6.0 ICE TEST METHOD ACCURACY

6.1 Accuracy of the ICE 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 ICE test method to correctly identify ocular corrosives and severe irritants, as defined by the GHS, EPA, and EU classification systems (EPA 1996; EU 2001; UN 2003)[1], was evaluated using two approaches. In the first approach, the performance of ICE was assessed separately for each *in vitro-in vivo* comparative study (i.e., publication) reviewed in **Sections 4.0** and **5.0**. In the second approach, an overall analysis of ICE test method accuracy was conducted by combining results from each study, and then an overall ocular irritancy classification was assigned for each substance. When the same substance was evaluated in multiple laboratories, the overall ICE ocular irritancy classification was based on the majority of calls among all of the studies. When there was an equal 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).

The three regulatory ocular hazard classification systems considered during this analysis use different decision criteria to identify ocular corrosives and severe irritants based on *in vivo* rabbit eye test results (see **Section 1.0**). All three classification systems are based on individual animal data in terms of the magnitude of the response and, for the EPA and GHS,

[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), as 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**).

on the extent to which induced ocular lesions fail to reverse by day 21. Thus, to evaluate the accuracy of the ICE test method for identifying ocular corrosives and severe irritants, individual rabbit data collected at the different observation times are needed for each substance. However, these data were not consistently available in the studies considered, which limited the number of results that could be used to assess test method accuracy. 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.

This evaluation of ICE test method performance included substances evaluated in Prinsen and Koëter (1993), Balls et al. (1995), Prinsen (1996), Prinsen (2000) and Prinsen (2005). Two studies (Prinsen and Koëter 1993; Prinsen 2000) provided, for each substance tested, summary *in vivo* rabbit eye data and the corresponding ocular irritancy classification according to the EU classification system (i.e., R41, R36, nonirritating [EU 2001]). The authors did not provide the individual rabbit *in vivo* data on which this classification was based (these data were requested but not provided). Thus, irritancy classification for some of the substances tested in these studies according to the EPA and GHS systems was not possible. However, for some nonsevere irritating substances, the summary information provided by the authors could be used to assign a nonsevere irritancy classification according to the GHS (Category 2A, 2B, non-irritant [UN 2003]) or EPA (Category II, III, IV [EPA 1996]) systems. Although not helpful for assessing sensitivity or the false negative rate, inclusion of these substances in the performance evaluation did increase the numbers of nonsevere substances included in calculating specificity and the false positive rate of the ICE test method.

For the remaining studies considered (Balls et al. 1995, Prinsen 1996, and Prinsen 2005), individual animal data for the substances screened with the ICE test method were available, so most of the test substances could be assigned an irritancy classification in each of the three regulatory ocular hazard classification systems. The number of substances analyzed for each classification system is noted in the section discussing the accuracy analysis for that system.

Accuracy of ICE for Individual Studies: For the *per study* accuracy analysis, two different analyses were used. For the first analysis, the ICE ocular irritancy potential of each substance in each study under consideration was determined (**Appendix C**). For the one study where the same substance was evaluated in more than one laboratory (see Balls et al. 1995 in **Appendix C**), the ICE ocular irritancy potential for each independent test result was determined. Subsequently, an overall ICE ocular irritancy classification was assigned for each substance in this study based on the majority of ocular irritancy classification calls, (e.g., if two tests classified a substance as a nonirritant and three tests classified a substance as a severe irritant; the overall *in vitro* irritancy classification for the substance was 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 each study under consideration, the ability of the ICE test method to identify ocular corrosives and severe irritants, as defined by the three different classification systems, was determined for each study. The *in vitro* and *in vivo* classifications assigned to each substance are provided in **Appendix D**.

In the second analysis used in the *per study* evaluation, each classification obtained when the same substance was evaluated in more than one laboratory (Balls et al. 1995) was used separately to assess test method accuracy (i.e., results were not combined across multiple tests to develop an overall ICE ocular irritancy classification). The ability of the ICE 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 tested substances.

Accuracy of ICE for Pooled Studies: For an overall analysis of ICE test method accuracy, results from all studies under consideration were combined and an ocular irritancy classification was determined for each substance. When the same substance was evaluated in more than one laboratory, the overall ICE ocular irritancy classification was based on the majority of calls among all of the laboratories in all studies under consideration (see **Appendix C**).

