Original Article

Adjustment of a no expected sensitization induction level derived from Bayesian network integrated testing strategy for skin sensitization risk assessment

Yuki Otsubo, Taku Nishijo, Hideyuki Mizumachi, Kazutoshi Saito, Masaaki Miyazawa
and Hitoshi Sakaguchi

Safety Science Research Laboratories, Kao Corporation, 2606 Akabane, Ichikai, Haga, Tochigi, 321-3497, Japan

(Received August 27, 2019; Accepted November 18, 2019)

ABSTRACT — Skin sensitization is a key adverse effect to be addressed during hazard identification and risk assessment of chemicals, because it is the first step in the development of allergic contact dermatitis. Multiple non-animal testing strategies incorporating *in vitro* tests and *in silico* tools have achieved good predictivities when compared with murine local lymph node assay (LLNA). The binary test battery of KeratinoSens™ and h-CLAT could be used to classify non-sensitizers as the first part of bottom-up approach. However, the quantitative risk assessment for sensitizing chemicals requires a *No Expected Sensitization Induction Level* (NESIL), the dose not expected to induce skin sensitization in humans. We used Bayesian network integrated testing strategy (BN ITS-3) for chemical potency classification. BN ITS-3 predictions were performed without a pre-processing step (selecting data from their physicochemical applicability domains) or post-processing step (Michael acceptor chemistry correction), neither of which necessarily improve prediction accuracy. For chemicals within newly defined applicability domain, all under-predictions fell within one potency class when compared with LLNA results, indicating no chemicals that were incorrectly classified by more than one class. Considering the potential under-prediction by one class, a worst case value to each class from BN ITS-3 was used to derive a NESIL. When *in vivo* and human data from suitable analogs cannot be used to estimate the uncertainty, adjusting the NESIL derived from BN ITS-3 may help perform skin sensitization risk assessment. The overall workflow for risk assessment was demonstrated by incorporating the binary test battery of KeratinoSens™ and h-CLAT.

Key words: Skin sensitization, Bayesian network, Integrated testing strategy, Potency prediction, Risk assessment

INTRODUCTION

Skin sensitizers can cause *allergic contact dermatitis* (ACD) upon contact with skin (Kimber *et al.*, 2011). Animal tests, such as the guinea pig maximization test, the Buehler test, and the murine local lymph node assay (LLNA) are traditionally used to assess skin sensitization of chemicals. ACD consists of a *sensitization phase* wherein antigen specific T-cells proliferate and an *elicitation phase* where subsequent contact with an allergen produces a skin reaction (Goebel *et al.*, 2012). Due to regulatory requirements and ethical concerns, non-animal test methods have been developed that focus on ear-

lier key events (KEs, KE1: protein binding, KE2: keratinocyte response and KE3: dendritic cell activation) of the adverse outcome pathway (AOP) during the induction phase of skin sensitization (OECD, 2012). The direct peptide reactivity assay (DPRA), examining KE1, reflects binding between proteins and the chemicals using model peptides containing cysteine or lysine residues (Gerberick *et al.*, 2004, 2007). The KeratinoSens™ examines KE2 by simulating antioxidant response elements (AREs) in HaCaT-derived human keratinocytes to detect activation of the Keap1–Nrf2–ARE pathway as an oxidative and electrophilic stress response (Emter *et al.*, 2010). The human Cell Line Activation Test (h-CLAT) addresses

KE3 by evaluating increased expression of surface markers CD86 and CD54 on THP-1 cells, a human monocytic leukemia cell line that serves as a dendritic cell surrogate (Ashikaga *et al.*, 2006, 2010; Sakaguchi *et al.*, 2006).

No single *in vitro* test can cover the entire AOP with higher accuracy than the LLNA. Consequently, integrated results from multiple *in vitro* tests, as well as *in silico* models, would be needed to assess the skin sensitizing potential of various chemicals. Recently, we reported on a binary test battery that combines KeratinoSens™ and h-CLAT (Otsubo *et al.*, 2017) as a hazard identification model, taking into account some limitations of the 2 out of 3 approach (Urbisch *et al.*, 2015). The binary test battery could determine if a test chemical was non-sensitizing, based on two negative results from KeratinoSens™ and h-CLAT. The battery has 93.4% (compared with the LLNA), or 94.4% sensitivity (compared with human data) for 203 chemicals. Considering its predictive limitations, the test battery achieved high sensitivity, similar to the 3 out of 3 ITS, requiring three negative results of the DPRA, KeratinoSens™, and h-CLAT. Therefore, the binary test battery consisting of KeratinoSens™ and h-CLAT could function as part of a “bottom-up” approach for skin sensitization hazard prediction.

On the other hand, when at least one positive result was obtained by any of three *in vitro* tests, a potency classification was necessary for risk assessment. Several promising potency prediction models were submitted to the OECD as case studies of defined approaches (OECD, 2016). The OECD guideline documents included defined approaches that used validated OECD *in vitro* test methods and simple, rule-based data interpretation procedures (2016). For example, the sequential testing strategy (STS) provides hazard and potency classification based on DPRA and h-CLAT data. In addition, the integrated testing strategy (ITS) provides hazard and potency classification by using the quantitative parameters of DPRA and h-CLAT, and the result from a commercial *in silico* tool (DEREK Nexus) that identifies structural alerts for sensitization (Nukada *et al.*, 2013; Takenouchi *et al.*, 2015; Kleinstreuer *et al.*, 2018). However, the STS and ITS do not address KE2. Over the past few years, some testing strategies, based on linear regression and machine learning system, were also developed (Natsch *et al.*, 2018; Jaworska *et al.*, 2015). Natsch *et al.* (2018) reported a prediction model based on a combination of kinetic peptide reactivity data and KeratinoSens™. The regression analysis was applied into the model and the construction is well-defined; however, this model does not consider KE3. On the other hand, the BN ITS-3 uses the quantitative results of DPRA, KeratinoSens™, and h-CLAT to repre-

sent KE1-3 of the skin sensitization AOP (Jaworska *et al.*, 2015). The model also uses physicochemical parameters (Log D, water solubility, ionization ratio and plasma protein binding) and structure-based prediction by TIMES. The BN ITS-3 can determine the sensitizing potency of chemicals as a probability distribution over 4 potency classes of LLNA [Non-sensitizer (NS), Weak (W), Moderate (M) or Strong to Extreme (S)]. BN ITS-3 predicted LLNA potency class in a training set (n = 147) or an external test set (n = 60) with 86.4% and 88.3% accuracy, respectively. However, there were still some mis-predictions, including over- and underestimation of the potencies of various chemicals.

