Research Article

Applying a Next Generation Risk Assessment Framework for Skin Sensitization to Inconsistent New Approach Methodology Information

Nicola Gilmour[1], Nathalie Alépée[2], Sebastian Hoffmann[3], Petra S. Kern[4], Erwin van Vliet[5], Dagmar Bury[6], Masaaki Miyazawa[7], Hayato Nishida[8] and Cosmetics Europe[9]

[1]Unilever, Colworth Science Park, Bedford, United Kingdom; [2]L'Oréal, Research & Innovation, Aulnay-sous-Bois, France; [3]seh consulting + services, Paderborn, Germany; [4]Procter & Gamble Services NV/SA, Strombeek-Bever, Belgium; [5]Innovitox Consulting & Services, Houten, The Netherlands; [6]L'Oréal, Research & Innovation, Clichy, France; [7]Kao Corporation, Tochigi, Japan; [8]Shiseido Global Innovation Center, Kanagawa, Japan; [9]Brussels, Belgium

Abstract

Cosmetic products must be safe for their intended use. Regulatory bans on animal testing for new ingredients have resulted in a shift towards the use of new approach methodologies (NAMs) such as *in silico* predictions and *in chemico / in vitro* data. Defined approaches (DAs) have been developed to interpret combinations of NAMs to provide information on skin sensitization hazard and potency, three having been adopted within OECD Test Guideline 497. However, the challenge remains as to how DAs can be used to derive a quantitative point of departure for use in next generation risk assessment (NGRA). Here we provide an update to our previously published NGRA framework and present two hypothetical consumer risk assessment scenarios (rinse-off and leave-on) on one case study ingredient. Diethanolamine (DEA) was selected as the case study ingredient based on the existing NAM information demonstrating differences with respect to the outcomes from *in silico* predictions and *in chemico / in vitro* data. Seven DAs were applied, and these differences resulted in divergent DA outcomes and reduced confidence with respect to the hazard potential and potency predictions. Risk assessment conclusion for the rinse-off exposure led to an overall decision of safe for all applied DAs. Risk assessment conclusion for the higher leave-on exposure was safe when based on some DAs but unsafe based on others. The reasons for this were evaluated as well as the inherent uncertainty from the use of each NAM and DA in the risk assessment, enabling further refinement of our NGRA framework.

1 Introduction

All cosmetic products that are placed onto the market must be safe for their intended use and as such must undergo a human health risk assessment (SCCS, 2021) (BPR, Regulation (EU) 528/2012). In Europe, a ban on animal testing for new cosmetic ingredients was implemented within the cosmetics legislation (Regulation (EC) No 1223/2009) in 2009, with a marketing ban implemented for skin sensitization from 2013. Other geographic regions have and continue to follow suit[1] (Daniel et al., 2018). Thus, risk assessment of cosmetics and their ingredients has shifted toward the use of new approach methodologies (NAMs) such as *in silico* predictions and *in chemico* / *in vitro* data.

The development of NAM within the field of skin sensitization has been particularly successful, aided by our understanding of the molecular mechanisms of skin sensitization and documentation of these within "The Adverse Outcome Pathway (AOP) for Skin Sensitisation" (OECD, 2014). A number of *in chemico* and *in vitro* NAM addressing key events (KEs) involved in the induction of skin sensitization have been developed (Ezendam et al., 2016), some of which have now been validated and adopted as Organisation for Economic Co-operation and Development (OECD) test guidelines (TG). OECD TG 442C describes the direct peptide reactivity assay (DPRA), amino acid derivative reactivity assay (ADRA), and the kinetic DPRA, which are test methods addressing KE 1 – the binding of haptens to

[1] https://www.afsacollaboration.org/cosmetics-global-regulatory-alignment/

Received November 16, 2022; Accepted February 24, 2023;
Epub March 14, 2023; © The Authors, 2023.

ALTEX 40(3), 439-451. doi:10.14573/altex.2211161

Correspondence: Nicola Gilmour
Unilever, Colworth Science Park
Bedford, MK44 1LQ, United Kingdom
(Nicola.Gilmour@Unilever.com)

proteins in the skin (OECD, 2020). OECD TG 442D describes KeratinoSens™ and LuSens, which are test methods addressing KE 2 – the activation of keratinocytes (OECD, 2018). OECD TG 442E describes h-CLAT, U-Sens™, IL-8 Luc assay, and GARDskin™, which are test methods addressing KE 3 – the activation of dendritic cells (OECD, 2022). Currently there are no NAMs addressing KE 4 – the activation and proliferation of a T cell response – that are sufficiently progressed for implementation in OECD TGs or for use in a next generation risk assessment (NGRA) (van Vliet et al., 2018). Due to the sequential nature of the AOP, the need for a NAM addressing KE 4 might not be a general requirement for all NGRAs.

NAMs can be used to provide information on skin sensitization hazard and potency. Defined approaches (DAs), which follow a specific data interpretation procedure, have been developed to interpret combinations of NAM information (Ezendam et al., 2016; Gilmour et al., 2020; Hoffmann et al., 2018; Kleinstreuer et al., 2018; Tollefsen et al., 2014). Three DAs that provide a skin sensitization hazard or UN GHS classification have now been adopted within OECD TG 497: Defined Approaches on Skin Sensitisation (OECD, 2021). Despite this progress in the identification of skin sensitizer hazard or UN GHS classes, there remains the challenge of how predictions from DAs can be used to derive a quantitative point of departure (PoD) for use in a NGRA to ensure consumer safety.

Cosmetics Europe has implemented a scientific research program, the Long-Range Science Strategy, to foster the development, assessment, and application of NAMs in human health risk assessments and to support their regulatory acceptance (Desprez et al., 2018). For skin sensitization, this work has included review/evaluation of available NAMs (Reisinger et al., 2015), prioritization of six NAMs to generate a database of NAM information (Hoffmann et al., 2022), which was used to assess the DAs and integrated approaches for testing and assessment (IATAs) for skin sensitization and ultimately propose a NGRA framework to allow application of a structured logic to the skin sensitization risk (Ezendam et al., 2016; Kleinstreuer et al., 2018; Gilmour et al., 2020; SCCS, 2021). Risk assessment case studies using only NAMs are being increasingly utilized to explore application of the different DAs aligned to such NGRA frameworks in hypothetical risk assessment scenarios (Assaf Vandecasteele et al., 2021; ENV/CBC/MONO(2022)32, 2022; Gautier et al., 2020; Gilmour et al., 2020, 2022; Natsch et al., 2018).

Here, diethanolamine (DEA) was selected as the case study ingredient based on the existing NAM information being available in the public domain and demonstrating differences with respect to the outcomes from *in silico* predictions and *in chemico* / *in vitro* data (Hoffmann et al., 2022). Whilst read-across, including the use of analogue data, could be used to reduce uncertainty in the risk assessment, this was considered out of scope for this case study to allow focus on how to deal with conflicting data in the absence of analogues. DEA is a pH adjuster, used in the manufacturing industry and in cosmetics, though not the EU and Canada where its use in cosmetics is prohibited (European Union Council Directive, 76/768 EEC) (Fiume et al., 2017). Hypothetical risk assessments using only NAM information were conducted for two consumer exposure scenarios: 0.8% DEA used in a shampoo (rinse-off product representing a low exposure) and 0.8% DEA used in an underarm deodorant (leave-on product representing a relatively high exposure). These are hypothetical exposures and do not represent realistic use levels. The impact of the inconsistent outcomes from the *in silico* predictions and the *in chemico* / *in vitro* data was analyzed with respect to the outcome of seven DA and the risk assessment decisions for the given exposure scenarios for each of the individual DA.

This case study represents an example of a complex risk assessment scenario that aims to address causes of and issues surrounding use of data sets with differing outcomes in risk assessment. Exploring the added complexity led to an update to the NGRA framework (Gilmour et al., 2020) by providing additional clarity on the hypothesis generation and how the available data will be used in risk assessment. Case studies such as the one presented here will enable us to continue to build our knowledge on the strengths and limitations of the NAM and confidence in the application of NAM within a NGRA framework for skin sensitization.

