



Assessment of a defined approach based on a stacking prediction model to identify skin sensitization hazard

Fleur Tourneix*, Nathalie Alépée, Ann Detroyer, Joan Eilstein, Silvia Martinuzzi Teissier, Laurent Nardelli, Hicham Noçairi, Thierry Pauloin, Cécile Piroird, Aurélia Del Bufalo

L'Oréal Research & Innovation, Aulnay-sous-Bois, France

ARTICLE INFO

Keywords:

Skin sensitization
In vitro
In silico
Defined approach
Hazard identification
Risk assessment

ABSTRACT

Skin sensitization is an important toxicological endpoint in the safety assessment of chemicals and cosmetic ingredients. Driven by ethical considerations and European Union (EU) legislation, its assessment has progressed from the reliance on traditional animal models to the use of non-animal test methods. It is generally accepted that the assessment of skin sensitization requires the integration of various non-animal test methods in defined approaches (DAs), to cover the mechanistic key events of the adverse outcomes pathway (AOP) (OECD, 2014). Several case studies for DAs predicting skin sensitization hazard or potency have been submitted to the OECD, including a stacking meta-model developed by L'Oréal Research & Innovation (OECD, 2017b; Del Bufalo et al., 2018; Noçairi et al., 2016). The present study evaluated the predictive performance of the defined approach integrating a stacking meta-model incorporating *in silico*, *in chemico* and *in vitro* assays, using the Cosmetics Europe (CE) skin sensitization database. Based on the optimized prediction cut-offs, the defined approach provided a hazard prediction for 97 chemicals with a sensitivity of 91%, a specificity of 76% and accuracy of 86% (kappa of 0.67) against human skin sensitization hazard data and a sensitivity of 85%, specificity of 91% and accuracy of 87% (kappa of 0.67) against Local Lymph Node Assay (LLNA) hazard data. A comparison of the *in vivo* LLNA with human hazard data for the same 97 chemicals showed a sensitivity of 92%, specificity of 51% and accuracy of 78% (kappa of 0.48). Thus, the defined approach showed a higher degree of concordance, as compared to the LLNA for predicting human skin sensitization hazard. Moreover, a comparison with the six DAs selected for evaluation of their predictivity in the study by Kleinstreuer et al. (2018) showed a similar high accuracy of 86% for 97 overlapping chemicals. The next step will be an independent evaluation of the DA for its integration in the performances based test guidelines (PBTG) for skin sensitization.

1. Introduction

Skin sensitization is an important toxicological endpoint in the safety assessment of chemicals and cosmetic ingredients. For several decades, skin sensitization potential has been assessed using animal models such as the guinea pig maximization test (Magnusson and Kligman, 1969), Buehler test (Buehler, 1965) or the more refined

murine local lymph node assay (LLNA) (Gerberick et al., 2000). However, ethical considerations, the ban on animal testing for cosmetic purposes (EU, 2003) and the REACH regulation (ECHA, 2017) have urged a transition towards the use of non-animal test methods and strategies (Reisinger et al., 2015; Basketter et al., 2015; Ezendam et al., 2016).

The development of non-animal test methods and strategies has

Abbreviations: AOP, Adverse Outcomes Pathway; CE, Cosmetics Europe; CLP, Classification Labelling Packaging; DAs, Defined Approaches; DMSO, Dimethyl Sulfoxide; DPRA, Direct Peptide Reactivity Assay; ECHA, European Chemicals Agency; EU, European Union; FN, False Negative; FP, False Positive; h-CLAT, human Cell Line Activation Test; HRIPT, The Human Repeat Insult Patch Test; IATA, The Integrated Approaches to Testing and Assessment; ICCVAM, The Interagency Coordinating Committee on the Validation of Alternative Methods; IL-8 Luc, Interleukin-8 Reporter Gene assay; INC, Inconclusive; ITS, The Integrated Testing Strategy; KE, Key Event; LLNA, Local Lymph Node Assay; N, Negative; NA, Not Applicable; NICEATM, The National Toxicology Program Interagency Center for Evaluation of Alternative Toxicological Methods; NOEL, No Observed Effect Level; NR, Not Reactive; NS, Non Sensitizer; OECD, The Organization for Economic Co-operation and Development; P, Positive; PBTG, The Performances Based Test Guidelines; QSAR, Quantitative Structure–Activity Relationship models; R, Reactive; REACH, Registration, Evaluation, Authorization and Restriction of Chemicals; S, Sensitizer; SVM, Support Vector Machine; TIMES-SS, Tissue Metabolism Simulator skin sensitization model; U-SENS™, U937 cell line activation Test; VP, Vapor Pressure

* Corresponding author at: L'Oréal Research & Innovation, 1 avenue Eugène Schueller, 93600 Aulnay Sous Bois, France.

E-mail address: fleur.tourneix@rd.loreal.com (F. Tourneix).

<https://doi.org/10.1016/j.tiv.2019.05.008>

Received 22 November 2018; Received in revised form 10 May 2019; Accepted 13 May 2019

Available online 14 May 2019

0887-2333/ © 2019 Elsevier Ltd. All rights reserved.

been facilitated by the documentation of the skin sensitization adverse outcome pathway (AOP), which identified four mechanistic key events (KEs): 1) the covalent binding of haptens to skin proteins, 2) keratinocyte activation, 3) dendritic cell activation and 4) T-cell proliferation (OECD, 2014). At present, the first three AOP KEs can be assessed by formally validated and OECD accepted non-animal test methods including: the Direct Peptide Reactivity Assay (DPRA) (KE 1), KeratinoSens™, Lu-SENS (KE 2), human Cell Line Activation Test (h-CLAT), U937 cell line activation Test (U-SENS™), and Interleukin-8 Reporter Gene Assay (IL-8 Luc assay) (KE 3), (OECD TG 442C, 442D, 442E) (OECD, 2015; OECD, 2018a; OECD, 2018b), whereas methods to assess T-cell proliferation (KE 4) still require further optimization (van Vliet et al., 2017).

It is generally accepted that the assessment of skin sensitization requires the integration of information from various non-animal test methods in Defined Approaches (DAs) to cover the AOP KEs (Jowsey et al., 2006; Vandebriel and van Loveren, 2010; Reisinger et al., 2015; Ezendam et al., 2016; Roberts and Patlewicz, 2017). Consequently, a number of case studies of DAs integrating *in chemico*, *in silico* and *in vitro* data for predicting skin sensitization hazard or potency have been submitted to the OECD according to a guidance document on reporting (OECD, 2017a, 2017b). Cosmetics Europe and the National Toxicology Program Interagency Center for Evaluation of Alternative Toxicological Methods (NICEATM) have evaluated the performance of six submitted DAs using a database with non-animal test method, LLNA and human skin sensitization data (Hoffmann et al., 2018). The results of this evaluation demonstrated that several DAs had a comparable or superior predictive capacity to the LLNA in predicting human skin sensitization hazard and potency (Kleinstreuer et al., 2018).

The defined approach integrating a stacking meta-model for skin sensitization hazard identification developed by L'Oréal Research & Innovation (Del Bufalo et al., 2018; Noçairi et al., 2016,) was one of the DA case studies submitted to the OECD (OECD, 2017b). The approach includes a statistical probabilistic model that combines 7 input variables including physicochemical properties (pH, volatility), *in silico* (TIMES, ToxTree) and *in vitro* (DPRA, KeratinoSens™, U-SENS™) parameters covering AOP KEs 1 to 3. All variables are analyzed through 5 statistical models (Boosting, Naive Bayes, Support Vector Machine, Sparse PLS-DA and Expert Scoring (Gomes et al., 2014). The outputs of each model are then linearly combined in a stacking model that provides a final probability for the test chemical to be a sensitizer allowing its hazard prediction (Sensitizer (S) / non sensitizer (NS)) (Noçairi et al., 2016). A key benefit of the stacking approach is that the distribution of the probabilities becomes more bimodal compared to the individual models, which enhances the confidence in the hazard predictions. The defined approach was built using a dataset of 165 chemicals with LLNA *in vivo* hazard data, for which it showed an overall accuracy of 93% (Del Bufalo et al., 2018).

As the stacking meta-model was not available for an external evaluation at the time of the Cosmetics Europe study (Kleinstreuer et al., 2018), this DA was only listed in the comparative analysis of all DAs based on qualitative evaluation criteria, but could not be part of the 6 DAs that were evaluated for their predictive performance on the CE database. Recently, the model was described by Del Bufalo et al., 2018 and became publically available. We present here the quantitative predictive performance of the defined approach for its ability to discriminate between sensitizers and non-sensitizers using the same CE skin sensitization database (Hoffmann et al., 2018). The final hazard (S/NS) predictions made by the stacking-model were compared to the human and LLNA skin sensitization hazard data to assess its performance in comparison to the *in vivo* “gold standards” and other DAs submitted as case studies to the OECD (OECD, 2017b; Kleinstreuer et al., 2018).

