



A defined approach for predicting skin sensitisation hazard and potency based on the guided integration of *in silico*, *in chemico* and *in vitro* data using exclusion criteria

Donna S. Macmillan*, Martyn L. Chilton

Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, UK

ARTICLE INFO

Keywords:

Skin sensitisation
Defined approach
Integrated testing strategy
Negative predictions
Expert system
(Q)SAR
In silico toxicology
Hazard
Risk
Potency

ABSTRACT

A decision tree-based defined approach (DA) has been designed using exclusion criteria based on applicability domain knowledge of *in chemico*/*in vitro* information sources covering key events 1–3 in the skin sensitisation adverse outcome pathway and an *in silico* tool predicting the adverse outcome (Derek Nexus). The hypothesis is that using exclusion criteria to de-prioritise less applicable assays and/or *in silico* outcomes produces a rational, transparent, and reliable DA for the prediction of skin sensitisation potential. Five exclusion criteria have been established: Derek Nexus reasoning level, Derek Nexus negative prediction, metabolism, lipophilicity, and lysine-reactivity. These are used to prioritise the most suitable information sources for a given chemical and results from which are used in a '2 out of 3' approach to provide a prediction of hazard. A potency category (and corresponding GHS classification) is then assigned using a k-Nearest Neighbours model containing human and LLNA data. The DA correctly identified the hazard (sensitiser/non-sensitiser) for 85% and 86% of a dataset with reference LLNA and human data. The correct potency category was identified for 59% and 68% of chemicals, and the GHS classification accurately predicted for 73% and 76% with reference LLNA and human data, respectively.

1. Introduction

Allergic contact dermatitis is a common skin reaction which may occur after exposure to a sensitising chemical. The skin sensitisation potential of such chemicals has traditionally been assessed using *in vivo* animal tests such as the murine local lymph node assay (LLNA) (OECD, 2010), the guinea pig maximisation test (GPMT) (OECD, 1992), and the Buehler test (OECD, 1992). However, there has been a considerable shift towards the use of non-animal alternatives due to animal welfare concerns and regulatory requirements which aim to restrict or even prohibit the use of animals to predict this endpoint (Environmental Protection Agency, 2018; European Union, 2009, 2006). Skin sensitisation has a well-understood adverse outcome pathway (AOP) consisting of four key events (KE) (OECD, 2012) and a number of *in chemico* and *in vitro* assays have been developed to measure individual KE. Several of these assays have been validated and adopted by the OECD with the publication of corresponding Test Guidelines. The direct peptide reactivity assay (DPRA) measures covalent binding of electrophilic substances to nucleophilic centres in skin proteins using synthetic model peptides containing either lysine or cysteine (OECD, 2015). This

equates to key event 1 (KE1), which is also the molecular initiating event (MIE) for the AOP. The KeratinoSens™ and LuSens assays measure key event 2 (KE2), the activation of keratinocytes and inflammatory responses. These assays use luciferase activity to determine the Nrf2-mediated activation of antioxidant response element (ARE)-dependent genes in a human keratinocyte-derived cell line with a luciferase reporter gene (OECD, 2018a). The human cell line activation test (h-CLAT) and the U-SENS™ assay measure key event 3 (KE3), dendritic cell activation by measuring changes in the expression of cell surface markers (CD86 and/or CD54) associated with the process in the human leukaemia cell line THP-1 (h-CLAT) or the human lymphoma cell line U937 (U-SENS™). The IL-8 Luc assay also measures KE3 but does this by quantifying changes in the expression of IL-8, a cytokine thought to be associated with dendritic cell activation (OECD, 2018b).

The current regulatory paradigm is that a single non-animal assay is unlikely to be a sufficient stand-alone replacement for the current 'gold standard' *in vivo* assays, the LLNA and the GPMT. Instead, the focus has been on combining multiple information sources (e.g. physicochemical properties, read-across, *in silico* expert/(Q)SAR predictions, *in chemico* and *in vitro* tests, historical *in vivo* data) to predict skin sensitisation

* Corresponding author.

