hour corneal opacity and swelling.

B = Mean corneal swelling at 0.5, 1, 2, 3 and 4 hours and a numerical irritation scale ranging from 0-4 based on the swelling measurements.

C = Maximum corneal opacity, mean corneal swelling, maximum fluorescein uptake and evaluation of epithelial integrity (1, 2, 3, 4 hours).

[1] Anal. = Analytical method used to transform the sample data into IRE classification.

[2] N = Number of substances included in this analysis/number of substances in the study.

[3] The data on which the percentage calculation is based.

[4] Performance calculated using the overall *in vitro* classification based on the majority and/or most severe classification among the four laboratories.

[5] Performance calculated using each individual *in vitro* classification from each of the four laboratories.

[6] Expanded Data Set includes substances classified as corrosives/severe irritants based on *in vitro* results by any single endpoint.

[7] Pooled Data Set includes data from Balls et al. (1995), Gettings et al. (1996), Guerriero et al. (2004). Consensus calls were used for substances tested more than once.

6.1.2.4 *Expanded Data Set*

Based on the available *in vivo* rabbit eye data and using the Expanded Data Set described in **Section 6.1,** 76 of 91 could be classified according to the EPA system (EPA 1996). For the 76 substances classified, the IRE test method has an accuracy of 66% (50/76), a sensitivity of 100% (31/31), a specificity of 42% (19/45), a false positive rate of 58% (26/45), and a false negative rate of 0% (0/31).

6.1.2.5 *Pooled Data Set*

An additional analysis using pooled data from the Balls et al. (1995), Gettings et al. (1996), and Guerriero et al. (2004) studies is shown in **Table 6-4**. In this pooled data set, consensus calls were used for *in vitro* results on two substances that were tested in more than one laboratory. In this pooled data set, an accuracy of 64% (68/107), a sensitivity of 69% (31/45), a specificity of 60% (37/62), a false positive rate of 40% (25/62), and a false negative rate of 31% (14/45) was obtained.

6.1.2.6 *Discordant Results According to the EPA Classification System*

In order to evaluate discordant responses of the IRE test method relative to the *in vivo* hazard classification, several accuracy subanalyses were performed. Due to a limited number of available substances using the decision criteria outlined in the IRE BRD, the subgroup analyses were based on both the Expanded Data Set (n = 76) shown in **Table 6-5** and the Pooled Data Set (n = 107) shown in **Table 6-6**. These included specific classes of chemicals with sufficiently robust numbers of substances (n ≥ 5), as well as certain properties of interest considered relevant to ocular toxicity testing (e.g., surfactants, pH, physical form).

As indicated in **Table 6-5,** using 76 substances in the Expanded Data Set, various subgroups of test substances impacted the performance of the IRE test method. For example, when substances were divided according to chemical class and there were at least 5 test substances included, false positive rates were greatest for alcohols (75% [6/8]), amines (67% [4/6]), esters (67% [4/6]), ethers (50% [3/6]), heterocycles (50% [4/8]), and ketones (67% [4/6]). There were no false negatives observed for any chemical class.

When physical properties were considered, liquids had a higher false positive rate (83% [20/24]) when compared to solids (29% [6/21]).

Of nine surfactants that were classified using the EPA classification system (EPA 1996), the false positive rate was 100% (3/3) with no false negative responses (0% [0/6]). Three nonionic surfactants produced a false positive rate of 100% (2/2) and a false negative rate of 0% (0/1). Six cationic surfactants produced a false positive rate of 100% (1/1) with no false negative responses 0% [0/5]). There were no anionic surfactants identified. Of 12 surfactant-based formulations, none (0/12) were overpredicted and none were underpredicted.