We examined the accuracy of potency classification (4 classes) of BN ITS-3 using a dataset of 175 chemicals (141 sensitizers and 34 NS) constructed in our previous work. Jaworska *et al.* (2015) performed BN ITS-3 predictions using both a pre-processing step of selecting data from their physicochemical applicability domains and a post-processing step of Michael acceptor (MA) chemistry correction, which could lead to less sensitizing predictions than those without processing steps. On the other hand, in this study, BN ITS-3 predictions were performed without either processing step in order to avoid potential under-predictions. While LLNA potency is generally expressed as percent weight per volume, Jaworska *et al.* (2015) converted all the values into molar units. However, substances with weak LLNA potency or low molecular weight (MW) could be under-predicted. Thus, in this study, LLNA potency was expressed as percent weight per volume. Then, we estimated uncertainty due to under-predictions in BN ITS-3 and assigned a worst value of EC3 to each class (e.g. EC3 = 10% for weak sensitizers) to derive a No Expected Sensitization Induction Level (NESIL) for use in skin sensitization quantitative risk assessment (QRA).

MATERIALS AND METHODS

Bayesian network integrated testing strategy (BN ITS-3)

We used the BN ITS-3, developed by Procter & Gamble (Jaworska *et al.*, 2015) using both the commercially-available BayesiaLab 5.4 platform (Bayesia SAS, Laval, Cedex, France) and the code which was kindly provided by Jaworska. The BN ITS-3 uses the following data: 1) bioavailability-related variables that rely on physicochemical properties (water solubility, log D, fraction ionized, plasma protein binding) calculated at pH = 7 using ACD labs Percepta (ACD Labs, Toronto, Canada); 2) structure-based *in silico* prediction, TIMES-SS

(Laboratory of Mathematical Chemistry, Bourgas, Bulgaria), which generates chemical alerts for sensitization using skin metabolism and autoxidation simulators; 3) quantitative data from DPRA, KeratinoSens™, and h-CLAT that represent skin sensitization KE1-3 of the AOP. In the BN ITS-3, the relevant in-domain evidence for DPRA, KeratinoSens™, and h-CLAT takes into account both the water solubility cutoffs and fraction ionized. Chemicals that were fully ionized at pH = 7 are deemed “not suitable” for cell-based assays due to poor bioavailability. In fact, in Jaworska *et al.* (2015), only squaric acid and tartaric acid, which feature ionization rates at pH = 7 of 99.997% and 99.998% respectively, were considered fully ionized chemicals. Thus, the results of KeratinoSens™ and h-CLAT were excluded from the model test set for these chemicals. On the other hand, the water solubility cutoffs and fraction ionized were considered in only the model test set, but not in the model training set. Additionally, the criteria for fraction ionized in Jaworska *et al.* (2015) remains unknown. Therefore, in this study, we did not consider water solubility cutoffs and fraction ionized for the applicability domain. The BN ITS-3 predicts the potency of skin sensitization as a probability distribution over 4 potency classes in LLNA. While the LLNA EC3 values, which correspond to the concentrations required to produce a positive response, are expressed as percent weight per volume, Jaworska *et al.* (2015) converted all the values into molar units. This conversion could result in underestimation of substances with EC3 values close to 90%, or for those with low MW (e.g., a W sensitizer or a M sensitizer in weight-based classification can be changed to a NS or a W sensitizer in molar-based classification, respectively). Therefore, in this study, LLNA potency was expressed as percent weight per volume. While BN ITS-3 predictions are post-processed to correct for MA chemistry, MA correction was only applied to the model test set and, therefore, changed the predicted potency into a weaker class (Jaworska *et al.*, 2015). Thus, in this study, to avoid prediction underestimation, MA correction was not considered. The predicted probability distribution was converted to Bayes factors (*B*), where a NS prediction can be accepted only when *B* > 3 (substantial strength of evidence), while a sensitizer prediction with *B* > 1 can be accepted. The highest *B* across potency classes drives the prediction, and a lower *B* indicates stronger uncertainty.

Chemical Dataset

The analysis was performed using the same dataset of 203 chemicals that were compiled by Otsubo *et al.* (2017). Of the 203 chemicals, seven metal salts, such

as potassium dichromate and sodium lauryl sulfate, and one mixture (methylchloroisothiazoline/methyl isothiazoline) were excluded from the analysis because they were not considered in the model training set and test set of Jaworska *et al.* (2015). Furthermore, bisphenol A glycerolate dimethacrylate (Bis-GMA) was excluded from the analysis because the data for lysine peptide are not available. Methyl pyruvate, clotrimazole, 1-cyclohexylethyl 2-butenate, 1-octen-3-yl acetate, methyl salicylate, sulfanilamide, 1-butanol, Isopropanol, propylene glycol, 4-hydroxybenzoic acid, glycerol, lactic acid, octanenitrile, vinylidene dichloride, N,N-diethyl-m-tol-uamide, sulphanilic acid, dimethyl formamide, and saccharin were not considered because they had negative results in all three *in vitro* assays and did not need potency prediction assessment. The remaining 175 chemicals (34 non-sensitizers and, 55 W, 56 M and 30 S sensitizers) were analyzed for potency prediction by the BN ITS-3. Supplementary Table 1 lists the data for the DPRA, KeratinoSens™, h-CLAT, LLNA, TIMES, physicochemical properties, and BN ITS-3.