2 Materials and methods

2.1 Selection of case study ingredient and scope

Diethanolamine (DEA) (CASRN# 111-42-2) was selected as the case study ingredient based on availability of existing NAM information where differences were evident with respect to the outcomes from the *in silico* predictions and the *in chemico* / *in vitro* data (Hoffmann et al., 2022). The aim of this case study was to explore the impact of inconsistent NAM information on the final risk assessment outcome based on two exposure scenarios, i.e., rinse-off and leave-on consumer use. The exposure scenarios are hypothetical, i.e., do not represent real consumer exposures, and were designed to exclude the application of exposure-based waiving and read-across within the NGRA framework.

2.2 Risk assessment framework

The previously published NGRA framework (Gilmour et al., 2020) was applied and then updated (Fig. 1) based on the learn-

Abbreviations: ADRA, amino acid derivative reactivity assay; AOP, adverse outcome pathway; BN, Bayesian network; CEL, consumer exposure level; DA, defined approach; DASS, defined approaches on skin sensitization; DEA, diethanolamine; DPRA, direct peptide reactivity assay; ED_{01}, dose predicted to sensitize 1% of a HRIPT population; HRIPT, human repeat insult patch test; IATA, integrated approaches for testing and assessment; ITS, integrated testing strategy; KE, key event; LLNA, local lymph node assay; MoE, margin of exposure; MW, molecular weight; NAM, new approach methodology; NGRA, next generation risk assessment; OECD, Organisation for Economic Co-operation and Development; PoD, point of departure; QRA, quantitative risk assessment; QSPR, quantitative structure-property relationship; SAF, safety assessment factors; STS, sequential testing strategy; WoE, weight of evidence

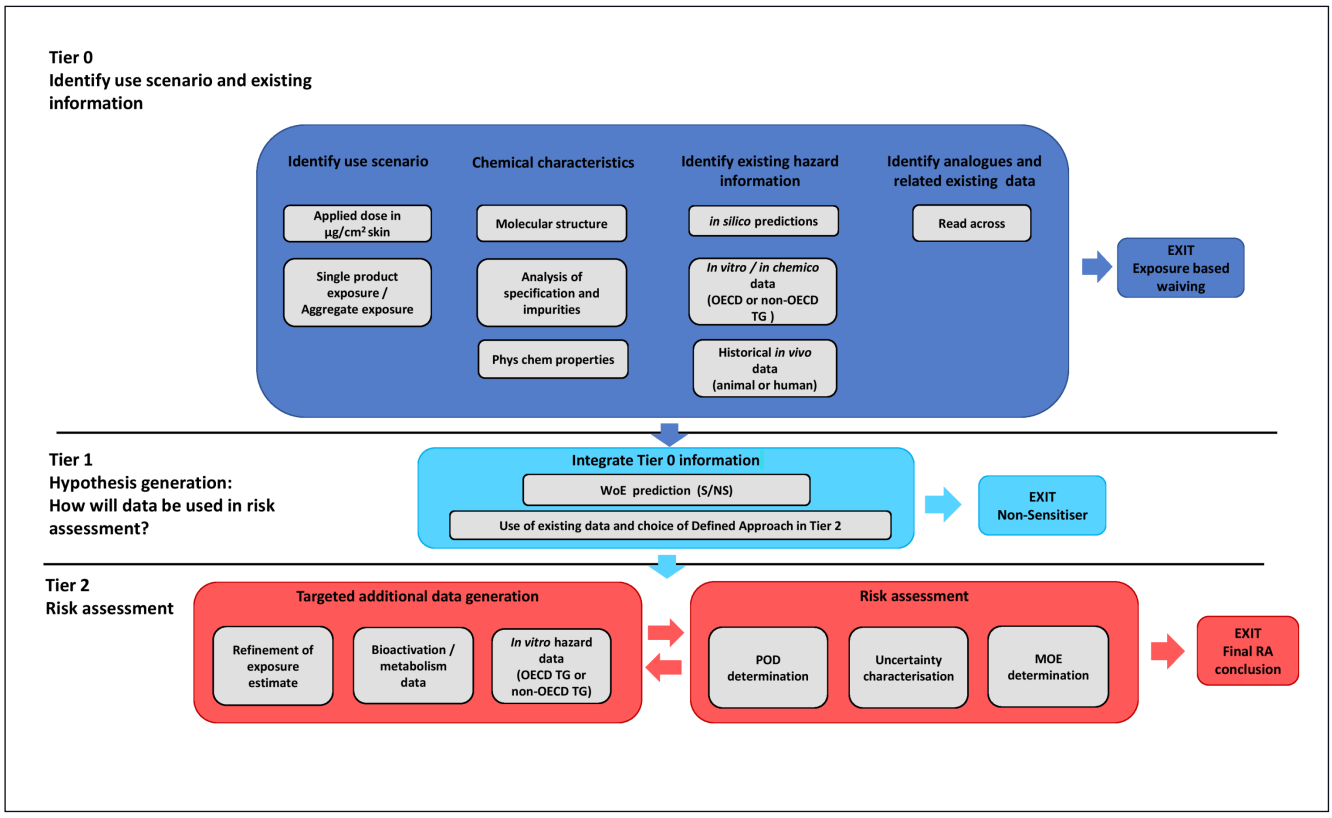

Fig. 1: Updated NGRA framework as applied to the presented case study

ings from this case study (described below). In short, the framework is a tiered and iterative approach, involving first the collation of all available information; thereafter hypothesis generation and consideration as to how this information can be used within the risk assessment; whether further data generation is required; and then the final risk assessment conclusion.

2.3 Available information applied within this case study

Exposure

The consumer exposure level (CEL) in dose per unit area (µg/cm^2) applied per day for DEA in each product type was calculated based on the 90th percentile consumer use provided in the SCCS Notes of Guidance (SCCS, 2021).

Chemical characteristics

Information on molecular weight (MW), LogP, LogS, LogVP, boiling point and melting point were sourced from a previously published dataset (Hoffmann et al., 2022). These values were calculated using the quantitative structure-property relationship (QSPR) (Zang et al., 2017). The pH was measured using an adaptation of the method described in OECD TG 122 (OECD, 2013), and information on volatility was obtained from the MPBPWIN™ model from the open source Episuite™ software version 4.1 (Spicer et al., 2002). The BN-ITS uses some additional physicochemical properties (LogD and plasma protein binding) calculated by ACD Lab software (version 2021.2.0). Fraction ionized is calculated via LogP/LogD.

Available (NAM) hazard information

Information from the NAMs listed below were sourced from a previously published dataset (Hoffmann et al., 2022):

- *Derek Nexus v. 6.1:* A prediction of skin sensitization potential was obtained from Derek Nexus, a commercial knowledge-based expert system with a set of rules coded as structural alerts developed by Lhasa Ltd (Chilton et al., 2018).
- *TIMES-SS V2.30.1.11:* Predictions of skin sensitization potency for parent chemical and its potential metabolites were obtained from The TImes MEtabolism Simulator platform, a hybrid commercial expert system developed by the Laboratory of Mathematical Chemistry that encodes structure-toxicity and structure-skin metabolism relationships through a number of transformations (Patlewicz et al., 2007, 2014).
- *ToxTree v. 3.1.0:* A prediction of the skin sensitization reactivity domain was obtained from ToxTree, a rule-based system for the assignment of substances to one or several skin sensitization reactivity domains (Enoch et al., 2008). Information using this version was collated in the reference dataset (Hoffmann et al., 2022).
- *OECD QSAR Toolbox v 4.5:* A skin sensitization prediction was obtained from the OECD QSAR Toolbox skin sensitiza-

tion automated workflow for DASS (defined approaches on skin sensitization), a software application for assessing the hazards of chemicals[2].