6.1.1 GHS Classification System: ICE Test Method Accuracy

The four studies Prinsen and Koëter (1993), Balls et al. (1995), Prinsen (1996), Prinsen (2005) contained ICE test method data on 171 substances, 144 of which had sufficient *in vivo* data to be assigned an ocular irritancy classification according to the GHS classification system (UN [2003])[2] (see **Appendix C**). Based on results from *in vivo* rabbit eye experiments, 30[3] of the 144 substances were classified as severe irritants (i.e., Category 1), the other 114 substances were classified as nonsevere irritants (either Category 2A, 2B) or nonirritants. The 27 substances that could not be classified according to the GHS classification system due to the lack of adequate animal data are so noted in **Appendix C**.

6.1.1.1 *Prinsen and Koëter (1993)*

Based on the available *in vivo* rabbit eye data, 10 of the 21 substances tested in this study could be assigned a GHS classification (**Table 6-1**). The remaining 11 substances had insufficient *in vivo* data for assigning a classification according to the GHS system (UN 2003). For the 10 substances that could be evaluated, the ICE test method has an accuracy of 80% (8/10), a sensitivity of 100% (2/2), a specificity of 75% (6/8), a false positive rate of 25% (2/8), and a false negative rate of 0% (0/2)

6.1.1.2 *Balls et al (1995)*

Based on the available *in vivo* rabbit eye data, 54 of the 59 substances tested in this study could be assigned a GHS classification (**Table 6-1**). The remaining five substances had

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

[3] One chemical (benzalkonium chloride, 1%) was tested *in vivo* twice in the same laboratory. The results were discordant with respect to GHS classification. According to one test, the classification was Category 1, while results from the other test yielded a Category 2B classification. The accuracy analysis was performed with the substance classified as Category 1.

Table 6-1. Evaluation of the Performance of the ICE 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 and Overall

Data Source	N[2]	Accuracy		Sensitivity		Specificity		Positive Predictivity		Negative Predictivity		False Positive Rate		False Negative Rate	
		%	No.[3]	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.
Prinsen and Koëter (1993)	10/21	80	8/10	100	2/2	75	6/8	50/2/4	3/4	100	6/6	25	2/8	0	0/2
Balls et al. (1995)[4,5]	54/59	69	37/54	50	11/22	81	26/32	65	11/17	70	26/37	19	6/32	50	11/22
Balls et al. (1995)[4]	215/235	70	150/215	46	40/87	86	110/128	69	40/58	70	110/157	14	18/128	54	47/87
Prinsen (1996)	36/44	97	35/36	50	1/2	100	34/34	100	1/1	97	34/35	0	0/34	50	1/2
Prinsen (2005)	46/50	89	41/46	0	0/4	98	41/42	0	0/1	91	41/45	2	1/42	100	4/4
Entire Data Set[5,6]	144/171	83	120/144	50	15/30	92	105/114	63	15/24	88	105/120	8	9/114	50	15/30

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

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

[3] No.· = Data used to calculate the percentage.

[4] One chemical (benzalkonium chloride, 1%) was tested *in vivo* twice within the same laboratory. The results were discordant with respect to GHS classification; the analysis was performed assuming Category 1 classification.

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

[6] Includes the data from Balls et al. (1995) using the overall *in vitro* classification based on the majority and/or most severe classification among the four laboratories

inadequate *in vivo* data for assigning a classification according to the GHS system (UN [2003]). Using the first accuracy analysis approach (single call per test substance), for the 54 substances assigned a GHS classification, the ICE test method has an accuracy of 69% (37/54), a sensitivity of 50% (11/22), a specificity of 81% (26/32), a false positive rate of 19% (6/32), and a false negative rate of 50% (11/22). Using the second accuracy analysis approach (results not combined across multiple tests to develop an overall ICE ocular irritancy classification) for the 215 substances considered, the ICE test method has an accuracy of 70% (150/215), a sensitivity of 46% (40/87), a specificity of 86% (110/128), a false positive rate of 14% (18/128), and a false negative rate of 54% (47/87).

