

Non-animal methods to predict skin sensitization (II): an assessment of defined approaches

Nicole C. Kleinstreuer, Sebastian Hoffmann, Nathalie Alépée, David Allen, Takao Ashikaga, Warren Casey, Elodie Clouet, Magalie Cluzel, Bertrand Desprez, Nichola Gellatly, Carsten Göbel, Petra S. Kern, Martina Klaric, Jochen Kühnl, Silvia Martinozzi-Teissier, Karsten Mewes, Masaaki Miyazawa, Judy Strickland, Erwin van Vliet, Qingda Zang & Dirk Petersohn

To cite this article: Nicole C. Kleinstreuer, Sebastian Hoffmann, Nathalie Alépée, David Allen, Takao Ashikaga, Warren Casey, Elodie Clouet, Magalie Cluzel, Bertrand Desprez, Nichola Gellatly, Carsten Göbel, Petra S. Kern, Martina Klaric, Jochen Kühnl, Silvia Martinozzi-Teissier, Karsten Mewes, Masaaki Miyazawa, Judy Strickland, Erwin van Vliet, Qingda Zang & Dirk Petersohn (2018) Non-animal methods to predict skin sensitization (II): an assessment of defined approaches, Critical Reviews in Toxicology, 48:5, 359-374, DOI: [10.1080/10408444.2018.1429386](https://doi.org/10.1080/10408444.2018.1429386)

To link to this article: <https://doi.org/10.1080/10408444.2018.1429386>



This work was authored as part of the Contributor's official duties as an Employee of the United States Government and is therefore a work of the United States Government. In accordance with 17 U.S.C. 105, no copyright protection is available for such works under U.S. Law.



View supplementary material [↗](#)



Published online: 23 Feb 2018.



Submit your article to this journal [↗](#)



Article views: 3262



View Crossmark data [↗](#)



Citing articles: 3 View citing articles [↗](#)

REVIEW ARTICLE



Non-animal methods to predict skin sensitization (II): an assessment of defined approaches**

Nicole C. Kleinstreuer^a, Sebastian Hoffmann^b, Nathalie Alépée^c, David Allen^d, Takao Ashikaga^{e*}, Warren Casey^a, Elodie Clouet^f, Magalie Cluzel^g, Bertrand Desprez^h, Nichola Gellatlyⁱ, Carsten Göbel^j, Petra S. Kern^k, Martina Klaric^h, Jochen Kühnl^l, Silvia Martinozzi-Teissier^c, Karsten Mewes^m, Masaaki Miyazawaⁿ, Judy Strickland^d, Erwin van Vliet^o, Qingda Zang^d and Dirk Petersohn^m

^aNIH/NIEHS/DNTP/NICEATM, Research Triangle Park, NC, USA; ^bSEH Consulting + Services, Paderborn, Germany; ^cL'Oréal Research & Innovation, Aulnay-sous-Bois, France; ^dILS, Research Triangle Park, NC, USA; ^eShiseido, Yokohama-shi, Kanagawa, Japan; ^fPierre Fabre, Toulouse, France; ^gLVMH, St Jean de Braye, France; ^hCosmetics Europe, Brussels, Belgium; ⁱUnilever, London, UK; ^jCoty, Darmstadt, Germany; ^kProcter & Gamble Services Company NV, Strombeek-Bever, Belgium; ^lBeiersdorf AG, Hamburg, Germany; ^mHenkel AG & Co. KGaA, Düsseldorf, Germany; ⁿKao Corporation, Haga, Tochigi, Japan; ^oServices & Consultations on Alternative Methods (SeCAM), Magliaso, Switzerland

ABSTRACT

Skin sensitization is a toxicity endpoint of widespread concern, for which the mechanistic understanding and concurrent necessity for non-animal testing approaches have evolved to a critical juncture, with many available options for predicting sensitization without using animals. Cosmetics Europe and the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods collaborated to analyze the performance of multiple non-animal data integration approaches for the skin sensitization safety assessment of cosmetics ingredients. The Cosmetics Europe Skin Tolerance Task Force (STTF) collected and generated data on 128 substances in multiple *in vitro* and *in chemico* skin sensitization assays selected based on a systematic assessment by the STTF. These assays, together with certain *in silico* predictions, are key components of various non-animal testing strategies that have been submitted to the Organization for Economic Cooperation and Development as case studies for skin sensitization. Curated murine local lymph node assay (LLNA) and human skin sensitization data were used to evaluate the performance of six defined approaches, comprising eight non-animal testing strategies, for both hazard and potency characterization. Defined approaches examined included consensus methods, artificial neural networks, support vector machine models, Bayesian networks, and decision trees, most of which were reproduced using open source software tools. Multiple non-animal testing strategies incorporating *in vitro*, *in chemico*, and *in silico* inputs demonstrated equivalent or superior performance to the LLNA when compared to both animal and human data for skin sensitization.

ARTICLE HISTORY

Received 6 July 2017
Revised 11 December 2017
Accepted 3 January 2018

KEYWORDS

Skin sensitization; defined approaches; hazard assessment; potency assessment; integrated testing strategy; human reference data; LLNA reference data; new approach methodologies; adverse outcome pathway

Table of contents

Introduction	360	JRC CCT	364
Methods	360	Givaudan ITS	364
Selection of defined approaches	360	Shiseido ANN-EC3	364
Qualitative evaluation criteria	361	Procter & Gamble BN-ITS 3	365
Database	361	Kao STS	365
Predictive performance assessment	361	Kao ITS	365
Results	362	Unilever SARA	365
Qualitative evaluation	362	Summary of qualitative evaluation	365
BASF "2 out of 3"	363	LLNA and human reference data	366
RIVM STS	363	Quantitative performance assessments	366
DuPont IATA-SS	363	BASF "2 out of 3"	367
L'Oréal Stacking Meta-model	364	Kao STS	367
ICCVAM SVM	364	Kao ITS	367
		ICCVAM SVM	368

CONTACT Nicole C. Kleinstreuer ✉ nicole.kleinstreuer@nih.gov NIH/NIEHS/DNTP/NICEATM, P.O. Box 12233, Mail Stop K2-16, Research Triangle Park, NC 27709, USA

*Present Address: Japanese Center for the Validation of Alternative Methods (JaCVAM), National Institute of Health Sciences (NIHS) 1-18-1 Kamiyoga, Setagaya, Tokyo, Japan.

**This manuscript is part of a set of papers; the companion manuscript can be found here: <https://doi.org/10.1080/10408444.2018.1429385>.

This work was authored as part of the Contributor's official duties as an Employee of the United States Government and is therefore a work of the United States Government. In accordance with 17 U.S.C. 105, no copyright protection is available for such works under U.S. Law.

This is an Open Access article that has been identified as being free of known restrictions under copyright law, including all related and neighboring rights (<https://creativecommons.org/publicdomain/mark/1.0/>). You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

<i>Shiseido ANN</i>	368
<i>P&G BN ITS-3</i>	368
<i>Common test set</i>	369
<i>Concordance across substances</i>	369
Discussion	369
Acknowledgements	372
Declaration of interest	372
Supplemental material	372
References	372

Introduction

Skin sensitization is a common toxicity endpoint of concern and accounts for 10–15% of known occupational illnesses in the U.S. and Europe (Anderson et al. 2011). It is estimated that 15–20% of the general population will become sensitized to one or more chemicals in commerce at some point in their lifetime (Bruckner et al. 2000; Thyssen et al. 2007). Skin sensitizers can cause a local skin reaction characterized by redness, swelling and itching, also known as allergic contact dermatitis (Murphy et al. 2012). There are two stages to allergic contact dermatitis: induction and elicitation. Induction (sensitization) occurs when allergen-specific T cells are generated in response to a substance exposure, and elicitation occurs when a previously sensitized individual is re-exposed to the substance, resulting in a pruritic rash. The latency period between exposure to a substance and appearance of the rash shortens with subsequent exposures. Due to its prevalence, persistence, and impact on quality of life, skin sensitization is recognized as an important occupational and environmental health issue (Kimber et al. 2011). There are a variety of national and regional regulatory requirements for chemical testing to identify skin sensitizers.

European, U.S., and other regulatory agencies from around the world have different requirements for skin sensitization data depending on the substance use category and the agency, but almost all have traditionally relied upon regulatory tests using animal models (Luebke 2012; Birnbaum 2013; Daniel et al. 2018). However, in March 2013, the European Union's 7th Amendment of the Cosmetic Directive placed a complete ban on animal testing for all cosmetics ingredients, necessitating the rapid development and use of non-animal testing methods (European Union 2003). Furthermore, the European Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) regulation calls for the use of non-animal methods. While REACH in general mandates that animal tests should be used as a last resort, a 2017 update of the regulation requires the uses of *in vitro* and *in silico* methods as the first choice for skin sensitization and permits animal testing under exceptional circumstances, at times including potency assessment (EU 2016; Sauer et al. 2016). Despite this, there are currently no stand-alone non-animal methods for identifying skin sensitizers (Natsch 2014).

Mechanisms of skin sensitization have been investigated intensively for many years and are documented by the Organization for Economic Co-operation and Development

(OECD) in its publication, "The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins" (OECD 2012). The adverse outcome pathway (AOP) includes four key mechanistic events: (1) binding of haptens to endogenous proteins in the skin, (2) keratinocyte activation, (3) dendritic cell activation, and (4) proliferation of antigen-specific T cells. The most commonly used animal test, the murine local lymph node assay (LLNA), is based on the understanding of this complex series of events underlying the immune response after exposure to a chemical sensitizer, and covers all key events (Gerberick et al. 2000; Williams et al. 2015). However, the construction of the AOP for skin sensitization has enabled the development of a multitude of non-animal test methods that are associated with one or more of the AOP key events (Mehling et al. 2012; Reisinger et al. 2015). While certain methods show great promise for the prediction of skin sensitization potential (Gerberick et al. 2004; Natsch and Emter 2008; Nukada et al. 2012), each method, when used individually or in combination, has its own set of limitations. Further, the complexity of the underlying biology indicates that no single measurement is yet sufficient to predict sensitizer potency (Rovida et al. 2015; Urbisch et al. 2015). Therefore, it is generally assumed that only a combination of several methods in an integrated testing strategy will obviate the need for animal testing, although methods that can serve as stand-alone replacements may be developed in the future (MacKay et al. 2013; Steiling 2016). Consequently, many testing strategies combining non-animal methods have been submitted to the OECD as "defined approaches" that form part of integrated approaches to testing and assessment (IATA) of skin sensitization (OECD 2016a, 2016b).

This paper presents an evaluation of defined approaches (DAs) representing non-animal skin sensitization testing strategies that were submitted to the OECD (OECD 2016a, 2016b). Here, the Cosmetics Europe skin tolerance task force (CE STTF) and the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) partnered to reproduce a subset of these DAs in a transparent manner, and evaluated them using a large, partly novel data set of 128 substances chosen based on the availability of both animal and human skin sensitization information (Hoffmann et al. 2018). This evaluation provides insight into the complexity and applicability of the DAs and harmonizes their assessment by using the same high-quality dataset.