2. Materials and methods

2.1. Data source

The CE skin sensitization database published by Hoffmann et al., 2018 was used as a data source to compile a comprehensive dataset to evaluate the predictive performance of the L'Oréal defined approach. In addition to the *in vitro* test methods data, the database contains curated LLNA and human potency categories skin sensitization reference data by Basketter et al., 2014. Of the 128 chemicals in the database, 105 chemicals fulfilled the data input criteria for the present defined approach. For the remaining 23 chemicals, no U-SENS™ data was available. Based on their human potency class (described in Basketter et al.), the human skin sensitization hazard categories (S/NS) for the 105 chemicals were designated according following rule (suggested by Basketter et al.): chemicals in categories 1–4 were designated as positives (S) and chemicals in categories 5 and 6 were grouped together as negatives (NS). This classification provided a balanced dataset of 67 sensitizers and 38 non-sensitizers. Of the 105 selected chemicals, 31 were overlapping with the learning test set in the L'Oréal internal database used to develop the defined approach integrating a stacking meta-model (Del Bufalo et al., 2018). The details for the 105 chemicals including the input data, LLNA hazard category and human potency category are presented in Table 1.

2.2. Physicochemical parameters

The volatility of the 105 chemicals was determined using the MPBPWIN™ model from the open source Episuite™ software version 4.1 (US EPA, 2012). The model estimates vapor pressure (VP) based on the structure of a given chemical. These VP values were converted into volatility classes according to Spicer, 2002: VP < 10–7 mmHg = non-volatile; VP between 10 and 7 and 10–1 mmHg = semi volatile; VP between 10 and 1 and 380 mmHg = volatile; VP > 380 mmHg = very volatile. The pH values were measured internally by L'Oréal laboratories using a method adapted from OECD Test Guideline No. 122 for the Testing of Chemicals (OECD, 2013). For 3 chemicals, the pH could not be determined and a pH of 7 was attributed (see Table 1).

2.3. In silico data

The *in silico* data was generated using the open source *in silico* prediction software ToxTree (version 2.6.13) and more specifically the module: Skin sensitisation reactivity domains (Ideacon Ltd). When nucleophilic aromatic substitution (S_NAr), bimolecular nucleophilic substitution (S_N2), acyl transfer agent, Michael acceptor and Schiff base formation alerts were identified, the chemical was classified as reactive (R), whereas in the absence of reactivity alerts, it was classified as non-reactive (NR), (Aptula and Roberts, 2006). Moreover, the *in silico* Tissue MEtabolism Simulator skin sensitization model (TIMES-SS version 2.27.18.3) (OASIS-LMC) was used to analyze the parent molecular structure as well as its simulated metabolites to predict skin sensitization potency based on structural alerts and 3D-QSARs. The final *in silico* prediction was based on the most potent structure among the parent chemical and its metabolites and was expressed as non-sensitizer, weak or strong sensitizer (Patlewicz et al., 2007). For the stacking meta-model input, the weak and strong sensitizers were grouped together as sensitizer.

2.4. In vitro data

The *in vitro* data including DPRA, KeratinoSens™ and U-SENS™ data were directly obtained from the CE database (Hoffmann et al., 2018).

Table 1
Dataset of 105 chemicals from the Cosmetics Europe' skin sensitization database.

Chemical name	CAS	L'Oreal (Learning/Test set)	pH	Volatility	Toxtree	TIMES- SS	DPRA	Keratino- Sens™	U-SENS™	Stacking Score ¹	Human Pot. Cat ²	Human Hazard Cat ³	LLNA EC3 median ⁴	LLNA Hazard Cat ⁵	Composite Hazard Cat ⁶
1,4-Phenylenediamine	106-50-3	L	9	Semi_volatile	R	P	R	P	P	93.7	1	S	0.15	S	S
tetrachlorosalicylanilide	1154-59-2	no	6.6	Non_volatile	R	P	R	P	P	83.8	1	S	0.04	S	S
methylisothiazolinone (act.9.7%)	2682-20-4	no	4.5	Semi_volatile	R	P	R	P	P	95.1	1	S	1.9	S	S
dimethyl fumarate	624-49-7	no	7.1	Semi_volatile	R	P	R	P	P	91.6	1	S	0.35	S	S
diphenylcyclopropanone	886-38-4	no	7.8	Semi_volatile	R	P	R	P	P	91.5	1	S	0.05	S	S
2,4-dinitrochlorobenzen DNCB	97-00-7	L	8.7	Semi_volatile	R	P	R	P	P	93.7	1	S	0.06	S	S
cinnamic aldehyde	104-55-2	no	2.7	Semi_volatile	R	P	R	P	P	95.2	2	S	1.15	S	S
glyoxal (act. 40%)	107-22-2	L	5	Very_volatile	NR	P	R	P	P	93.8	2	S	0.6	S	S
3-dimethylaminopropylamine	109-55-7	L	11.5	Very_volatile	R	INC	NR	P	P	81.6	2	S	2.2	S	S
methyl heptine carbonate	111-12-6	T	5.4	Very_volatile	R	P	R	P	P	95.6	2	S	0.5	S	S
glutaraldehyde (act. 50%)	111-30-8	L	4.1	Very_volatile	R	P	R	P	P	95.6	2	S	0.09	S	S
lauryl gallate	1166-52-5	no	8.0	Non_volatile	R	P	R	P	P	82.7	2	S	0.3	S	S
propyl gallate	121-79-9	T	6	Semi_volatile	R	P	R	P	P	91.7	2	S	0.32	S	S
diethyl maleate	141-05-9	L	8.8	Very_volatile	R	P	R	P	P	94.5	2	S	4.7	S	S
1,2-benzisothiazolin-3-one	2634-33-5	L	6.3	Semi_volatile	R	P	R	P	P	91.6	2	S	4.8	S	S
lyral	31906-04-4	no	7.1	Semi_volatile	R	P	R	P	P	91.6	2	S	17.1	S	S
methyldibromo glutaronitrile	35691-65-7	L	8	Semi_volatile	R	P	R	P	P	91.5	2	S	0.9	S	S
formaldehyde (act. 37%)	50-00-0	T	8	Very_volatile	R	P	R	P	P	92.4	2	S	0.85	S	S
2-nitro-1,4-phenylenediamine	5307-14-2	L	8.2	Semi_volatile	R	P	R	P	P	91.4	2	S	3.95	S	S
toluene diamine sulphate	615-50-9	L	2.5	Semi_volatile	R	P	R	P	P	95.2	2	S	0.4	S	S
2-aminophenol	95-55-6	no	8.4	Semi_volatile	R	P	R	P	P	91.4	2	S	0.45	S	S
thioglycerol	96-27-5	L	6.9	Semi_volatile	NR	P	R	N	P	71.4	2	S	3.6	S	S
iso Eugenol	97-54-1	L	7.3	Semi_volatile	R	P	R	P	P	91.5	2	S	1.35	S	S
methyl 2-nonynoate (Methyl octine carbonate)	111-80-8	no	6.7	Semi_volatile	R	P	R	P	P	91.6	2	S	2.5	S	S
trans-2-hexenal	6728-26-3	L	2.9	Very_volatile	R	P	R	P	P	95.5	2	S	4.05	S	S
2-methoxy-4-methylphenol	93-51-6	no	7.5	Semi_volatile	R	P	NR	N	P	70.4	2	S	5.6	S	S
cinnamic alcohol	104-54-1	no	4.1	Semi_volatile	R	P	R	P	P	95.1	3	S	20.6	S	S
ethylene diamine	107-15-3	L	11.4	Very_volatile	R	INC	NR	P	P	81.7	3	S	3.4	S	S
1,4-dihydroquinone	123-31-9	L	7.2	Semi_volatile	R	P	R	P	P	91.6	3	S	0.12	S	S
5-methyl-2,3-hexanedione	13706-86-0	no	3.7	Very_volatile	R	P	R	P	P	95.6	3	S	26	S	S
tetramethylthiuramdisulfide	137-26-8	L	8.7	Semi_volatile	NR	P	R	P	P	87.1	3	S	2.93	S	S
2-mercaptobenzothiazole	149-30-4	T	5.4	Semi_volatile	R	INC	R	P	P	93.5	3	S	2.6	S	S
bisphenol A-diglycidyl ether	1675-54-3	no	7.5	Semi_volatile	R	P	R	P	P	91.5	3	S	1.5	S	S
butyl glycidyl ether	2426-08-6	T	8.2	Very_volatile	R	P	R	P	P	92.4	3	S	30.9	S	S
glyceryl monothioglycolate	30618-84-9	no	3.3	NA	NA	NA	R	P	P	84.4	3	S	4.7	S	S
imidazolidinyl urea	39236-46-9	no	7.2	Non_volatile	R	P	R	P	P	83.4	3	S	24	S	S
farnesol	4602-84-0	no	6.8	Semi_volatile	R	P	NR	P	P	83.8	3	S	4.8	S	S
chlorpromazine	50-53-3	no	8.0	Semi_volatile	R	P	NA	N	P	80.2	3	S	1	S	S
abietic acid	514-10-3	no	5.4	Semi_volatile	NR	P	R	P	P	92.3	3	S	11	S	S
citral	5392-40-5	no	3.5	Semi_volatile	R	P	R	P	P	95.2	3	S	5.7	S	S
metol	55-55-0	L	7.7	Semi_volatile	R	P	R	P	P	91.5	3	S	0.8	S	S
phenyl benzoate	93-99-2	no	7.4	Semi_volatile	R	P	R	N	P	87.8	3	S	18.33	S	S