6.1.1.3 *Prinsen (1996)*

Based on the *in vivo* rabbit eye data, 36 of the 44 substances tested in this study could be assigned a GHS classification (**Table 6-1**). The remaining eight substances had inadequate *in vivo* data for assigning a classification according to the GHS system (UN 2003). For the 36 substances that could be evaluated, the ICE test method has an accuracy of 97% (35/36), a sensitivity of 50% (1/2), a specificity of 100% (34/34), a false positive rate of 0% (0/34), and a false negative rate of 50% (1/2).

6.1.1.4 *Prinsen (2005)*

Based on the available *in vivo* rabbit eye data provided in this submission, 46 of the 50 substances tested in this study could be assigned a GHS classification (**Table 6-1**). The remaining four substances had inadequate *in vivo* data for assigning a classification according to the GHS system. For the 46 substances that could be evaluated, the ICE test method has an accuracy of 89% (41/46), a sensitivity of 0% (0/4), a specificity of 98% (41/42), a false positive rate of 2% (1/42), and a false negative rate of 100% (4/4).

6.1.1.5 *Entire Data Set*

A total of 144 substances had sufficient *in vivo* data among the four studies to perform an accuracy analysis, based on the GHS classification system (**Table 6-1**). Twenty-two substances lacked sufficient *in vivo* information on which to assign a GHS classification. Based on these 144 substances, the ICE test method has an accuracy of 83% (120/144), a sensitivity of 50% (15/30), a specificity of 92% (105/114), a false positive rate of 8% (9/114), and a false negative rate of 50% (15/30).

6.1.1.6 *Discordant Results According to the GHS Classification System*

In order to evaluate discordant responses of the ICE test method relative to the *in vivo* hazard classification, several accuracy sub-analyses were performed. 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., pesticides, surfactants, pH, physical form).

As indicated in **Table 6-2**, there were some notable trends in the performance of the ICE test method. According to the GHS classification system, the most consistently overpredicted

Table 6-2. False Positive and False Negative Rates of the ICE Test Method, by Chemical Class and Properties of Interest, for the GHS[1] Classification System

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	144	8	9/114	50	15/30
***Chemical Class*[6]**					
Alcohol	12	50	5/10	50	1/2
Amine/Amidine	5	0	0/2	33	1/3
Carboxylic acid	10	0	0/3	43	3/7
Ester	9	13	1/8	0	0/1
Heterocyclic	9	0	0/3	33	2/6
Onium compound	8	0	0/2	33	2/6
Properties of Interest					
Liquids	108	10	9/90	44	8/18
Solids	36	0	0/24	58	7/12
Pesticide	11	0	0/6	60	3/5
Surfactant – Total	21	0	0/12	56	5/9
-nonionic	4	0	0/3	100	1/1
-anionic	2	0	0/1	100	1/1
-cationic	7	0	0/1	33	2/6
pH – Total[7]	20	-	-	40	8/20
- acidic (pH < 7.0)	12	-	-	33	4/12
- basic (pH > 7.0)	8	-	-	50	4/8
Category 1 Subgroup[8]					
- Total	23[10]	-	-	35	8/23
- 4 (CO=4 at any time)	12	-	-	33	4/12
- 3 (severity/persistence)	2	-	-	50	1/2
- 2 (severity)	4	-	-	0	0/4
- *2-4 combined*[9]	18	-	-	28	5/18
- 1 (persistence)	5	-	-	60	3/5

[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 ICE test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh) as defined in **Appendix B**.
[7]Total number of GHS Category 1 substances for which pH information was obtained.
[8]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 CO=4); 3: based on lesions that are severe (not including CO=4) and persistent; 4: corneal opacity (CO) = 4 at any time.
[9]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.
[10]The number of substances evaluated in the Category 1 subgroup analysis may be less than the total number of *in vivo* Category 1 substances evaluated since some substances could not be classified into the subgroups used in the evaluation.

(i.e., false positive[4]) substances were alcohols, which accounted for five out of nine overpredicted substances overall. Other chemical classes represented among overpredicted

[4] False positive in this context refers to a substance classified as a nonsevere (mild or moderate) irritant or nonirritant based on *in vivo* data, but as a severe irritant by the ICE test method.

substances were one each of alkalis, ketones, esters, and an unclassified substance. Regarding the physical form of overpredicted substances, eight were liquids and one (the unclassified substance) was an emulsion (which was counted as a liquid in this analysis). No solid test substances were overpredicted by the ICE test method.