Methods

Selection of defined approaches

Our evaluation considered 12 case studies submitted as examples for the OECD reporting guidance (OECD 2016a, 2016b), covering DAs with fixed data interpretation procedures as well as IATA incorporating expert judgment. Five of the DAs are consensus and decision-tree models that either predict skin sensitization hazard (sensitizer versus non-sensitizer) or assign the test substance to one of three skin sensitization potency categories. Within this group are "BASF '2

out of 3' – Sens ITS", "KAO STS", "KAO ITS", "RIVM STS", and the "JRC CCT". Another group of approaches apply more complex statistical techniques ("P&G BN-ITS-3", "Shiseido ANN-EC3", "ICCVAM-SVM", the "L'Oreal stacking model", and "Givaudan ITS") that provide categorical or continuous predictions of either hazard or potency. The remaining two approaches are the "DuPont IATA-SS", which is a decision framework integrating all possibly relevant information for hazard identification, and the "Unilever SARA", which provides a framework for risk assessment.

Qualitative evaluation criteria

To allow for a harmonized and systematic evaluation, qualitative evaluation criteria were predefined. We evaluated the most important technical and practical aspects of the 12 approaches in a manner that facilitated comparison and selection for the second stage of the evaluation, which would assess the approaches' predictive performance. Six criteria categories were used (Table 2). The category "characteristics" included basic information, such as the purpose of the approach and pivotal information sources used for the evaluation. The category "input data" described the information sources, e.g. *in vitro*, *in chemico*, *in silico*, and expert systems, used by the approach. The category "prediction algorithm" described the procedure used by the approach to process input data to predict skin sensitization hazard potential or potency. The "mechanistic relevance" of the approach was characterized with respect to the OECD AOP and the relevant key event(s). Information on the "applicability domain" included chemical space coverage and assay limitations. Finally, "practical aspects" addressed characteristics such as relative cost factors and availability through contract research organizations (CROs).

Using these criteria, and placing particular emphasis on the availability of input data and the feasibility of reproducing the algorithm using open-source software, we chose six of the 12 DAs for evaluation of their predictivity. This evaluation used the database described in the companion manuscript (Hoffmann et al. 2018).

Database

To evaluate the predictivity of these six DAs, the CE STTF collected or generated data on 128 substances with high-quality human and animal data using previously evaluated non-animal test methods (Reisinger et al. 2015). LLNA data in the NICEATM Integrated Chemical Environment resource (Bell et al. 2017) and human potency classifications according to (Basketter et al. 2014) were curated for all substances as *in vivo* reference standards for comparison, described in detail in (Hoffmann et al. 2018). The *in vitro* assays and *in silico* prediction tools used here are described briefly here and in more detail in Hoffmann et al. (2018). The direct peptide reactivity assay (DPRA) is an *in chemico* test that measures the ability of a substance to form a hapten–protein complex (Gerberick et al. 2004, 2007; OECD 2015a), and corresponds to the molecular initiating event in the skin sensitization AOP as

described by OECD (2012). An *in vitro* cell-based assay available commercially under the trade name KeratinoSensTM (Givaudan) [an equivalent method called LuSens is also available (Ramirez et al. 2014)] assesses the activation of the Nrf2 pathway in keratinocytes, indicating a substance's ability to induce cytoprotective responses and release of cytokines by keratinocytes, the second key event in the AOP (Emter et al. 2010; OECD 2015b). *In vitro* cell activation assays (e.g. the Human Cell Line Activation Test, h-CLAT; and U937 cell line activation Test, U-SENSTM) measure induction of the third key event in the AOP, the ability of a substance to activate and mobilize dendritic cells in the skin (Ashikaga et al. 2006; Alepee et al. 2015; Piroird et al. 2015; OECD 2017). Software tools used to generate structure-based predictions of skin sensitization hazard included the proprietary software tools DEREK (Macmillan et al. 2016) and TIMES (Patlewicz et al. 2007), and the open-source read-across tool OECD Toolbox (OECD 2014a). Finally, experimental data on six physicochemical properties relevant to skin exposure and penetration (octanol:water partition coefficient (log *P*), water solubility, vapor pressure, melting point, boiling point, and molecular weight) were collected where available. If experimental values were not available, these properties were predicted using validated prediction models (Jaworska et al. 2011, 2013; Patlewicz et al. 2014; Strickland et al. 2016, 2017; Zang et al. 2017a).

Predictive performance assessment

There were twelve strategies submitted to the OECD as case studies for predicting skin sensitization potential that relied on DAs using different combinations of these *in vitro/in chemico/in silico* data. Six of the DAs, two of which yielded two models each, were selected based on the above criteria, reproduced and assessed for predictive performance (indicated in Table 1, far right column). The predictive performance of these six DAs, covering eight models, was assessed across all 128 substances against LLNA and human endpoints. The DAs were first evaluated for their ability to discriminate between sensitizers and non-sensitizers, and then for their ability to characterize skin sensitization potency if that ability was included in the stated purpose of the approach (as indicated in Table 1). Due to differing degrees of overlap between the training set for each DA and the 128-substance set, the DAs were also individually evaluated on the substances that were not part of their respective training sets, as well as on a common test set.

The potency cutoffs for the LLNA data were based on recommendations developed by the European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) (Loveless et al. 2010). In this classification scheme, sensitizers with summary LLNA EC3 less than 0.1 were classified as extreme, sensitizers with EC3 between 0.1 and 1 classified as strong, sensitizers with EC3 between 1 and 10 classified as moderate, and sensitizers with EC3 greater than 10 classified as weak. The five potency classes were also combined into three classes (strong = extreme + strong; weak = moderate + weak; and non-sensitizing) to facilitate evaluation of the DAs providing corresponding three-class predictions. The human potency classifications were assigned according to

Table 1. Twelve defined and/or integrated approaches to testing and assessment for assessing skin sensitization potential.

OECD case study title (submitter)	Purpose	Data inputs	Reference	Quantitative evaluation?
An AOP-based "2 out of 3" Integrated Testing Strategy Approach to Skin Hazard Identification (BASF)	Hazard ID	DPRA, h-CLAT, KeratinoSens TM , U-SENS TM	Urbisch et al. (2015)	Yes
Sequential Testing Strategy (STS) for Hazard Identification of Skin Sensitisers (RIVM)	Hazard ID	DPRA, h-CLAT, KeratinoSens TM , HaCaT gene signature, MultiCASE, CAESAR, DEREK, OECD QSAR toolbox	van der Veen et al. (2014)	No
A non-testing pipeline approach for skin sensitisation (DuPont/G. Patlewicz)	Hazard ID	Existing data, protein binding profile, physicochemical properties, TIMES-SS, expert judgment	Patlewicz et al. (2014)	No
Stacking Meta-model for Skin Sensitisation Hazard Identification (L'Oréal)	Hazard ID	DPRA, KeratinoSens TM , U-SENS TM , TIMES-SS, ToxTree, volatility, pH	Del Bufalo et al. (2018)	No
Integrated decision strategy for skin sensitisation hazard (ICCVAM)	Hazard ID	DPRA, h-CLAT, KeratinoSens TM , OECD QSAR Toolbox, physicochemical properties	Strickland et al. (2016)	Yes
Consensus of Classification Trees for Skin Sensitisation Hazard Prediction (EC-JRC)	Hazard ID	TIMES-SS, DRAGON descriptors	Asturiol et al. (2016)	No
Sensitizer Potency Prediction Based on Key Event 1 + 2: Combination of Kinetic Peptide Reactivity Data and KeratinoSens [®] Data (Givaudan)	Potency (continuous)	Cor1C420 (kinetic peptide reactivity), KeratinoSens TM , TIMES-SS	Natsch et al. (2015)	No
The Artificial Neural Network Model for Predicting LLNA EC3 (Shiseido)	Potency class/EC3	DPRA, h-CLAT, ARE (or KeratinoSens TM)	Hirota et al. (2015)	Yes
Bayesian Network DIP (BN-ITS-3) for Hazard and Potency Identification of Skin Sensitizers (P&G)	Potency class	DPRA, h-CLAT, KeratinoSens TM , TIMES-SS, bioavailability (solubility at pH 7, log <i>D</i> at pH 7, plasma protein binding, fraction ionized)	Jaworska et al. (2015)	Yes
Sequential Testing Strategy (STS) for Sensitising Potency Classification Based on <i>In Chemico</i> and <i>In Vitro</i> Data (Kao)	Potency class	DPRA, h-CLAT	Takenouchi et al. (2015)	Yes
ITS for Sensitising Potency Classification Based on <i>In Silico</i> , <i>In Chemico</i> , and <i>In Vitro</i> Data (Kao)	Potency class	DPRA, h-CLAT, DEREK	Takenouchi et al. (2015)	Yes
Data Interpretation Procedure for Skin Allergy Risk Assessment (SARA) (Unilever)	Sensitization probability	Bioavailability, skin protein kinetics, ordinary differential equation model	MacKay et al. (2013)	No

(Basketter et al. 2014; Api et al. 2017), where category 1 was categorized as extreme, 2 was strong, 3 was moderate, 4 was weak, and 5 and 6 were grouped together as negative. Similar to the approach taken for the LLNA, the five human potency classes were combined into three classes (strong = extreme + strong; weak = moderate + weak; and non-sensitizing) to facilitate evaluation of the DAs providing corresponding three-class predictions. For hazard identification, these human categories were dichotomized by considering categories 1–4 as sensitizer and categories 5 and 6 as non-sensitizers. The hazard analyses were also repeated with categories 1–5 designated as sensitizers (4 and 5 grouped together as weak) and category 6 designated as non-sensitizer.

Five of the six DAs chosen for predictive performance assessment (Table 1) were reproduced in an open-source manner by writing R code to apply each prediction method as described in the respective publications (and summarized in the results section). The following R packages were used for analysis and visualization: *e1071*, *RSNNS*, *gdata*, *caret*, and *ggplot2*.

The artificial neural network (ANN) models proposed by Shiseido were originally built using commercial software QwikNet v2.23 (Hirota et al. 2015). The open-source R code written to replace the commercial software QwikNet was able to reproduce the published results from both ANN models, with an R^2 of 0.62 and root-mean-squared error of 0.64 [matching Figure 2A from Hirota et al. (2015)] and an R^2 of

0.72 and root-mean-squared error of 0.64 [matching Figure 5A from Hirota et al. (2015)] between predicted and observed EC3 values. The inputs and output were log transformed and the ANN models had two hidden layers with five and two nodes respectively; logistic activation functions were used for the hidden and output layers; 10 000 iterations were used for training, and learning rate; scaling functions, and momentum parameters were inferred from (Hirota et al. 2015) and communication with model developers.

Due to its dependence on commercial BayesiaLab software and a post-processing probability distribution correction, the Bayesian ITS-3 network was implemented in BayesiaLab as in (Jaworska et al. 2015) and was applied here after generating the necessary input features. Predictions for substances that had been published previously in Jaworska et al. (2015) were rerun for the current evaluation, as some input parameters had changed slightly due to software updates or variability in assay results. The code reproducing each DA (with the exception of the Bayesian ITS-3), the input data files, the resulting output predictions, and code to run detailed performance statistics are available in the Supplemental Materials.

Results

Qualitative evaluation

The results of the qualitative evaluation of the DAs/IATA are summarized below, particularly with respect to complexity,

Table 2. Qualitative evaluation categories and criteria.