- *Direct peptide reactivity assay (DPRA):* Information on reactivity (KE 1), measured as percentage depletion of synthetic cysteine and lysine-containing peptides, was obtained from the DPRA, OECD TG 442C (OECD, 2020; Gerberick et al., 2004).
- *KeratinoSens™:* Information on cellular activation (KE 2), measured as activation of the Keap1-Nrf2-ARE-pathway, was obtained from the KeratinoSens™ assay, OECD TG 442D (Natsch and Emter, 2016; OECD, 2018).
- *h-CLAT:* Information on dendritic cell activation (KE 3), measured as expression of the cell surface markers CD86 and CD54 in the human monocytic leukemia cell line THP-1, was obtained from the h-CLAT assay, OECD TG 442E (OECD, 2022; Sakaguchi et al., 2010).
- *U-SENS™:* Information on dendritic cell activation (KE 3), measured as expression of the cell surface marker CD86 in the human histiocytic lymphoma cell line U937, was obtained from the U-SENS™ assay, OECD TG 442E (OECD, 2022; Piroird et al., 2015).

2.4 DA applied in case study

The following seven DA were explored within this case study based on a combination of factors including the authors' working knowledge of the DA; accessibility of the DA; availability of input data; ability to derive a PoD; and previous use within NGRA case studies for cosmetic ingredients (ENV/CBC/MONO(2022)32, 2022; Gilmour et al., 2020). Since each DA is described in detail elsewhere, only a brief description is provided below:

Integrated testing strategy (ITSv1 and ITSv2)

The ITSv1 and ITSv2 DAs were recently adopted as OECD TG 497 (OECD, 2021). The DAs require information from the DPRA, h-CLAT, and Derek Nexus (ITSv1) or the OECD automated workflow (ITSv2). All inputs are converted to a score of 0 to 3; then the summed score, ranging between 0 and 7, can be used to predict the skin sensitizing potential and a sub-categorization according to the UN GHS: Cat. 1A (6-7), Cat. 1B (2-5), not classified (non-sensitizers) (0-1), or inconclusive (0-1, due to either the *in silico* prediction or one of the OECD test methods being out of domain).

Artificial neural network (ANN) model (ANN-TIMES and ANN-ToxTree)

The ANN is an ITS for prediction of the skin sensitization potential and potency (predicted EC3) (Hirota et al., 2018). The model uses information from the DPRA, KeratinoSens™, h-CLAT, and TIMES-SS or ToxTree (ANN-TIMES and ANN-ToxTree). The ANN is a nonlinear statistical data modelling tool that predicts local lymph node assay (LLNA) EC3 values that can be used as (continuous) potency estimations or to subcategorize skin sensitizers into UN GHS Cat. 1A and 1B. Whilst individual inputs can be used to make a prediction using the ANN, the performance improves when all inputs are utilized.

Sequential testing strategy (STS) DA

The STS DA is a tiered approach providing hazard identification and potency category information (UN GHS Cat. 1A/1B/not classified). The first tier (hazard identification) requires information from DPRA, U-SENS™, KeratinoSens™, TIMES-SS, and physicochemical parameters (pH and volatility). The second tier (hazard potency) requires information from DPRA, KeratinoSens™, U-SENS™ (and optionally SENS-IS (Cottrez et al., 2016)), and physicochemical parameters (MW, volatility, and clogP). The STS is a meta-model stacking five statistical methods: boosting, naïve Bayes, support vector machine, sparse PLS-DA, and expert scoring (Gomes et al., 2014). Each statistical method provides a probability of the chemical being a skin 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 skin sensitizer (Gomes et al., 2012; Nocairi et al., 2016; Wolpert, 1992), providing a probability to belong to the group of interest (UN GHS Cat. 1A, UN GHS Cat. 1B, not classified (non-sensitizers)). Thresholds of $\geq 70\%$ probability of being a skin sensitizer are applied for UN GHS Cat. 1 classification and $\leq 30\%$ for categorization as not classified. Based on comparison to historical data, it was shown that the level of confidence in the hazard prediction was high when these thresholds were applied (Del Bufalo et al., 2018).

Bayesian network ITS (BN-ITS) DA

The BN-ITS DA is a Bayesian network (BN) that provides skin sensitizer hazard identification and allows potency derivation distinguishing four categories (ENV/CBC/MONO(2022)32, 2022; Jaworska et al., 2015). The model uses information from DPRA, KeratinoSens™, h-CLAT, TIMES-SS, and physicochemical properties (water solubility at pH7 (M), protein binding (% bound), log D (pH7), fraction ionised (calculated using LogKOW and LogD pH7)). The model prediction is given as a potency probability distribution, the pEC3, over 4 potency classes: non-sensitizer, weak, moderate, strong/extreme. The BN-ITS can reason and provide potency information using partial datasets as input. In general, using all input parameters will lower uncertainty in the prediction, and by adding additional data sequentially one can explore the impact on confidence in the prediction. Expressing the prediction as a probability distribution allows quantification of uncertainty in the pEC3. In a subsequent step, the probability is converted into Bayes factors, estimating strength of evidence, also for partial input datasets (Goodman, 1999), and the final decision of the predicted potency category is based on the highest Bayes factor in a category.

SARA model

The SARA model utilizes a Bayesian statistical approach to infer a human-relevant metric of sensitizer potency and a measure of consumer risk for any given consumer exposure to a chemical of interest (Reynolds et al., 2022). It can utilize any combination

[2] https://qsartoolbox.org/

of data from human repeat insult patch tests (HRIPT) (Politano and Api, 2008), LLNA (OECD, 2010), DPRA, KeratinoSens™, h-CLAT or U-SENS™ to derive sensitizer potency, which is expressed as the ED_{01} (the dose in μg/cm^2 which is predicted to sensitize 1% of a HRIPT population), with explicit quantification of the uncertainty associated with the prediction. Whilst the SARA model can derive a potency and risk prediction based on a single NAM input, the precision of the prediction depends on which NAM is used and the experimental values obtained from the NAM. Generally, an increased number of inputs will reduce uncertainty in the SARA predictions.

In addition to a potency assessment, the SARA model also provides the probability of whether a given exposure is low-risk (SARA risk metric). The margin of exposure (MoE) is calculated for the given exposure within the SARA model by dividing the ED_{01} by the CEL (in μg/cm^2). This MoE is then regressed against the MoE derived for benchmark consumer exposures. The benchmark exposures are historical/current exposures to consumer products that have been designated as high- or low-risk for induction of skin sensitization based on publicly available clinical/patch test data.

2.5 Risk assessment

Determination of point of departure (PoD)

The predictions from the DAs were converted to PoD, based on each unique DA and outcome. The purpose here was not to use the information incorporated within the DA and the DA predictions in isolation but to apply them in the context of all available information within the NGRA. As described below in the hypothesis, uncertainty was introduced from the different outcomes observed in the available information for DEA. Therefore, whilst not necessarily required for all purposes, in this case study scenario a DA prediction of non-sensitizer was converted into a PoD with the aim to further increase confidence in the risk assessment. The PoD derivation was conducted for this case study as follows:

- For DAs that predicted a non-sensitizer outcome for DEA, an EC3 value of 100% (negative LLNA) was utilized to derive a PoD.
- For DAs that predicted a sensitizer UN GHS Cat. 1B for DEA, a default LLNA EC3 value of > 2% was applied (GHS, 2021).
- The ANN output is a predicted EC3 value (in %).
- The SARA model makes a prediction of the human sensitization potency, the ED_{01}; this value is used within the model as the PoD (Reynolds et al., 2022).

Whilst several proposals have been published on how to convert LLNA EC3 values into sensitization potency categories or PoD values for risk assessment (Api et al., 2008; Griem et al., 2003), for this case study we applied a unified approach, converting the DA-predicted LLNA threshold values (EC3%) into a PoD as dose per unit area (in μg/cm^2) by using a factor of 250 (Robinson et al., 2000). This is based on the standard LLNA protocol where 25 μL test solution is distributed over a surface of 1 cm^2 per mouse ear (Griem et al., 2003).

Margin of exposure

The margin of exposure (MoE) is calculated using the following equation:

$$MoE = PoD / CEL$$

Risk assessment outcome

The overall risk assessment outcome is evaluated as a weight of evidence (WoE) considering the calculated MoE (and in the case of SARA, the corresponding p(low risk)), the confidence in use of NAM input data within the DA, and the relative conservatism in the transformation of the DA outcome to a PoD.