According to the GHS classification system, the most consistently underpredicted (i.e., false negative[5]) substances were carboxylic acids (3), followed closely by heterocyclics (2) and onium compounds (2). Other chemical classes represented among underpredicted substances included one each of alcohols, amines/amidines/polycyclics, imides/organic sulfur compounds, inorganic salts/boron compounds and five unclassified substances. Underpredicted substances were evenly distributed regarding physical form, with seven each of solids and liquids, along with one emulsion (which was counted as a liquid in this analysis). For eight underpredicted substances for which pH data was available, four had a pH less than 7.00, ranging from 3.34 to 5.72 and four had a pH greater than 7.00, ranging from 7.18 to 9.98. Finally, for the eight underpredicted substances classified as severe irritants (GHS Category 1) for which such information was available, three were classified as severe irritants based on persistent lesions (3/5; 60%) while four were classified as severe irritants based on severe lesions (5/18; 28%).

Table 6.3 shows the effects on the ICE test method performance characteristics of excluding from the data set problematic classes (i.e., that gave the most discordant results, according to the GHS classification system). In general, exclusion of alcohols, surfactants or solids individually resulted in small changes in the performance statistics, with the exception that the exclusion of alcohols from the data set caused a two-fold decrease in the false positive rate from 8% (9/114) to 4% (4/104). Similarly, when both alcohols and surfactants were excluded from the data set, changes in the performance statistics were small, again with the exception of the effect on the false positive rate, which decreased two-fold, from 8% (9/114) to 4% (4/92). The largest changes in almost all of the performance statistics were observed when all three discordant classes were excluded from the data set; accuracy increased from 83% (120/144) to 92% (69/75), and the false negative rate decreased from 50% (15/30) to 29% (2/7). The false positive rate decreased from 8% (9/114) to 6% (4/68), but the decrease was not as large as that observed when alcohols alone or alcohols plus surfactants were removed from the data set.

6.1.2 EPA Classification System: ICE Test Method Accuracy

The four studies (Prinsen and Koëter 1993; Balls et al. 1995; Prinsen 1996; Prinsen 2005) contained ICE test method data on 171 substances, 145 of which had sufficient *in vivo* data to be assigned an ocular irritancy classification according to the EPA classification system (EPA 1996)[6] (see **Appendix C**). Based on results from the *in vivo* rabbit eye test, 29 of these 145 substances were classified as severe irritants (i.e., Category I), while the other 116

[5] False negative in this context refers to a substance classified as a nonsevere (mild or moderate) irritant or nonirritant by the ICE test method, but as a severe irritant based on *in vivo* data.

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

Table 6-3. Effect of Exclusion of Discordant Classes on False Negative and False Positive Rates of the ICE Test Method, for the GHS[1] Classification System

Data Set	Accuracy		False Positive Rate[2]		False Negative Rate[3]	
	%	No.[4]	%	No.	%	No.
Overall	83	120/144	8	9/114	50	15/30
w/o Alcohols	86	114/132	4	4/104	50	14/28
w/o Surfactants	85	104/123	9	9/102	48	8/18
w/o Solids	84	91/108	10	9/90	44	8/18
w/o Alcohols & Surfactants	86	96/111	4	4/92	47	9/19
w/o Alcohols & Surfactants & Solids	92	69/75	6	4/68	29	2/7

[1] GHS =- Globally Harmonized System (UN 2003).
[2] False Positive Rate = the proportion of all negative substances that are falsely identified as positive *in vitro*
[3] False Negative Rate = the proportion of all positive substances that are falsely identified as negative *in vitro*
[4] Data used to calculate the percentage.

substances were classified as nonsevere irritants or nonirritants (Categories II, III, or IV). The 26 substances that could not be classified according to the EPA classification system are so noted in **Appendix C**.