(continued on next page)

Table 1 (continued)

benzoyl peroxide	94-36-0	no	7.2	Semi_volatile	R	P	R	N	P	87.8	3	S	0.06	S	S
eugenol	97-53-0	no	7.5	Semi_volatile	R	P	R	N	P	87.8	3	S	12.9	S	S
phenylacetaldehyde	122-78-1	L	3.4	Very_volatile	R	P	R	P	P	95.6	3	S	4.7	S	S
penicillin G	61-33-6	no	6.9	Non_volatile	R	P	R	N	N	19.9	3	S	30	S	S
2-Hydroxyethyl acrylate	818-61-1	no	7.3	Semi_volatile	R	INC	R	P	P	89.8	3	S	1.56	S	S
coumarin	91-64-5	L	8.9	Semi_volatile	R	N	NR	P	P	14.8	3	S	100	NS	S
amylcinnamyl alcohol	101-85-9	no	7.7	Semi_volatile	R	P	NR	N	P	70.2	4	S	100	NS	S
geraniol	106-24-1	no	8.3	Semi_volatile	R	P	NR	P	P	82.8	4	S	23.2	S	S
hydroxycitronellal	107-75-5	T	3.4	Semi_volatile	R	P	R	P	P	95.2	4	S	22.2	S	S
resorcinol	108-46-3	L	6.3	Semi_volatile	R	P	NR	N	P	71.6	4	S	6.3	S	S
amyl cinnamic aldehyde	122-40-7	no	6.9	Semi_volatile	R	P	NR	P	P	83.8	4	S	11.2	S	S
ethyl acrylate	140-88-5	no	7.0	Very_volatile	R	INC	R	P	P	90	4	S	32.75	S	S
iodopropenyl butylcarbamate	55406-53-6	no	6.7	Semi_volatile	R	INC	R	P	P	88.3	4	S	0.9	S	S
carvone	6485-40-1	no	6.9	Very_volatile	R	P	R	P	P	92.5	4	S	12.9	S	S
lillial	80-54-6	no	7.0*	Semi_volatile	R	P	R	N	P	87.9	4	S	11.35	S	S
ethyleneglycol dimethacrylate	97-90-5	T	8.3	Very_volatile	R	P	R	P	P	92.4	4	S	35	S	S
α -methyl cinnamic aldehyde	101-39-3	L	8.6	Semi_volatile	R	P	R	P	P	93.8	4	S	4.5	S	S
benzyl cinnamate	103-41-3	no	8.1	Semi_volatile	R	P	NR	P	P	83	4	S	18.4	S	S
neomycin sulphate	1405-10-3	no	10.0	Non_volatile	R	N	NA	N	N	6.6	4	S	100	NS	S
hexyl salicylate	6259-76-3	no	6.6	Semi_volatile	NR	N	NR	N	P	5.6	4	S	0.18	S	S
kanamycin sulfate	70560-51-9	no	6.3	Non_volatile	R	N	R	N	N	5.7	4	S	100	NS	S
benzocaine	94-09-7	no	8.5	Semi_volatile	NR	N	R	P	P	11.9	4	S	100	NS	S
aniline	62-53-3	no	8.3	Very_volatile	NR	P	NR	N	P	55	4	S	89	S	S
linalool	78-70-6	no	8.7	Semi_volatile	NR	P	NR	N	P	48.6	4	S	30.4	S	S
methylmethacrylate	80-62-6	no	5.9	Very_volatile	R	N	R	P	P	30.7	4	S	75	S	S
propyl paraben	94-13-3	L	7.0	Semi_volatile	NR	N	NR	P	P	7.1	5	NS	100	NS	NS
anisyl alcohol	105-13-5	no	7.8	Semi_volatile	NR	P	R	N	N	28	5	NS	5.91	S	S
isopropyl myristate	110-27-0	T	8.4	Semi_volatile	NR	N	NR	N	N	3.4	5	NS	44	S	S
pyridine	110-86-1	no	5.4	Very_volatile	NR	N	NR	N	N	6	5	NS	72	S	S
methyl salicylate	119-36-8	T	7.8	Semi_volatile	NR	N	NR	N	N	3.4	5	NS	100	NS	NS
phenoxyethanol	122-99-6	no	7.7	Semi_volatile	NR	N	R	N	P	9.5	5	NS	100	NS	NS
4-aminobenzoic acid	150-13-0	no	3.7	Semi_volatile	NR	N	NR	N	N	4.2	5	NS	100	NS	NS
hydrocortisone	50-23-7	L	8.3	Non_volatile	R	INC	R	P	N	17.1	5	NS	100	NS	NS
propylene glycol	57-55-6	L	9.1	Very_volatile	NR	N	NR	N	N	6.4	5	NS	100	NS	NS
isopropanol	67-63-0	L	8.8	Very_volatile	NR	N	NR	N	N	6.4	5	NS	100	NS	NS
benzalkonium chloride	8001-54-5	L	6.3	NA	NA	NA	NR	N	P	20	5	NS	0.1	S	S
hexyl cinnamic aldehyde	101-86-0	no	6.7	Semi_volatile	R	P	NR	P	P	83.9	5	NS	10.6	S	S
benzyl salicylate	118-58-1	no	5.4	Semi_volatile	R	P	NR	P	P	92.3	5	NS	2.85	S	S
benzyl benzoate	120-51-4	T	8.8	Semi_volatile	R	P	NR	P	P	86.9	5	NS	17	S	S
vanillin	121-33-5	no	5.1	Semi_volatile	R	P	NR	N	P	87.9	5	NS	100	NS	NS
pentachlorophenol	87-86-5	no	7.0*	Semi_volatile	R	P	R	N	P	87.9	5	NS	20	S	S
α -iso-methylionone	127-51-5	no	4	Semi_volatile	R	P	NR	N	P	88.3	5	NS	21.8	S	S
benzaldehyde	100-52-7	no	2.9	Very_volatile	R	N	NR	P	P	31.5	5	NS	100	NS	NS
triethanolamine	102-71-6	no	9.8	Semi_volatile	NR	P	NR	N	P	45.6	5	NS	100	NS	NS

(continued on next page)

Table 1 (continued)

anethole	104-46-1	no	7.9	Semi_volatile	R	P	NR	N	P	70	5	NS	2.3	S	S
citronellol	106-22-9	no	8.2	Semi_volatile	NR	P	R	N	P	70	5	NS	43.5	S	S
diethanolamine	111-42-2	no	10.3	Semi_volatile	NR	P	NR	N	P	44.3	5	NS	40	S	S
(R)-(+)-limonene	5989-27-5	no	7.3	Very_volatile	NR	P	NR	N	P	56.7	5	NS	52.5	S	S
phenol	108-95-2	no	7.1	Very_volatile	NR	N	R	N	P	14.3	6	NS	100	NS	NS
hexane	110-54-3	L	8.9	Very_volatile	NR	N	NR	N	N	6.4	6	NS	100	NS	NS
octanoic acid	124-07-2	L	3.7	Semi_volatile	NR	N	NR	N	P	10.2	6	NS	100	NS	NS
xylene	1330-20-7	no	7.7	Semi_volatile	NR	N	NR	N	N	3.4	6	NS	95.8	S	NS
diethyl toluamide	134-62-3	no	7.4	Semi_volatile	NR	N	NR	N	P	5.5	6	NS	100	NS	NS
sodium lauryl sulfate	151-21-3	T	8.3	Non_volatile	R	N	NR	N	P	7.2	6	NS	3.55	S	NS
dextran	3371-50-4	no	7.0*	Non_volatile	R	INC	NA	N	N	6.9	6	NS	100	NS	NS
lactic acid	50-21-5	T	2.4	Semi_volatile	NR	N	NR	N	N	4.3	6	NS	100	NS	NS
glycerol/glycerin	56-81-5	L	8.3	Semi_volatile	NR	N	NR	N	N	3.4	6	NS	100	NS	NS
tocopherol	59-02-9	no	7.8	Non_volatile	NR	P	NR	P	N	6.3	6	NS	8.1	S	NS
DMSO	67-68-5	no	7.5	Very_volatile	NR	N	NR	N	N	4.8	6	NS	72	S	NS
salicylic acid	69-72-7	L	2.5	Semi_volatile	NR	N	R	N	N	6	6	NS	12.2	S	NS
1-butanol	71-36-3	T	8.2	Very_volatile	NR	N	NR	N	N	4.7	6	NS	100	NS	NS
diethyl phthalate	84-66-2	L	8.4	Semi_volatile	NR	N	NR	N	P	5.4	6	NS	100	NS	NS
Tween 80	9005-65-6	no	6.7	NA	NA	NA	R	P	P	68.4	6	NS	100	NS	NS

NA = Not applicable; R = Reactive; NR = Not reactive; NS = Non-sensitizer; S = Sensitizer; INC = Inconclusive.