Direct Peptide Reactivity Assay (DPRA): OECD test guideline 442C

The DPRA evaluates KE1, the molecular initiating event in the skin sensitization AOP. The protein reactivity was reported as percent depletion using two synthetic model hepta-peptides, one containing cysteine (Ac-RFAACAA-COOH) and the other containing lysine (Ac-RFAAKAA-COOH) (Gerberick *et al.*, 2004, 2007). The test chemical was mixed with the cysteine or lysine peptides at 10:1 or 50:1, and the depletion of the peptides after 24 hr (as a result of binding to the test chemical) was measured using HPLC. Each of percent peptide depletion was calculated for individual peptides, and they were averaged. If the average depletion value was > 6.38%, the chemicals were classified as positive.

KeratinoSens™: OECD test guideline 442D

The KeratinoSens™ addresses KE2 or keratinocyte activation by measuring the activation of the keap1–Nrf2–ARE pathway (Emter *et al.*, 2010). The KeratinoSens™ is a reporter cell-based assay measuring luciferase gene induction as an indicator of ARE activation in HaCaT. The cells were incubated with the test chemical for 48 hours, and then luciferase induction was measured by luminometer. Test chemicals were classified as positive if the luciferase gene induction showed a statistically significant increase >1.5-fold over the vehicle control at a concentration <1000 µM, with cell viability >70%. The estimated concentration of 1.5-fold luciferase induc-

tion (EC1.5) was calculated by linear interpolation.

Human cell line activation test (h-CLAT): OECD test guideline 442E

The h-CLAT addresses KE3 or dendritic cell activation. Dendritic cells are known to up-regulate the expression of CD86 and CD54 cell surface proteins following exposure to sensitizers. The h-CLAT measures CD86 and CD54 expressions on THP-1 cells that serve as a dendritic cell surrogate (Ashikaga *et al.*, 2006, 2010; Sakaguchi *et al.*, 2006). The cells were treated with a test chemical for 24 hr at the concentration yielding 75% cell viability (CV75) and stained with fluorescence-labeled antibodies for CD86 or CD54. Changes in CD86 and CD54 expressions, induced by the test chemical, were then measured by flow cytometry. Relative fluorescence intensity (RFI) to vehicle control was measured by flow cytometry. When CD86 RFI was equal to or greater than 150% or CD54 RFI was equal to or greater than 200%, the test chemical was classified as positive. The estimated concentration inducing 150% of CD86 RFI and/or 200% of CD54 RFI (EC150 or EC200) was calculated, similar to the EC3 value determination in LLNA.

KOWWIN ver.4.1 in EPI suite™

The log Kow of each chemical was estimated using the program KOWWIN ver.4.1 in the EPI suite™ (Environmental Protection Agency, Washington, DC, USA) as an indicator of each chemical's water solubility.

Tissue Metabolism Simulator for predicting skin sensitization (TIMES) ver.2.27.17

The TIMES (Laboratory of Mathematical Chemistry, Bourgas, Bulgaria) is a licensed software package with prediction modules for several endpoints, including skin sensitization. The TIMES-SS generates chemical alerts for sensitization using a skin metabolism and autoxidation simulator and classifies test chemicals into three potency categories: non-sensitizers, weak sensitizers and strong sensitizers (Patlewicz *et al.*, 2007, 2014). The TIMES-SS was used to assign the pre- and pro-haptens that require abiotic (e.g., auto-oxidation) or biotic (e.g., metabolic transformation) activation before becoming sensitizers.

ACD/Percepta

ACD/Percepta (ACD Labs) was used to calculate physicochemical properties (LogKow, LogD, water solubility, and plasma protein binding). Ionization ratios were calculated using the formula below:

$$\text{ionization ratio} = |1 - 10^{\text{LogD}} / 10^{\text{LogKow}}|$$

where || means an absolute value.

Statistical analysis

The predictive performances of the test batteries were calculated according to Cooper statistics as sensitivity, specificity, and accuracy (Cooper *et al.*, 1979). Sensitivity was the proportion of sensitizers predicted as positive. Specificity was the proportion of non-sensitizers predicted as negative. Accuracy was the overall proportion of correct predictions.

Case studies of fragrance ingredients to derive NESILs

The potency assessment is important to apply the QRA prior to use in consumer products. For a sensitizing chemical, the QRA requires a NESIL, expressed in $\mu\text{g}/\text{cm}^2$, along with safety assessment factors to derive an acceptable exposure level (AEL) of a chemical in a particular product (Api *et al.*, 2008; Felter *et al.*, 2003; Basketter and Safford, 2016). Generally, a NESIL is a dose that is not expected to cause induction of skin sensitization in humans. The NESIL is derived from an EC3 value in LLNA (e.g., 1% of EC3 in LLNA is 250 $\mu\text{g}/\text{cm}^2$ of NESIL) or from no observed effect levels (NOELs) in the Human Repeated Insult Patch Test or Human Maximization Test (McNamee *et al.*, 2008). The Consumer Exposure Level (CEL) ($\mu\text{g}/\text{cm}^2$) for the product is calculated according to the SCCS Notes of Guidance 10th revision (European Commission, 2018). If $\text{AEL} > \text{CEL}$, the chemical can be safely used in the product type. In this study, we selected a limited number of fragrance ingredients listed in the IFRA standard, 48th amendment of IFRA Code of Practice (IFRA, 2015), considering acylating chemicals (amine reactive chemicals), pre/pro-haptens, and lipophilic chemicals ($\log \text{Kow} > 3.5$), which are deemed outside of BN ITS-3 applicability domain. Then, we examined how close a NESIL derived by BNITS-3 is a NESIL from either LLNA or human data.