6.1.2.1 *Prinsen and Koëter (1993)*

Based on the available *in vivo* rabbit eye data, 10 of the 21 substances tested in this study could be assigned an EPA classification (**Table 6-4**). The remaining 11 substances had inadequate *in vivo* data for assigning a classification according to the EPA system (EPA 1996). For the 10 substances that could be evaluated, the ICE test method has an accuracy of 80% (8/10), a sensitivity of 100% (2/2), a specificity of 75% (6/8), a false positive rate of 25% (2/8), and a false negative rate of 0% (0/2).

6.1.2.2 *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**). The remaining six substances had inadequate *in vivo* data for assigning a classification according to the EPA system (1996).

Table 6-4. Evaluation of the Performance of the ICE Test Method In Predicting Ocular Corrosives and Severe Irritants Compared to the *In Vivo* Rabbit Eye Test Method, as Defined by the EPA[1] Classification System, by Study and Overall

Data Source	N[2]	Accuracy		Sensitivity		Specificity		Positive Predictivity		Negative Predictivity		False Positive Rate		False Negative Rate	
		%	No.[3]	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.
Prinsen and Koëter (1993)	10/21	80	8/10	100	2/2	75	6/8	50	2/4	100	6/6	25	2/8	0	0/2
Balls et al. (1995)[4,5]	53/59	72	38/53	53	10/19	82	28/34	63	10/16	76	28/37	18	6/34	47	9/19
Balls et al. (1995)[4]	211/235	74	156/211	51	38/75	87	118/136	68	38/56	76	118/155	13	18/136	49	37/75
Prinsen (1996)	36/44	97	35/36	50	1/2	100	34/34	100	1/1	97	34/35	0	0/34	50	1/2
Prinsen (2005)	46/50	89	41/46	0	0/4	98	41/42	0	0/1	91	41/45	2	1/42	100	4/4
Entire Data Set[5,6]	145/171	84	122/145	52	15/29	92	107/116	63	13/24	89	107/121	8	9/116	48	14/29

[1] EPA = U.S. Environmental Protection Agency (EPA 1996).

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

[3] Data used to calculate the percentage.

[4] One chemical (benzalkonium chloride, 1%) was tested *in vivo* twice within the same laboratory. The results were discordant with respect to EPA classification; the analysis was performed assuming Category I classification.

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

[6] Includes the data from Balls et al. (1995) using the overall *in vitro* classification based on the majority and/or most severe classification among the four laboratories

Using the first accuracy analysis approach (single call per test substance), for the 53 substances assigned an EPA classification, the ICE test method has an accuracy of 72% (38/53), sensitivity of 53% (10/19), a specificity of 82% (28/34), a false positive rate of 18% (6/34), and a false negative rate of 47% (9/19). Using the second accuracy analysis approach (results not combined across multiple tests to develop an overall ICE ocular irritancy classification), for the 211 substances considered, the ICE test method has an accuracy of 74% (156/211), a sensitivity of 51% (38/75), a specificity of 87% (118/136), a false positive rate of 13% (18/136), and a false negative rate of 49% (37/75).

6.1.2.3 *Prinsen (1996)*

Based on the *in vivo* rabbit eye data, 36 of the 44 substances tested in this study could be assigned an EPA classification (**Table 6-4**). The remaining eight substances had inadequate *in vivo* data for assigning a classification according to the EPA system (1996). For the 36 substances that could be evaluated, the ICE test method has an accuracy of 97% (35/36), a sensitivity of 50% (1/2), a specificity of 100% (34/34), a false positive rate of 0% (0/34), and a false negative rate of 50% (1/2).

6.1.2.4 *Prinsen (2005)*

Based on the available *in vivo* rabbit eye data, 46 of the 50 substances tested in this study could be assigned an EPA classification (**Table 6-4**). The remaining four substances had inadequate *in vivo* data for assigning a classification according to the EPA system (1996). For the 46 substances that could be evaluated, the ICE test method has an accuracy of 89% (41/46), a sensitivity of 0% (0/4), a specificity of 98% (41/42), a false positive rate of 2% (1/42), and a false negative rate of 100% (4/4).

6.1.2.5 *Entire Data Set*

A total of 145 substances had sufficient *in vivo* data among the four studies to perform an accuracy analysis, based on the EPA classification system (**Table 6-4**). Twenty-six substances lacked sufficient *in vivo* information on which to assign an EPA classification (EPA [1996]). Based on these 145 substances, the ICE test method has an accuracy of 84% (122/145), a sensitivity of 52% (15/29), a specificity of 92% (107/116), a false positive rate of 8% (9/116) and a false negative rate of 48% (14/29).