Evaluation category	Evaluation criteria
Characteristics	Principle Prediction (i.e. hazard versus potency [categories or continuous]) Publication Information sources
Input data	Test method (<i>in vitro</i> and <i>in chemico</i>) - read-out used - validation status - reproducibility - issues (e.g. IP, availability) <i>In silico</i> /expert system data/physicochemical properties - read-out used - availability - reliability - issues (e.g. IP, availability) Expert knowledge - input used - availability Principle Prediction (i.e. hazard vs. potency [categories or continuous]) Publication Information sources
Prediction algorithm	Type Availability Transparency Requirements for implementation (specific software) Self-learning Complexity Sequential information generation All inputs required? Predictivity: Sample size (total and for categories) Predictivity: Parameters (sensitivity, specificity, concordance)
Mechanistic relevance	OECD AOP key events covered Sequence of OECD AOP events considered Justification/discussion of the mechanistic relevance
Applicability domain	Chemical spectrum tested Limitations (solubility, surfactants, ...) Potential limitations for cosmetic ingredients (e.g. natural extracts cannot be processed by <i>in silico</i> approaches)
Practical aspects	Costs Can be conducted by CRO? Time required (per substance)

AOP coverage, limitations, and selection for quantitative evaluation. Detailed information for all criteria listed in Table 2 is provided in [Supplementary Table S1](#).

BASF “2 out of 3”

The DA BASF “2 out of 3” was first described in Bauch et al. (2012) and subsequently applied to a larger dataset by Urbisch et al. (2015). The approach predicts skin sensitization hazard by sequential testing in up to three internationally accepted non-animal methods with OECD TGs (DPRA, KeratinoSensTM or LuSens, and h-CLAT or U-SENSTM) in an undefined order. Applying the standard prediction models of the individual methods, a substance is considered a skin sensitizer if at least two methods predict the substance to be a sensitizer, and vice versa. In this way, at least two AOP key events are addressed via a simple consensus approach/decision tree. So far, mainly industrial chemicals and cosmetic

ingredients have been tested with this approach. The limitations of this approach are specific to the individual test methods, such as reduced applicability to poorly water-soluble substances. Costs are correlated to the number of *in vitro* assays required. All of the test methods used in the approach are currently available through CROs and can be performed without licensing fees. This approach was selected for quantitative performance assessment based on the availability of the input features.

RIVM STS

The sequential testing strategy proposed by the RIVM (van der Veen et al. 2014) predicts human skin sensitization hazard using a tiered combination of *in silico* and *in vitro* methods. First, the test substance is evaluated using a battery of four QSAR models (MultiCASE, CAESAR, DEREK, and OECD QSAR Toolbox), which are combined to derive a sensitizer probability prediction (addressing KE 4, as the individual QSAR methods predict LLNA hazard). Depending on the probability from the consensus QSAR model, either the DPRA is performed (completing Tier 1) or the substance is tested in a Tier 2 test method (KeratinoSensTM, if Tier 1 indicates “sensitizer”, or HaCaT gene signature if Tier 1 indicates “non-sensitizer”). If the Tier 1 and 2 results are concordant, the test substance is classified accordingly. If the Tier 1 and 2 results are discordant, the h-CLAT is performed and the test substance is classified according to the standard h-CLAT prediction model. If all *in vitro* test methods are conducted, the RIVM STS covers all key events in the skin sensitization AOP. The approach can be readily applied if the user has access to all four QSAR models, and costs increase accordingly based on the *in vitro* methods that need to be run. The applicability domain of this approach is limited by the respective applicability domains of the *in vitro* methods required. This approach was not selected for quantitative performance assessment due to lack of HaCaT gene signature data.

DuPont IATA-SS

The “DuPont IATA-SS” approach is an example of a comprehensive IATA (requiring an expert judgment step), rather than a DA with a fixed data interpretation procedure. In this IATA, described in detail in (Patlewicz et al. 2014), a variety of data are considered using an ordered decision and integration process in a weight-of-evidence approach to predict skin sensitization potential. Expert knowledge is required to integrate existing *in vivo* and non-animal experimental data, information on protein binding profiles, existing mutagenicity and genotoxicity data, existing skin corrosion and irritation data, information on possible metabolites and some physicochemical properties. If data are insufficient to support a prediction of skin sensitization potential, additional non-animal and possible *in vivo* testing is conducted. Being an IATA, the approach is highly complex, requiring detailed understanding of the many various inputs and outsourcing to a CRO seems unlikely. This approach was not selected for quantitative performance assessment primarily due to reliance on expert judgment.

L'Oreal Stacking Meta-model

The L'Oreal "Stacking Meta-model" (Gomes et al. 2014; Del Bufalo et al. 2018) combines multiple *in vitro* and *in silico* parameters covering key events 1 to 3 of the skin sensitization AOP and gives a probability-based sensitizer hazard prediction. Specifically, this approach integrates *in silico* predictions from TIMES-SS and Toxtree; *in vitro* predictions from DPRA, KeratinoSensTM, and U-SENSTM; and physicochemical properties of the test substance (such as pH and volatility); using five different statistical methods (Boosting, Naïve Bayes, SVM, Sparse PLS-DA, and Expert Scoring) to create a stacking model (Gomes et al. 2014) that provides a probability of the test substance being a sensitizer or non-sensitizer based on defined thresholds. The Stacking Meta-model can tolerate missing individual data sources and can be run, for example, on substances with no defined structure lacking *in silico* predictions or on pigments that give inconclusive outcomes in *in chemico* or *in vitro* assays. The applicability domain of the Stacking Meta-model includes several classes of cosmetic chemicals and non-cosmetic organic substances that include pre- or pro-haptens. The approach can be applied to multi-constituent substances, substances of unknown or variable composition, complex reaction products, biological materials, or mixtures. Costs are driven by the necessity to run three *in chemico/in vitro* test methods, all of which should be available through CROs. This approach was not selected for quantitative performance assessment due to lack of data for several input features at the time of this study.

ICCVAM SVM

The ICCVAM skin sensitization working group has published several open-source models to predict skin sensitization hazard and potency based on evaluating multiple machine learning approaches and feature combinations (Strickland et al. 2016, 2017; Zang et al. 2017b). The ICCVAM SVM model to predict LLNA hazard is based on h-CLAT results, OECD Toolbox (version 3.2) predictions of hazard, and six physicochemical properties. The model to predict human hazard is based on KeratinoSensTM, h-CLAT, DPRA, log *P*, and OECD Toolbox predictions. These cover one or more of the AOP key events depending on the model. Based on the inclusion of physicochemical properties and structure-based predictions, these models are limited to substances with defined structures and those that are compatible with the *in vitro* test methods. Costs vary depending on the choice of model and associated number of *in vitro* tests; all can easily be performed by a CRO. This approach was selected for quantitative performance assessment based on the availability of the input features.

JRC CCT

The DA proposed by the Joint Research Center (JRC) of the European Commission is a classification trees consensus (CCT) model that predicts skin sensitization hazard based on structural features or protein reactivity descriptors (Asturiol et al. 2016). The underlying database for the JRC CCT approach consists of 269 organic substances with *in vitro* and *in*

chemico assay data and approximately 4500 descriptors generated with seven freely available or commercial software packages. Classification trees used in the approach were built by machine learning using 80% of the substances as a training set, resulting in a consensus model of two classification trees using descriptors from the TIMES-SS and Dragon software packages. This model should be understood primarily as a correlative approach, as the mechanistic relevance of the descriptors for skin sensitization is unclear. The approach is limited to organic substances with a well-defined identity, so that they can be processed in both TIMES-SS and Dragon. The costs are those of the two commercial software packages. This approach was not selected for quantitative performance assessment due to the lack of availability of proprietary structural descriptors for this substance set.

Givaudan ITS

The "Givaudan ITS" as described by Natsch et al. (2015) is an integrated testing strategy that predicts continuously scaled human and LLNA potency by multivariate regression. Various key events of the skin sensitization AOP are addressed using the following features: the KeratinoSensTM assay, information on adduct formation and peptide reactivity rate constants from the non-validated Cor1C420-assay, mechanistic alerts for reactivity of the test substance and key theoretical metabolites, log *P*, and vapor pressure calculated using TIMES-SS. Global regression models and models for specific domains, e.g. epoxides and aldehydes, are available. The approach cannot be used for substances with log *P* greater than 5, substances without defined structures, or substances that are insoluble in standard solvents. Costs are driven by the *in vitro* tests and requirement for commercial software; the Cor1C420-assay is not currently available through a CRO. This approach was not selected for quantitative performance assessment due to absence of Cor1C420-assay data for this substance set.

Shiseido ANN-EC3

The "ANN-EC3" model developed by Shiseido (Hirota et al. 2015) is a non-linear statistical model that combines multiple *in vitro* and *in silico* parameters covering Key Events 1–3 of the skin sensitization AOP. The model predicts the LLNA EC3 value using three *in vitro* methods (SH protein reactivity test, Antioxidant response element (ARE) test, and h-CLAT). Information sources can be interchangeable; for example, DPRA and KeratinoSensTM can be used as replacements for the SH test and ARE assay, respectively, as they are mechanistically and technologically equivalent. Multiple ANN models were built, with physicochemical properties of the test substance and QSAR predictions added as additional descriptors. The ANN models consist of an input layer (descriptors from *in vitro* or *in silico* results), a hidden layer, and an output layer (EC3 predictions), where predictive performance was evaluated via a 10-fold cross-validation procedure (Hirota et al. 2015). Costs and limitations are specific to the *in vitro* methods, e.g. poorly water soluble substances are outside of the applicability domain. This approach was selected for

quantitative performance assessment based on the availability of the input features.

Procter & Gamble BN-ITS 3

The Bayesian network integrated testing strategy (BN ITS-3) published by Procter & Gamble (P&G) in Jaworska et al. (2015) is the latest iteration in a series of decision support systems (Jaworska et al. 2011, 2013) intended to provide quantitative weight of evidence for potency categorization. The BN ITS-3 structure uses the quantitative data (rather than binary outcomes) from the three assays, DPRA, KeratinoSensTM, and h-CLAT that represent Key Events 1–3 of the skin sensitization AOP. The model also includes structure-based predictions from TIMES-SS, and a bioavailability calculation that relies on physicochemical properties calculated at pH = 7 (water solubility, log *D*, fraction ionized, and plasma protein binding; calculated using ACD labs software). The relevant in-domain evidence is integrated using a multistep process, and a skin sensitization potency prediction is provided as a probability distribution over four potency classes. The prediction is post-processed to correct for Michael acceptor chemistry and subsequently converted to Bayes factors, where the highest Bayes factor across potency classes drives the prediction, and a lower Bayes factor indicates stronger uncertainty. Costs for the approach are based on the necessity of running three *in vitro* assays and obtaining commercial software licenses. The BN ITS-3 can also be used with limited missing data to suggest which experiment should be conducted next to achieve maximum information and minimize cost or time investment. This approach was selected for quantitative performance assessment based on the availability of the input features and the willingness of P&G to run the proprietary BN ITS-3 model.