RESULTS

Predictive performance of BN ITS-3 compared to four class potency in the LLNA

Supplementary Table 1 shows the dataset of 175 chemicals, including physicochemical properties, and the data for DPRA, KeratinoSens™, h-CLAT, LLNA, TIMES, and BN ITS-3, sorted according to their EC3 values. In this study, EC3 values were expressed as percent weight per volume, while Jaworska *et al.* (2015) converted all the values into molar units. BN ITS-3 generated a probability distribution over four potency classes without pre-processing step of selecting data from the physicochemical applicability domains or post-processing step of the

Table 1. Predictive performance of BN ITS-3 for positive chemicals in any of three *in vitro* tests.

LLNA category	Number	BN ITS-3			
	175	strong	moderate	weak	NS
extreme-strong	30	22	6	2	0
moderate	56	14	27	13	2
weak	55	5	6	38	6
NS	34	1	2	2	29
Hazard	Sensitivity	94.3% (133/141)			
	Specificity	85.3% (29/34)			
	Accuracy	92.6% (162/175)			
Potency	Accuracy	66.3% ((22+27+38+29)/175)			
	Accuracy for S	73.3% (22/30)			
	Accuracy for M	48.2% (27/56)			
	Accuracy for W	69.1% (38/55)			
	Under-predictive rate	20.6% ((6+2+13+2+6)/141)			
	Over-predictive rate	17.1% ((14+5+6+1+2+2)/175)			

MA correction. Then, the probability distribution was converted into a B for each potency category. The prediction of B (NS) > 3 and B (W, M, or S) < 1 indicated a non-sensitizer. On the other hand, the prediction of B (W, M, or S) > 1 indicated a sensitizer. The highest B indicated the predicted potency of the test chemical. As Table 1 presents, in the predictive performance for hazard identification in the LLNA, the sensitivity was 94.3% (133/141), the specificity was 85.3% (29/34) and the accuracy was 92.6% (162/175). BN ITS-3 could identify the sensitizing potential of test chemicals with high reliability. Table 1 also summarizes the predictive performance of the four classes of potency (NS, W, M, or S) of BN ITS-3, compared with LLNA. The four class potency accuracy was 66.3% (116/175). Accuracy for strong, moderate, or weak sensitizers in LLNA was 73.3% (22/30), 48.2% (27/56), or 69.1% (38/55), respectively. Thus, for 20.6% (29/141 sensitizers), the potency was under-predicted, and for 17.1% (30/175 chemicals), it was over-predicted. On the other hand, for 163 of 175 chemicals examined, the mis-prediction fell within one potency class of the true potency, indicating that there were almost no chemicals that were incorrectly classified by more than one class. Only four under-predicted chemicals resulted in two potency class under-predictions: hexyl salicylate and benzoyl peroxide were classified as strong potency on the LLNA, but weak potency on the BN ITS-3; diethylenetriamine and squaric acid were classified as moderately potent on the LLNA, but negative on the BN ITS-3.

Under-predicted chemicals by BN ITS-3

Table 2 summarizes the 29 chemicals where potency was under-predicted on the BN ITS-3 compared with

the LLNA. Of these, phthalic anhydride, maleic anhydride, and 1,2-cyclohexane dicarboxylic anhydride, which are acylating agents (Natsch *et al.*, 2013, 2015), were classified as strong on the LLNA, but as moderate on the BN ITS-3. Benzoyl peroxide, which is an acylating agent or an amine reactive chemical (Piroird *et al.*, 2015; Natsch *et al.*, 2013), was classified as strong on the LLNA, but as weak on the BN ITS-3. Squaric acid diethyl ester and squaric acid, which are amine reactive chemicals (Natsch *et al.*, 2013), were classified as strong and moderate (respectively) on the LLNA, but as moderate and non-sensitizing on the BN ITS-3. Undec-10-enal, p-isobutyl- α -methyl hydrocinnamaldehyde, hexyl salicylate, and farnesol, which are lipophilic sensitizers with log Kow > 3.5 , were under-predicted as weak in the BN ITS-3. Compared with LLNA, 2-Nitro-4-phenylenediamine, dihydroeugenol, dibenzyl ether, 4-chloroaniline, and diethylenetriamine, which are pre/pro-haptens, were under-predicted. Benzoyl peroxide, squaric acid, hexyl salicylate, and diethylenetriamine, which produced two potency class under-predictions between the LLNA and BN ITS-3, are either an acylating agent, an amine reactive chemical, a lipophilic chemical, or a pre/pro-hapten. Therefore, these are all potential under-predicted chemicals on the BN ITS-3. By excluding potentially under-predicted chemicals, all the predictions fell within one potency class under-prediction when compared with LLNA. Table 3 summarizes the predictive performance of the BN ITS-3 for 94 chemicals by excluding 81 chemicals corresponding to acylating agents, amine reactive chemicals, lipophilic chemicals, and pre/pro-haptens from Table 1. The four class potency accuracy was 71.2% (67/94). Accuracy for strong, moderate, weak sen-

Table 2. Under-predicted chemicals by BN ITS-3.