6.1.2.6 *Discordant Results According to the EPA Classification System*

In order to evaluate discordant responses of the ICE test method relative to the *in vivo* hazard classification, several accuracy sub-analyses were performed. 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., pesticides, surfactants, pH, physical form).

As indicated in **Table 6-5**, there were some notable trends in the performance of the ICE test method. According to the EPA classification system, the most consistently overpredicted (i.e., false positive) substances were alcohols, which accounted for five out of nine overpredicted substances overall. Other chemical classes represented among overpredicted substances, with one instance each, were alkalis, esters, ketones and one unclassified

Table 6-5. False Positive and False Negative Rates of the ICE Test Method, by Chemical Class and Properties of Interest, for the EPA[1] Classification System

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	143	8	9/116	52	14/27
***Chemical Class*[6]**					
Alcohol	12	50	5/10	50	1/2
Amine/Amidine	5	0	0/3	50	1/2
Carboxylic acid	10	0	0/3	43	3/7
Ester	9	11	1/9	0	0/0
Heterocyclic	8	0	0/3	40	2/5
Onium compound	7	0	0/2	40	2/5
Properties of Interest					
Liquids	109	10	9/92	41	7/17
Solids	34	0	0/24	70	7/10
Pesticide	11	0	0/7	50	2/4
Surfactant – Total	20	0	0/13	57	4/7
-nonionic	4	0	0/4	0	0/0
-anionic	2	0	0/1	100	1/1
-cationic	6	0	0/1	40	2/5
pH – Total[7]	16	-	-	44	7/16
- acidic (pH < 7.0)	10	-	-	40	4/10
- basic (pH > 7.0)	6	-	-	50	3/6

[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 ICE test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh) as defined in **Appendix B**.
[7] Total number of EPA Category I substances for which pH information was obtained.

substance. Regarding the physical form of overpredicted substances, nine were liquids and none were solids.

According to the EPA classification system, the most consistently underpredicted (i.e., false negative) substances were carboxylic acids, which accounted for three out of 14 overpredicted substances overall. Other chemical classes represented among overpredicted substances included heterocyclics (2), onium compounds (2), imides (1), inorganic boron compounds (1), and polycyclics (1). Regarding the physical form of underpredicted substances, seven were liquids and seven were solids. For the seven underpredicted substances classified as severe irritants (EPA Category I) for which pH data was available, four had a pH less than 7.00, ranging from 3.34 to 5.72 and three had a pH greater than 7.00, ranging from 7.95 to 9.98.

6.1.3 EU Classification System: ICE Test Method Accuracy

The five studies (Prinsen and Koëter 1993; Balls et al. 1995; Prinsen 1996; Prinsen 2000; Prinsen 2005) contained ICE test method data on 175 substances, 154 of which had sufficient *in vivo* data to be assigned an ocular irritancy classification according the EU classification system (EU 2001)[7] (see **Appendix C**). Based on results from the *in vivo* rabbit eye test, 32[8] of the 154 substances were classified as severe irritants (i.e., R41) and the other 122 substances were classified as nonsevere irritants (i.e., R36) or nonirritants. The 21 substances that could not be classified according to the EU classification system are so noted in **Appendix C**.

6.1.3.1 *Prinsen and Koëter (1993)*

All 21 substances tested in this study were included in an analysis of accuracy (**Table 6-6**). Based on the available *in vivo* rabbit eye data or the EU ocular irritancy classification for each substance provided in the published study (individual rabbit eye test data was not available for all of the substances) and using the first accuracy analysis approach (single call per test substance), the ICE test method has an accuracy of 95% (20/21), a sensitivity of 100% (7/7), a specificity of 93% (13/14), a false positive rate of 7% (1/14), and a false negative rate of 0% (0/7).