Overall, the false positive rate for 27 substances with pH information that were assigned a classification according to the EPA system (EPA 1996) was 24% (4/17) with no false negatives (0% [0/10]). Eighteen acidic substances (pH < 7.0) produced a false positive rate of 20% (2/10) with no false negative responses (0% [0/8]). Seven basic substances (pH >

7.0) produced a higher false positive rate (33% [2/6]) than the acidic substances with no false negative responses (0% [0/1]). Two neutral substances (pH = 7.0) did not produce any false positive or false negative responses.

Table 6-5. False Positive and False Negative Rates of the IRE Test Method, by Chemical Class and Properties of Interest, for the EPA[1] Classification System (Analysis Based on the Expanded Data Set)

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	76	58	26/45	0	0/31
Chemical Class					
Alcohol	10	75	6/8	0	0/2
Amide	5	0	0/3	0	0/2
Amine	10	67	4/6	0	0/4
Carboxylic acid	6	67	2/3	0	0/3
Ester	6	67	4/6	-	0
Ether	8	50	3/6	0	0/2
Formulation	12	100	2/2	0	0/10
Heterocycle	15	50	4/8	0	0/7
Ketone	6	67	4/6	-	0
Onium compound	11	67	4/6	0	0/5
Sulfur compound	7	20	1/5	0	0/2
Properties of Interest					
Liquid/Solution	43	83	20/24	0	0/19
Solid	33	29	6/21	0	0/12
Surfactants – Total	9	100	3/3	0	0/6
-nonionic	3	100	2/2	0	0/1
-anionic	-	-	-	-	-
-cationic	6	100	1/1	0	0/5
Surfactant-based formulations	12	0	0/12	-	-
pH – Total[7]	27	24	4/17	0	0/10
- acidic (pH < 7.0)	18	20	2/10	0	0/8
- basic (pH > 7.0)	7	33	2/6	0	0/1
- neutral (pH = 7.0)	2	0	0/1	0	0/1

[1] EPA = U.S. Environmental Protection Agency (EPA 1996).
[2] N = Number of substances.
[3] False Positive Rate = the proportion of all negative substances that are falsely identified as positive *in vitro*.
[4] False Negative Rate = the proportion of all positive substances that are falsely identified as negative *in vitro*.
[5] Data used to calculate the percentage.
[6] Chemical classes included in this table are represented by at least five substances tested in the IRE test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh). See **Appendix B**.
[7] Total number of EPA Category I substances for which pH information was available.

Table 6-6 False Positive and False Negative Rates of the IRE Test Method, by Chemical Class and Properties of Interest, for the EPA[1] Classification System (Analysis Based on the Pooled Data Set)

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	107	40	25/62	31	14/45
***Chemical Class*[6]**					
Alcohol	13	55	6/11	50	1/2
Amide	5	0	0/3	0	0/2
Amine	12	57	4/7	20	1/5
Carboxylic acid	12	50	3/6	50	3/6
Ester	10	30	3/10	-	0/0
Ether	9	43	3/7	0	0/2
Formulation	25	25	2/8	41	7/17
Heterocycle	17	44	4/9	13	1/8
Ketone	6	67	4/6	-	0/0
Onium compound	9	33	1/3	0	0/6
Organic	13	29	2/7	50	3/6
Sulfur compound	8	20	1/5	33	1/3
Properties of Interest					
Liquid/Solution	66	50	19/38	32	9/28
Solids	41	25	6/24	29	5/17
Surfactants[7] - Total	12	50	3/6	17	1/6
-nonionic	4	50	2/4	-	0/0
-anionic	2	0	0/1	100	1/1
-cationic	6	100	1/1	0	0/5
Surfactant-based formulations	25	25	2/8	41	7/17
pH - Total[8]	27	24	4/17	0	0/10
-acidic (pH < 7.0)	18	20	2/10	0	0/8
-basic (pH > 7.0)	7	33	2/6	0	0/1
-neutral (pH = 7.0)	2	0	0/1	0	0/1

[1] EPA = U.S. Environmental Protection Agency (EPA 1996).
[2] N = Number of substances.
[3] False Positive Rate = the proportion of all negative substances that are falsely identified as positive *in vitro*.
[4] False Negative Rate = the proportion of all positive substances that are falsely identified as negative *in vitro.*
[5] Data used to calculate the percentage.
[6] Chemical classes included in this table are represented by at least five substances tested in the BCOP test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh)
[7] Combines single chemicals labeled as surfactants along with surfactant-containing formulations.
[8] Total number of EPA Category I substances for which pH information was obtained.