Chemical name	LLNA		BN ITS-3		Discussion
	EC3 (%)	Potency category	Potency category		
Phthalic anhydride	0.16	strong	moderate	Acyl transfer agent, amine reactive chemical (Natsch <i>et al.</i> , 2015)	
Maleic anhydride	0.16	strong	moderate	Acyl transfer agent, amine reactive chemical (Natsch <i>et al.</i> , 2015)	
1,2-cyclohexane dicarboxylic anhydride	0.84	strong	moderate	Acyl transfer agent, amine reactive chemical (Natsch <i>et al.</i> , 2015)	
Benzoyl peroxide	0.22	strong	weak	Acyl transfer agent, amine reactive chemical (Roiroird <i>et al.</i> , 2015; Natsch <i>et al.</i> , 2015)	
Squaric acid diethyl ester	0.9	strong	moderate	Amine reactive chemical (Natsch <i>et al.</i> , 2013)	
Squaric acid	4.3	moderate	NS	Amine reactive chemical (Natsch <i>et al.</i> , 2013)	
Undec-10-enal	6.8	moderate	weak	LogKow=4.12	
p-Isobutyl- α -methyl hydrocinnamaldehyde	9	moderate	weak	LogKow=4.4	
Hexyl salicylate	0.18	strong	weak	LogKow=5.06	
Farnesol	4.1	moderate	weak	LogKow=5.77, Pre-hapten	
2-Nitro-4-phenylenediamine	0.5	strong	moderate	pre/pro-hapten	
Dihydroeugenol	6.8	moderate	weak	pre/pro-hapten	
Dibenzyl ether	6.3	moderate	weak	pre-hapten	
4-chloroaniline	6.5	moderate	weak	pro-hapten	
Diethylenetriamine	5.8	moderate	NS	pro-hapten	
Formaldehyde	0.61	strong	moderate		
1-Phenyl-1,2-propanedione	1.3	moderate	weak		
Allyl phenoxacetate	3.1	moderate	weak		
6-Methyl-3,5-heptadien-2-one	5	moderate	weak		
trans-2-Hexenal	5.5	moderate	weak		
Perillaaldehyde	8.1	moderate	weak		
Methyl methanesulphonate	8.1	moderate	weak		
3-Methyl-1-phenylpyrazolone	8.5	moderate	weak		
Oxalic acid anhydrous	15	weak	NS		
Benzocaine	22	weak	NS		
Pyridine	72	weak	NS		
Diethyl acetaldehyde	76	weak	NS		
Aniline	89	weak	NS		
Methylmethacrylate	90	weak	NS		

Adjustment of NESIL by BN ITS3 for skin sensitization risk assessment

Table 3. Predictive performance of BN ITS-3 after excluding potentially under-predicted chemicals.

LLNA category	Number	BN ITS-3			
	94	strong	moderate	weak	NS
extreme-strong	13	12	1	0	0
moderate	30	7	16	7	0
weak	25	2	3	14	6
NS	26	0	0	1	25
Potency	Accuracy	71.2% ((12+16+14+25)/94)			
	Accuracy for S	92.3% (12/13)			
	Accuracy for M	53.3% (16/30)			
	Accuracy for W	56.0% (14/25)			
	Under-predictive rate	20.6% ((1+7+6)/68)			
	Over-predictive rate	13.8% ((7+2+3+1)/94)			

Table 4. Case studies on fragrance ingredients.

	LLNA weighted mean EC3 values; LLNA NESIL ($\mu\text{g}/\text{cm}^2$)	Human NOEL; Human NESIL ($\mu\text{g}/\text{cm}^2$)	BN ITS-3 prediction (Lowest EC3)	NESIL ($\mu\text{g}/\text{cm}^2$) from BN ITS-3 prediction
Methyl heptine carbonate	<125	118	strong (not defined)	Not defined
3-Propylenephthalide	350	945	moderate (1%)	250
Safranal	<250	29.5	moderate (1%)	250
2-Phenylpropionaldehyde	1575	388	moderate (1%)	250
Cyclamen aldehyde	5413	5905	weak (10%)	2500
α -Amyl cinnamic aldehyde	2942	23622	weak (10%)	2500
Carvone	2675	2657	weak (10%)	2500
Perillaaldehyde	2175	709	weak (10%)	2500
Benzyl alcohol	>12500	5906	NS (100%)	25000
Coumarin	>12500	3543	NS (100%)	25000

sensitizers or non-sensitizers in LLNA was 92.3% (12/13), 53.3% (16/30), 56.0% (14/25), or 96.2% (25/26), respectively. Potency was under-predicted in 20.6% (14/68 sensitizers) and over-predicted in 13.8% (13/94 chemicals). One approach to minimize uncertainty from mis-prediction as one class lower (20.6% of under-predictive rate) is to assign the lowest EC3 to each class from BN ITS-3 as the point of departure (PoD) for risk assessment as shown in Fig. 1 (Gerberick *et al.*, 2001). A weak potency was adjusted to 10% of EC3. Moderate potency was adjusted to 1% of EC3. For strong potency, EC3 was not determined since the strong prediction in the BN ITS-3 indicates strong ($\text{EC3} \leq 0.1\%$) or extreme ($\text{EC3} \leq 0.01\%$) sensitizers. When a test chemical was predicted as a non-sensitizer by the BN ITS-3, 100% of EC3 was assigned to derive NESIL, considering mis-prediction as one class lower. This indicates that the usage limit would be set. Fig. 2 presents the workflow of sensitizing potential and potency classification by incorporating our previously published testing strategy (Otsubo *et al.*, 2017). When at

least one positive result was obtained by any of three *in vitro* tests, the prediction of the test chemical would be considered a sensitizer. Therefore, it should be noted that even the negative prediction by the BN ITS-3 represents 100% of EC3 that can be used for the risk assessment.

Case studies on fragrance ingredients

Ten fragrance ingredients with LLNA and human data were used as case studies, considering acylating chemi-

BN ITS-3 Decision		Worst value of EC3
NS	⇒	EC3 = 100%
Weak	⇒	EC3 = 10%
Moderate	⇒	EC3 = 1%
Strong	⇒	Not defined

Fig. 1. EC3 from the BN ITS-3 predictions.

Fig. 2. Workflow to evaluate hazard and potency classification for skin sensitization.

cals (amine reactive chemicals), pre/pro-haptens, and lipophilic chemicals (log Kow > 3.5), which are deemed outside of BN ITS-3 applicability domain. Table 4 indicates that the LLNA- and human-derived NESIL values are quite different (more than an order of magnitude) in some cases and very similar in others. On the other hand, the NESIL from BN ITS-3 is close to or lower than LLNA-derived NESIL in most cases.