- The value associated with an acceptable MoE for risk assessments based on NAMs is yet to be defined. For the purposes of this case study, the MoE was considered high if > 100 and considered low if < 100 in order to illustrate the decision-making process.
- Note, this does not apply to SARA, which incorporates clinical benchmarks to provide empirical support for the size of the MoE.
- Confidence in use of the NAM data informing each DA was defined by whether the chemical was predicted to be directly or indirectly reactive (i.e., a pre- or pro-hapten), which indicates whether the test method should be considered as in or out of domain and whether the *in silico* prediction was in domain. For the purposes of this case study, this was considered high if all NAM utilized in the DA were within domain and low if all NAM utilized in the DA were out of domain.
- Relative conservatism in transformation of DA outcome to PoD was considered high when the PoD was derived from a non-sensitizer outcome from the DA. It was considered low when the PoD was provided as a quantitative output of the DA and considered as unknown when the PoD was derived from CLP Cat. 1B or 1A DA outputs.

3 Results

3.1 Updated NGRA framework

In the process of constructing this case study, it became evident that the original NGRA framework (Gilmour et al., 2020) could be refined. The updated framework is shown in Figure 1. The modifications are as follows:

- The elements of Tier 0 (gathering existing information) have been grouped together as they are applied without any specific order, e.g., not in a sequence as indicated in the earlier version.
- Within Tier 0, the gathering of chemical characteristics was expanded to explicitly include the physicochemical parameters, which are used in several DA.
- Tier 1 (hypothesis generation: How will data be used in risk assessment) was simplified to better reflect the logic applied when the integration of the information allows (or does not allow) a WoE conclusion of non-sensitizer with high confidence (and thus exits) from Tier 1. This is described in more detail in the following section.

Tab. 1: Summary of which NAM information are used within the individual DA applied within this case study

NAM	NAM information	ITSv1	ITSv2	ANN (TIMES-SS)	ANN (ToxTree)	STS		BN-ITS	SARA
						Tier 1	Tier 2		
PC properties	MW: 105.14 Da						√		
	LogP: -1.46						√		
	Fraction ionized: 0							√	
	LogD @ pH 7: -3.38							√	
	LogKoW							√	
	Volatility: semi-volatile					√	√		
	pH: 10.3					√			
	H_2O solubility @ pH 7: 3.0 g/L							√	
	Plasma protein binding (% bound): 11.3							√	
TIMES-SS	Parent: non-sensitizer; Metabolite: non-sensitizer			√		√		√	
TOXTREE	Protein binding alert: Schiff base				√	√			
OECD Toolbox	OASIS protein binding alerts for skin sensitization: negative/no alerts; Skin sensitization automated workflow for DASS: negative/non-sensitizer		√						
DEREK Nexus	Positive/sensitizer (equivocal)	√							
Mechanistic domain: expert review	Pro-Schiff base								
DPRA	Negative/minimal					√			
	Cys depl.: 5.9%	√	√	√	√		√	√	√
	Lys depl.: 2.2%	√	√	√	√		√	√	√
KeratinoSens™	Negative					√			
	EC1.5: > 2000 μM			√	√		√	√	√
	EC3: > 2000 μM							√	
	Imax: 1.0								
	IC50%: > 2000 μM							√	
U-SENS™	Positive					√			
	CD86 EC150: 26.9 μg/mL						√		√
	CV70: > 200 μg/mL								
h-CLAT	Positive								
	CD86 EC150: 1242.5 μg/mL	√	√	√	√			√	√
	CD54 EC200: 1280.9 μg/mL	√	√	√	√			√	√
	CV75: 2277 μg/mL			√	√			√	

3.2 DEA – Summary of available information

Use scenario

The case study exposure scenarios were selected to represent one relatively low consumer exposure (0.8% DEA in a shampoo) and one relatively high consumer exposure via a product to remain on the skin (0.8% DEA in an underarm deodorant). The exposure was calculated based on 90th percentile consumer use provided in the SCCS Notes of Guidance (SCCS, 2021).

The consumer exposure from use of 0.8% DEA in a shampoo was calculated to be 0.6 µg/cm^2 (11 g shampoo/day x 0.8% use concentration × 0.01 skin retention factor / 1440 cm^2 skin surface).

The consumer exposure from use of 0.8% DEA in an underarm deodorant was calculated to be 60 µg/cm^2 (1.5 g deodorant roll-on/day × 0.8% use concentration × 1 skin retention factor / 200 cm^2 skin surface).

Chemical characteristics

For the purposes of this case study, it was assumed that DEA was 100% pure and that risk assessment of impurities was not required. Physicochemical properties are summarized in Table 1.

Hazard information

Table 1 summarizes which NAMs and values are used within each DA applied in the case study.

3.3 Hypothesis generation and how the data will be used in risk assessment

DEA was assumed to be a novel/new ingredient that is proposed for use in two cosmetic products: 0.8% in a shampoo and 0.8% in an underarm deodorant.

Two of the four applied *in silico* tools predicted no reactivity or skin sensitization potential, all predictions for DEA were in domain for the *in silico* tools. Derek Nexus predicted that DEA was a skin sensitizer, although this outcome was reported as equivocal. A detailed review of the report revealed that this alert (aminophenol) was triggered by differences with respect to sensitizer/non-sensitizer outcome in the historical *in vivo* data underpinning the alert, which can vary from moderate/weak sensitizer potency in the LLNA and negative skin sensitization potential in guinea pig maximization test. The human patch test data also demonstrated that materials within this chemical class are rarely human sensitizers (Basketter et al., 2014; Lessmann et al., 2009). ToxTree triggered a protein binding alert and reported DEA could be reactive via a Schiff base mechanism; this was corroborated by expert review, which concluded that DEA has the potential to form a Schiff base by probiotic activation (i.e., is a pro-hapten). The Derek Nexus report also supports this chemistry hypothesis, indicating that DEA is likely a pro-Schiff base, and as an aminoethanol it is thought to sensitize by reaction pathways involving either abiotic or enzymatic processes in which this class of compounds is oxidized into imine/aldehydes and further hydrolyzed into glycolaldehyde (CASRN# 141-46-8), which has an EC3 = 1.8% (Anderson et al., 2007). Whist the metabolite is a contact allergen of moderate potency, 100% enzymatic conversion is not plausible/likely to happen *in vivo* since biotic transformation requires access to cells within the skin but skin penetration has been shown to be low (below 3% over 24 h in human skin *in vitro*); most of the penetrating DEA remained in the skin reservoir (Brain et al., 2005; Kraeling et al., 2004).

The available NAM data demonstrate differing outcomes. DPRA and KeratinoSens™ were negative while U-SENS™ and h-CLAT were positive, according to the prediction models specified in the respective OECD TGs. DEA was predicted to potentially be a pro-hapten, thus some caution should be applied in interpreting the negative outcomes in the DPRA and KeratinoSens™ due to the theoretical lack of enzymatic metabolic capability in these test systems (OECD, 2018, 2020). A positive outcome was evident in the h-CLAT and U-SENS™, which have been shown to detect pro-haptens due to respective enzyme activities (OECD, 2022). Note that when used in combination, the skin sensitization NAMs can detect most pre- and pro-haptens (Urbisch et al., 2016, Patlewicz et al., 2016).

Based on the above information, it is not possible to reach the conclusion that DEA is a non-sensitizer with high certainty. Thus, the NGRA framework cannot be exited at Tier 1 and should be progressed to Tier 2 (Fig. 1). The next step was the application of DA to combine the NAM information and generate skin sensitization potential and potency predictions.

Whilst in principle for a DA predicting a non-sensitizer the determination of a PoD is not required as a quantitative risk assessment (QRA) is usually not done, based on the reduced confidence in use of some of the NAM test data and in the case of the deodorant, it being associated with high consumer exposure and prolonged occlusive skin contact, a non-sensitizer outcome from a DA should be converted to a PoD. The WoE risk assessment should thus consider the calculated MoE and, in the case of SARA the p(low risk), the confidence in use of NAM input data on DA outcome, and the relative conservatism in transformation of DA outcome to PoD.