As indicated in **Table 6-6,** using 107 substances in the Pooled Data Set, various subgroups of test substances impacted the performance of the IRE test method. For example, when substances were divided according to chemical class and there were at least 5 test substances included, false positive rates were greatest for alcohols (55% [6/11]), amines (57% [4/7]), carboxylic acids (50% [3/6]), and ketones (67% [4/6]). False negative rates were greatest for carboxylic acids (50% [3/6]) and organic compounds (50% [3/6]).

When physical properties were considered, liquids had a higher false positive rate (50% [19/38]) when compared to solids (25% [6/24]). The false negative rate of liquids was 32% (9/28) and 29% (5/17) for solids.

Of 12 surfactants that were classified using the EPA classification system (EPA 1996), the false positive rate was 50% (3/6) and the false negative rate was 17% (1/6). Four nonionic surfactants produced a false positive rate of 50% (2/4) with no false negative responses. Of two anionic surfactants identified, there were no false positives (0% [0/1]) and one false negative outcome (100% [1/1]). Six cationic surfactants produced a false positive rate of 100% (1/1) and a false negative rate of 0% (0/5). Of 25 surfactant-based formulations, 25% (2/8) were overpredicted and 41% (7/17) were underpredicted.

Overall, the false positive rate for 27 substances with pH information that were assigned a classification according to the EPA system (EPA (1996) was 24% (4/17) with no false negatives (0% [0/10]). Eighteen acidic substances (pH < 7.0) produced a false positive rate of 20% (2/10) with no false negative responses (0% [0/8]). Seven basic substances (pH > 7.0) produced a higher false positive rate (33% [2/6]) than the acidic substances with no false negative outcomes (0% [0/1]).

6.1.3 EU Classification System: IRE Test Method Accuracy

Accuracy analyses using the EU regulatory classification system[4] (EU 2001) were conducted on data obtained from CEC (1991), Balls et al. (1995), Gettings et al. (1996), and Guerriero et al. (2004). To the extent possible, severe ocular irritants were identified from the *in vitro* data in these reports using the Guerriero et al. (2004) IRE test method scoring system described in **Section 6.1**. The EU classification (EU 2001) assigned to each test substance is presented in **Appendix D**. To the extent possible, severe ocular irritants were identified from the *in vitro* data in these reports using the Guerriero et al. (2004) IRE test method scoring system described in **Section 6.1**. The performance characteristics of the four studies are shown in **Table 6-7** and are based on the available *in vivo* reference data for each study. Of the four studies, CEC (1991) and Balls et al. (1996) provided IRE data for substances tested in multiple laboratories; the first set of accuracy calculations for these studies in **Table 6-7** (n = 21 and n = 59, respectively) represents the results obtained using the consensus call for each test substance, while the second set of accuracy calculations for each study represents the results obtained when each independent test result from each laboratory was considered separately (n = 63 and n = 236, respectively).

6.1.3.1 *CEC Collaborative Study (1991)*

In this collaborative study, 15 of 21 substances tested had sufficient information to assign a EU classification (EU 2001). Of the 15 substances that could be evaluated, the IRE test method had an accuracy of 87% (13/15), a sensitivity of 100% (5/5), a specificity of 80% (8/10), a false positive rate of 20% (2/10), and a false negative rate of 0% (0/5) (**Table 6-7**).