Shading line = false predicted chemical in the defined approach for skin sensitization hazard categorization with regard to human data.

*pH of 7 was attributed. ¹Stacking probability scale is 0–100; ²From Basketter et al., 2014; ³Human potency cat. 1–4 = S and 5–6 = NS; ⁴From Hoffmann et al., 2018;

⁵According to ECETOC, 2003; ⁶Composite ref. for Human potency cat.5 + LLNA neg. = NS and cat.5 + LLNA pos. = S.

For the meta-model input, the binary hazard predictions (R/NR or P/N) were used.

2.5. Stacking-meta model

The stacking meta-model runs the physicochemical parameters (pH, volatility), *in silico* (Toxtree, TIMES-SS) and *in vitro* data (DPRA, KeratinoSens™ and U-SENS™) for a test chemical in five different statistical models (Boosting, Naïve Bayes, Support Vector Machine (SVM), Sparse PLS-DA and Expert Scoring (Gomes et al., 2014), each providing a probability of the chemical being a sensitizer. These intermediate probabilities are then linearly combined in the stacking meta-model which determines a final overall probability (a stacking score) of the chemical being a sensitizer (Wolpert, 1992; Gomes et al., 2012; Noçairi et al., 2016). The stacking meta-model from L'Oréal Research & Innovation is available online as an open software tool at www.liebertpub.com/aivt.

2.6. Predictive performance assessment

The predictive performance for the DA compared both to LLNA and human reference data was quantitatively assessed with the Cooper's statistics in terms of accuracy (number of correctly classified substances over total number of substances assessed), sensitivity (the proportion of sensitizers that are correctly identified as such), specificity (the proportion of non-sensitizers that are correctly identified as such) and the Kappa coefficient is calculated to test interrater reliability. The Cohen's Kappa can range from −1 to +1. It suggested that a Kappa result be interpreted as follows: values ≤0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate,

0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement (Landis and Koch, 1977). The false negatives (FN), and false positives (FP) was also highlighted.

3. Results

A comprehensive dataset for 105 test chemicals from the CE skin sensitization database was analyzed by the defined approach using the stacking meta-model to evaluate its predictive performance for human and *in vivo* LLNA skin sensitization hazard data. The output parameter given by the stacking meta model is a probability to be a sensitizer, the cut-offs have been made in order to calculate the sensitivity capacity. This DA allows to move the cut-offs, depending on the sensitivity rate and end-users application of this defined approach. To be able to calculate Coopers' statistics, the previously optimized probability cut-offs to predict a chemical as a sensitizer or a non-sensitizer were applied as reported in Del Bufalo et al., 2018. A test chemical was predicted as a sensitizer (S) if the stacking score was ≥70% and as a non-sensitizer (NS) if the stacking score was ≤30%. When the stacking score was > 30% and < 70%, the uncertainty of the prediction was considered too high and no prediction could be made. The call was defined as inconclusive (INC). The data analysis results showed that, based on these probability cut-offs, the meta-model provided a final hazard prediction (S/NS) for 97 chemicals and the outcome was inconclusive for 8 chemicals (Table 1 and Fig. 1).

A comparison of the present defined approach hazard predictions with the designated human hazard categories showed a sensitivity of 91%, specificity of 76% and accuracy of 86% (kappa of 0.67) (Table 2). In total, 14 chemicals were mis-predicted (Table 3), including 8 false positives (Hexyl cinnamic aldehyde, Anethole, Citronellol, Benzyl

salicylate, Benzyl benzoate, Vanillin, Pentachlorophenol, α -iso-Methylionone) and 6 false negatives (Penicillin G, Coumarin, Neomycin sulphate, Hexyl salicylate, Kanamycin sulfate, Benzocaine). When compared with the LLNA hazard data, the defined approach predictions showed a sensitivity of 85%, specificity of 91% and accuracy of 87% (kappa of 0.67) (Table 2). In this case, 13 chemicals were mis-predicted (Table 4), including 2 false positives (Amylcinnamyl alcohol, Vanillin) and 11 false negatives (Penicillin G, Hexyl salicylate, Anisyl alcohol, Isopropyl myristate, Pyridine, Benzalkonium chloride, Xylene, Sodium lauryl sulfate, Tocopherol, DMSO, Salicylic acid). A comparison between the *in vivo* LLNA and human hazard reference data for the 97 chemicals showed that the LLNA had a sensitivity of 92%, specificity of 51% and accuracy of 78% (kappa of 0.48) for human skin sensitization hazard (Table 2). The LLNA showed 23 mis-predictions against human hazard data, which included 18 false positives (Anisyl alcohol, Isopropyl myristate, Pyridine, Benzalkonium chloride, Hexyl cinnamic aldehyde, Benzyl salicylate, Benzyl benzoate, Pentachlorophenol, α -iso-Methylionone, Anethole, Citronellol, Diethanolamine, (R)-(+)-Limonene, Xylene, Sodium lauryl sulfate, Tocopherol, DMSO, Salicylic acid) and 5 false negatives (Coumarin, Amylcinnamyl alcohol, Neomycin sulphate, Kanamycin sulfate, Benzocaine). In addition, for the complete dataset of 105 chemicals, the LLNA showed a sensitivity of 92%, specificity of 53% and accuracy of 78% (kappa of 0.49) to predict human skin sensitization hazard.

The 8 chemicals classified as inconclusive included the following sensitizers (Aniline, Linalool, Methylmethacrylate) and non-sensitizers (Benzaldehyde, Triethanolamine, Diethanolamine, (R)-(+)-Limonene, Tween 80) according to their designated human hazard category.

4. Discussion

Driven by ethical considerations and EU legislation (EU, 2003; ECHA, 2017), the assessment of skin sensitization has progressed from the reliance on traditional animal models to the use of non-animal test methods (Vandebriel and van Loveren, 2010; Basketter et al., 2015; Reisinger et al., 2015; Ezendam et al., 2016). To meet the regulatory requirements for skin sensitization assessment, it is generally agreed that information from multiple non-animal test methods representing the AOP KEs needs to be integrated in DAs (Roberts and Patlewicz, 2017). L'Oréal Research & Innovation has developed a defined approach with a stacking meta-model for the prediction of skin sensitization hazard (Del Bufalo et al., 2018). The meta-model input variables include physicochemical properties (pH, volatility), *in silico* (Toxtree, TIMES-SS) and *in vitro* DPRA, KeratinoSens™ and U-SENS™ data for a test chemical, which represent KEs 1–3 of the AOP (OECD, 2014). The meta-model linearly combines the binary probabilities of 5 statistical models to generate a final overall probability or stacking score for the test chemical being a sensitizer (Gomes et al., 2012; Noçairi et al., 2016). The statistical probabilistic stacking meta-model thereby maximizes the use of evidence from five individual predictive models, while minimizing the impact of the limitations of any particular model. The meta-model showed an improved performance and enhanced concordance for predicting LLNA skin sensitization hazard data, when compared to the five individual models (Del Bufalo et al., 2018).

The defined approach with a stacking meta-model has been submitted to the OECD as a DA case study for skin sensitization (OECD, 2017b). Moreover, the meta-model was part of a systematic evaluation study by CE and NICEATM, that qualitatively reviewed all DAs submitted to OECD and assessed their ability to discriminate between sensitizers and non-sensitizers of the available subset of 6 DAs (Kleinstreuer et al., 2018). Since the open source stacking meta-model was not available at the beginning of the external evaluation, the DA was not included in the quantitative predictive performance assessment presented by Kleinstreuer et al. (2018). As a complement to this study, we therefore present here the predictive performance of our defined approach using the same CE skin sensitization database (Hoffmann

et al., 2018). As the ultimate goal of safety assessment is to predict skin sensitization in humans, the human potency data in the CE database was used as the key reference standard for the present evaluation. The defined approach predictions were also compared with LLNA data, as the prediction model of the individual *in vitro* methods (DPRA, KeratinoSens™ and U-SENS™) were developed using the LLNA as a reference (Gerberick et al., 2004; Emter and Ellis, 2010; Piroird et al., 2015; Alépée et al., 2017). The 105 chemicals that fulfilled the meta-model data input criteria were designated to a human skin sensitization hazard category based on their human potency category as defined by Basketter et al., 2014. The chemicals in potency categories 1–4 were designated as sensitizers and categories 5 and 6 were grouped together as non-sensitizers in agreement with Kleinstreuer et al. (2018) and Hoffmann et al. (2018). The grouping of category 5 and 6 was based on the high NOELs ($\geq 10,000 \mu\text{g}/\text{cm}^2$) and very low intrinsic ability of category 5 chemicals to cause skin sensitization (Basketter et al., 2014). The induction of skin sensitization in the general population by category 5 chemicals is extremely rare, which justifies no classification e.g. according to EU CLP (EU, 2008) and their grouping with category 6 chemicals as non-sensitizers. With the grouping of category 5 and 6 chemicals, the dataset comprised 67 sensitizers and 38 non-sensitizers (Table 1). Of the 105 chemicals, 31 were overlapping with the learning test set belonging to 165 chemicals used to develop the current defined approach (Del Bufalo et al., 2018). The present evaluation study included 74 naïve test chemicals not previously being evaluated.