Kao STS

The sequential testing strategy (STS) developed by Kao (Nukada et al. 2013; Takenouchi et al. 2015) is a straightforward decision tree based on DPRA and h-CLAT data (addressing Key Events 1 and 3 of the AOP). The approach predicts three LLNA potency classes: strong, weak, and non-sensitizing. First, the h-CLAT is conducted; if the response is positive, the test substance is classified as “strong” or “weak” based on the lowest concentration that led to a positive response and no further testing is needed. If a negative h-CLAT result is obtained, the DPRA is conducted. The substance is classified as “weak” if the DPRA is positive (according to the standard prediction model), and “non-sensitizing” if the DPRA is negative. The limitations of this approach are specific to the individual test methods, such as reduced predictivity for poorly water-soluble substances. Costs are driven by the two test methods, both of which should be available through CROs. This approach was selected for quantitative performance assessment based on the availability of the input features.

Kao ITS

The integrated testing strategy (ITS) developed by Kao (Nukada et al. 2013; Takenouchi et al. 2015) is a scoring-

based decision tree that uses data from the DPRA and h-CLAT assays and a skin sensitization hazard prediction generated by the Derek Nexus software package, and covers Key Events 1 and 3 of the AOP. Similar to the Kao STS, the Kao ITS predicts three LLNA potency classes: strong, weak, and non-sensitizing. The two *in vitro* tests produce scores of 0–3, based on the strength of the observed responses, and the DEREK Nexus alert is either 0 (no alert) or 1 (alert). The sum of the scores yields the predicted LLNA category (0–1: non-sensitizer; 2–6: weak; 7: strong). The limitations of this approach are specific to the individual test methods, and costs are based on accessing the commercial software DEREK and running the two test methods, both of which should be available through CROs. This approach was selected for quantitative performance assessment based on the availability of the input features.

Unilever SARA

The SARA model developed by Unilever uses ordinary differential equations to provide a qualitative (mechanistic) and quantitative explanation of the induction of the naïve CD8⁺ T cell response. The SARA model evaluates Key Events 1, 3, and 4 of the skin sensitization AOP and also assesses diffusion and partitioning of sensitizing chemicals within human skin (MacKay et al. 2013; Maxwell et al. 2014). The model generates a prediction of skin sensitization potential using chemical-specific input parameters from a modification of an accepted test method for skin penetration (OECD 2014b) that generates data on skin bioavailability kinetics and protein haptentation (Davies et al. 2011). Uncertainty analysis of the SARA model has enabled parameter uncertainty to be directly visualized, and additional risk assessment case studies are ongoing to further evaluate the predictive capacity of the approach. Costs are based on the required assays, which have potentially limited CRO availability. This approach was not selected for quantitative performance assessment due to lack of data for several input features and the complexity of the approach.

Summary of qualitative evaluation

The qualitative evaluation of the 12 case studies submitted to OECD resulted in the selection of six DAs for quantitative performance assessment (Table 1, far right column). The six approaches constituted eight models, as the evaluation included two versions of the ICCVAM SVM approach (one predicting the LLNA and one predicting human hazard) and two versions of the Shiseido ANN approach (differing in input parameters used). Key to the selection of the six approaches were the ready availability of input data for these approaches, either through curation or generation by the CE STTF. In general, the six approaches selected for quantitative performance assessment relied upon the three *in vitro* methods with OECD guidelines (DPRA, KeratinoSensTM, and h-CLAT), did not require expert judgment, and had data interpretation procedures that could be reproduced using open-source code. The exception to this was the Bayesian ITS-3, for which generation of the probabilities and Bayes factor

calculations were reproduced by P&G using the commercially available BayesiaLab software package. Data analysis and statistical evaluation for all approaches was conducted by NICEATM to ensure consistency.

LLNA and human reference data

The LLNA and human data used for the quantitative performance assessment are described in detail elsewhere (Hoffmann et al. 2018). Briefly, NICEATM collected data from literature sources for 128 substances having high-quality human studies and one or more high-quality LLNA studies. There were multiple LLNA study results for approximately half the substances, providing an indication of reproducibility, similar to Dumont et al. (2016), Hoffmann (2015), Roberts et al. (2016) and others, and necessitating the calculation of a modified median that required sufficiently high testing concentrations for negative results (Hoffmann et al. 2018). Using this highly curated dataset, the ability of the LLNA to accurately predict human test results was 74% for hazard prediction (sensitizer versus non-sensitizer), 59% for potency prediction using three classes (strong, weak, and non-sensitizer) and 45% for potency prediction using five classes (extreme, strong, moderate, weak, and non-sensitizer). These numbers provided the

baseline for comparison when assessing the predictive performance of the DAs.

Quantitative performance assessments

Following the qualitative evaluation of the DAs shown in Table 1, the predictive performance of each of the six methods, encompassing eight models, was quantitatively assessed against the maximum subset of substances (out of 128) that had a complete data matrix of input features. Depending on the DA, this ranged from 120 to 127 substances (indicated in Supplementary Table S2). The overall statistics in terms of accuracy (number of correctly classified substances over total number of substances assessed), false negatives (FN), and false positives (FP), sensitivity, specificity, and balanced accuracy are reported for each DA compared both to LLNA and human reference data. For potency categorization, the numbers of under- and over-predicted substances are also reported. The quantitative performance results for LLNA and human endpoints are shown in Tables 3 and 4 (human and LLNA hazard, respectively) and Tables 5 and 6 (human and LLNA potency, respectively) for every substance eligible for evaluation by each DA. In Tables 3 and 5, the LLNA performance against the human endpoint is shown in bold. The predictive performance of each DA was also assessed uniquely

Table 3. Defined Approach (DA) performance in predicting human hazard (sensitizer/non-sensitizer).

Predicting Human Hazard								
Defined Approach:	BASF 2/3 (DKH)	Kao STS	Kao ITS	ICCVAM SVM (Human)	Shiseido ANN (D_hC)	Shiseido ANN (D_hC_KS)	P&G BN ITS-3	LLNA
N	127	126	120	120	126	126	119	128
Accuracy (%)*	77.2	80.2	85.0	81.7	78.6	78.6	75.6	74.2
Sensitivity (%)	79.3	97.7	93.8	86.4	95.4	100	81.3	85.2
Specificity (%)	72.5	41.0	66.7	71.8	41.0	30.8	64.1	50.0
BA (%)	75.9	69.4	80.3	79.1	68.2	65.4	72.7	67.6

*Performance is shown against the maximum subset (N) out of 128 substances with all necessary DA features.

BA: balanced accuracy; STS: sequential testing strategy; ITS: integrated testing strategy; SVM: support vector machine; ANN: artificial neural network; BN: Bayesian network; DKH and D_hC_KS: DPRA/h-CLAT/KeratinoSensTM; D_hC: DPRA/h-CLAT.

Table 4. Defined approach (DA) performance in predicting LLNA hazard (sensitizer/non-sensitizer).

Predicting LLNA Hazard							
Defined Approach:	BASF 2/3 (DKH)	Kao STS	Kao ITS	ICCVAM SVM (LLNA)	Shiseido ANN (D_hC)	Shiseido ANN (D_hC_KS)	P&G BN ITS-3
N	127	126	120	120	126	126	119
Accuracy (%)*	70.1	77.8	79.2	88.3	76.2	81.0	83.2
Sensitivity (%)	72.3	92.6	85.6	93.3	90.4	97.9	83.2
Specificity (%)	63.6	34.4	60.0	73.3	34.4	31.3	83.3
BA (%)	68.0	63.5	72.8	83.3	62.4	64.6	83.3

*Performance is shown against the maximum subset (N) out of 128 substances with all necessary DA features.

LLNA: local lymph node assay; BA: balanced accuracy; STS: sequential testing strategy; ITS: integrated testing strategy; SVM: support vector machine; ANN: artificial neural network; BN: Bayesian network; DKH and D_hC_KS: DPRA/h-CLAT/KeratinoSensTM; D_hC: DPRA/h-CLAT.

Table 5. Defined Approach (DA) performance in predicting human sensitizing potency.

Predicting Human Potency (Strong, Weak, Non-sensitizers)						
Defined approach:	Kao STS	Kao ITS	Shiseido ANN (D_hC)	Shiseido ANN (D_hC_KS)	P&G BN ITS-3	LLNA
N	126	120	126	126	115	128
Accuracy (%)*	63.5	69.2	61.1	62.7	54.8	59.4
Over-predicted (%)	22.2	13.3	22.2	25.4	20.0	19.5
Under-predicted (%)	14.3	17.5	16.7	11.9	25.2	21.1

*Performance was assessed for prediction of three potency classes as described in the main text, and is shown against the maximum subset (N) out of 128 substances with all necessary DA features. With the exception of the P&G BN ITS-3, all misclassifications varied by one class only (i.e. no non-sensitizers were predicted as strong sensitizers or vice versa).

STS: sequential testing strategy; ITS: integrated testing strategy; SVM: support vector machine; ANN: artificial neural network; BN: Bayesian network; DKH and D_hC_KS: DPRA/h-CLAT/KeratinoSensTM; D_hC: DPRA/h-CLAT.

Table 6. Defined Approach (DA) performance in predicting LLNA sensitizing potency.

Predicting LLNA Potency (Strong, Weak, Non-sensitizers)					
Defined Approach: N	Kao STS 126	Kao ITS 120	Shiseido ANN (D_hC) 126	Shiseido ANN (D_hC_KS) 126	P&G BN ITS-3 115
Accuracy (%)*	67.5	66.7	65.1	69.8	67.8
Over-predicted (%)	21.4	14.2	21.4	23.0	12.2
Under-predicted (%)	11.1	19.2	13.5	7.1	20.0

*Performance was assessed for prediction of three potency classes as described in the main text, and is shown against the maximum subset (N) out of 128 substances with all necessary DA features. With the exception of the P&G BN ITS-3, all DA human potency predictions were off by one class only (i.e. no non-sensitizers predicted as strong or vice versa).

LLNA: local lymph node assay; STS: sequential testing strategy; ITS: integrated testing strategy; SVM: support vector machine; ANN: artificial neural network; BN: Bayesian network; DKH and D_hC_KS: DPRA/h-CLAT/KeratinoSensTM; D_hC: DPRA/h-CLAT.

against each non-overlapping test set (ranging from 36 to 66 substances, depending on the DA), and is reported in the following sections. The detailed predictions on a substance-by-substance basis are provided in [Supplementary Table S2](#), including indications of missing data and overlap with training sets where applicable, and detailed statistics (e.g. 95% confidence intervals, no-information rates, *p* values) can be obtained by running the code provided in the [Supplementary Materials](#).

BASF “2 out of 3”

The straightforward consensus approach proposed by BASF (Urbisch et al. 2015) relies on a majority rule from a set of three assays combining the DPRA, KeratinoSensTM, and the h-CLAT (referred to as “2/3 DKH”). Another consensus approach substituted the U-SENSTM for the h-CLAT, but there were too few substances with U-SENSTM data in this dataset to allow for a similar evaluation of that approach. When compared to the human data, the BASF 2/3 DKH yielded 77.2% accuracy for the binary hazard prediction (18 FN, 11 FP, *n* = 127). When compared to the LLNA, the BASF 2/3 DKH was 70.1% accurate in predicting hazard (26 FN, 12 FP, *n* = 127). This approach did not use any algorithm to optimize the decision rule to reference data, so there was no training set; however, there was a 91 substance overlap with the published dataset (Urbisch et al. 2015). The performance against only the non-overlapping set of 36 substances was 66.7% in predicting human hazard (7 FN, 5 FP, *n* = 36) and 58.3% in predicting LLNA hazard (7 FN, 8 FP, *n* = 36).