When the performance was calculated on each individual test substance based on availability of *in vivo* rabbit eye test data (n = 44/63), the IRE test method had an accuracy of 77% (34/44), a sensitivity of 86% (12/14), a specificity of 73% (22/30), a false positive rate of 27% (8/30), and a false negative rate of 14% (2/14) (**Table 6-7**).

[4] For the purpose of this accuracy analysis, *in vivo* rabbit study results were used to identify R41 irritants (i.e., severe irritants); substances classified as R36 were defined as nonsevere irritants.

Table 6-7 Evaluation of the Performance of the IRE Test Method In Predicting Ocular Corrosives and Severe Irritants Compared to *In Vivo* Findings, as Defined by the EU Classification System, by Study

Data Source	Anal.[1]	N[2]	Accuracy		Sensitivity		Specificity		Positive Predictivity		Negative Predictivity		False Positive Rate		False Negative Rate	
			%	No.[3]	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.
CEC 1991[4]	A	15/21	87	13/15	100	5/5	80	8/10	71	5/7	100	8/8	20	2/10	0	0/5
CEC 1991[5]	A	44/63	77	34/44	86	12/14	73	22/30	60	12/20	92	22/24	27	8/30	14	2/14
Balls et al. 1995	B	49/59	55	27/49	74	14/19	43	13/30	45	14/31	72	13/18	57	17/30	26	5/19
Balls et al. 1995 [e]	B	196/236	62	121/196	76	58/76	53	63/120	50	58/115	78	63/81	48	57/120	24	18/76
Gettings et al. 1996	C	24/25	67	16/24	63	10/16	75	6/8	83	10/12	50	6/12	25	2/8	38	6/16
Guerriero et al. 2004	D[f]	38/44	79	30/38	100	11/11	70	19/27	58	11/19	100	19/19	30	8/27	0	0/11
Expanded Data Set[6]	-	80/90	70	56/80	100	37/37	44	19/43	61	37/61	100	19/19	56	24/43	0	0/37
Pooled Data Set[7]	-	114/149	69	79/114	76	37/49	65	42/65	62	37/60	78	42/54	35	23/65	24	12/49

A = Corneal opacity; corneal swelling, fluorescein retention at 0.5, 1, 1.25, 2, 3 and 4 hours.

B = 1 and 4 hour corneal opacity and swelling.

C = Mean corneal swelling at 0.5, 1, 2, 3 and 4 hours and a numerical irritation scale ranging from 0-4 based on the swelling measurements.

D = Corneal opacity/area; fluorescein penetration, corneal swelling, epithelial integrity at 0.5, 1, 2, 3, and 4 hours using assigned cut-off values.

[1] Anal. = Analytical method used to transform the sample data into IRE classification.

[2] N = Number of substances included in the study/number of substances in the study.

[3] The data on which the percentage calculation is based.

[4] Performance calculated using the overall *in vitro* classification based on the majority and/or most severe classification among the three or four testing laboratories.

[5] Performance calculated using each individual *in vitro* classification from each of the four laboratories.

[6] Expanded Data Set includes substances classified as corrosives/severe irritants based on *in vitro* results by any single endpoint.

[7] Pooled Data Set includes data from CEC (1991), Balls et al. (1995), Gettings et al. (1996), Guerriero et al. (2004). Consensus calls were used for substances tested more than once.

6.1.3.2 *Balls et al. (1995)*

In this validation study, 49 of 59 substances tested could be classified according to the EU system (EU 2001). Using these data, the IRE test method had an accuracy of 55% (27/49), a sensitivity of 74% (14/19), a specificity of 43% (13/30), a false positive rate of 57% (17/30), and a false negative rate of 26% (5/19) (**Table 6-7**).

Using the second approach, in which the result of each IRE test method experiment was considered separately (n = 196/236), the IRE test method had an accuracy of 62% (121/196), a sensitivity of 76% (58/76), a specificity of 53% (63/120), a false positive rate of 48% (57/120) and a false negative rate of 24% (18/76), for identifying ocular corrosives and severe irritants as classified by the EU (EU 2001).