3.4 Risk assessment

3.4.1 DA outcome and derivation of PoD

The ITSv1 predicts DEA to be a UN GHS Cat. 1B based on an overall score of 2, derived as described in Section 2.5. This equates to a default LLNA value of > 2%, which when converted to dose per unit area results in a PoD of > 500 µg/cm^2.

The ITSv2 predicts an overall score of 1; however the DA outcome is considered inconclusive based on DEA being predicted to be a pro-hapten and thus out of domain of the DPRA (OECD, 2021). In accordance with OECD TG 497 for a DA prediction with low confidence (inconclusive), a WoE was applied within the context of the IATA (OECD, 2021). As outlined above in the hypothesis, DEA must undergo metabolism to induce skin sensitization. Whist the metabolite is a contact allergen of moderate potency, 100% enzymatic conversion is not plausible/likely to happen *in vivo* since biotic transformation requires access to cells within the skin and skin penetration has been shown to be low with weak responses observed in the h-CLAT and U-SENS™ (Tab. 1). Thus, in the case study risk assessment using the ITSv2, DEA is treated as a UN GHS Cat. 1B. This equates to a default LLNA value of > 2%, which when converted to dose per unit area results in a PoD of > 500 µg/cm^2.

Tab. 2: Summary of NAM risk assessment outcomes based on the 7 DA for the use of 0.8% DEA in a shampoo

DA	ITSv1	ITSv2	ANN (TIMES-SS)	ANN (ToxTree)	STS	BN-ITS	SARA
DA output							
	Cat. 1B	inconclusive	EC3 = 81.5%	EC3 = 59.1%	NS P(NS) = 87%	NS P(NS) = 99% Bayes factor (> 30%)	ED_{01} = 13,000 $\mu g/cm^2$ (5^{th}-95^{th} percentile 530-370,000 $\mu g/cm^2$)
PoD ($\mu g/cm^2$)							
	> 500	> 500	20,375	14,775	25,000	25,000	13,000
Calculate MoE for 0.8% shampoo							
Consumer exposure level ($\mu g/cm^2$)	0.6	0.6	0.6	0.6	0.6	0.6	0.6
MoE (PoD/CEL) **p(low risk)*SARA ONLY**	> 833	> 833	33,958	24,625	41,667	41,667	24,000 p(low risk) = 0.98
Weight of evidence assessment / characterize uncertainty							
Confidence in NAM input	moderate[a]	moderate[a]	moderate[b]	moderate[b]	moderate[b]	moderate[b]	moderate[b]
Conservatism in transformation of DA outcome to PoD	unknown[c]	unknown[c]	low[d]	low[d]	high[e]	high[e]	low[f]
MoE **p(low risk)*SARA ONLY**	high[g]	high[g]	high[g]	high[g]	high[g]	high[g]	p(low risk) = 0.98 high[h]
Risk assessment							
Risk assessment outcome	**SAFE**	**SAFE**	**SAFE**	**SAFE**	**SAFE**	**SAFE**	**SAFE**

[a] Based on use of DPRA. [b] Based on use of DPRA and KeratinoSens™. [c] Based on the PoD being derived based on the use of GHS Cat. 1B, categorizing DEA as a weak/moderate sensitizer with an EC3 value of > 2%, which is converted to a PoD of > 500 $\mu g/cm^2$. The exact PoD value is undetermined. [d] Based on the DA outcome being a quantitative measure of potency (EC3%), which is converted to dose per unit area. [e] The DA outcome was non-sensitizer. [f] The DA outcome is a quantitative measure of potency (ED_{01}), which is converted to p(low risk) for a given exposure. [g] MoE > 100. [h] MoE = 24,000 and the p(low) risk is 0.98, i.e., it is 98% certain that the exposure is low risk.

The different outcomes of the two versions of the ITS were attributable to the different *in silico* tools applied, i.e., Derek Nexus (ITSv1) predicted sensitizer and OECD automated workflow (ITSv2), and the use of an out of domain test method having greater impact and decreasing confidence when the overall DA outcome is non-sensitizer compared to when the outcome is sensitizer.

The ANN (TIMES-SS) predicts an EC3 value of 81.5%, which when converted to dose per unit area results in a PoD of 20,375 $\mu g/cm^2$.

The ANN (ToxTree) predicts an EC3 value of 59.1%, which when converted to dose per unit area results in a PoD of 14,775 $\mu g/cm^2$.

The STS predicts DEA to be a non-sensitizer with a high probability of 87% (pCat1 = 13%), which, as described in Section 2.5, equates to a default LLNA value of 100%, which converted to dose per unit area results in a PoD of 25,000 $\mu g/cm^2$.

The BN-ITS predicts DEA to be a non-sensitizer with a high probability (> 99%) and a high Bayes factor (> 30), which again equates to a default LLNA value of 100%, which converted to dose per unit area results in a PoD of 25,000 $\mu g/cm^2$.

The SARA model predicts an expected ED_{01} of 13,000 $\mu g/cm^2$ (95^{th}% confidence interval of 530-370,000 $\mu g/cm^2$), which is consistent with a prediction of a weak/moderate skin sensitizer potency.

3.4.2 Derivation of a PoD and calculation of MoE

For use of 0.8% DEA in a shampoo, the exposure was calculated to be 0.6 $\mu g/cm^2$, and the MoE obtained from the 7 DA ranged from 833 to 41,667. Individual values are shown in Table 2.

Tab. 3: Summary of NAM risk assessment outcomes based on the 7 DA for the use of 0.8% DEA in an underarm deodorant product

DA	ITSv1	ITSv2	ANN (TIMES-SS)	ANN (ToxTree)	STS	BN-ITS	SARA
DA output							
	Cat. 1B	inconclusive	EC3 = 81.5%	EC3 = 59.1%	NS P(NS) = 87%	NS P(NS) = 99% Bayes factor (> 30%)	ED_{01} = 13,000 $\mu g/cm^2$ (5^{th}-95^{th} percentile 530-370,000 $\mu g/cm^2$)
PoD ($\mu g/cm^2$)							
	> 500	> 500	20,375	14,775	25,000	25,000	13,000
Calculate MoE for 0.8% deodorant							
Consumer exposure level ($\mu g/cm^2$)	60	60	60	60	60	60	60
MoE (PoD/CEL) **p(low risk)*SARA ONLY**	> 8	> 8	340	246	416	416	217 (8.8-617) p(low risk) = 0.5
Weight of evidence assessment / characterize uncertainty							
Confidence in NAM input	moderate[a]	moderate[a]	moderate[b]	moderate[b]	moderate[b]	moderate[b]	moderate[b]
Conservatism in transformation of DA outcome to PoD	unknown[c]	unknown[c]	low[d]	low[d]	high[e]	high[e]	low[f]
MoE **p(low risk)*SARA ONLY**	low[g]	low[g]	high[h]	high[h]	high[h]	high[h]	p(low risk) = 0.5 low[i]
Risk assessment							
Risk assessment outcome	**UNSAFE**	**UNSAFE**	**SAFE**	**SAFE**	**SAFE**	**SAFE**	**UNSAFE**

[a] Based on use of DPRA. [b] Based on use of DPRA and KeratinoSens™. [c] Based on the PoD being derived based on the use of GHS Cat. 1B, categorizing DEA as a weak/moderate sensitizer with an EC3 value of > 2%, which is converted to a PoD of > 500 $\mu g/cm^2$. The exact PoD value is undetermined. [d] Based upon the DA outcome being a quantitative measure of potency (EC3%), which is converted to dose per unit area. [e] The DA outcome was non-sensitizer. [f] The DA outcome is a quantitative measure of potency (ED_{01}), which is converted to p(low risk) for a given exposure. [g] MoE < 100; [h] MoE > 100; [i] MoE = 217 and the p(low) risk is 0.5, i.e., it is highly uncertain whether the exposure is high or low risk.