Kao STS

The Kao STS DA is a sequential decision tree relying solely upon the DPRA and h-CLAT assays to provide both hazard and potency (three class) predictions. The approach predicted human hazard with accuracy of 80.2% (2 FN, 23 FP, *n* = 126). The three-class potency prediction was 63.5% accurate, with 18 substances under-predicted and 28 over-predicted. The Kao STS predicted LLNA hazard outcomes with 77.8% accuracy (7 FN, 21 FP, *n* = 126), and classified substances as LLNA strong, weak, and non-sensitizers with 67.5% accuracy, with 14 substances under-predicted and 27 substances over-predicted. This DA also did not use a training set in the machine learning context, but its decision rules were optimized based on a set of published data covering 101 substances (Nukada et al. 2013), and later applied to an additional set of 38

substances (Takenouchi et al. 2015). The combined Kao STS dataset of 139 substances had an overlap of 74 substances with the set used here. When considering the non-overlapping set of 52 substances, the approach predicted human hazard with an accuracy of 65.4% (0 FN, 18 FP, *n* = 52), and human potency with an accuracy of 53.9% for three classes, with five substances under-predicted and 19 over-predicted. The approach predicted LLNA hazard of the non-overlapping substance set with 67.3% accuracy (1 FN, 16 FP, *n* = 52) and predicted LLNA potency with 61.5% accuracy, with three under-predicted and 17 over-predicted.

Kao ITS

The Kao ITS is a test battery-based DA that relies upon the quantitative results from the DPRA and the h-CLAT as well as *in silico* hazard predictions from the structure-based DEREK software. Therefore, natural extracts without defined structure (Jasmine grandiflorum, Jasmine sambac, Oakmoss, Treemoss, Tea leaf, and Ylang ylang) were omitted from the assessment of this DA due to inability to generate DEREK predictions for these substances. Two versions of the Kao ITS were evaluated; the results presented here are from the most recent version which includes the consideration that under the condition that the lysine peptide co-elutes with the test substance, one should only consider cysteine peptide depletion (Hirota et al. 2015). The prediction results from the previous ITS (Nukada et al. 2013) differ for only a few substances, and can be found in [Supplementary Table S2](#).

The Kao ITS predicted human skin sensitization hazard with 85% accuracy (5 FN, 13 FP, *n* = 120), and the 3-classes potency prediction was 69.2% accurate, with 21 substances under-predicted and 16 over-predicted. The approach predicted LLNA hazard outcomes with 79.2% accuracy (13 FN, 12 FP, *n* = 120) and predicted three-class LLNA potency outcomes with 66.7% accuracy, with 23 substances under-predicted and 17 substances over-predicted. This DA did not have a training set, rather an optimization set (Takenouchi et al. 2015), although DEREK had a training set. The accuracies against only the 46 non-overlapping substances for human hazard were 69.6% (3 FN, 11 FP, *n* = 46) and for human potency 63%, with six substances under-predicted and 11 over-predicted. The approach predicted LLNA hazard outcome for the non-overlapping set with 63% accuracy (7 FN, 10 FP, *n* = 46) and LLNA potency outcomes with 58.7% accuracy, with nine substances under-predicted and 10 over-predicted.

ICCVAM SVM

We evaluated two ICCVAM SVM models to predict skin sensitization hazard, one that was trained on human data (Strickland et al. 2017) and another that was trained on LLNA data (Strickland et al. 2016).

The human SVM model uses a prediction of skin sensitization hazard generated by OECD Toolbox, log *P*, the binary results from the KeratinoSensTM and the h-CLAT, and the quantitative result (average of lysine and cysteine depletion) from the DPRA. The model resulted in 81.7% accuracy for human hazard prediction (11 FN, 11 FP, *n* = 120). This DA was trained using a machine learning algorithm on a set that had a 54-substance overlap with the current evaluation set. The performance against the external test set was 71.2% (8 FN, 11 FP, *n* = 66) in predicting human hazard.

The LLNA SVM model uses the binary (positive/negative) result from the h-CLAT, the OECD Toolbox hazard prediction, and six physicochemical properties as inputs (Strickland et al. 2016; Zang et al. 2017a). There were 120 substances those had all the necessary input features to run the SVM model. The model was 88.3% accurate for LLNA hazard prediction (6 FN, 8 FP, *n* = 120). This DA was based on a machine learning algorithm, and used a training set that had 57 substances in common with set used here. The performance was slightly less accurate for the external test set, providing 79.4% accuracy for LLNA hazard prediction (5 FN, 8 FP, *n* = 63).

Shiseido ANN

Two of the four ANN models described in Hirota et al. (2015) were evaluated here, chosen based on availability of the input data and published performance of the models. The first model (ANN_D_hC) used quantitative values from the DPRA (average of lysine and cysteine depletion) and the h-CLAT (minimum of CD86 EC150, CD54 EC200, and CV75) to predict the EC3 value that would be produced in the LLNA. The second model (ANN_D_hC_KS) used the same structure with an additional maximum response input from an in-house ARE assay that was mechanistically and functionally similar to the KeratinoSensTM assay. For this analysis, the *I*_{max} value from the KeratinoSensTM was used as the third input. In both cases, the substance was predicted as negative if all the inputs were negative; therefore the ANN models only produced quantitative LLNA EC3 predictions for substances that were positive in at least one of the input assays.

There were 126 substances in the current database that had the necessary inputs to run the ANN models. The ANN_D_hC model was 78.6% predictive of human hazard (4 FN, 23 FP, *n* = 126), and 61.1% predictive for three potency classes, with 21 substances under-predicted and 28 over-predicted. The ANN_D_hC model was 76.2% predictive for LLNA hazard overall (9 FN, 21 FP, *n* = 126), and 65.1% predictive for 3 potency classes, with 17 substances under-predicted and 27 over-predicted. The ANN_D_hC model training set had a 74-substance overlap with the current set, leaving 52 substances to serve as an external test set. For human hazard, the test set accuracy was 63.5% (1 FN, 18 FP, *n* = 52) and 51.9% for the three potency classes, with six substances under-predicted and 19 over-predicted. The model was 65.4%

accurate in predicting LLNA sensitization hazard against only that test set (2 FN, 16 FP, *n* = 52) and 59.6% accurate when predicting the three potency classes, with four substances under-predicted and 17 over-predicted. The ANN_D_hC_KS model was 78.6% predictive for human hazard overall (0 FN, 27 FP, *n* = 126), and 62.7% predictive for three potency classes, with 15 substances under-predicted and 32 over-predicted. The ANN_D_hC_KS model was 81% predictive for LLNA hazard overall (2 FN, 22 FP, *n* = 126), and 69.8% predictive for three potency classes, with nine substances under-predicted and 29 over-predicted. The ANN_D_hC_KS model training set had a 44-substance overlap with the current set, resulting in an 82-substance test set. For human hazard, the test set accuracy was 73.2% (0 FN, 22 FP, *n* = 82) and 56.1% for the three potency classes, with 11 substances under-predicted and 25 over-predicted. The ANN_D_hC_KS model was 78.1% accurate in predicting test set LLNA sensitization hazard (0 FN, 18 FP, *n* = 82) and 65.1% accurate when predicting the three LLNA potency classes, with five substances under-predicted and 23 over-predicted.

P&G BN ITS-3

There were 120 substances with the necessary inputs for the BN ITS-3 model to calculate Bayes factors corresponding to probabilistic predictions for four potency classes (non-sensitizer and weak, moderate, and strong sensitizers). Of the 120 substances run through the model, 23 substances had Bayes factors less than 3, indicating a low prediction reliability; these substances were nonetheless included in the comparison. In five cases, two potency classes had an almost identical Bayes factors; in one case the prediction was borderline non-sensitizer versus weak, in four other cases it was borderline within sensitizer categories (e.g. weak/moderate). For the purpose of the hazard evaluation, the one substance that was borderline non-sensitizer versus weak sensitizer was not included in the statistical analysis. For the potency evaluation all five borderline cases were removed. Thus, the ITS-3 model yielded 119 substances with hazard predictions and 115 substances with potency predictions that could be compared to the LLNA and human data. For potency prediction comparisons, moderate and weak were combined into one category labeled "weak".

The BN ITS-3 model predicted human skin sensitization hazard with 75.6% accuracy (15 FN, 14 FP, *n* = 119), and 54.8% predictive for three potency classes, with 29 substances under-predicted and 23 over-predicted. The BN ITS-3 model predicted LLNA hazard outcomes with 83.2% accuracy (15 FN, 5 FP, *n* = 119), and 67.8% accuracy for three potency classes, with 23 substances under-predicted and 14 over-predicted. The BN ITS-3 model training set had a 66-substance overlap with the current set, leaving 54 substances with data to serve as an external test set. The model was 66.7% accurate in predicting human sensitization hazard against only that test set (9 FN, 9 FP) and 52.8% accurate when predicting the three potency classes, with 12 substances under-predicted and 13 over-predicted. For LLNA hazard, the test set accuracy was 75.9% (9 FN, 4 FP, *n* = 54) and 66% for the

three potency classes, with 11 substances under-predicted and seven over-predicted.

Common test set

The eight models from the six DAs had varying degrees of substance overlap between the 128-substance dataset used for the current analysis and their respective optimization or training sets. There were 28 substances (3 strong human sensitizers, 15 weak, and 10 non-sensitizers) that were common to the non-overlapping external test sets of all the DAs. Should the reader wish to compare DAs in a pairwise or sub-set fashion with a larger number of substances, all the necessary information is provided in [Supplementary Table S2](#). The performance of each of the *in vitro* assays and the DAs, in terms of sensitivity, specificity, balanced accuracy, and accuracy, against the 28-substance common test set, for both LLNA and human data, is shown in [Supplementary Table S3](#). The LLNA predictive performance against the human endpoint for these 28 substances was quite low, at 50% accuracy for hazard prediction. The performance was uniformly higher (61–75%) across all the DAs when predicting human hazard. The performance ranged from 50% to 79% when comparing the DAs to the LLNA endpoint. Several of the DAs (e.g. Kao STS and Shiseido ANN models) were skewed toward sensitivity. Similar to the performance against the entire 128-substance dataset and their respective external test sets, all the DAs demonstrated superior performance to the LLNA in predicting the human endpoint for the common test set.