6.1.3.3 *Gettings et al. (1996)*

Based on the available *in vivo* rabbit eye data, 24 of the 25 substances tested could be assigned a classification according to the EU system (EU 2001). For these substances, the IRE test method had an accuracy of 67% (16/24), a sensitivity of 63% (10/16), a specificity of 75% (6/8), a false positive rate of 25% (2/8), and a false negative rate of 38% (6/16) (**Table 6-7**).

6.1.3.4 *Guerriero et al. (2004)*

In this study, 38 of 44 substances tested could be classified using the EU classification system (EU 2001). Five substances were excluded from EU classification based on the use of *in vitro* data only (e.g., pH extremes) to classify the ocular irritancy of these substances according to the EU system (EU 2001). Furthermore, although the EU classification (EU 2001) of one substance (allyl alcohol) was based on *in vivo* rabbit eye data, the raw *in vivo* scores were not available for subsequent analysis. For these 38 substances, the IRE test method had an accuracy of 79% (30/38), a sensitivity of 100% (11/11), a specificity of 70% (19/27), a false positive rate of 30% (8/27), and a false negative rate of 0% (0/11) (**Table 6-7**).

6.1.3.5 *Expanded Data Set*

Based on the available *in vivo* rabbit eye data and using the Expanded Data Set described in **Section 6.1,** 80 of 90 substances could be classified according to the EU system (EU 2001). For the 80 substances classified, the IRE test method has an accuracy of 70% (56/80), a sensitivity of 100% (37/37), a specificity of 44% (19/43), a false positive rate of 56% (24/43), and a false negative rate of 0% (0/37).

6.1.3.6 *Pooled Data Set*

An additional analysis using pooled data from the CEC (1991), Balls et al. (1995), Gettings et al. (1996), and Guerriero et al. (2004) studies is shown in **Table 6-7**. In this pooled data set, consensus calls were used for *in vitro* results on eight substances that were tested in more than one laboratory. In this pooled data set, an accuracy of 69% (79/114), a sensitivity of 76% (37/49), a specificity of 65% (42/65), a false positive rate of 35% (23/65), and a false negative rate of 24% (12/49) was obtained.

6.1.3.7 *Discordant Results According to the EU Classification System*

In order to evaluate discordant responses of the IRE test method relative to the *in vivo* hazard classification, several accuracy subanalyses were performed using the Expanded Data Set (n = 80 substances) and the Pooled Data Set (n = 114 substances). These included specific classes of chemicals with sufficiently robust numbers of substances (n ≥ 5), as well as certain properties of interest considered relevant to ocular toxicity testing (e.g., surfactants, pH, physical form).

As indicated in **Table 6-8** using the 80 substances in the Expanded Data Set, various subgroups of test substances impacted the performance of the IRE test method. For example, when substances were divided according to chemical class and there were at least 5 test substances included, false positive rates were greatest for alcohols (60% [6/10]), amines (60% [3/5]), carboxylic acids (60% [3/5]), esters (67% [4/6]), heterocycles (50% [4/8]), and ketones (67% [4/6]). There were no false negatives observed for any chemical class.

When physical properties were considered, liquids had a high false positive rate (82% [18/22]) when compared to solids (25% [5/20]).

Of nine surfactants that were classified using the EU system (EU 2001), the false positive rate was 100% (3/3) with no false negatives (0% [0/6]). For three nonionic surfactants, the false positive rate was 100% (2/2), while the false positive rate was 0% (0/1). For six cationic surfactants the false positive rate was 100% (1/1) with no false negatives identified (0% [0/5]). There were no anionic surfactants identified. None of 12 surfactant-based formulations (0/12) produced false positive responses and none produced false negative responses.

