Extension of the performance statistics of defined approaches to distinguish between the three UN GHS categories for eye hazard identification and beyond

Poster #1037

Els Adriaens[1], Takayuki Abo[2], Nathalie Alépée[3], Dan Bagley[4], Jalila Hibatallah[5], Karsten R Mewes[6], Arianna Giusti[7]

[1] Adriaens Consulting bvba, [2] Kao Corporation, [3] L'Oréal Research & Innovation, [4] Colgate-Palmolive Co, [5] Chanel Parfums Beauté, [6] Henkel AG & Co. KGaA, [7] Cosmetics Europe - The Personal Care Association

Introduction

In recent decades, considerable progress has been made in the field of alternatives to animal testing to assess the safety of chemicals. In order to determine the performance of these methods, the results are compared with reference data (animal or human) based on a confusion matrix, a cross table that reports the predicted classes against the reference classes. In case of a binary response, sensitivity and specificity are widely used statistical measures of performance. In case of a multi-class response (three or more categories), sensitivity and specificity can be calculated based on a one-vs-all approach and provide information on the class-specific performance.

Recently, Cosmetics Europe developed two defined approached (DAs) for eye hazard identification that distinguish between the three UN GHS categories, including Cat. 2 defined as chemicals causing reversible effects on the eye (Alépée et al., 2019a, 2019b). The overall performance of the DAs was assessed based on 3x3 confusion matrices. To evaluate the class-specific performance, the 3x3 matrix is converted into three 2x2 matrices and sensitivity, specificity, and balanced accuracy are calculated for each matrix separately.

Evaluation of the predictive capacity for two categories (binary response)

Several *in vitro* test methods have been OECD-adopted to identify chemicals causing serious eye damage (Cat.1), or to identify chemicals not requiring classification and labelling for eye irritation or serious eye damage (No Cat.). Their predictive capacities (accuracy, sensitivity, false negatives (FN), specificity, and false positives (FP)) are generally computed based on a 2x2 confusion matrix (Table 1).

Table 1. *2x2 confusion matrix used for assessing the performance of test methods that can be used to identify chemicals inducing serious eye damage [UN GHS Cat. 1; left table] or to identify chemicals not requiring classification and labelling for eye irritation or serious eye damage [UN GHS No Cat., right table]*

		Actual UN GHS categories	
		in vivo Cat. 1	*in vivo* Cat. 2/ No Cat.
Pred. Cat.	Cat. 1	TP	FP
	Cat. 2 + No Cat.	FN	TN

		Actual UN GHS categories	
		in vivo Cat. 1/Cat. 2	*in vivo* No Cat.
Pred. Cat.	Cat. 1 + Cat. 2	TP	FP
	No Cat.	FN	TN

Accuracy = (TP+TN)/(TP+FN+FP+TN), Sensitivity (Se) = TP/(TP+FN), Specificity (Sp) = TN/(FP+TN), and Balanced Accuracy (BA) = (Se + Sp)/2

Evaluation of the predictive capacity for the three UN GHS categories (multi-class response)

The confusion matrix (Table 2) shows the agreement in predictions between the reference test (Draize eye test) and a defined approach (DA) with a total number of 94 liquid reference chemicals (DAL). Detailed information on the DA is presented in poster # 1043 (DAL-1 with Validated Reference Method 1: VRM1).

The 3x3 matrix provides complete information on the proportion of correct (cells A, E, I; green background), under- (cells B, C, F; yellow background) and over- (cells D, G, H; blue background) predictions for each UN GHS category.

Since the dataset is imbalanced (17 Cat. 1, 22 Cat. 22, and 55 No Cat.) reporting accuracy is inappropriate therefore the balanced accuracy is reported, this is the average of the proportion of correct predictions.

Note that for a single chemical multiple results were available for the different *in vitro* test methods (sometimes resulting in different predictions) and therefore a weighted calculation approach was used so that each chemical has the same weight of 1 (sum of all fractions = 1, this is illustrated in the Figure).