For the use of 0.8% DEA in an underarm deodorant product, the exposure was calculated to be 60 µg/cm^2, and the MoE obtained from the 7 DA ranged from 8 to 658. Individual values are shown in Table 3.

3.4.3 WoE assessment and risk assessment outcomes

The overall risk assessment outcomes were evaluated as WoE considering the calculated MoE and in the case of SARA the p(low risk), the confidence in use of NAM input data, and the relative conservatism in transformation of DA outcome to PoD.

Table 2 summarizes the NAM risk assessment outcomes based on the 7 DAs for the use of 0.8% DEA in a shampoo and Table 3 for the use of 0.8% DEA in an underarm deodorant. A detailed summary of the risk assessment outcome for each applied DA is given below.

- *ITSv1:* The overall conclusion was that for risk assessment based on the ITSv1, use of 0.8% DEA in a shampoo was SAFE and use of 0.8% DEA in a deodorant was considered UNSAFE.
- *ITSv2:* The overall conclusion was that for risk assessment based on the ITSv2, use of 0.8% DEA in a shampoo was SAFE and use of 0.8% DEA in a deodorant was considered UNSAFE.
- *ANN (TIMES-SS):* The overall conclusion was that for the risk assessment based on the ANN (TIMES-SS), use of 0.8% DEA in a shampoo and 0.8% DEA in a deodorant were considered SAFE.
- *ANN (ToxTree):* The overall conclusion was that for the risk assessment based on the ANN (ToxTree), use of 0.8% DEA in a shampoo and 0.8% DEA in a deodorant were considered SAFE.

– *STS:* The overall conclusion was that for the risk assessment based on the STS, use of 0.8% DEA in a shampoo and 0.8% DEA in a deodorant were considered SAFE.
– *BN-ITS:* The overall conclusion was that for the risk assessment based on the BN-ITS, use of 0.8% DEA in a shampoo and 0.8% DEA in a deodorant were considered SAFE.
– *SARA model:* The overall conclusion was that for the risk assessment based on the SARA model, use of 0.8% DEA in a shampoo was considered SAFE and use of 0.8% DEA in a deodorant was considered UNSAFE.

4 Discussion

Consumer safety risk assessments for new cosmetic ingredients are no longer based on hazard characterization relying on *in vivo* animal tests but on NAMs. Previously, we published a NGRA framework for skin sensitization to aid the construction of risk assessments based on NAMs (Gilmour et al., 2020), and an increasing number of case studies have been applied aligned to this framework, building our knowledge and confidence in application of these new information sources (Assaf Vandecasteele et al., 2021; ENV/CBC/MONO(2022)32, 2022; Gautier et al., 2020; Gilmour et al., 2020, 2022; Natsch et al., 2018).

The present case study was selected based on observed differences in outcome for the existing NAM information set (Hoffmann et al., 2022), allowing exploration of how these differences impact the DA and the risk assessment outcomes. The exposure scenarios were hypothetical and selected to represent a relatively high and relatively low consumer exposure. These exposures do not represent real use levels of DEA in cosmetics (European Union Council Directive 76/768 EEC) (Fiume et al., 2017). We considered DEA as a "new ingredient" without any existing *in vivo* or human data. Whilst the use of read-across analogues with historical data has previously been shown to reduce uncertainty (Gautier et al., 2020), for the purposes of this case study it was considered out of scope to allow a focus on how to deal with the inconsistent NAM information – a scenario of relevance for risk assessment of novel ingredients. It is intended that read-across will form the topic of a subsequent case study.

Information regarding the chemistry, i.e., the mechanism by which a chemical can interact with protein, is a critical element to understand the applicability domain of a NAM. It also is one element that defines the confidence in using the NAM information within the risk assessment. DEA was predicted to be a potential pro-hapten by two of the four *in silico* tools and an expert chemistry review, i.e., to become a hapten, it would require metabolic activation to convert to a reactive intermediate before it can then react with protein. The potential inability of DPRA and KeratinoSens™ to detect pro-haptens, as the test systems lack metabolic capacity, is well documented. However, when used in combination, the majority of the skin sensitization NAMs are reported to be able to detect most pre- and pro-haptens (Patlewicz et al., 2016; Urbisch et al., 2016). The negative response observed in both DPRA and KeratinoSens™ combined with the positive responses observed in U-SENS™ and h-CLAT (the cell-based assays addressing KE 3) for DEA introduced some uncertainty in the assessment, resulting in a conclusion that it was not possible to confidently define DEA as a non-sensitizer based on the WoE (and exit at Tier 1). Whilst some of the DAs applied did predict DEA to be a non-sensitizer, the reduced confidence in the use of some of the NAM information utilized within the DAs and positive outcomes in test methods not used within the DAs but available within the IATA led within this particular risk assessment case study to a more conservative approach being applied and a PoD being derived using an LLNA EC3 of 100% (or 25,000 μg /cm^2). Furthermore, due to this uncertainty and the *in silico* tools applied, an inconclusive result was obtained from the ITSv2. In this instance, in accordance with OECD TG 497 and as exemplified in case studies published elsewhere, a WoE can be applied to derive an outcome that can be used in hazard and risk assessment (Macmillan et al., 2022; OECD, 2021).

Many of the DAs (all but the SARA) applied within this case study utilize information from *in silico* tools. Uncertainty can also be introduced when different *in silico* tools which provide the same information for use within a DA are used. This is best illustrated here by the different outcomes of the two versions of the ITS DA. This was in part due to different *in silico* tools being applied, i.e., ITSv1 (Derek Nexus) outcome was that DEA was a skin sensitizer (UN GHS Cat. 1B) and the ITSv2 (OECD automated workflow for DASS) for DEA was inconclusive. The OECD automated workflow for DASS is an open access software application, whereas the Derek Nexus software requires a commercial license for use. This does raise the challenge when the safety assessor has access to both *in silico* tools and the predictions are a) in domain but b) differ in outcome and subsequently result in different UN GHS categories, which approach should be applied? It is not only the use of different tools in a DA: The *in silico* tools are regularly updated with new expert knowledge resulting in updated predictions, which may influence the output of DAs that utilize *in silico* predictions. Thus, it is important to document the versions of *in silico* tools used in the risk assessment.

Within the NGRA framework, there is always the possibility to generate additional NAM information, e.g., on potential metabolites (Reynolds, 2021); however, this was not considered in this case study. Another way to account for uncertainty within a risk assessment is by using safety assessment factors (SAFs). For example, the QRA for skin sensitization utilizes SAF to account for uncertainty in the extrapolation from the PoD no expected sensitization induction level (NESIL), which is commonly derived from *in vivo* data, i.e., either HRIPT or LLNA data, to a consumer product use scenario. Uncertainties considered within these SAFs include human variability (increased population size) and the way the product is used compared to the HRIPT exposure (frequency, anatomical site) (Api et al., 2008, 2020; Basketter and Safford, 2016). With the evolution of the skin allergy risk assessment paradigm and the use of NAM, it is yet to be determined whether simply applying the same SAFs as the historical QRA is appropriate, particularly due to the different uncertainties associated with use of historical *in vivo* data versus use of NAM information. In the present case study, we have applied a different approach, i.e., we derived a MoE and then evaluated

possible areas of uncertainty within the risk assessment process, namely the confidence in the use of NAM input data within the DA and the relative conservatism in the transformation of the DA outcome to a PoD.

The SARA model translated the MoE into the risk metric p(low risk) based on the model regressing the MoE for the case study ingredient DEA against the MoE for established high/low risk benchmark exposures. This feature means the benchmarks determine whether the MoE is sufficiently high. For the risk assessments using the information from the other 6 DAs, an arbitrary value of 100 was first assigned to see whether the MoE was sufficiently high. It should, however, be noted that it remains to be determined as to whether this value should be refined for skin sensitization risk assessments based on NAMs. This needs to be re-visited and a systematic approach taken to ensure that all the appropriate uncertainties are addressed. The WoE uncertainty assessment outlined here is a very simplistic framework that was applied to explore how such an approach could be further developed to increase transparency in the decision-making as to whether an exposure should be considered as safe or unsafe. Our work is ongoing to expand upon this type of uncertainty assessment to ensure that our NGRA reaches the desired level of transparency and adequately addresses all associated uncertainties.