E-mail address: donna.macmillan@lhasalimited.org (D.S. Macmillan).

<https://doi.org/10.1016/j.yrtph.2018.11.001>

Received 21 August 2018; Received in revised form 2 November 2018; Accepted 6 November 2018

Available online 12 November 2018

0273-2300/ © 2018 Elsevier Inc. All rights reserved.

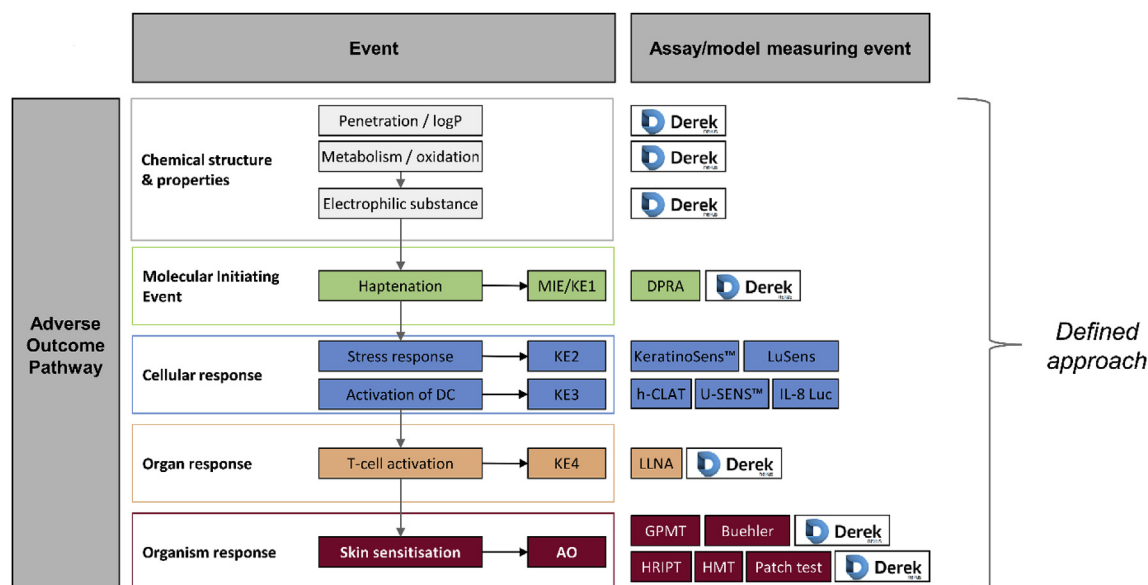


Fig. 1. The adverse outcome pathway (AOP) for skin sensitisation initiated by covalent binding to proteins, reproduced from (OECD, 2012). Key events (KE) in the AOP and corresponding assays/models which measure these are shown. Information from these assays/models can be combined in a defined approach to predict skin sensitisation without the use of animal tests.

potential in what is known as a defined approach (DA) (OECD, 2016a) (Fig. 1). Many diverse DA have been published and are reported to predict skin sensitisation potential with an accuracy comparable to that of the traditional *in vivo* tests (OECD, 2016b).

The DA described herein builds on previous work (Macmillan et al., 2016) which used a Derek Nexus (Derek) prediction in combination with up to two additional *in chemico/in vitro* assays, and a global applicability domain (excluding pre/prohaptens and/or chemicals with a $\log P > 3.5$) to predict skin sensitisation hazard.

Derek, an expert knowledge-based system for toxicity predictions, is considered to cover multiple KE in the skin sensitisation AOP:

- Pre-MIE & MIE/KE1 - Derek can fire an alert for directly acting (haptens) and indirectly acting sensitisers (prehaptens and prohaptens). Derek can also provide a prediction for a chemical regardless of $\log P$.
- Key event 4 and the adverse outcome - Derek skin sensitisation alerts are created from public and proprietary animal and human toxicity data. The animal data consists primarily of mouse assays (e.g. LLNA) and guinea pig assays (e.g. GPMT and Buehler test), which measure KE4 and the adverse outcome, respectively. Data from human assays, which also assess the adverse outcome, are also used in Derek alerts.