DISCUSSION

The purpose of this study was to derive the NESIL of the chemical predicted as a sensitizer using a binary test battery of KeratinoSens™ and h-CLAT, or the 3 out of 3 ITS. The BN ITS-3, developed by Procter & Gamble, could provide a quantitative WoE for potency categorization, allowing for risk assessment of chemicals (Jaworska *et al.*, 2015). We analyzed 175 chemicals with the necessary data for the BN ITS-3 to calculate *B*, corresponding to the probabilistic prediction for four potency classes (NS, W, M, and S). The BN ITS-3 predicted LLNA hazard outcomes with 92.6% (162/175) accuracy and the four potency classes with 66.3% (116/175) accu-

racy. For 20.6% (29/141) of sensitizers, the potency was under-predicted and in 17.1% (30/175) it was over-predicted. Importantly, while these predictions were performed without either the pre-processing step of selecting data from their physic-chemical applicability domains, or the post-processing step of MA correction, these results were better than, or equal to, those of the smaller dataset obtained using both processing steps (Kleinstreuer *et al.*, 2018); LLNA hazard outcomes demonstrated 83.2% (99/119) accuracy, and potency accuracy was 67.8% (78/115). These results suggested that the processing steps did not necessarily improve prediction accuracy. Moreover, for 163 of 175 chemicals examined, the mis-prediction fell within one potency class of the LLNA potency, indicating that there were almost no chemicals that were incorrectly classified by more than one class. A detailed analysis of the under-predicted chemicals showed that acylating chemicals, amine reactive chemicals, lipophilic chemicals, and pre/pro-haptens had predictive limitations. By excluding these potentially under-predicted chemicals, the four class potency accuracy was 71.2% and all the predictions fell within one potency class lower than the LLNA-predicted class. Furthermore, to minimize uncer-

tainty from under-predictions (20.6% of under-predictive rate) during skin sensitization risk assessment, the lowest EC3 to each class from BN ITS-3 was assigned as the PoD as shown in Fig. 1.

LLNA EC3 values are generally used to derive the NESIL. However, use of animals for testing cosmetic ingredients and products has been banned in EU since 2013 (European Union, 2009). Thus, alternative approaches for deriving NESIL for chemical risk assessment are urgently required. The complexity of the underlying mechanisms for induction of skin sensitization indicates that no existing *in vitro* test methods are appropriate for potency prediction; however, use of previously-developed testing strategies consisting of several test results may be a potential way to predict sensitizing potency. The BN ITS-3 achieved the high performance to predict the sensitizing potency in LLNA (Jaworska *et al.*, 2015; Kleinstreuer *et al.*, 2018). However, it should be noted that the four potency class prediction by the BN ITS-3 might be insufficient for the purposes of precise risk assessment, given its under-prediction rate of 20.6% (29/141 sensitizers). Chemicals with associated *in vivo* and human data to support suitable read across are very important references for risk assessment. Unless such supporting information like read across is available, the risk assessment would use potency predictions of BN ITS-3 to set a NESIL. When the under-predictions are not negligible, the conservative risk assessment would need to be done by estimating the degree of under-predictions in the BN ITS-3.

According to Kleinstreuer *et al.* (2018), 10 out of 35 MA chemicals (dimethyl fumarate, methyl heptane carbonate, toluene diamine sulfate, resorcinol, hexyl cinnamic aldehyde, ethyleneglycol dimethacrylate, carvone, ethyl acrylate, methylmethacrylate, and trans-2-hexenal) were under-predicted by the BN ITS-3 as compared with three class potency predictions by LLNA. The chemicals within the domain of MA were generally predicted with high accuracy of at least 80% by all of the non-animal test methods. The human and LLNA data are also concordant for most of the chemicals (Urbisch *et al.*, 2015). Under-predictions that unexpectedly occur by MA correction need to be avoided, since MA correction will shift a chemical to a less potency class. For BN ITS-3, the validity of MA correction might need further investigation; however, MA correction was not considered as a post-processing step in this study.

As compared with the sensitizing potency in LLNA, the 29 chemicals that were under-predicted are presented in Table 2. Only four chemicals (benzoyl peroxide, squaric acid, hexyl salicylate, and diethylenetriamine) resulted in two potency class under-predictions. First, benzoyl

peroxide and squaric acid are known to be amine reactive and show a higher reactivity for lysine, compared with cysteine reactivity (Piroird *et al.*, 2015; Natsch *et al.*, 2013). Of the 29 chemicals, phthalic anhydride, maleic anhydride, 1,2-cyclohexane dicarboxylic anhydride, and squaric acid diethyl ester were also amine reactive. Although this feature can be observed in acylating chemicals which predominantly transfer acyl groups onto lysine residues (Natsch *et al.*, 2013), benzoyl peroxide, phthalic anhydride, maleic anhydride, and 1,2-cyclohexane dicarboxylic anhydride are also assigned as acylating chemicals. On the other hand, the anhydrides and reactive esters quickly hydrolyze in water and might not be applicable to existing *in vitro* test methods. Since acyl transfer chemicals or amine reactive chemicals are potentially under-predicted by the BN ITS-3, these chemicals would need to undergo risk assessment evaluation using associated *in vivo* and human data of suitable analogs, to support read across. Secondary, hexyl salicylate is a lipophilic sensitizer with log Kow = 5.06. Of the 29 chemicals, undec-10-enal, p-isobutyl- α -methyl hydrocinnamaldehyde, and farnesol were also lipophilic sensitizers with log Kow > 3.5. Chemicals with log Kow > 3.5, with the h-CLAT, or log Kow > 5, with the KeratinoSensTM, tend to produce false-negative results due to the aqueous nature of the cell culture system and solubility issues. Thus, the BN ITS-3, as proposed by Jaworska *et al.* (2015), used water solubility cutoffs for selecting data from the DPRA, KeratinoSensTM, and h-CLAT in only the analysis of the model test set, but not training set. The utility of cutoffs may need further investigation using an expanded dataset. Since lipophilic chemicals with high log Kow are potentially under-predicted chemicals by the BN ITS-3, these chemicals would need to be evaluated for risk assessment using suitable analogs to support read across. Finally, diethylenetriamine was predicted a pro-hapten by the TIMES. In addition, of the 29 chemicals, farnesol and dibenzyl ether are predicted pre-haptens and 4-chloroaniline is a pro-hapten by the TIMES. 2-Nitro-4-phenylenediamine and dihydroeugenol are also pre/pro-haptens by the TIMES. Pro-haptens need to be metabolically activated to act as sensitizers. Pre-haptens need to be abiotically (through autoxidation) activated to act as sensitizers. Jaworska *et al.* (2015) required more careful interpretation of BN ITS-3 for pre/pro-haptens, but the related processing steps have not been determined. Associated *in vivo* and human data from suitable analogs might be useful for risk assessment, and to minimize such uncertainty associated with predictions for pre/pro-haptens in the BN ITS-3. Taken together, most chemicals that were under-predicted by the BN ITS-3 were acylating chem-