Overall, the false positive rate of 27 substances with pH information that could be classified according to the EU (EU 2001) system was 24% (4/17) with a false negative rate of 0% (0/8). Eighteen acidic substances (pH < 7.0) produced a false positive rate of 20% (2/10) and a false negative rate of 0% (0/8). Seven basic substances (pH > 7.0) produced a higher false positive rate of 33% (2/6) than the acidic substances with no false negatives (0/1). Neutral substances (pH = 7.0; n = 2) did not produce any false positive (0% [0/1]) or false negative responses (0% [0/1]).

As indicated in **Table 6-9** using the 114 substances in the Pooled Data Set, various subgroups of test substances impacted the performance of the IRE test method. For example, when substances were divided according to chemical class and there were at least 5 test substances included, false positive rates were greatest for alcohols (46% [6/13]), heterocycles (44% [4/9]), and ketones (67% [4/6]). False negative rates were greatest for formulations (38% [6/16]).

When physical properties were considered, liquids had a high false positive rate (43% [18/42]) when compared to solids (22% [5/23]). The false negative rate for liquids was 22% (7/32) and 29% (5/17) for solids.

Table 6-8 False Positive and False Negative Rates of the IRE Test Method, by Chemical Class and Properties of Interest, for the EU[1] Classification System (Analysis Based on the Expanded Data Set)

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	80	56	24/43	0	0/37
Chemical Class[6]					
Alcohol	11	60	6/10	0	0/1
Amide	5	0	0/3	0	0/2
Amine	9	60	3/5	0	0/4
Carboxylic acid	7	60	3/5	0	0/4
Ester	6	67	4/6	-	0
Ether	8	40	2/5	0	0/3
Formulation	12	100	2/2	0	0/10
Heterocycle	16	50	4/8	0	0/8
Ketone	6	67	4/6	-	0
Onium compound	10	33	1/3	0	0/7
Sulfur compound	7	20	1/5	0	0/2
Properties of Interest					
Liquid/Solution	48	82	18/22	0	0/26
Solid	32	25	5/20	0	0/12
Surfactants – Total	9	100	3/3	0	0/6
-nonionic	3	100	2/2	0	0/1
-anionic	-	-	-	-	-
-cationic	6	100	1/1	0	0/5
Surfactant-based formulations	12	0	0/12	-	-
pH – Total[7]	27	24	4/17	0	0/10
- acidic (pH < 7.0)	18	20	2/10	0	0/8
- basic (pH > 7.0)	7	33	2/6	0	0/1
- neutral (pH = 7.0)	2	0	0/1	0	0/1

[1] EU = European Union (EU 2001).

[2] N = Number of substances.

[3] False Positive Rate = the proportion of all negative substances that are falsely identified as positive *in vitro*.

[4] False Negative Rate = the proportion of all positive substances that are falsely identified as negative *in vitro*.

[5] Data used to calculate the percentage.

[6] Chemical classes included in this table are represented by at least five substances tested in the IRE test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh). See **Appendix B**.

[7] Total number of EU R41 substances for which pH information was available.

Table 6-9 False Positive and False Negative Rates of the IRE Test Method, by Chemical Class and Properties of Interest, for the EU[1] Classification System (Analysis Based on the Pooled Data Set)

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	114	35	23/65	24	12/49
***Chemical Class*[6]**					
Alcohol	15	46	6/13	50	1/2
Amide	5	0	0/3	0	0/2
Amine	12	43	3/7	20	1/5
Carboxylic acid	12	33	2/6	33	2/6
Ester	12	25	3/12	-	0/0
Ether	9	33	2/6	0	0/3
Formulation	24	25	2/8	38	6/16
Heterocycle	18	44	4/9	11	1/9
Ketone	6	67	4/6	-	0/0
Onium compound	11	33	1/3	0	0/8
Organic	12	17	1/6	33	2/6
Sulfur compound	8	20	1/5	33	1/3
Properties of Interest					
Liquid/Solution	74	43	18/42	22	7/32
Solids	40	22	5/23	29	5/17
Surfactant - Total	13	40	2/5	0	0/8
-nonionic	4	33	1/3	0	0/1
-anionic	1	0	0/1	-	0/0
-cationic	8	100	1/1	0	0/7
Surfactant-based formulations	24	25	2/8	38	6/16
pH – Total[7]	27	24	4/17	0	0/10
-acidic (pH < 7.0)	18	20	2/10	0	0/8
-basic (pH > 7.0)	7	33	2/6	0	0/1
-neutral (pH = 7.0)	2	0	0/1	0	0/1