With the probability cut-offs set at $\leq 30\%$ for NS and $\geq 70\%$ for S, the defined approach including the stacking meta-model provided a hazard prediction for 97 chemicals whereas 8 chemicals fell in the inconclusive zone ($30 < p < 70$). These 8 inconclusive included one human category 6 chemical, four human category 5 chemicals and three chemicals of human category 4. The defined approach was therefore able to predict a hazard category for all human potency category 1–2 sensitizers and inconclusive chemicals were mainly weak and extremely rare human sensitizers.

For regulatory purposes the most important issue is to reduce false negatives and therefore to increase sensitivity. For that purpose cut-offs could be set with the following rules: if the $p \leq 10\%$, we conclude NS and if $p \geq 90\%$ we conclude S. With these cut-offs the defined approach's predictive performance for human skin sensitization hazard data show a sensitivity which goes from 91% to 92%, a specificity from 76% to 95%, accuracy from 86% to 93% (kappa from 0.67 to 0.86) and was therefore able to predict a hazard category for all human potency category 1–3 sensitizers and the false negative decrease at three chemicals of human category 4. However, the defined approach provided a hazard prediction for only 60 chemicals whereas 45 chemicals fall in the inconclusive zone ($10 < p < 90$), which could be the main limitation to use these probability cut-offs.

The rate of inconclusive predictions depends on the cut-off values that are chosen. When probability cut-offs are set at < 50 and $\geq 50\%$, the defined approach's predictive performance for human skin sensitization hazard data show a sensitivity which goes from 91% to 88%, a specificity from 76% to 74% and accuracy from 86% to 83% (kappa from 0.67 to 0.62). Also the defined approach's predictive performance for LLNA skin sensitization hazard data show a sensitivity which goes from 85% to 83%, a specificity from 91% to 88% and accuracy from 87% to 84% (kappa from 0.67 to 0.61). So, the defined approach provides a hazard prediction for all 105 chemicals with only a slight reduction in quantitative predictive performance (around 3% of decrease) and concordance levels (decrease of kappa around 0.05). In the previously reported study (Del Bufalo et al., 2018), a range of probability cut-off values was investigated to find the optimal balance between the degree of concordance (kappa value) and rate of inconclusive prediction. When the gap between the probability cut-offs was widened, the degree of concordance increased (up to 0.82), however, this also increased the number of inconclusive predictions (up to 10%). Based on the study results, the ≤ 30 and $\geq 70\%$ cut-off values were identified as

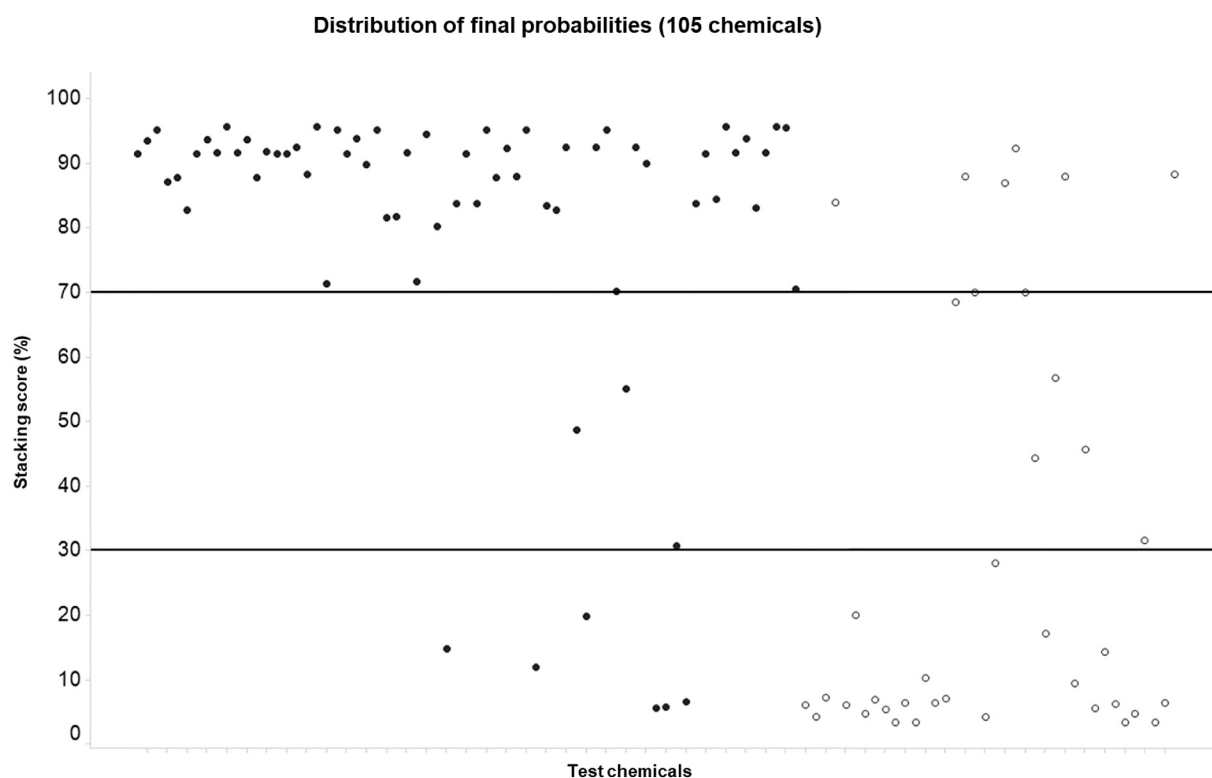


Fig. 1. Distribution of final probabilities predicted by the stacking meta-model for the 105 test chemicals, with the human skin sensitizers shown as “black dots” and the human non-sensitizers shown as “open circles”. The circles in the higher area ($\geq 70\%$) are predicted as sensitizer (S) and in the lower area ($\leq 30\%$) as non-sensitizer (NS). The circles in the middle area ($> 30\%$ and $< 70\%$) are predicted as inconclusive (INC). The order of test chemicals (left to right) is according to Table 1.

optimally balanced to obtain a high degree of concordance (kappa value: 0.82) in the hazard predictions, while limiting the number of inconclusive predictions (17/165 chemicals).

The obtained hazard predictions for the 97 chemicals were compared to human and LLNA hazard reference data in the CE database to evaluate the predictive performance of this defined approach. Results showed that the DA has a similar accuracy for predicting human (86%) as well as LLNA (87%) hazard reference data. According to Table 2, sensitivity is higher for predicting for human compared to LLNA, whereas specificity is higher to predict LLNA. A comparison of the LLNA and human data in the CE database showed that the LLNA has an accuracy of 78% for predicting human hazard, evidencing that the DA outperformed the LLNA, in particular with regard to the identification of non-sensitizers, since the LLNA showed a low specificity of 51% compared to 76% for the DA. Moreover, among the 18 false positives 13 are categorized as human potency category 5, which could be considered as sensitizers since their HRIPT are equivocal (Basketter et al., 2014). The accuracy of the LLNA for human hazard data at 78%

($n = 97$) is comparable to previous studies, reporting 82% ($n = 111$) (Urbisch et al., 2015) and 74% ($n = 128$) (Kleinstreuer et al., 2018). In both these studies, the LLNA also had a low specificity of 64% and 50%, respectively to identify human non-sensitizers.

The evaluation study by Kleinstreuer et al. (2018) showed accuracy levels for the six predictive performance evaluated DAs ranging from 76 to 85% for human hazard and 70 to 84% for LLNA hazard data. The best performing DAs for human hazard were the integrated testing strategy (ITS) developed by Takenouchi et al. (Takenouchi et al., 2015) with a sensitivity of 94%, specificity of 67% and accuracy of 85% and the support vector machine (SVM) model developed by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) (Strickland et al., 2017) with a sensitivity of 86%, specificity of 72% and accuracy of 82% for an identical set of 120 chemicals in the CE database. These DAs also integrate *in vitro*, *in chemico* and *in silico* data inputs covering AOP KE 1–3 as in the present DA. Although the performance of the DA cannot be directly compared with the ITS and SVM published DAs due to some little differences in the chemical sets,

Table 2

Predictive performance of the defined approach using a stacking meta-model for human and LLNA hazard data with ≤ 30 and $\geq 70\%$ probability cut-offs.