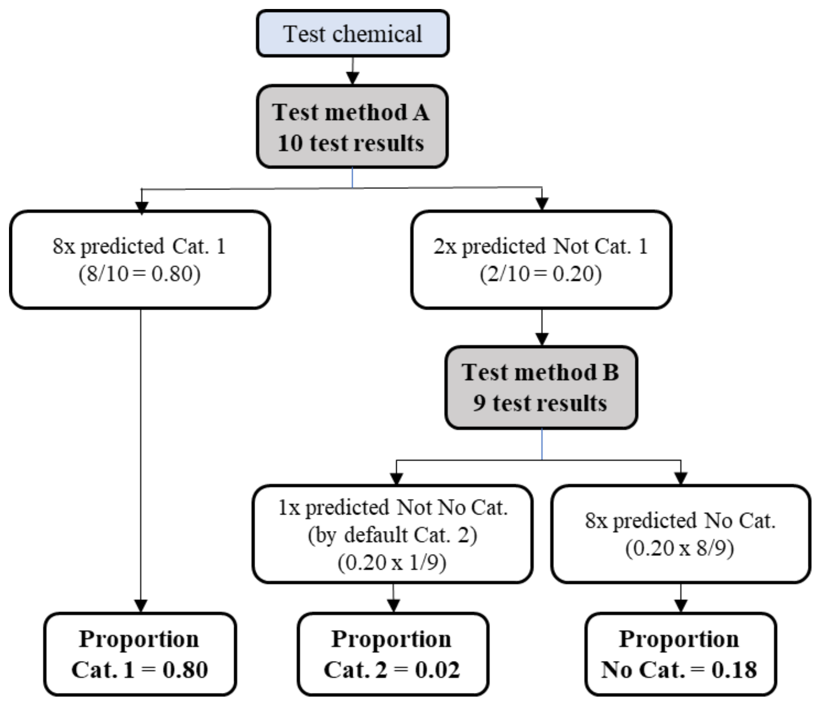

Table 2. *3x3 confusion matrix showing the data of DAL-1 with VRM1 (see poster # 1043)*

		Actual UN GHS categories - reference test		
		in vivo Cat. 1	*in vivo* Cat. 2	*in vivo* No Cat.
Predicted categories **DAL-1 with VRM1**	Cat. 1	**A = 13.0** 76.5% (13.0/17)	**D = 6.0** 27.3% (6.0/22)	**G = 0.0** 0% (0.0/55)
	Cat. 2	**B = 4.0** 23.5% (4.0/17)	**E = 13.0** 59.1% (13.0/22)	**H = 15.3** 27.9% (15.3/55)
	No Cat.	**C = 0.0** 0.0% (0.0/17)	**F = 3.0** 13.6% (3.0/22)	**I = 39.7** 72.1% (39.7/55)
	Total = 94	A+B+C =17	D+E+F =22	G+H+I =55

Accuracy = (A+E+I)/Total, Balanced Accuracy (BA) = [A/(A+B+C)+E/(D+E+F)+I/(G+H+I)]/3

69.2% balanced accuracy

Performance metrics per UN GHS category (class-specific performance)

Method

For the class specific performance metrics of each UN GHS category, the 3x3 matrix is converted into three 2x2 matrices (Table 3). Based on these matrices the sensitivity and specificity for each UN GHS category is calculated Table 4).

Note on terminology

TP and TN are related to the statistical definition and not the biological effect.

For example, the class-specific performance of UN GHS No Cat. (Table 4, last 2 columns) is calculated based on a 2x2 matrix which consists of *in vivo* No Cat. (true within class) and all other classes (Cat. 1 and Cat. 2, true outside class). As such, sensitivity (true positive rate) is defined as the proportion of correct predictions within the class (e.g. No Cat.), specificity (true negative rate) is defined as the proportion of correct predictions outside the class (e.g. Cat.1 and Cat. 2), and balanced accuracy is the average of sensitivity and specificity.