6.1.3.2 *Balls et al. (1995)*

Based on the available *in vivo* rabbit eye data, 50 of the 59 substances tested in this study could be assigned an EU classification (**Table 6-6**). Nine substances lacked sufficient *in vivo* information on which to assign an EU classification (EU 2001). For the 50 substances assigned an EU classification, the ICE test method has an accuracy of 72% (36/50), sensitivity of 53% (10/19), a specificity of 84% (26/31), a false positive rate of 16% (5/31), and a false negative rate of 47% (9/19). Using the second accuracy analysis approach (results not combined across multiple tests to develop an overall ICE ocular irritancy classification), for the 199 substances considered, the ICE test method has an accuracy of 73% (145/199), a sensitivity of 48% (36/75), a specificity of 88% (109/124), a false positive rate of 12% (15/124), and a false negative rate of 52% (39/75).

6.1.3.3 *Prinsen (1996)*

Based on the *in vivo* rabbit eye data, 36 of the 44 substances tested in this study could be assigned an EU classification (**Table 6-6**). Eight substances lacked sufficient *in vivo* information on which to assign an EU classification (EU 2001). For the 36 substances that could be evaluated, the ICE test method has an accuracy of 97% (35/36), a sensitivity of 50% (1/2), a specificity of 100% (34/34), a false positive rate of 0% (0/34), and a false negative rate of 50% (1/2).

6.1.3.4 *Prinsen (2000)*

The EU classifications were provided by the author for the four substances tested in this study that were used for the accuracy analysis (**Table 6-6**). For these substances, the ICE test method has an accuracy (4/4), sensitivity (1/1), and specificity (3/3) of 100%, and false positive (0/3) and false negative (0/1) rates of 0%.

[7] 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.

[8] One chemical (benzalkonium chloride, 1%) was tested *in vivo* twice in the same laboratory. The results were discordant with respect to EU classification. According to one test, the classification was R41, while results from the other test yielded an R36 classification. The accuracy analysis was performed with the substance classified as R41.

Table 6-6. Evaluation of the Performance of the ICE Test Method In Predicting Ocular Corrosives and Severe Irritants Compared to the *In Vivo* Rabbit Eye Test Method, as Defined by the EU[1] Classification System, by Study and Overall

Data Source	N[2]	Accuracy		Sensitivity		Specificity		Positive Predictivity		Negative Predictivity		False Positive Rate		False Negative Rate	
		%	No.[3]	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.
Prinsen and Koëter (1993)	21/21	95	20/21	100	7/7	93	13/14	88	7/8	100	13/13	7	1/14	0	0/7
Balls et al. (1995)[4,5]	50/59	72	36/50	53	10/19	84	26/31	67	10/15	74	26/35	16	5/31	47	9/19
Balls et al. (1995)[4]	199/235	73	145/199	48	36/75	88	109/124	71	36/51	74	109/148	12	15/124	52	39/75
Prinsen (1996)	36/44	97	35/36	50	1/2	100	34/34	100	1/1	97	34/35	0	0/34	50	1/2
Prinsen (2000)	4/4	100	4/4	100	1/1	100	3/3	100	1/1	100	3/3	0	0/3	0	0/1
Prinsen (2005)	46/50	89	41/46	0	0/4	98	41/42	0	0/1	91	41/45	2	1/42	100	4/4
Entire Data Set[5,6]	154/175	87	134/154	59	19/32	94	115/122	73	19/26	90	115/128	6	7/122	41	13/32

[1] EU =- European Union System (EU 2001).
[2] N = Number of substances included in this analysis/the total number of substances in the study.
[3] Data used to calculate the percentage.
[4] One chemical (benzalkonium chloride, 1%) was tested *in vivo* twice within the same laboratory. The results were discordant with respect to EU classification; the analysis was performed assuming Category 1 classification.
[5] Performance calculated using the overall *in vitro* classification based on the majority and/or most severe classification among the four laboratories.
[6] Includes the data from Balls et al. (1995) using the overall *in vitro* classification based on the majority and/or most severe classification among the four laboratories

6.1.3.5 *Prinsen (2005)*

Based on the available *in vivo* rabbit eye data, 46 of the 50 substances tested in this study could be assigned an EU classification (**Table 6-6**). The remaining four substances had inadequate *in vivo* data for assigning a classification according to the EU system. For the 46 substances that could be evaluated, the ICE test method has an accuracy of 89% (41/46), a sensitivity of 0% (0/4), a specificity of 98% (41/42), a false positive rate of 2% (1/42), and a false negative rate of 100% (4/4).