Probability cut-offs: ≤ 30 and $\geq 70\%$ applied to the stacking meta-model	Chemicals with hazard prediction	Sensitivity (%)	Specificity (%)	Accuracy (%)	Kappa
DA predictions vs. human	97	91 (58/64)	76 (25/33)	86 (83/97)	0.67
DA predictions vs. LLNA	97	85 (64/75)	91 (20/22)	87 (84/97)	0.67
DA predictions vs. composite reference ^a	97	87 (65/75)	95 (21/22)	89 (86/97)	0.72
LLNA vs. human	97	92 (59/64)	51 (17/33)	78 (76/97)	0.48
LLNA vs. human	105	92 (62/67)	53 (20/38)	78 (82/105)	0.49
DA predictions vs. human	66 ^b	88 (38/43)	65 (15/23)	80 (53/66)	0.55
LLNA vs. human	66 ^b	91 (39/43)	39 (9/23)	73 (48/66)	0.33

^a Composite ref. for Human potency cat.5 + LLNA neg. = NS and cat.5 + LLNA pos. = S.

^b Test set of naïve chemicals not used to develop nor challenge the defined approach for which a hazard prediction was generated using the 30/70% probability cut-offs (Del Bufalo et al., 2018).

Table 3

Table of the false positives (FP) and false negative (FN) in the defined approach for skin sensitization hazard categorization with regard to human data.

Chemical name	CAS	Stacking score ¹	Human pot. cat ²	Human hazard cat ³	LLNA EC3 median ⁴	LLNA hazard cat ⁵	False positive and false negative
Penicillin G	61-33-6	19,9	3	S	30	S	FN
Coumarin	91-64-5	14,8	3	S	100	NS	FN
Neomycin sulphate	1405-10-3	6,6	4	S	100	NS	FN
Hexyl salicylate	6259-76-3	5,6	4	S	0.18	S	FN
Kanamycin sulfate	70,560-51-9	5,7	4	S	100	NS	FN
Benzocaine	94-09-7	11,9	4	S	100	NS	FN
Hexyl cinnamic aldehyde	101-86-0	83,9	5	NS	10.6	S	FP
benzyl salicylate	118-58-1	92,3	5	NS	2.85	S	FP
Benzyl benzoate	120-51-4	86,9	5	NS	17	S	FP
Vanillin	121-33-5	87,9	5	NS	100	NS	FP
Pentachlorophenol	87-86-5	87,9	5	NS	20	S	FP
α -iso-Methylionone	127-51-5	88,3	5	NS	21.8	S	FP
Anethole	104-46-1	70	5	NS	2.3	S	FP
Citronellol	106-22-9	70	5	NS	43.5	S	FP

NS = Non-sensitizer; S = Sensitizer; FN = False negative; FP = False positive.

¹ Stacking probability scale is 0–100.² From Basketter et al., 2014.³ Human potency cat. 1–4 = S and 5–6 = NS.⁴ From Hoffmann et al., 2018.⁵ According to ECETOC, 2003.

the DA showed a similar high accuracy of 86% for 97 overlapping chemicals (30/70% cut-offs) and 83% for 105 overlapping chemicals (50/50% cut-off) for the human data. In our set of 105 there were no < 74 naïve chemicals not used in the learning set of our DA (Del Bufalo et al., 2018). Using the ≤ 30 and $\geq 70\%$ probability cut-offs of the meta-model from the 74 chemicals, 8 fell into the inconclusive. For the 66 chemicals with a hazard prediction, the DA showed a sensitivity of 88%, specificity of 65% and accuracy of 80% (kappa of 0.55) for human data, which is better than the LLNA performance with a sensitivity of 91%, specificity of 39% and accuracy of 73% (kappa of 0.33) (Table 2). This also holds true when the < 50 and $\geq 50\%$ cut-offs are applied providing a hazard prediction for all 74 chemicals.

To further improve the confidence in the DA using a stacking meta-model, the mis-predictions were carefully examined with the objective to understand the potential reasons. In general, the DA did not mis-predict any of the human potency category 1 and 2 (extreme/strong sensitizers) or human potency category 6 (non-sensitizers), which indicates the model is reliable for these chemicals with unequivocal human evidence. As the safety assessment is based on the precautionary principle, the false negative predictions for the human potency of category 3 chemicals penicillin G and coumarin are of most concern.

Penicillin G showed *in silico* positive results, and negative results in the U-SENS™ and KeratinoSens™ and, moreover was predicted as a non-sensitizer in the SENS-IS assay (Hoffmann et al., 2018). Because penicillin is a drug chemical dosed orally or *via* injection, the route of exposure by which it caused skin sensitization reactions in humans could play a role. Skin testing showed that the incidence of sensitization by penicillin is very low (Nugent et al., 2003). Penicillin G was classified with a lower potency (weak) in the LLNA where it is applied topically, therefore the exposure route may affect the available dose and related potency. Coumarin was part of the learning set to develop the meta-model, and has been already mis-predicted as a non-sensitizer (Del Bufalo et al., 2018). The previous study argued that highly purified coumarin (99.9%) is not a sensitizer (Vocanson et al., 2006, 2007), but that the positive predictions are due to impurities. In any cases, it is crucial to have an expert analysis on the quality/ relevance of each input parameter in order to form a conviction on the DA's prediction. Additional false negative predictions include sensitizers of the human potency category 4 neomycin sulphate, hexyl salicylate and kanamycin sulfate. Neomycin sulphate and kanamycin sulfate are antibiotics that showed discordant input data and were also predicted as negative in the LLNA (Gerberick et al., 1992, 2005). Hexyl salicylate seems to be a

Table 4

Table of the false positives (FP) and false negative (FN) in the defined approach for skin sensitization hazard categorization with regard to LLNA data:

Chemical name	CAS	Stacking score ¹	Human pot. cat ²	Human hazard cat ³	LLNA EC3 median ⁴	LLNA hazard cat ⁵	False positive and false negative
Penicillin G	61-33-6	19,9	3	S	30	S	FN
Hexyl salicylate	6259-76-3	5,6	4	S	0.18	S	FN
Anisyl alcohol	105-13-5	28	5	NS	5.91	S	FN
Isopropyl myristate	110-27-0	3,4	5	NS	44	S	FN
Pyridine	110-86-1	6	5	NS	72	S	FN
Benzalkonium chloride	8001-54-5	20	5	NS	0.1	S	FN
Xylene	1330-20-7	3,4	6	NS	95.8	S	FN
Sodium lauryl sulfate	151-21-3	7,2	6	NS	3.55	S	FN
Tocopherol	59-02-9	6,3	6	NS	8.1	S	FN
DMSO	67-68-5	4,8	6	NS	72	S	FN
Salicylic acid	69-72-7	6	6	NS	12.2	S	FN
Amylcinnamyl alcohol	101-85-9	70,2	4	S	100	NS	FP
Vanillin	121-33-5	87,9	5	NS	100	NS	FP

NS = Non-sensitizer; S = Sensitizer; FN = False negative; FP = False positive.

¹ Stacking probability scale is 0–100.² From Basketter et al., 2014.³ Human potency cat. 1–4 = S and 5–6 = NS.⁴ From Hoffmann et al., 2018.⁵ According to ECETOC, 2003.

challenging chemical for both the *in silico* and *in vitro* methods, as all input data were negative except for the U-SENS™ (Piroird et al., 2015). Thus, the positive identification by the U-SENS™ was not sufficient to identify the chemical as a sensitizer in the DA. Benzocaine showed positive results in all the *in vitro* methods and negative results in the *in silico* approaches, pointing out a substantial weight of *in silico* input data in the final hazard prediction. However, Benzocaine is a mis-predicted LLNA like false negative (Warbrick et al., 2000) and the *in silico* model could be based on this LLNA result. Although of less concern, the DA using a stacking meta-model provided false positive predictions for chemicals of the human potency category 5 Hexyl cinnamic aldehyde, Anethole, Citronellol, Benzyl salicylate, Benzyl benzoate, Vanillin, Pentachlorophenol and α -Iso-methylionone. For all these category 5 chemicals, the U-SENS™, TIMES-SS and TOXTREE inputs were positive with the exception of the TOXTREE input for Citronellol. Nonetheless, all the 8 false positive category 5 chemicals listed above, with the exception of vanillin, were also predicted as skin sensitizers in the LLNA (Gerberick et al., 2005) (Table 3). This indicates that these false-positive chemicals are challenging not only for the *in vitro* and *in silico* methods, but also for the animal test. As reported by Basketter et al., 2014, the induction of skin sensitization in the general population by category 5 chemicals is extremely rare but may occur in highly selected patients (based on individual case reports of sensitization). Therefore, the predictive performance was re-analyzed using a “composite” hazard reference for the category 5 chemicals. This “composite” hazard reference proposed in Johansson et al., 2017, considers a human category 5 chemical as sensitizer if it is also positive in the LLNA and as a non-sensitizer if it is negative in the LLNA (Table 1). The data re-analysis results showed an improved sensitivity of 87%, specificity of 95% and accuracy of 89% (kappa of 0.72) of the meta-model for the 97 chemicals (Table 2). It should be taken into account that the human potency categories are based on expert judgment of human evidence (including exposure information and human repeat insult patch test (HRIPT) data) and are not entirely transparent and reproducible. This especially applies to the category 5 chemicals for which the HRIPT no observed effect levels are variable or absent due to the inherent inaccuracy of threshold determination in a very small panel size (Basketter et al., 2014). Therefore, the use of a “composite” hazard reference combining human and *in vivo* evidence is a possible way to increase confidence in the hazard classification for the category 5 chemicals. Although they are based on expert judgment, the human potency categories are the only available human references at that moment and were proven to be useful in various skin sensitization test method evaluation studies (Urbisch et al., 2015; Reisinger et al., 2015; Kleinstreuer et al., 2018).