[1] EU = European Union (EU 2001).
[2] N = Number of substances.
[3] False Positive Rate = the proportion of all negative substances that are falsely identified as positive *in vitro*.
[4] False Negative Rate = the proportion of all positive substances that are falsely identified as negative *in vitro.*
[5] Data used to calculate the percentage.
[6] Chemical classes included in this table are represented by at least five substances tested in the BCOP test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh)
[7] Total number of substances for which pH information was obtained.

Of 13 surfactants that were classified using the EU system (EU 2001), the false positive rate was 40% (2/5) with a false negative rate of 0% (0/8). For four nonionic surfactants, the false positive rate was 33% (1/3), while the false negative rate was 0% (0/1). One anionic surfactant was identified that produced no false positive (0% [0/1]) or false negative (0/0) responses. For eight cationic surfactants the false positive rate was 100% (1/1) with no false negatives identified (0% [0/7]). For 25 surfactant-based formulations, the overprediction rate was 25% (2/8) and the false negative rate was 38% (6/16).

Overall, the false positive rate for substances with pH information that were classified according to the EU system (EU 2001) was 24% (4/17) with no false negatives (0% [0/10]). Eighteen acidic substances (pH < 7) produced a false positive rate of 20% (2/10) and a false negative rate of 0% (0/8). Seven basic substances (pH > 7) produced a higher false positive rate of 33% (2/6) than the acidic substances with no false negative outcomes (0/1). Two neutral substances (pH = 7) did not produce any false positive (0% [0/1]) or false negative responses (0% [0/1]).

6.2 Accuracy of the IRE Test Method for Identifying Ocular Corrosives and Severe Irritants - Summary of Results

While there were some differences in results among the three hazard classification systems evaluated (i.e., GHS [UN 2003], EPA [1996], EU [2001]), the accuracy analysis revealed that IRE test method performance was comparable among the three systems. As can be seen in **Tables 6-1**, **6-4**, and **6-7**, the overall accuracy of the IRE test method ranged from 51 to 87%, depending on the classification system used. For example, in the Balls et al. (1995) study the accuracy among the three regulatory classifications systems (GHS [UN2003]; EPA [1996], and EU [EU 2001]) ranged from 51 to 55%, sensitivity ranged from 65 to 74%, specificity ranged from 42 to 44%, the false positive rate ranged from 56 to 58%, and the false negative rate ranged from 26 to 35%. For the Guerriero et al. (2004) study, the overall accuracy, sensitivity, specificity, false positive rate, and false negative rate was 79%, 100%, 70%, 30%, and 0%, respectively, across the three regulatory systems. Given the relatively homogeneous performance of the IRE test method among the three classification systems, the discussion below encompasses the three hazard classification systems, unless otherwise indicated.

6.2.1 Discordance Among Chemical Classes

The accuracy analysis based on chemical class was performed using the Expanded Data Set (n = 76 to 80) and the Pooled Data Set (n = 107 to 114) depending on the regulatory classification system used, because each data set presented advantages and disadvantages. For the purposes of these analyses, chemical classes represented by fewer than five substances were not considered.