Interpretation Class-specific performance metrics

Table 5 shows the class-specific performance metrics for DAL-1 with VRM1. The within and outside class performance of Cat. 1 and No Cat. results in a balanced accuracy which ranges from 82.2-84.4%. About 59% of the *in vivo* Cat. 2 liquids were correctly identified and about 73% of the *in vivo* Cat. 1 and No Cat. liquids were correctly identified as not Cat. 2 resulting in a balanced accuracy of 66.1%.

Table 3. *Class-specific performance metrics: conversion of the 3x3 matrix (Table 2) into three 2x2 matrices*

	Cat. 1 versus remaining categories			Cat. 2 versus remaining categories			No Cat. versus remaining categories		
		in vivo Cat. 1	*in vivo* Cat. 2 + No Cat.		*in vivo* Cat. 2	*in vivo* Cat. 1 + No Cat.		*in vivo* No Cat.	*in vivo* Cat. 1 + Cat. 2
Predicted categories	Cat. 1	TP (A = 13.0)	FP (D+G = 6.0)	Cat. 2	TP (E = 13.0)	FP (B+H = 19.3)	No Cat.	TP (I = 39.7)	FP (C+F = 3.0)
	Cat. 2 + No Cat.	FN (B+C = 4.0)	TN (E+F+H+I = 71.0)	Cat. 1 + No Cat.	FN (D+F = 9.0)	TN (A+C+G+I = 52.7)	Cat. 1 + Cat. 2	FN (G+H = 15.3)	TN (A+B+D+E = 36.0)

Table 4. *Calculation of the class-specific performance metrics*

Class specific Statistic	Cat. 1 versus remaining categories		Cat. 2 versus remaining categories		No Cat. versus remaining categories	
	Categories	Formula	Categories	Formula	Categories	Formula
Sensitivity (True within class)	True Cat. 1	$\frac{A}{A+B+C}$	True Cat. 2	$\frac{E}{D+E+F}$	True No Cat.	$\frac{I}{G+H+I}$
Specificity (True outside class)	True Cat. 2 and No Cat.	$\frac{E+F+H+I}{D+E+F+G+H+I}$	True Cat. 1 and No Cat.	$\frac{A+C+G+I}{A+B+C+G+H+I}$	True Cat. 1 and Cat. 2	$\frac{A+B+D+E}{A+B+C+D+E+F}$
Balanced accuracy		(sensitivity + specificity)/2		(sensitivity + specificity)/2		(sensitivity + specificity)/2

Table 5. *Performance metrics per UN GHS category for DAL-1 with VRM1*

Statistic	Cat. 1	Cat. 2	No Cat.
Sensitivity (True within class)	13/17*100 = 76.5	13/22*100 = 59.1	39.7/55*100 = 72.1
Specificity (True outside class)	71/77*100 = 92.2	52.7/72*100 = 73.2	36/39*100 = 92.3
Balanced accuracy	(76.5+92.2)/2 = 84.4	(59.1+73.2)/2 = 66.1	(72.1+92.3)/2 = 82.2

Conclusion

Performance metrics such as sensitivity, specificity, and accuracy are well known to assess the predictive capacity of alternative methods based on a binary response.

This is the first time that this approach is applied to eye hazard identification considering the three UN GHS categories. While the focus is on the performance of alternative~~s~~ methods/defined approaches for eye hazard identification, the approach shown here with a numerical example is also applicable to other domains of hazard identification and to more than three categories.

The computational approach presented here was discussed with the OECD expert group on Skin & Eye Irritation/ Phototoxicity and generally accepted for assessing the performance of DAs and/or stand-alone test method.

Cosmetics Europe
the personal care association

References

Alépée N at al., 2019a. Toxicol In Vitro; 59:100-114. doi: 10.1016/j.tiv.2019.04.011.
Alépée N at al., 2019b. Toxicol In Vitro; 57:154-163. doi: 10.1016/j.tiv.2019.02.019.