As in other combinatory statistical models e.g. Bayesian or artificial neural network (Valkova et al., 2004; Kjaerulff and Madsen, 2013), determining the weight of each input parameter in the final predictions by a stacking meta-model is not an easy task. An ongoing work aims at determining the influence of the input parameters on the prediction of the DA using an experimental plan approach. It is important to emphasize that the DA is a predictive tool meant to support risk assessment decisions in a weight of evidence approach. Therefore, the output can be considered as scientific evidence with a statistically indicated level of confidence, and the expert judgment of the risk assessor can still be added to the weight of evidence. In particular, when the input data shows discrepancies it is crucial to take into account other sources of information and expertise.

DAs integrating non-animal data generated with the regulatory adopted methods have been proposed as valid components of Integrated Approaches to Testing and Assessment (IATA). The predictions generated using this DA could be used, where applicable and adequate, instead of LLNA data or in conjunction with such data if they already exist, in the context of IATA for assessing skin sensitization hazard and for hazard classification purposes. DA has been developed and therefore tested with a substantial number of chemicals for which *in vivo* skin sensitization data are available, therefore it is likely that its

domain of applicability covers the main reaction mechanisms relevant to skin sensitization.

5. Conclusions

The evaluation of the defined approach with a stacking meta-model developed by L'Oréal R&I, using the CE skin sensitization database demonstrated that the DA has a superior performance with a higher degree of confidence (kappa value) as compared to the LLNA in predicting human skin sensitization hazard as determined by human test data. The accuracy of the DA applying either a 30/70% cut-off approach was found to be in the range of the best performing DA's submitted to the OECD as case studies. Moreover, the DA did not mis-predict any of the human potency category 1 and 2 (extreme/strong sensitizers) or human potency category 6 (non-sensitizers), which indicates the model is reliable for these chemicals with unequivocal human evidence. The next step will be an independent evaluation of the DA for its integration in the performances based test guidelines (PBTG) for skin sensitization.

The present DA is a first solid tier in the hazard evaluation process proposed by L'Oréal Research & Innovation starting by the hazard identification (S/NS classification). The next step is to gather quantitative informations from *in silico* / *in vitro* tools in order to address the human skin sensitization potency. We are currently working on such a Tier 2 DA which will allow to discriminate (UN GHS) cat. 1A/1B sensitizers.

Acknowledgements

We thank Erwin van Vliet who supported this study.

Funding

No external funding was received for this study. The organization of the study and evaluation of the data was funded by L'Oréal Research & Innovation.

Conflict of interest

All authors are L'Oreal employees and declare there are no conflicts of interest.

References

- Alépée, N., Piroird, C., Nardelli, L., 2017. U-SENS™: A U937 cell line activation test for skin sensitization. In: Book Alternatives for Dermal Toxicity Testing. Springer, Cham, pp. 311–330. <https://doi.org/10.1007/978-3-319-50353-0>.
- Aptula, A.O., Roberts, D.W., 2006. Mechanistic applicability domains for non-animal-based prediction of toxicological end points: general principles and application to reactive toxicity. Chem. Res. Toxicol. 19, 1097–1105. <https://doi.org/10.1021/tx0601004>.
- Basketter, D.A., Alépée, N., Ashikaga, T., Barroso, J., Gilmour, N., Goebel, C., Hibatallah, J., Hoffmann, S., Kern, P., Martinozzi-Teissier, S., Maxwell, G., Reisinger, K., Sakaguchi, H., Schepky, A., Tailhardat, M., Templier, M., 2014. Categorization of chemicals according to their relative human skin sensitizing potency. Dermatitis 25, 11–21. <https://doi.org/10.1097/DER.0000000000000003>.
- Basketter, D.A., Ashikaga, T., Casati, S., Hubesch, B., Jaworska, J., de Knecht, J., Landseid, R., Manou, I., Mehling, A., Petersohn, D., Rorije, E., Rossi, L., Steiling, W., Teissier, S., Worth, A., 2015. Alternatives for skin sensitization testing and assessment. Regul. Toxicol. Pharmacol. 73, 660–666. <https://doi.org/10.1016/j.yrtph.2015.10.005>.
- Buehler, E.V., 1965. Delayed contact hypersensitivity in the Guinea pig. Arch. Dermatol. 91, 171.
- Del Bufalo, A., Pauloin, T., Alépée, N., Clouzeau, J., Detroyer, A., Eilstein, J., Gomes, C., Nocairi, H., Piroird, C., Rousset, F., Tourneix, F., Basketter, D., Martinozzi Teissier, S., 2018. Alternative integrated testing for skin sensitization: assuring consumer safety. Appl In Vitro Toxicol 4, 30–41. <https://doi.org/10.1089/aivt.2017.0023>.
- EC, 2008. Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on Classification, Labelling and Packaging of Substances and Mixtures, Amending and Repealing Directives 67/548/EEC and 1999/45/EC, and Amending Regulation (EC) No 1907/2006. Official Journal No L 353. pp. 0001–1355. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32008R1272>.