Concordance across substances

Out of the 128-substance dataset, 127 substances could be run in at least one of the DAs (2-Hexylidene cyclopentanone did not have data for DPRA or h-CLAT). Out of the 127 substances, there were 50 human skin sensitizers that were correctly predicted by all the DAs, all of which were also positive in the LLNA, and nine non-sensitizers that were also correctly predicted as negative by all the DAs and the LLNA. The remaining 68 substances had some degree of discordance when compared across all the DAs, as shown in [Figure 1](#). The dendrogram on the left shows clusters of substances (by complete linkage method) that exhibit varying patterns across the DA predictions when compared to LLNA and human hazard data. [Supplementary Figures S1 and S2](#) show similar heatmaps for all substances, for hazard and potency respectively. Patterns in DA predictivity across physicochemical properties, mechanistic reaction domains, and ability to predict pre- and pro-haptens were examined (comparisons given in [Supplementary Table S4](#)). Similar to the results for the individual test methods presented in (Hoffmann et al. 2018), few significant differences from overall predictive performance across the substance set were observed. As was observed with the DPRA alone, the substances reacting by acyl transfer, which are all human sensitizers, were uniformly predicted as positive across the DAs, with the exceptions of Benzoyl peroxide (FN in the BASF 2 out of 3) and Penicillin G (FN in the ICCVAM_SVM_Human model). Of the 22 human

sensitizing pre- and pro-haptens, all were predicted correctly by the majority (five or more) of the DAs, and 17 (77%) were predicted correctly by every DA. Limonene was designated as a pre-hapten by (Patlewicz et al. 2016), and was predicted as a sensitizer by five DAs, but was categorized as a human non-sensitizer based on Basketter et al. (2014).

Discussion

Given the marketing ban on cosmetics ingredients tested on animals that has been implemented or is pending in many parts of the world, there is a critical need to identify and validate non-animal testing strategies for skin sensitization hazard and potency. Validated alternative testing strategies will enable rapid, effective risk assessment of cosmetics products and other types of substances. The field of skin sensitization is ahead of other areas of toxicology in this regard because the biological steps comprising the adverse outcome pathway for skin sensitization are relatively well understood, and corresponding test methods have been developed to address these key cellular and molecular events. In addition, the existence of reliable human data, i.e. clinical reports from occupational exposures and standardized data collection from human patch tests, is a rare asset that facilitates development and assessment of new methods for this endpoint.

Evaluating non-animal testing strategies that combine these novel approaches requires an appropriate basis for comparison. The regulatory “reference standard” animal model, the LLNA, provides a benchmark, both in terms of predictive performance against the human response and reliability/reproducibility of the animal test. When compared to the human data for 128 substances used in this analysis, the LLNA was 74% predictive for the binary hazard endpoint, 59% predictive for three potency classes, and 45% predictive for five potency classes. Other studies have shown similar results, such as the 1999 ICCVAM evaluation of the LLNA, where accuracy with respect to human hazard classification was 72% ($n=57$) or more recently as high as 82% ($n=111$) (Urbisch et al. 2015). For three potency categories (strong, weak, and non-sensitizers), ICCVAM (2010) found the accuracy of the LLNA to be 61% ($n=56$). Certain regulatory agencies, such as the U.S. Food and Drug Administration, still accept guinea pig data; the accuracy of this animal test in predicting human outcomes is similar to the LLNA, with 72% accuracy in hazard classification ($n=57$) and 59% accuracy in potency classification ($n=56$). The baseline reproducibility of the LLNA provides another metric for comparison. The concordance between individual LLNA tests for a given substance tested multiple times and the summary statistic for that substance ranges from 63% to 73% (depending on the summary statistic used). In contrast, the inter-laboratory reproducibilities of the OECD-validated *in vitro* tests are all greater than 80% (OECD 2015a, 2015b, 2017).

In the current study, multiple DAs and IATA submitted to the OECD as case studies for predicting skin sensitization hazard or potency were first evaluated in a qualitative manner, i.e. by characterizing the input requirements, feasibility of application by naïve users, transparency, etc. Those that could be reproduced using open-source software (or in the

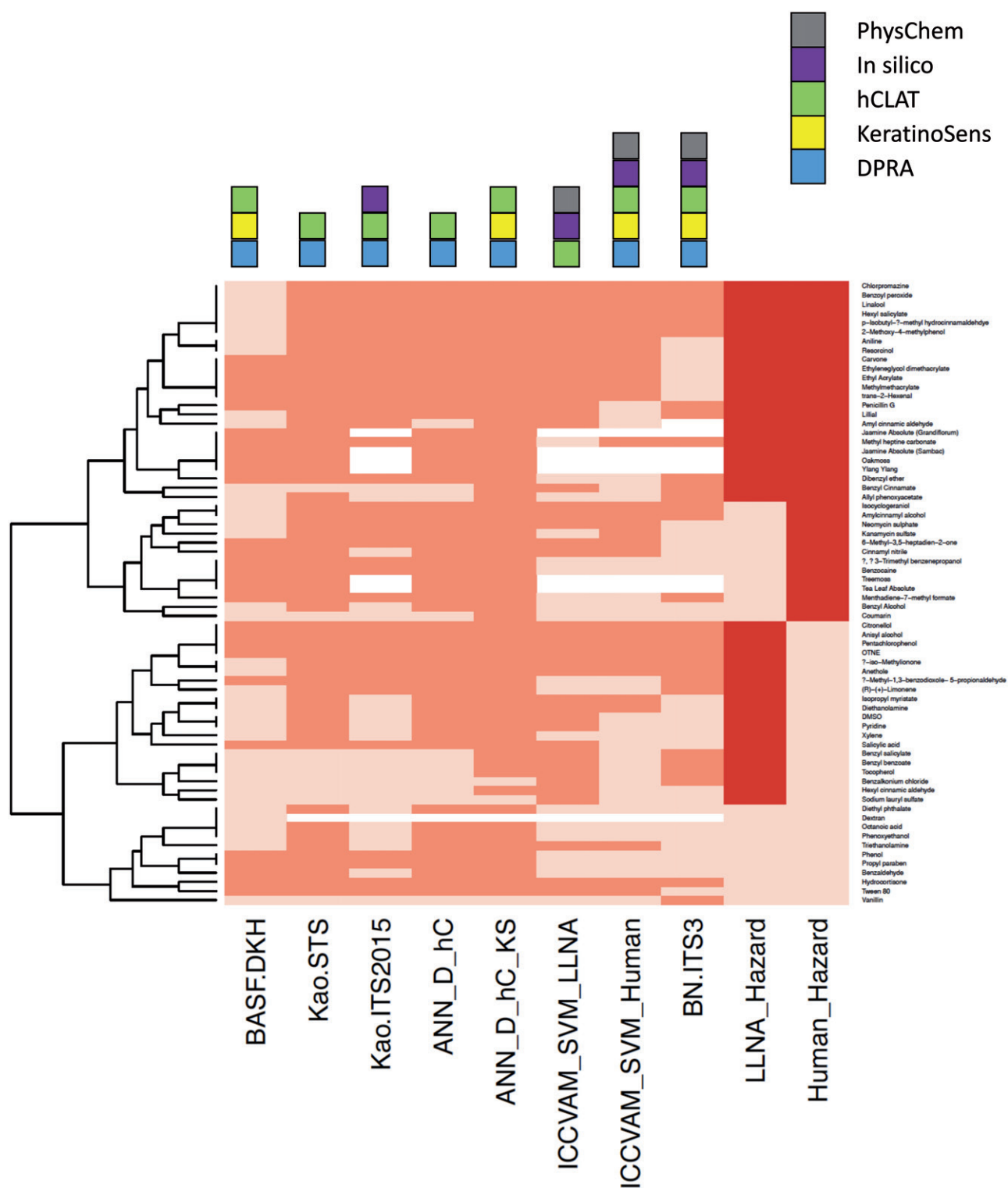


Figure 1. The heatmap shows defined approach (DA) predictions and LLNA/Human hazard data for the 68 substances with some degree of discordance across the results. Orange or red indicates sensitizer based on DA predictions or *in vivo* data, respectively, and tan cells are non-sensitizer predictions/data. White indicates that not all required features were present to run the model for the DA. The dendrogram on the left shows clusters of substances (complete linkage method) and the color coding above the plot indicates which features were used by each model. Abbreviations for each DA name are defined in the text.

case of the Bayesian ITS-3, software provided by CE STTF members) were quantitatively evaluated against a large set of substances having both LLNA and human data available, including many cosmetics ingredients. For this entire substance set, the accuracy of the eight models, derived from six DAs, in predicting the LLNA skin sensitization hazard endpoint ranged from 70% to 88%. The five potency prediction

models evaluated here showed 65–70% predictive performance in discriminating strong, weak, and non-sensitizers based on the LLNA endpoint. As indicated previously, when examining multiple LLNA tests for the same substance there is a fairly low concordance between the potency categories that would be assigned if one relied exclusively on an individual LLNA result versus the weight-of-evidence provided by

multiple tests, regardless of the summary metric used for the EC3 (e.g. median, mean, etc.). For hazard identification, the LLNA displays 78% concordance with itself, while potency categorization is 63–73% depending on the summary metric (Hoffmann et al. 2018). This reflects the variability in the animal data and provides an indication of the uncertainty that exists when comparing DA predictions to the LLNA endpoint, especially in the case of substances that only have one LLNA study. Further, as the goal is ultimately to predict skin sensitization in humans, comparing the DA outcomes to the animal results obscures the true accuracies of the models, which may be, and in many cases are, superior to the animal data in predicting the human endpoint.

In fact, all the non-animal DAs evaluated here performed as well as or better than the LLNA in predicting human skin sensitization endpoints for both hazard and potency. The performance of the DAs evaluated here in predicting human skin sensitization hazard endpoint ranged from 76% to 85% (compared to the LLNA, which was 74% accurate). The DAs that predicted three potency classes showed 55–69% predictive performance in discriminating strong, weak, and non-sensitizers based on the human categories outlined in Basketter et al. (2014), (compared to 59% for the LLNA). No strong human sensitizers were falsely predicted as non-sensitizers by any of the DAs, except for the BN ITS-3 predictions for trans-2-hexenal and 6-methyl-3,5-heptadien-2-one. Certain DAs maximize sensitivity at the expense of specificity (e.g. Kao STS and both Shiseido ANN models). The highest balanced accuracy (80%) against the human hazard endpoint was produced by the Kao ITS model, which depends on DPRA, h-CLAT, and DEREK predictions. However, as shown in Figure 1, the DAs that include *in silico* predictions cannot be applied to natural products without defined structures, such as Treemoss, Jasmine, and Ylang ylang. Certain compounds were consistently mispredicted as sensitizers across all the DAs, including those that were also false positives in the LLNA (Anisyl alcohol, Citronellol, Pentachlorophenol, and OTNE), and Hydrocortisone, which was negative in the LLNA. None of the human sensitizers were missed by every DA, but Coumarin was a false negative in all DAs except the ANN_D_hC_KS. Moreover, three other substances (Allyl phenoxyacetate, Benzyl alcohol, and Benzyl Cinnamate) were correctly identified as sensitizers by only three out of the eight models.

Each of the data interpretation procedures used by the DAs has a unique set of advantages and limitations. The test battery methods require very limited computational power and are transparent and mechanistically interpretable. The machine learning methods can combine features in unique ways that may increase the predictive power, with a tradeoff in loss of transparency. Interestingly, the hazard predictions for the Kao STS and the ANN_D_hC models (which both depend on DPRA and h-CLAT) are almost identical, despite having extremely different data interpretation procedures, indicating at least in this case that complex approaches may not provide added benefit over a simple decision tree. However, the more complex DAs, like the ANNs, also have the ability to provide a quantitative prediction of potency potentially to be used for skin sensitization risk assessment.