The results of this analysis on the Expanded Data Set indicated that alcohols are often overpredicted in the IRE test method (60% to 75% [6/8 to 6/10] false positive rate, depending on the classification system used). Amines (60 to 67% [3/5 to 4/6]), carboxylic acids (60 to 67% [2/3 to 3/5]), esters (67% [4/6]), heterocycles (50% [4/8]), ketones (67% [4/6]) and onium compounds (33% to 67% [1/3 to 4/6]) also were overpredicted.

There were no underpredicted substances in the Expanded Data Base.

The results of this analysis on the Pooled Data Set indicated that alcohols are often overpredicted in the IRE test method (46 to55%[6/11 to 6/13]) false positive rate, depending on the classification system used). Amines (43 to 57% [3/7 to 4/7]), carboxylic acids (33 to 50% [2/6 to 3/6]), heterocycles (44% [4/9 across all classifications]), and ketones (67% [4/6]) also had high false positive rates. Liquid substances produced a false positive rate of 43% (18/42), and solid substances produced a false positive rate of 22% (5/23).

There were false negative responses produced in the Pooled Data Set by alcohols (50% [1/2]), carboxylic acids (33 to 67% [2/6 to 4/6]), formulations (38 to 41% [6/16 to 7/17]), and organic compounds (33 to 50% [2/6 to 3/6]).

6.2.2 Discordance Among Physical or Chemical Properties of Interest

With regard to physical form of the substances overpredicted by the IRE test method using the Expanded Data Set, 19 to 20 were liquids or solutions and five to six were solids. Considering the proportion of the total available database, liquids (19/23 to 20/24) appear more likely than solids (5/20 to 6/21) to be overpredicted by the IRE test method.

Of nine to 13 surfactants evaluated, 40 to 100% (2/5 to 3/3) were overpredicted across the three regulatory classification systems. One or both (50 to 100%) of two surfactants that could be identified as nonionic surfactants were overpredicted depending on the classification system used. One substance identified as a cationic surfactant was overpredicted across the three regulatory classification systems. Of the 12 surfactant-based formulations evaluated across regulatory classification systems, the overprediction rate was 0% (0/12) and no substances were underpredicted.

Of 27 substances with pH information, 24% (4/17) were overpredicted across the three regulatory classification systems. Basic substances (pH > 7) appear to contribute the highest false positive rate (33% [4/6]) across the three regulatory classification systems.

Of the twenty-five substances categorized as GHS Category 1 (UN 2003) severe irritants, 12 were subgrouped as producing persistent lesions (Subgroup 1), whereas 13 were subgrouped as producing severe lesions (subgroup 2 to 4). There were no underpredicted substances in these subgroups.

With regard to physical form of the substances overpredicted by the IRE test method using the Pooled Data Set, 18 to 19 were liquids or solutions and 5 to 6 were solids. Considering the proportion of the total available database, liquids (18/42 to 19/38) appear more likely than solids (5/23 to 6/24) to be overpredicted by the IRE test method.

Of the 17 to 25 surfactants evaluated, 25 to 36% (2/8 to 4/11) were overpredicted across the three regulatory classification systems. The actual number of overpredicted substances for any specific form of surfactant (nonionic, cationic, or anionic) ranged from 0 to 2 and was not adequate to draw any significant conclusions on these subclasses from the data. Of the 25 surfactant-based formulations, 25% were overpredicted (2/8) and 38% (6/16) were underpredicted.

Of 27 substances with pH information, 24% (4/17) were overpredicted across the three regulatory classification systems. Basic substances (pH > 7) appear to contribute the highest false positive rate (33%; 4/6) across the three regulatory classification systems.

Of the 37 substances categorized as GHS Category 1 (UN 2003) severe irritants, 19 were subgrouped as producing persistent lesions (Subgroup 1), whereas 18 were subgrouped as producing severe lesions (subgroup 2 to 4), while underpredicted substances in the Pooled

Data Set (25 to 37% [1/4 to 7/19]),. However, the underprediction rate was relatively uniform across all subgroups and was independent of persistence or severity.

[This Page Intentionally Left Blank]