icals (amine reactive chemicals), pre/pro-haptens, and lipophilic chemicals ($\log K_{ow} > 3.5$), which can be identified by using the TIMES or EPI suite. The prediction models of BN ITS-3 rely heavily on *in vitro* test methods, and also possess their corresponding limitations. Therefore, factors associated with the applicability domain of the individual methods would affect the performance of BN ITS-3.

When acylating chemicals (amine reactive chemicals), pre/pro-haptens, and lipophilic chemicals ($\log K_{ow} > 3.5$) were excluded from the applicability domain, the BN ITS-3 prediction slightly improved. However, the uncertainty that the BN ITS-3 results might fall within one class mis-prediction of LLNA results would remain unsolved. Therefore, a worst value of EC3 to each potency class of the BN ITS-3 (e.g. EC3 = 10% for weak sensitizers) would serve as NESIL for the QRA, accounting for being misclassified as one class lower. Unless associated *in vivo* and human data on suitable analogs are available, adjustment of NESILs derived from BN ITS-3 would be useful for skin sensitization risk assessment.

The LLNA results in our database were collected from the published available information. However, the analysis of inherent variability of *in vivo* reference data (i.e. LLNA) is an important aspect requiring further considerations when evaluating testing strategies. Dumont *et al.* (2016) reported that the number of discordant results increases when a chemical is tested in more than one solvent. Therefore, further characterization of variability and uncertainty in the LLNA data is needed.

The case studies on fragrance ingredients within the applicability domain of BN ITS-3 indicate that the BN ITS-3 could predict human NESIL with accuracy similar to that of LLNA, but the predictions of BN ITS-3 and LLNA seem a bit less conservative than human NESIL. Historically, a human study to support actual use levels for a chemical has been conducted at a dose close to or below the EC3 of the LLNA. The confirmatory human studies have been tested at a maximal concentration well below the expected NOEL. In contrast, the LLNA- and BN ITS-3-derived NESIL are not actual NOELs but concentrations at which an effect is observed.

In conclusions, our results indicated that, even when mis-predictions by the BN ITS-3 happened, almost all of them fell within one potency class under-prediction when compared with LLNA. For chemicals within newly defined applicability domain, no chemicals were incorrectly classified by more than one class. Thus, a worst value of EC3 to each potency class from the BN ITS-3 was assigned to derive a NESIL, taking into account uncertainty by being misclassified as one class lower.

When *in vivo* and human data obtained from suitable analogs are not used to estimate the prediction uncertainty, adjustment of the NESIL based on information derived from BN ITS-3 would be useful for skin sensitization risk assessment.

ACKNOWLEDGMENTS

The authors are grateful to Petra S. Kern and Joanna S. Jaworska (Procter & Gamble) for supporting BN ITS-3 accessibility and discussing the data generated by the BN ITS-3.

Conflict of interest— The authors declare that there is no conflict of interest.

REFERENCES

- Api, A.M., Basketter, D.A., Cadby, P.A., Cano, M.F., Ellis, G., Gerberick, G.F., Griem, P., McNamee, P.M., Ryan, C.A. and Safford, R. (2008): Dermal sensitization quantitative risk assessment (QRA) for fragrance ingredients. *Regul. Toxicol. Pharmacol.*, **52**, 3-23.
- Ashikaga, T., Yoshida, Y., Hirota, M., Yoneyama, K., Itagaki, H., Sakaguchi, H., Miyazawa, M., Ito, Y., Suzuki, H. and Toyoda, H. (2006): Development of an *in vitro* skin sensitization test using human cell lines: the human Cell Line Activation Test (h-CLAT). I. Optimization of the h-CLAT protocol. *Toxicol. In Vitro*, **20**, 767-773.
- Ashikaga, T., Sakaguchi, H., Sono, S., Kosaka, N., Ishikawa, M., Nukada, Y., Miyazawa, M., Ito, Y., Nishiyama, N. and Itagaki, H. (2010): A comparative evaluation of *in vitro* skin sensitisation tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). *Altern. Lab. Anim.*, **38**, 275-284.
- Basketter, D. and Safford, B. (2016): Skin sensitization quantitative risk assessment: A review of underlying assumptions. *Regul. Toxicol. Pharmacol.*, **74**, 105-116.
- Cooper, J.A. 2nd, Saracci, R. and Cole, P. (1979): Describing the validity of carcinogen screening tests. *Br. J. Cancer*, **39**, 87-89.
- Dumont, C., Barroso, J., Matys, I., Worth, A. and Casati, S. (2016): Analysis of the Local Lymph Node Assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches. *Toxicol. In Vitro*, **34**, 220-228.
- Emter, R., Ellis, G. and Natsch, A. (2010): Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers *in vitro*. *Toxicol. Appl. Pharmacol.*, **245**, 281-290.
- European Commission. (2018): THE SCCS NOTES OF GUIDANCE FOR THE TESTING OF COSMETIC INGREDIENTS AND THEIR SAFETY EVALUATION 10TH REVISION. https://ec.europa.eu/health/sites/health/files/scientific_committees/consumer_safety/docs/sccs_o_224.pdf
- European Union (2009): Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. *Official Journal of the European Union*, **342**, 59-209.
- Felter, S.P., Ryan, C.A., Basketter, D.A., Gilmour, N.J. and