Most of the DAs also rely heavily upon one or more of three *in vitro* test methods. The prediction models for the individual test methods were optimized in different fashions, to increase accuracy or to reduce false negatives, which is demonstrated via the performance of the individual test methods and propagated to the DAs that depend on those methods and the associated prediction models. Factors affecting applicability of the individual methods (e.g. technological limitations) need to be considered in application of the DAs. There is also a need for careful tracking of software versions for all DAs that rely on *in silico* tools, in case retraining of the model is warranted.

There exist particular challenges concerning the application of the potency categorizations, for both the LLNA and the human data. The LLNA potency prediction models evaluated here were trained on the ECETOC cutoff of an EC3 of 1% for strong sensitizers, rather than the EU-CLP cutoff of 2% for Categories 1 A/1B. The database presented in the companion paper (Hoffmann et al. 2018) could be used to develop DAs that specifically target these subcategories. In the case of the human data, substances with individual case reports of sensitization have been considered here as non-sensitizers, matching the current European regulatory approach, but in certain high-exposure scenarios for cosmetics ingredients very weak potencies may play a role. Therefore, the current analysis was repeated under a reclassification where only human category 6 substances (Basketter et al. 2014) were considered negatives (category 5 were designated weak positives), and the accuracies were very similar, with a tradeoff resulting in very few false positives, but more false negatives from category 5 (see Supplementary Material for code to run performance statistics with alternate human data classification). An example from category 5 is isopropanol, a substance used in many cosmetic products. No relevant intrinsic sensitization potential has been reported [recently reviewed by Zhang et al. (2017)]. However, some positive patch test results under specific medical treatment conditions have been reported (Garcia-Gavin et al. 2011) that are not considered relevant for cosmetics. Consequently, there is an intrinsic potency to induce skin sensitization in humans, but cosmetic exposure is highly unlikely to induce skin sensitization.

The system used by Basketter et al. (2014) enables human potency categorization by a combination of direct human induction (e.g. HRIPT) and pre-clinical induction (LLNA) data with human clinical data (diagnostic patch test). Consequently, the categorization is based on the existence of intrinsic hazard potency of the substance in question (HRIPT, LLNA) and its contact allergy prevalence, particularly referring to the subset of the general population attending dermatology clinics (Thyssen et al. 2007). This is considered to permit better (expert) judgment of sensitization rates in the context of real life exposure relevant for consumers. While human potency categorization based on expert judgment is very valuable, the transparency and reproducibility of this classification system needs to be increased, and work is ongoing in this area (Api et al. 2017). Another point to consider is that many of the DAs, and indeed the individual test methods they rely upon, use prediction models that were trained on LLNA data, which may be negatively influencing their

predictive performance against human sensitization categories. As discussed in Hoffmann et al. (2018), prediction models could be developed for subsets of substances based on *a priori* information such as physicochemical properties, rather than attempting to apply global models across chemical space.

The DAs and IATA that we were not able to quantitatively evaluate here have also demonstrated strong performance in predicting skin sensitization potential, and the current work is not meant to exclude them from consideration in regulatory risk assessment. Integrated assessments will inherently incorporate all reliable and relevant information, such as physicochemical data, intended substance function (e.g. dye) or origin (e.g. natural extracts), information on suspected transformation, skin penetration information, etc. Additional considerations that could be the subject of future efforts include strategy optimization for potency predictions (to specifically identify strong sensitizers, for example), and development of testing strategy amendments for specific purposes (e.g. explicit confirmation of non-sensitization). Further characterization of variability and uncertainty in the human data and the non-animal data, as well as expansion of the human reference dataset, are also needed. These and other aspects are being addressed by the International Cooperation on Alternative Test Methods (ICATM) in the context of the OECD Test Guidelines Program to facilitate international regulatory harmonization (Casati et al. 2017), and also by Cosmetics Europe (Hoffmann et al. 2018). Furthermore, these results can be combined with data on other types of substances to expand and evaluate the applicability domain of the DAs. The U.S. National Toxicology Program is working with ICCVAM federal agencies to generate *in vitro* data on nominated substances and formulation products that will facilitate further evaluation of these DAs for non-cosmetics, e.g. pesticides and industrial chemicals.

In this study, we qualitatively and quantitatively evaluated several non-animal DAs for skin sensitization testing that used data from validated *in vitro* assays, non-testing resources such as *in silico* predictions and physicochemical properties, and algorithms that were in most cases reproduced using open-source code. The code provided here will enable others to apply the DAs to additional substances as new datasets are generated. All of the DAs demonstrated equivalent or superior performance to existing animal tests and were successful in predicting human skin sensitization outcomes for both hazard and potency. These DAs rely heavily upon *in vitro* assays mapped to key events in the AOP, and in some cases require input from commercial or open-source *in silico* tools. This work represents significant progress towards formalizing performance criteria that can be applied to any new non-animal testing strategy, and identifies DAs that could serve as effective replacements to the LLNA in predicting human skin sensitization hazard or indeed for risk assessment.

Acknowledgements

The authors are grateful to Eileen Phillips (Integrated Laboratory Systems, Inc.) for data curation, Cindy Ryan and Greg Dameron (P&G) for

generating all BN ITS3 data, Jason Pirone (formerly Sciome, LLC) for assistance in ANN function coding, and Catherine Sprankle (Integrated Laboratory Systems, Inc.) for editorial support. In addition, the authors gratefully acknowledge the constructive comments of the anonymous reviewers, which were very helpful in improving the manuscript.

Declaration of interest

This work has been conceived, planned and executed by Cosmetic Europe's Skin Tolerance Task Force (CE STTF) in collaboration with NIH/NIEHS/DNTP/NICEATM (Nicole Kleinstreuer, Warren Casey), supported by ILS (Dave Allen, Judy Strickland, Qingda Zang).

Cosmetics Europe is the European trade association for the cosmetics and personal care industry. The members include cosmetics and personal care manufacturers, and also associations representing our industry at national level, right across Europe (for more information see: <https://www.cosmeticseurope.eu/>). Cosmetic Europe facilitated scientific meetings of the CE STTF, coordinated the overall project management and administrative tasks relating to the completion of this work, and contributed to the writing of the manuscript. The CE STTF is composed of cosmetic company experts (see affiliations on cover page), who are not paid for their work on the task force, external consultants (Sebastian Hoffmann and Erwin van Vliet paid by Cosmetics Europe) and Cosmetics Europe staff (Bertrand Desprez and Martina Klaric employed by Cosmetics Europe).

Nicole Kleinstreuer is employed by NIH/NIEHS/DNTP/NICEATM, and conducted the analyses presented herein, wrote the code to reproduce the defined approaches, and drafted the manuscript. Warren Casey is employed by NIH/NIEHS/DNTP/NICEATM and contributed editorial input to the manuscript. Dave Allen, Judy Strickland, and Qingda Zang are employed by ILS, and contributed editorial input to the manuscript, assistance with coding, and feature inputs to defined approaches. ILS employees were supported with federal funds from the National Institute of Environmental Health Sciences, National Institutes of Health under Contract No. HHSN273201500010C in support of NICEATM.

The opinions expressed herein and the conclusions of this publication are those of the authors and do not necessarily represent the views of any government agency, Cosmetics Europe, nor those of its member companies.

None of the authors has appeared in any legal or regulatory proceedings in the last five years related to the contents of this paper.

The NIEHS and the companies of the CE STTF members have approved this work for publication.

Supplemental material

Supplemental material for this article is available online [here](#).

References

- Alepee N, Piroird C, Aujoulat M, Dreyfuss S, Hoffmann S, Hohenstein A, Meloni M, Nardelli L, Gerbeix C, Cotovio J. 2015. Prospective multi-centre study of the U-SENS test method for skin sensitization testing. *Toxicol In Vitro*. 30:373–382.
- Anderson SE, Siegel PD, Meade BJ. 2011. The LLNA: a brief review of recent advances and limitations. *J Allergy (Cairo)*. 2011:424203.
- Api AM, Parakhia R, O'Brien D, Basketter DA. 2017. Fragrances categorized according to relative human skin sensitization potency. *Dermatitis*. 28:299–307.
- Ashikaga T, Yoshida Y, Hirota M, Yoneyama K, Itagaki H, Sakaguchi H, Miyazawa M, Ito Y, Suzuki H, 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.
- Asturiol D, Casati S, Worth A. 2016. Consensus of classification trees for skin sensitisation hazard prediction. *Toxicol In Vitro*. 36:197–209.