- ECETOC, 2003. Contact sensitisation: classification according to potency. In: Technical Report No. 87. European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels, Belgium.
- ECHA, 2017. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.7a: Endpoint Specific Guidance, Version 6.0 July 2017. https://echa.europa.eu/documents/10162/13632/information_requirements_r7a_en.pdf (last accessed Nov. 1, 2018).
- Emter, R., Ellis, G., Natsch, A., 2010. Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. *Toxicol. Appl. Pharmacol.* 15 (245), 281–290. <https://doi.org/10.1016/j.taap.2010.03.009>.
- EU, 2003. Directive 2003/15/EC of the European parliament and the council of 27 February 2003 amending council directive 76/768/EEC on the approximation of the laws of the members states relating to cosmetic products. *Off. J. Eur. Union* L66, 26–35.
- Ezendam, J., Braakhuis, H.M., Vandebruiel, R.J., 2016. State of the art in non-animal approaches for skin sensitization testing: from individual test methods towards testing strategies. *Arch. Toxicol.* 90, 2861–2883. <https://doi.org/10.1007/s00204-016-1842-4>.
- Gerberick, G.F., House, R.V., Fletcher, E.R., Ryan, C.A., 1992. Examination of the local lymph node assay for use in contact sensitization risk assessment. *Fundam. Appl. Toxicol.* 19, 438–445.
- Gerberick, G.F., Ryan, C.A., Kimber, I., Dearman, R.J., Lea, L.J., Basketter, D.A., 2000. Local lymph node assay validation assessment for regulatory purposes. *Am. J. Contact Dermat.* 11, 3–18. <https://doi.org/10.1053/ajcd.2000.0003>.
- Gerberick, G.F., Vassallo, J.D., Bailey, R.E., Chaney, J.G., Morrall, S.W., Lepoittevin, J.P., 2004. Development of a peptide reactivity assay for screening contact allergens. *Toxicol. Sci.* 81, 332–343. <https://doi.org/10.1093/toxsci/kfh213>.
- Gerberick, G.F., Ryan, C.A., Kern, P.S., Schlatter, H., Dearman, R.J., Kimber, I., Patlewicz, G.Y., Basketter, D.A., 2005. Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods. *Dermatitis* 16, 157–202.
- Gomes, C., Noçairi, H., Thomas, M., Collin, J.F., Ibanez, F., Saporta, G., 2012. Stacking prediction for a binary outcome. In: COMPSTAT, 20th International Conference on Computational Statistics, Limassol, Cyprus, pp. 271–282.
- Gomes, C., Noçairi, H., Thomas, M., Collin, J.F., Saporta, G., 2014. A simple and robust scoring technique for binary classification. *J. Artif. Intell. Res.* 3, 52–58.
- Hoffmann, S., Kleinstreuer, N., Alépée, N., Allen, D., Api, A.M., Ashikaga, T., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Göbel, C., Kern, P.S., Klaric, M., Kühnl, J., Lalko, J., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Parakhia, R., van Vliet, E., Zang, Q., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database. *Crit. Rev. Toxicol.* 48, 344–358. <https://doi.org/10.1080/10408444.2018.1429385>.
- Johansson, H., Gradin, R., Forreryd, A., Agemark, M., Zeller, K., Johansson, A., Larne, O., van Vliet, E., Borrebaeck, C., Lindstedt, M., 2017. Evaluation of the GARD assay in a blind Cosmetics Europe study. *ALTEX* 34, 515–523. <https://doi.org/10.14573/altex.1701121>.
- Jowsey, I.R., Basketter, D.A., Westmoreland, C., Kimber, I., 2006. A future approach to measuring relative skin sensitizing potency: a proposal. *J. Appl. Toxicol.* 26, 341–350. <https://doi.org/10.1002/jat.1146>.
- Kjaerulff, U.B., Madsen, A.L., 2013. Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis, 2nd edn. Springer, New York. <https://doi.org/10.1007/978-0-387-74101-7>.
- Kleinstreuer, N., Hoffmann, S., Alépée, N., Allen, D., Ashikaga, T., Casey, W., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Göbel, C., Kern, P.S., Klaric, M., Kühnl, J., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Strickland, J., van Vliet, E., Zang, Q., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (II): an assessment of defined approaches. *Crit. Rev. Toxicol.* 48, 359–374. <https://doi.org/10.1080/10408444.2018.1429386>.
- Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometric* 33, 159–174. <https://www.ncbi.nlm.nih.gov/pubmed/843571>.
- Magnusson, B., Kligman, A.M., 1969. The identification of contact allergens by animal assay. The Guinea pig maximisation test. *J. Invest. Dermatol.* 52, 268–276.
- Noçairi, H., Gomes, C., Thomas, M., Saporta, G., 2016. Improving stacking methodology for combining classifiers; applications to cosmetic industry. *Electron. J. Appl. Stat. Anal.* 9, 340–361. <https://doi.org/10.1285/i20705948v9n2p340>.
- Nugent, R.S., Quinn, J.M., McGrath, C.M., Hrcir, D.E., Boleman, W.T., Freeman, T.M., 2003. Determination of the incidence of sensitization after penicillin skin testing. *Ann. Allergy Asthma Immunol.* 90, 398–403. [https://doi.org/10.1016/S1081-1206\(10\)61823-6](https://doi.org/10.1016/S1081-1206(10)61823-6).
- OECD, 2013. OECD Guidelines for the Testing of Chemicals, Section 1 Physical-chemical Properties, Test No. 122: Determination of pH, Acidity and Alkalinity. OECD Publishing, Paris. <https://doi.org/10.1787/20745753>.
- OECD, 2014. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. OECD Series on Testing and Assessment. No. 168. OECD Publishing, Paris. <https://doi.org/10.1787/9789264221444-en>.
- OECD, 2015. Test no. 442C: In Chemico skin sensitisation. In: Direct Peptide Reactivity Assay (DPRA), OECD Guidelines for the Testing of Chemicals, Section 4 Ed. OECD, Paris. <https://doi.org/10.1787/20745788>.
- OECD, 2017a. Guidance Document on the Reporting of Defined Approaches to Be Used within Integrated Approaches to Testing and Assessment, OECD Series on Testing and Assessment, No. 255. OECD Publishing, Paris. <https://doi.org/10.1787/9789264274822-en>.
- OECD, 2017b. Guidance Document on the Reporting of Defined Approaches and Individual Information Sources to Be Used within Integrated Approaches to Testing and Assessment (IATA) for Skin Sensitisation, OECD Series on Testing and Assessment, No. 256. OECD Publishing, Paris. <https://doi.org/10.1787/9789264279285-en>.
- OECD, 2018a. Test No. 442E: In Vitro Skin Sensitisation: In Vitro Skin Sensitisation Assays Addressing the Key Event on Activation of Dendritic Cells on the Adverse Outcome Pathway for Skin Sensitisation, OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris. <https://doi.org/10.1787/9789264264359-en>.
- OECD, 2018b. Test No. 442D: In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method, OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris. <https://doi.org/10.1787/9789264229822-en>.
- Patlewicz, G., Dimitrov, S.D., Low, L.K., Kern, P.S., Dimitrova, G.D., Comber, M.I., Aptula, A.O., Phillips, R.D., Niemelä, J., Madsen, C., Wedebye, E.B., Roberts, D.W., Bailey, P.T., Mekenyan, O.G., 2007. TIMES-SS—a promising tool for the assessment of skin sensitization hazard. A characterization with respect to the OECD validation principles for (Q)SARs and an external evaluation for predictivity. *Regul. Toxicol. Pharmacol.* 48, 225–239. <https://doi.org/10.1016/j.yrtph.2007.03.003>.
- Piroid, C., Ovigne, J.M., Rousset, F., Martinozzi Teissier, S., Gomes, C., Cotovio, J., Alépée, N., 2015. The myeloid U937 skin sensitization test (U-SENS) addresses the activation of dendritic cell event in the adverse outcome pathway for skin sensitization. *Toxicol. in Vitro* 29, 901–916. <https://doi.org/10.1016/j.tiv.2015.03.009>.
- Reisinger, K., Hoffmann, S., Alépée, N., Ashikaga, T., Barroso, J., Elcombe, C., Gellatly, N., Galbiati, V., Gibbs, S., Groux, H., Hibatallah, J., Keller, D., Kern, P., Klaric, M., Kolle, S., Kuehn, J., Lambrechts, N., Lindstedt, M., Millet, M., Martinozzi-Teissier, S., Natsch, A., Petersohn, D., Pike, L., Sakaguchi, H., Schepky, A., Tailhardat, M., Templier, M., van Vliet, E., Maxwell, G., 2015. Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. *Toxicol. in Vitro* 29, 259–270. <https://doi.org/10.1016/j.tiv.2014.10.018>.
- Roberts, D., Patlewicz, G., 2017. Non-animal assessment of skin sensitization hazard: is an integrated testing strategy needed, and if so what should be integrated? *J. Appl. Toxicol.* 38, 41–50. <https://doi.org/10.1002/jat.3479>.
- Spicer, C.W., 2002. Hazardous Pollutant Handbook, Measurements, Properties and Fate in Ambient Air. CRC Press LLC, FL, USA.
- Strickland, J., Zang, Q., Paris, M., Lehmann, D.M., Kleinstreuer, N., Allen, D., Choksi, N., Matheson, J., Jacobs, A., Casey, W., 2017. Multivariate models for prediction of human skin sensitization hazard. *J. Appl. Toxicol.* 37, 347–360. <https://doi.org/10.1002/jat.3366>.
- Takenouchi, O., Fukui, S., Okamoto, K., Kurotani, S., Imai, N., Fujishiro, M., Kyotani, D., Kato, Y., Kasahara, T., Fujita, M., Toyoda, A., Sekiya, D., Watanabe, S., Seto, H., Hirota, M., Ashikaga, T., Miyazawa, M., 2015. Test battery with the human cell line activation test, direct peptide reactivity assay and DEREK based on a 139 chemical data set for predicting skin sensitizing potential and potency of chemicals. *J. Appl. Toxicol.* 35, 1318–1332. <https://doi.org/10.1002/jat.3127>.
- Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., Landsiedel, R., Jaworska, J., Kern, P.S., Gerberick, F., Natsch, A., Emter, R., Ashikaga, T., Miyazawa, M., Sakaguchi, H., 2015. Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul. Toxicol. Pharmacol.* 71, 337–351. <https://doi.org/10.1016/j.yrtph.2014.12.008>.
- US EPA, 2012. EPI SuiteTM, Version 4.1. www.epa.gov/sites/production/files/2015-05/documents/05.pdf (last accessed Nov. 1, 2018).
- Valkova, L., Vračko, M., Basak, S.C., 2004. Modeling of structure–mutagenicity relationships: counter propagation neural network approach using calculated structural descriptors. *Anal. Chim. Acta* 509, 179–186. <https://doi.org/10.1016/j.aca.2003.12.035>.
- van Vliet, E., Kühnl, J., Goebel, C., Martinozzi-Teissier, S., Alépée, N., Ashikaga, T., Blömeke, B., Del Bufalo, A., Cluzel, M., Corsini, E., Delrue, N., Desprez, B., Gellatly, N., Giese, C., Gribaldo, L., Hoffmann, S., Klaric, M., Maillere, B., Naisbitt, D., Pallardy, M., Vocanson, M., Petersohn, D., 2017. State-of-the-art and new options to assess T cell activation by skin sensitizers: Cosmetics Europe Workshop. *ALTEX* 35, 179–192. <https://doi.org/10.14573/altex.1709011>.
- Vandebruiel, R.J., van Loveren, H., 2010. Non-animal sensitization testing: state-of-the-art. *Crit. Rev. Toxicol.* 40, 389–404. <https://doi.org/10.3109/10408440903524262>.
- Vocanson, M., Goujon, C., Chabeau, G., Castelain, M., Valeyrie, M., Floch, F., Maliverney, C., Gard, A., Nicolas, J.F., 2006. The skin allergic properties of chemicals may depend on contaminants—evidence from studies on coumarin. *Int. Arch. Allergy Immunol.* 140, 231–238. <https://doi.org/10.1159/000093248>.
- Vocanson, M., Valeyrie, M., Rozières, A., Hennino, A., Floch, F., Gard, A., Nicolas, J.F., 2007. Lack of evidence for allergenic properties of coumarin in a fragrance allergy mouse model. *Contact Dermatitis* 57, 361–364. <https://doi.org/10.1111/j.1600-0536.2007.01276.x>.
- Warbrick, E.V., Dearman, R.J., Basketter, D.A., Kimber, I., 2000. Failure of vehicle to influence local lymph node assay response to benzocaine. *Contact Dermatitis* 42, 164–165.
- Wolpert, D.H., 1992. Stacked generalization. *Neural Netw.* 5, 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).