6.1.3.6 *Entire Data Set*

A total of 154 substances had sufficient *in vivo* data among the five studies to perform an accuracy analysis, based on the EU classification system (**Table 6-6**). For these 154 substances, the ICE test method has an accuracy of 87% (134/154), a sensitivity of 59% (19/32), a specificity of 94% (115/122), a false positive rate of 6% (7/122), and a false negative rate of 41% (13/32).

6.1.3.7 *Discordant Results According to the EU Classification System*

As indicated in **Table 6-7**, there were some notable trends in the performance of the ICE test method. According to the EU classification system, the most consistently overpredicted (i.e., false positive) substances were alcohols, which accounted for three out of seven overpredicted substances overall. Other chemical classes represented among overpredicted substances, with one instance each, were alkalis, esters, ketones and one unclassified substance. Regarding the physical form of overpredicted substances, seven were liquids and none were solids.

According to the EU classification system, the most consistently underpredicted (i.e., false negative) substances were heterocyclics and onium compounds, with two representatives each out of 13 total underpredicted substances. Other chemical classes represented among underpredicted substances included one each of alcohols, amines/amidines, carboxylic acids, imides/organic sulfur compounds, polycyclics and polyethers. Underpredicted substances were evenly distributed with regard to physical form with six each of liquids and solids and one emulsion (counted as a liquid in this analysis). For the seven underpredicted substances classified as severe irritants (EU Category R41) for which pH data was available, three had a pH less than 7.00, ranging from 3.77 to 5.72 and four greater than 7.00, ranging from 7.18 to 9.98.

6.2 Accuracy of the ICE Test Method for Identifying Ocular Corrosives and Severe Irritants – Summary of Results

While differences in results among the three hazard classification systems evaluated occurred (i.e., EPA [1996], EU [2001], and GHS [UN 2003]), the accuracy analysis revealed that the ICE test method performance was comparable among the three systems. As can be seen in **Tables 6-1**, **6-4**, and **6-6**, depending on the classification system, the overall accuracy of the ICE test method ranged from 83% to 87%. Sensitivity ranged from 50% to 59% and specificity ranged from 92% to 94%. The false positive rate ranged from 6% to 8%, while the false negative rate ranged from 41% to 50%. Given the relatively homogeneous performance of the ICE test method among the three classification systems, the discussion below encompasses all three of them, unless otherwise indicated.

Table 6-7. False Positive and False Negative Rates of the ICE Test Method, by Chemical Class and Properties of Interest, for the EU[1] Classification System

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	154	6	7/122	41	13/32
Chemical Class[6]					
Alcohol	14	27	3/11	33	1/3
Carboxylic acid	10	0	0/4	17	1/6
Ester	9	13	1/8	0	0/1
Heterocyclic	9	0	0/3	33	2/6
Inorganics	5	0	0/3	50	1/2
Onium compound	8	0	0/2	33	2/6
Polyether	5	0	0/4	100	1/1
Properties of Interest					
Liquids	116	7	7/97	39	7/18
Solids	38	0	0/25	46	6/13
Pesticide	13	0	0/8	40	2/5
Surfactant – Total	24	0	0/15	44	4/9
-nonionic	5	0	0/5	0	0/0
-anionic	3	0	0/2	0	0/1
-cationic	7	0	0/1	33	2/6
pH – Total[7]	18	-	-	39	7/18
- acidic (pH < 7.0)	11	-	-	27	3/11
- basic (pH > 7.0)	7	-	-	57	4/7

[1] EU =- European Union System (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 ICE test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh) as defined in **Appendix B**.
[7] Total number of EU Category R41 substances for which pH information was obtained.

6.2.1 Discordance Among Chemical Classes

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. For the purposes of these analyses, NICEATM considered five substances per chemical class to be the threshold number for consideration, and thus classes represented by fewer than five substances were not considered.

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).

6.2.2 Discordance Among Physical or Chemical Properties of Interest

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.

Exclusion of three discordant classes (i.e., alcohols, surfactants and solids) from the data set resulted in an increased accuracy (from 83% to 92%), a decreased false positive rate (from 8% to 6%) and a decreased false negative rate (from 50% to 29%).

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 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]) (**Table 6-2**)