Whilst the different outcomes observed in the NAM information led to differences in the DA outputs, there was less impact on the risk assessment outcome. In the case of the rinse-off scenario, the use of 0.8% DEA in a shampoo was considered as safe regardless of the DA used within the risk assessment. For the leave-on scenario of 0.8% DEA in an underarm deodorant product, 4 of 7 applied DA resulted in a conclusion of safe (STS, BN-ITS, ANN-TIMES, and ANN-ToxTree) whilst 3 of 7 (ITSv1, ITSv2, and SARA) resulted in a conclusion of unsafe. The most likely explanation for this observation is that this is largely dependent on the rules that were applied within this case study. For example, in the process of applying DAs that were developed to derive UN GHS categories, a conservative approach such as demonstrated here in the case of the ITSv1 and ITSv2 was applied. A PoD value of > 500 μg/cm^2 was applied as a most conservative estimate of a true threshold since it is not possible to determine where the exact threshold would lie. Furthermore, whilst the SARA model has integrated high/low risk benchmarks, which provide empirical support for whether a MoE is sufficient and provides a p(low risk), the other DAs do not incorporate this functionality. Thus, an arbitrary value of 100 was set as the "acceptable MoE" so that the NGRA process could be illustrated. If, for example, a higher or lower value was applied as the "acceptable MoE", then the risk assessment outcomes for both exposures could be different. As noted above, this is all work in progress. It is envisaged that this will become evident as we evolve a systematic WoE uncertainty assessment approach, and it may ultimately transpire that the "acceptable MoE" value depends on the DA applied, the uncertainty associated with the NAM data used within the DA, and available additional NAM information not used in the DA.

The purpose of this case study was to demonstrate that DAs can be used in an IATA. It is not required to apply all the DAs to conduct a NGRA. In fact, not all DAs are publicly available for use at this moment. Which DA to apply will depend on accessibility of the DA, experience with the use of a certain DA, the information required from a DA (e.g., GHS classification or quantitative potency prediction), and the existing NAM information. A safety assessor will first apply a DA for which the required NAM input data are available before generating new data.

Overall, we demonstrated that our NGRA framework can be successfully applied to more complex cases. Most importantly, the framework is transparent enough to explain exactly what has been done and why. This NGRA framework will continue to evolve and thus be adaptable to different scenarios. Other case studies will follow to further challenge our NGRA framework, build our knowledge in application of NAM, and increase confidence in the risk assessment for skin sensitization of cosmetic ingredients based on NAM.

References

Anderson, S. E., Wells, J., Fedorowicz, A. et al. (2007). Evaluation of the contact and respiratory sensitization potential of volatile organic compounds generated by simulated indoor air chemistry. *Toxicol Sci 97*, 355-363. doi:10.1093/toxsci/kfm043

Api, A. M., Basketter, D. A., Cadby, P. A. et al. (2008). Dermal sensitization quantitative risk assessment (QRA) for fragrance ingredients. *Regul Toxicol Pharmacol 52*, 3-23. doi:10.1016/j.yrtph.2007.10.008

Api, A. M., Basketter, D., Bridges, J. et al. (2020). Updating exposure assessment for skin sensitization quantitative risk assessment for fragrance materials. *Regul Toxicol Pharmacol 118*, 104805. doi:10.1016/j.yrtph.2020.104805

Assaf Vandecasteele, H., Gautier, F., Tourneix, F. et al. (2021). Next generation risk assessment for skin sensitisation: A case study with propyl paraben. *Regul Toxicol Pharmacol 123*, 104936. doi:10.1016/j.yrtph.2021.104936

Basketter, D. A., Alépée, N., Ashikaga, T. et al. (2014). Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis 25*, 11-21. doi:10.1097/der.0000000000000003

Basketter, D. and Safford, B. (2016). Skin sensitization quantitative risk assessment: A review of underlying assumptions. *Regul Toxicol Pharmacol 74*, 105-116. doi:10.1016/j.yrtph.2015.11.013

Brain, K. R., Walters, K. A., Green, D. M. et al. (2005). Percutaneous penetration of diethanolamine through human skin in vitro: Application from cosmetic vehicles. *Food Chem Toxicol 43*, 681-690. doi:10.1016/j.fct.2004.12.021

Chilton, M. L., Macmillan, D. S., Steger-Hartmann, T. et al. (2018). Making reliable negative predictions of human skin sensitisation using an in silico fragmentation approach. *Regul Toxicol Pharmacol 95*, 227-235. doi:10.1016/j.yrtph.2018.03.015

Cottrez, F., Boitel, E., Ourlin, J. C. et al. (2016). SENS-IS, a 3D reconstituted epidermis based model for quantifying chemical sensitization potency: Reproducibility and predictivity results from an inter-laboratory study. *Toxicol In Vitro 32*, 248-260. doi:10.1016/j.tiv.2016.01.007

Daniel, A. B., Strickland, J., Allen, D. et al. (2018). International

regulatory requirements for skin sensitization testing. *Regul Toxicol Pharmacol 95*, 52-65. doi:10.1016/j.yrtph.2018.03.003

Del Bufalo, A., Pauloin, T., Alépée, N. et al. (2018). Alternative integrated testing for skin sensitization: Assuring consumer safety. *Appl In Vitro Toxicol 4*, 30-43. doi:10.1089/aivt.2017.0023

Desprez, B., Dent, M., Keller, D. et al. (2018). A strategy for systemic toxicity assessment based on non-animal approaches: The cosmetics Europe long range science strategy programme. *Toxicol In Vitro 50*, 137-146. doi:10.1016/j.tiv.2018.02.017

Enoch, S. J., Madden, J. C. and Cronin, M. T. (2008). Identification of mechanisms of toxic action for skin sensitisation using a smarts pattern based approach. *SAR QSAR Environ Res 19*, 555-578. doi:10.1080/10629360802348985

Ezendam, J., Braakhuis, H. M. and Vandebriel, 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. doi:10.1007/s00204-016-1842-4

Fiume, M. M., Heldreth, B., Bergfeld, W. F. et al. (2017). Safety assessment of diethanolamine and its salts as used in cosmetics. *Int J Toxicol 36*, 89s-110s. doi:10.1177/1091581817707179

Gautier, F., Tourneix, F., Assaf Vandecasteele, H. et al. (2020). Read-across can increase confidence in the next generation risk assessment for skin sensitisation: A case study with resorcinol. *Regul Toxicol Pharmacol 117*, 104755. doi:10.1016/j.yrtph.2020.104755

Gerberick, G. F., Vassallo, J. D., Bailey, R. E. et al. (2004). Development of a peptide reactivity assay for screening contact allergens. *Toxicol Sci 81*, 332-343. doi:10.1093/toxsci/kfh213

GHS, U. (2021). Globally Harmonised System of Classification and Labelling of Chemicals (GHS). 9th revision.

Gilmour, N., Kern, P. S., Alépée, N. et al. (2020). Development of a next generation risk assessment framework for the evaluation of skin sensitisation of cosmetic ingredients. *Regul Toxicol Pharmacol 116*, 104721. doi:10.1016/j.yrtph.2020.104721

Gilmour, N., Reynolds, J., Przybylak, K. et al. (2022). Next generation risk assessment for skin allergy: Decision making using new approach methodologies. *Regul Toxicol Pharmacol 131*, 105159. doi:10.1016/j.yrtph.2022.105159

Gomes, C., Nocairi, H., Thomas, M. et al. (2012). Stacking prediction for a binary outcome. *Compstat 2012*, 271-282.