- Basketter DA, Alepee N, Ashikaga T, Barroso J, Gilmour N, Goebel C, Hibatallah J, Hoffmann S, Kern P, Martinozzi-Teissier S, et al. 2014. Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis*. 25:11–21.
- Bauch C, Kolle SN, Ramirez T, Eltze T, Fabian E, Mehling A, Teubner W, van Ravenzwaay B, Landsiedel R. 2012. Putting the parts together: combining in vitro methods to test for skin sensitizing potentials. *Regul Toxicol Pharmacol*. 63:489–504.
- Bell S, Phillips J, Sedykh A, Tandon A, Sprankle C, Morefield S, Shapiro A, Allen D, Shah R, Maull E, et al. 2017. An integrated chemical environment to support 21st century toxicology. *Environ Health Perspect*. 125:054401.
- Birnbaum LS. 2013. 15 years out: reinventing ICCVAM. *Environ Health Perspect*. 21:a40.
- Bruckner AL, Weston WL, Morelli JG. 2000. Does sensitization to contact allergens begin in infancy?. *Pediatrics*. 105:e3.
- Casati S, Aschberger K, Barroso J, Casey W, Delgado I, Kim TS, Kleinstreuer N, Kojima H, Lee JK, Lowit A, et al. 2017. Standardisation of defined approaches for skin sensitisation testing to support regulatory use and international adoption: position of the International Cooperation on Alternative Test Methods. *Arch Toxicol*. <https://doi.org/10.1007/s00204-017-2097-4>
- Daniel A, Strickland J, Allen D, Casati S, Zuang V, Barroso J, Whelan M, Regimbald-Krnel MJ, Kojima H, Nishikawa A, et al. 2018. International regulatory requirements for skin sensitization testing. *Regul Toxicol Pharmacol*.
- Davies M, Pendlington RU, Page L, Roper CS, Sanders DJ, Bourner C, Pease CK, MacKay C. 2011. Determining epidermal disposition kinetics for use in an integrated nonanimal approach to skin sensitization risk assessment. *Toxicol Sci*. 119:308–318.
- Del Bufalo A, Pauloin T, Alépée N, Clouzeau J, Detroyer A, Eilstein J, Gomes C, Nocairi H, Piroird C, Rousset F, et al. 2018. Alternative integrated testing for skin sensitisation: assuring consumer safety. *Appl In Vitro Toxicol*. <https://doi.org/10.1089/aivt.2017.0023>
- Dumont C, Barroso J, Matys I, Worth A, 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, Natsch A. 2010. Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. *Toxicol Appl Pharmacol*. 245:281–290.
- EU. 2016. Commission regulation (EU) 2016/1688 of 20 September 2016 amending Annex VII to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards skin sensitisation. *Off J Eur Union*. 255:1–3.
- European Union. 2003. Directive 2003/15/EC of the European Parliament and of the Council of 27 February 2003 amending Council Directive 76/768/EEC on the approximation of the laws of the Member States relating to cosmetic products. *Off J Eur Union*. 66:26–35.
- Garcia-Gavin J, Lissens R, Timmermans A, Goossens A. 2011. Allergic contact dermatitis caused by isopropyl alcohol: a missed allergen? *Contact Dermat*. 65:101–106.
- Gerberick GF, Ryan CA, Kimber I, Dearman RJ, Lea LJ, Basketter DA. 2000. Local lymph node assay: validation assessment for regulatory purposes. *Am J Contact Dermat*. 11:3–18.
- Gerberick GF, Vassallo JD, Bailey RE, Chaney JG, Morrall SW, Lepoittevin JP. 2004. Development of a peptide reactivity assay for screening contact allergens. *Toxicol Sci*. 81:332–343.
- Gerberick GF, Vassallo JD, Foertsch LM, Price BB, Chaney JG, Lepoittevin JP. 2007. Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. *Toxicol Sci*. 97:417–427.
- Gomes C, Nocairi H, Thomas M, Collin J, Saporta G. 2014. A simple and robust scoring technique for binary classification. *AIR*. 3:52–58.
- Hirota M, Fukui S, Okamoto K, Kurotani S, Imai N, Fujishiro M, Kyotani D, Kato Y, Kasahara T, Fujita M, et al. 2015. Evaluation of combinations of in vitro sensitization test descriptors for the artificial neural network-based risk assessment model of skin sensitization. *J Appl Toxicol*. 35:1333–1347.
- Hoffmann S. 2015. LLNA variability: an essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. *Altex*. 32:379–383.
- Hoffmann S, Kleinstreuer N, Alepee N, Allen D, Api AM, Ashikaga T, Clouet E, Cluzel M, Desprez B, Gellatly N, et al. 2018. Non-animal methods to predict skin sensitisation (I): the Cosmetics Europe database. *Crit Rev Toxicol*. <https://doi.org/10.1080/10408444.2018.1429385>
- Jaworska J, Dancik Y, Kern P, Gerberick F, Natsch A. 2013. Bayesian integrated testing strategy to assess skin sensitization potency: from theory to practice. *J Appl Toxicol*. 33:1353–1364.
- Jaworska J, Harol A, Kern PS, Frank Gerberick G. 2011. Integrating non-animal test information into an adaptive testing strategy – skin sensitization proof of concept case. *Altex*. 28:211–225.
- Jaworska JS, Natsch A, Ryan C, Strickland J, Ashikaga T, 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 DA, Gerberick GF, Ryan CA, Dearman RJ. 2011. Chemical allergy: translating biology into hazard characterization. *Toxicol Sci*. 120:S238–S268.
- Loveless SE, Api AM, Crevel RW, Debruyne E, Gamer A, Jowsey IR, Kern P, Kimber I, Lea L, Lloyd P, et al. 2010. Potency values from the local lymph node assay: application to classification, labelling and risk assessment. *Regul Toxicol Pharmacol*. 56:54–66.
- Luebke R. 2012. Immunotoxicant screening and prioritization in the twenty-first century. *Toxicol Pathol*. 40:294–299.
- MacKay C, Davies M, Summerfield V, Maxwell G. 2013. From pathways to people: applying the adverse outcome pathway (AOP) for skin sensitization to risk assessment. *Altex*. 30:473–486.
- Macmillan DS, Canipa SJ, Chilton ML, Williams RV, Barber CG. 2016. Predicting skin sensitisation using a decision tree integrated testing strategy with an in silico model and in chemico/in vitro assays. *Regul Toxicol Pharmacol*. 76:30–38.
- Maxwell G, MacKay C, Cubberley R, Davies M, Gellatly N, Glavin S, Gouin T, Jacquilleot S, Moore C, Pendlington R, et al. 2014. Applying the skin sensitisation adverse outcome pathway (AOP) to quantitative risk assessment. *Toxicol In Vitro*. 28:8–12.
- Mehling A, Eriksson T, Eltze T, Kolle S, Ramirez T, Teubner W, van Ravenzwaay B, Landsiedel R. 2012. Non-animal test methods for predicting skin sensitization potentials. *Arch Toxicol*. 86:1273–1295.
- Murphy K, Travers P, Walport M, Janeway C. 2012. *Janeway's immunobiology*. 8th ed. New York: Garland Science.
- Natsch A. 2014. Integrated approaches to safety testing: general principles and skin sensitization as a test case. In: Allen D, Waters MD, editors. *Reducing, refining and replacing the use of animals in toxicity testing*. London: Royal Society of Chemistry. p. 364–288.
- Natsch A, Emter R. 2008. Skin sensitizers induce antioxidant response element dependent genes: application to the in vitro testing of the sensitization potential of chemicals. *Toxicol Sci*. 102:110–119.
- Natsch A, Emter R, Gfeller H, Haupt T, 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.
- Nukada Y, Ashikaga T, Miyazawa M, Hirota M, Sakaguchi H, Sasa H, Nishiyama N. 2012. Prediction of skin sensitization potency of chemicals by human Cell Line Activation Test (h-CLAT) and an attempt at classifying skin sensitization potency. *Toxicol In Vitro*. 26:1150–1160.
- Nukada Y, Miyazawa M, Kazutoshi S, Sakaguchi H, 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. OECD Series on Testing and Assessment No. 168. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Part 1: Scientific Assessment. Paris: OECD Publishing.
- OECD. 2014a. The OECD QSAR Toolbox. Paris: OECD Publishing. <http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm>.

- OECD. 2014b. Test No. 428. Skin absorption: in vitro method. In: OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. Paris: OECD Publishing.
- OECD. 2015a. Test No. 442C. *In Chemico* skin sensitization: direct peptide reactivity assay (DPRA). In: OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. Paris: OECD Publishing.
- OECD. 2015b. Test No. 442D. *In Vitro* Skin Sensitisation: ARE-Nrf2 Luciferase Test Method. In: OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. Paris: OECD Publishing.
- OECD. 2016a. Guidance document on the reporting of defined approaches to be used within integrated approaches to testing and assessment. Series on Testing and Assessment No. 255. Paris: OECD Publishing.
- OECD. 2016b. Guidance Document on the Reporting of Defined Approaches to be used within Integrated Approaches to Testing and Assessment. Series on Testing and Assessment No. 256. Paris: OECD Publishing.
- OECD. 2017. Test No. 442E. *In vitro* skin sensitisation: assays addressing the key event on activation of dendritic cells on the adverse outcome pathway for skin sensitisation. In: OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. Paris: OECD Publishing.
- Patlewicz G, Casati S, Basketter DA, Asturiol D, Roberts DW, Lepoittevin J-P, Worth AP, Aschberger K. 2016. Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? *Regul Toxicol Pharmacol.* 82:147–155.
- Patlewicz G, Dimitrov SD, Low LK, Kern PS, Dimitrova GD, Comber MI, Aptula AO, Phillips RD, Niemela J, Madsen C, 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.
- Patlewicz G, Kuseva C, Kesova A, Popova I, Zhechev T, Pavlov T, Roberts DW, Mekenyan O. 2014. Towards AOP application-implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization. *Regul Toxicol Pharmacol.* 69:529–545.
- Piroird C, Ovigne JM, Rousset F, Martinozzi-Teissier S, Gomes C, Cotovio J, Alepee 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.
- Ramirez T, Mehling A, Kolle SN, Wruck CJ, Teubner W, Eltze T, Aumann A, Urbisch D, van Ravenzwaay B, Landsiedel R. 2014. LuSens: a keratinocyte based ARE reporter gene assay for use in integrated testing strategies for skin sensitization hazard identification. *Toxicol In Vitro.* 28:1482–1497.
- Reisinger K, Hoffmann S, Alepee N, Ashikaga T, Barroso J, Elcombe C, Gellatly N, Galbiati V, Gibbs S, Groux H, et al. 2015. Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. *Toxicol In Vitro.* 29:259–270.
- Roberts DW, Api AM, Aptula AO. 2016. Chemical applicability domain of the Local Lymph Node Assay (LLNA) for skin sensitisation potency. Part 2. The biological variability of the murine Local Lymph Node Assay (LLNA) for skin sensitisation. *Regul Toxicol Pharmacol.* 80:255–259.
- Rovida C, Alepee N, Api AM, Basketter DA, Bois FY, Caloni F, Corsini E, Daneshian M, Eskes C, Ezendam J, et al. 2015. Integrated Testing Strategies (ITS) for safety assessment. *Altex.* 32:25–40.
- Sauer UG, Hill EH, Curren RD, Raabe HA, Kolle SN, Teubner W, Mehling A, Landsiedel R. 2016. Local tolerance testing under REACH: accepted non-animal methods are not on equal footing with animal tests. *Altern Lab Anim.* 44:281–299.
- Steiling W. 2016. Safety evaluation of cosmetic ingredients regarding their skin sensitization potential. *Cosmetics.* 3:14.
- Strickland J, Zang Q, Kleinstreuer N, Paris M, Lehmann DM, Choksi N, Matheson J, Jacobs A, Lowit A, Allen D, et al. 2016. Integrated decision strategies for skin sensitization hazard. *J Appl Toxicol.* 36:1150–1162.
- Strickland J, Zang Q, Paris M, Lehmann DM, Kleinstreuer N, Allen D, Choksi N, Matheson J, Jacobs A, Casey W. 2017. Multivariate models for prediction of human skin sensitization hazard. *J Appl Toxicol.* 37:347–360.
- Takenouchi O, Fukui S, Okamoto K, Kurotani S, Imai N, Fujishiro M, Kyotani D, Kato Y, Kasahara T, Fujita M, et al. 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.
- Thyssen JP, Linneberg A, Menne T, Johansen JD. 2007. The epidemiology of contact allergy in the general population—prevalence and main findings. *Contact Dermat.* 57:287–299.
- Urbisch D, Mehling A, Guth K, Ramirez T, Honarvar N, Kolle S, Landsiedel R, Jaworska J, Kern PS, Gerberick F, et al. 2015. Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul Toxicol Pharmacol.* 71:337–351.
- van der Veen JW, Rorije E, Emter R, Natsch A, van Loveren H, Ezendam J. 2014. Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals. *Regul Toxicol Pharmacol.* 69:371–379.
- Williams WC, Copeland C, Boykin E, Quell SJ, Lehmann DM. 2015. Development and utilization of an ex vivo bromodeoxyuridine local lymph node assay protocol for assessing potential chemical sensitizers. *J Appl Toxicol.* 35:29–40.
- Zang Q, Mansouri K, Williams AJ, Judson RS, Allen DG, Casey WM, Kleinstreuer NC. 2017a. In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *J Chem Inf Model.* 57:36–49.
- Zang Q, Paris M, Lehmann DM, Bell S, Kleinstreuer N, Allen D, Matheson J, Jacobs A, Casey W, Strickland J. 2017b. Prediction of skin sensitization potency using machine learning approaches. *J Appl Toxicol.* 37:792–805.
- Zhang H, Shi Y, Wang C, Zhao K, Zhang S, Wei L, Dong L, Gu W, Xu Y, Ruan H, et al. 2017. An improvement of LLNA:DA to assess the skin sensitization potential of chemicals. *J Toxicol Sci.* 42:129–136.