Adjustment of NESIL by BN ITS3 for skin sensitization risk assessment

- Gerberick, G.F. (2003): Application of the risk assessment paradigm to the induction of allergic contact dermatitis. *Regul. Toxicol. Pharmacol.*, **37**, 1-10.
- Gerberick, G.F., Robinson, M.K., Felter, S.P., White, I.R. and Basketter, D.A. (2001): Understanding fragrance allergy using an exposure-based risk assessment approach. *Contact Dermat.*, **45**, 333-340.
- Gerberick, G.F., Vassallo, J.D., Bailey, R.E., Chaney, J.G., Morrall, S.W. and Lepoittevin, J.P. (2004): Development of a peptide reactivity assay for screening contact allergens. *Toxicol. Sci.*, **81**, 332-343.
- Gerberick, G.F., Vassallo, J.D., Foertsch, L.M., Price, B.B., Chaney, J.G. and Lepoittevin, J.P. (2007): Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. *Toxicol. Sci.*, **97**, 417-427.
- Goebel, C., Aeby, P., Ade, N., Alépée, N., Aptula, A., Araki, D., Dufour, E., Gilmour, N., Hibatallah, J., Keller, D., Kern, P., Kirst, A., Marrec-Fairley, M., Maxwell, G., Rowland, J., Safford, B., Schellauf, F., Schepky, A., Seaman, C., Teichert, T., Tessier, N., Teissier, S., Weltzien, H.U., Winkler, P. and Scheel, J. (2012): Guiding principles for the implementation of non-animal safety assessment approaches for cosmetics: skin sensitisation. *Regul. Toxicol. Pharmacol.*, **63**, 40-52.
- IFRA. (2015): NOTIFICATION OF IFRA STANDARDS No. 13, 48th Amendment to the IFRA Code of Practice. <https://ifrafragrance.org/docs/default-source/ifra-code-of-practice-and-standards/ifra-standards---48th-amendment/ifra-standards-in-full---booklet.pdf>
- Jaworska, J.S., Natsch, A., Ryan, C., Strickland, J., Ashikaga, T. and Miyazawa, M. (2015): Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. *Arch. Toxicol.*, **89**, 2355-2383.
- Kimber, I., Basketter, D.A., Gerberick, G.F., Ryan, C.A. and Dearman, R.J. (2011): Chemical allergy: translating biology into hazard characterization. *Toxicol. Sci.*, **120** (Suppl 1), S238-S268.
- Kleinstreuer, N.C., 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. and Petersohn, D. (2018): Non-animal methods to predict skin sensitization (II): an assessment of defined approaches. *Crit. Rev. Toxicol.*, **48**, 359-374.
- McNamee, P.M., Api, A.M., Basketter, D.A., Frank Gerberick, G., Gilpin, D.A., Hall, B.M., Jowsey, I. and Robinson, M.K. (2008): A review of critical factors in the conduct and interpretation of the human repeat insult patch test. *Regul. Toxicol. Pharmacol.*, **52**, 24-34.
- Natsch, A., Ryan, C.A., Foertsch, L., Emter, R., Jaworska, J., Gerberick, F. and Kern, P. (2013): A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. *J. Appl. Toxicol.*, **33**, 1337-1352.
- Natsch, A., Emter, R., Gfeller, H., Haupt, T. and Ellis, G. (2015): Predicting skin sensitizer potency based on *in vitro* data from KeratinoSens and kinetic peptide binding: global versus domain-based assessment. *Toxicol. Sci.*, **143**, 319-332.
- Natsch, A., Emter, R., Haupt, T. and Ellis, G. (2018): Deriving a No Expected Sensitization Induction Level for Fragrance Ingredients Without Animal Testing: An Integrated Approach Applied to Specific Case Studies. *Toxicol. Sci.*, **165**, 170-185.
- Nukada, Y., Miyazawa, M., Kazutoshi, S., Sakaguchi, H. and Nishiyama, N. (2013): Data integration of non-animal tests for the development of a test battery to predict the skin sensitizing potential and potency of chemicals. *Toxicol. In Vitro*, **27**, 609-618.
- OECD. (2012): The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins Part I: Scientific evidence. OECD Environment, Health and Safety Publications Series on Testing and Assessment, 168, 1-46. [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2012\)10/part1&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2012)10/part1&doclanguage=en)
- OECD. (2016): OECD 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 Sensitization. Organisation for Economic Cooperation and Development. <https://one.oecd.org/document/ENV/JM/MONO%282016%2929/ANN1/en/pdf>
- Otsubo, Y., Nishijo, T., Miyazawa, M., Saito, K., Mizumachi, H. and Sakaguchi, H. (2017): Binary test battery with KeratinoSens™ and h-CLAT as part of a bottom-up approach for skin sensitization hazard prediction. *Regul. Toxicol. Pharmacol.*, **88**, 118-124.
- 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. and 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.
- Patlewicz, G., Kuseva, C., Mehmed, A., Popova, Y., Dimitrova, G., Ellis, G., Hunziker, R., Kern, P., Low, L., Ringeissen, S., Roberts, D.W. and Mekenyan, O. (2014): TIMES-SS--recent refinements resulting from an industrial skin sensitisation consortium. *SAR QSAR Environ. Res.*, **25**, 367-391.
- Piroird, C., Ovigne, J.M., Rousset, F., Martinozzi-Teissier, S., Gomes, C., Cotovio, J. and 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.
- Sakaguchi, H., Ashikaga, T., Miyazawa, M., Yoshida, Y., Ito, Y., Yoneyama, K., Hirota, M., Itagaki, H., Toyoda, H. and Suzuki, H. (2006): Development of an *in vitro* skin sensitization test using human cell lines; human Cell Line Activation Test (h-CLAT). II. An inter-laboratory study of the h-CLAT. *Toxicol. In Vitro*, **20**, 774-784.
- 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. and 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.
- 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. and Sakaguchi, H. (2015): Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul. Toxicol. Pharmacol.*, **71**, 337-351.