Gomes, C., Nocairi, H., Thomas, M. et al. (2014). A simple and robust scoring technique for binary classification. *J Artif Intell Res 3*, 52-58.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med 130*, 1005-1013. doi:10.7326/0003-4819-130-12-199906150-00019

Griem, P., Goebel, C. and Scheffler, H. (2003). Proposal for a risk assessment methodology for skin sensitization based on sensitization potency data. *Regul Toxicol Pharmacol 38*, 269-290. doi:10.1016/j.yrtph.2003.07.001

Hirota, M., Ashikaga, T. and Kouzuki, H. (2018). Development of an artificial neural network model for risk assessment of skin sensitization using human cell line activation test, direct peptide reactivity assay, KeratinoSens™ and in silico structure alert parameter. *J Appl Toxicol 38*, 514-526. doi:10.1002/jat.3558

Hoffmann, S., Kleinstreuer, N., Alépée, N. et al. (2018). Non-animal methods to predict skin sensitization (I): The cosmetics Europe database. *Crit Rev Toxicol 48*, 344-358. doi:10.1080/10408444.2018.1429385

Hoffmann, S., Alépée, N., Gilmour, N. et al. (2022). Expansion of the cosmetics Europe skin sensitisation database with new substances and PPRA data. *Regul Toxicol Pharmacol 131*, 105169. doi:10.1016/j.yrtph.2022.105169

Jaworska, J. S., Natsch, A., Ryan, C. et al. (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. doi:10.1007/s00204-015-1634-2

Kleinstreuer, N. C., Hoffmann, S., Alépée, N. et al. (2018). Non-animal methods to predict skin sensitization (II): An assessment of defined approaches. *Crit Rev Toxicol 48*, 359-374. doi:10.1080/10408444.2018.1429386

Kraeling, M. E., Yourick, J. J. and Bronaugh, R. L. (2004). In vitro human skin penetration of diethanolamine. *Food Chem Toxicol 42*, 1553-1561. doi:10.1016/j.fct.2004.04.016

Lessmann, H., Uter, W., Schnuch, A. et al. (2009). Skin sensitizing properties of the ethanolamines mono-, di-, and triethanolamine. Data analysis of a multicentre surveillance network (IVDK) and review of the literature. *Contact Dermatitis 60*, 243-255. doi:10.1111/j.1600-0536.2009.01506.x

Macmillan, D. S., Chilton, M. L., Gao, Y. et al. (2022). How to resolve inconclusive predictions from defined approaches for skin sensitisation in OECD Guideline No. 497. *Regul Toxicol Pharmacol 135*, 105248. doi:10.1016/j.yrtph.2022.105248

Natsch, A. and Emter, R. (2016). Nrf2 activation as a key event triggered by skin sensitisers: The development of the stable KeratinoSens reporter gene assay. *Altern Lab Anim 44*, 443-451. doi:10.1177/026119291604400513

Natsch, A., Emter, R., Haupt, T. et al. (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. doi:10.1093/toxsci/kfy135

Nocairi, H., Gomes, C., Thomas, M. et al. (2016). Improving stacking methodology for combining classifiers: Applications to cosmetic industry. *Electronic Journal of Applied Statistical Analysis 9*, 340-361. doi:10.1285/i20705948v9n2p340

OECD (2013). Test No. 122: Determination of pH, Acidity and Alkalinity. *OECD Guidelines for the Testing of Chemicals, Section 1*. OECD Publishing, Paris. doi:10.1787/9789264203686-en

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. doi:10.1787/9789264221444-en

OECD (2018). Test No. 442D: In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method. *OECD Guidelines for the Testing of Chemicals, Section 1*. OECD Publishing, Paris. doi:10.1787/9789264229822-en

OECD (2020). Test No. 442C: In Chemico Skin Sensitisation: Assays addressing the Adverse Outcome Pathway key event on covalent binding to proteins. *OECD Guidelines for the Testing of Chemicals, Section 4*. OECD Publishing, Paris. doi:10.1787/9789264229709-en

OECD (2021). Guideline No. 497: Defined Approaches On Skin Sensitisation. *OECD Guidelines for the Testing of Chemicals, Section 4*. OECD Publishing, Paris. doi:10.1787/b92879a4-en

OECD (2022). 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. doi:10.1787/9789264264359-en

Patlewicz, G., Dimitrov, S. D., Low, L. K. et al. (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. doi:10.1016/j.yrtph.2007.03.003

Patlewicz, G., Kuseva, C., Mehmed, A. et al. (2014). TIMES-SS – Recent refinements resulting from an industrial skin sensitisation consortium. *SAR QSAR Environ Res 25*, 367-391. doi:10.1080/1062936x.2014.900520

Patlewicz, G., Casati, S., Basketter, D. A. et al. (2016). Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? *Regul Toxicol Pharmacol 82*, 147-155. doi:10.1016/j.yrtph.2016.08.007

Piroird, C., Ovigne, J. M., Rousset, F. et al. (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. doi:10.1016/j.tiv.2015.03.009

Reisinger, K., Hoffmann, S., Alépée, N. et al. (2015). Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. *Toxicol In Vitro 29*, 259-270. doi:10.1016/j.tiv.2014.10.018

Reynolds, J., Gilmour, N., Baltazar, M. T. et al. (2022). Decision making in next generation risk assessment for skin allergy: Using historical clinical experience to benchmark risk. *Regul Toxicol Pharmacol 134*, 105219. doi:10.1016/j.yrtph.2022.105219

Reynolds, G. R. J., Gilmour, N., Cubberley, R. et al. (submitted). A skin sensitisation next generation risk assessment case study for coumarin in hypothetical cosmetic products.

Robinson, M. K., Gerberick, G. F., Ryan, C. A. et al. (2000). The importance of exposure estimation in the assessment of skin sensitization risk. *Contact Dermatitis 42*, 251-259. doi:10.1034/j.1600-0536.2000.042005251.x

Sakaguchi, H., Ryan, C., Ovigne, J. M. et al. (2010). Predicting skin sensitization potential and inter-laboratory reproducibility of a human cell line activation test (h-CLAT) in the European cosmetics association (COLIPA) ring trials. *Toxicol In Vitro 24*, 1810-1820. doi:10.1016/j.tiv.2010.05.012

Spicer, C. W., Gordon, S. M., Holdren, M. W. et al. (2002). *Hazardous Air Pollutant Handbook: Measurements, Properties, and Fate in Ambient Air*. Boca Raton, FL, USA: CRC Press. doi:10.1201/9781420032352

Tollefsen, K. E., Scholz, S., Cronin, M. T. et al. (2014). Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA). *Regul Toxicol Pharmacol 70*, 629-640. doi:10.1016/j.yrtph.2014.09.009

Urbisch, D., Becker, M., Honarvar, N. et al. (2016). Assessment of pre- and pro-haptens using nonanimal test methods for skin sensitization. *Chem Res Toxicol 29*, 901-913. doi:10.1021/acs.chemrestox.6b00055

van Vliet, E., Kühnl, J., Goebel, C. et al. (2018). State-of-the-art and new options to assess T cell activation by skin sensitizers: Cosmetics Europe workshop. *ALTEX 35*, 179-192. doi:10.14573/altex.1709011

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks 5*, 241-259.

Zang, Q., Mansouri, K., Williams, A. J. et al. (2017). In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *J Chem Inf Model 57*, 36-49. doi:10.1021/acs.jcim.6b00625

Conflict of interest

This work was conducted under the project management of the Cosmetics Europe Long Range Science Strategy. The following authors are employed by consumer goods companies; Nicola Gilmour (Unilever), Nathalie Alépée and Dagmar Bury (L'Oréal), Petra S. Kern (Proctor& Gamble), Masaaki Miyazawa (Kao corporation) and Hayato Nishida (Shiseido). The authors Sebastian Hoffmann and Erwin van Vliet are consultants funded by Cosmetics Europe to support this project.

Data availability

The source data used to inform these case study risk assessments is publicly available and can be found in Hoffmann et al. (2022). All novel findings are reported in this manuscript.

Acknowledgements

The authors would like to thank the following colleagues for supporting this work: Nora Aptula for expert chemistry advice; Els DeConinck, Françoise Gautier and Kanako Nakayama for input on risk assessment using the BN, STS and ITS respectively, and Fanny Boisleve for manuscript review.