Accordingly, the hypothesis is that Derek can be a valuable addition to a defined approach and complement *in chemico* and *in vitro* data. In the DA, five specific exclusion criteria have now been established, based on the applicability domain of Derek predictions and the *in chemico/in vitro* assays used, and are used to prioritise the information sources in Fig. 1. Results from the prioritised information sources are then used in a decision tree to predict the hazard (sensitiser/non-sensitiser) and subsequently the potency (potency category/GHS classification) of a given chemical.

2. Material and methods

2.1. Derek Nexus

Derek Nexus 6.0.1 (using Knowledge Base 2018 1.1), an expert knowledge-based system for toxicity predictions containing structural

alerts derived from both public and proprietary data, was used to provide information and generate predictions for each chemical:

- (1) **Derek outcome:** If a skin sensitisation alert was activated with a reasoning level of certain, probable, or plausible, the chemical was considered to be a sensitiser. Chemicals firing an alert with a reasoning level of doubted, improbable, or impossible and chemicals with a negative prediction of non-sensitiser with no misclassified or unclassified features were considered as non-sensitisers. Where a chemical activated an alert with a reasoning level of equivocal, or gave a negative prediction of non-sensitiser with misclassified features, or a non-sensitiser with unclassified features, the *in silico* outcome was not considered further in the DA.
- (2) **Metabolic activation:** Alerting chemicals were assigned as exclusively prohaptens when the corresponding potential mechanism summary in the Derek Nexus alert description contained only a prohaptent mechanism for skin sensitisation (Fig. 2). Where multiple mechanisms were possible (prehaptens or prohaptens) or where the mechanism was undefined (e.g. pre/prohaptent) no assignment was made as transformation to the hapten may be achieved through the prehaptent mechanism.
- (3) **LogP:** The $\log P$ was used as calculated from Derek Nexus using the ClogP method from BioByte Corp., version 5.9.
- (4) **Lysine-reactivity:** Alerting chemicals were deemed to be lysine-reactive when the potential mechanism summary in Derek Nexus contained a mechanism in the acyl transfer domain (i.e. acylating, thioacylating, carbamylating, thiocarbamylating, and sulphonylating agents (Aptula and Roberts, 2006)) (Fig. 3).

2.2. Potency prediction model used in the DA

2.2.1. Human data

The human data used in the potency prediction model was taken

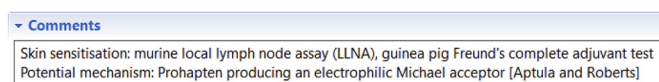


Fig. 2. Screenshot of Derek alert comments showing the class of sensitiser (hapten/prehaptent/prohaptent).

▼ Comments

Skin sensitisation: murine local lymph node assay (LLNA), guinea pig maximisation test (GPMT)
Potential mechanism: Hapten acting as an electrophilic acylating agent [Aptula and Roberts]

Fig. 3. Screenshot of Derek alert comments showing the mechanism believed to induce skin sensitisation.

from Basketter et al., (2014) ($n = 131$) and Api et al., (2017) ($n = 89$). The datasets were combined, and chemicals known to be undefined mixtures and/or polymers (Basketter et al., 2014 - peru balsam, petrolatum, dextran, tween 80; Api et al., 2017 - tea leaf absolute, oak-moss, jasmine absolute (jasminum grandiflorum), treemoss, jasmine absolute (jasminum sambac), ylang-ylang) were removed. The dataset was standardised (e.g. by ensuring consistent aromaticity perception and removing any tetrahedral stereochemistry) and curated (e.g. by removing counterions from salts) before being grouped by InChi string to give 203 unique chemicals with human potency categories.

2.2.2. LLNA data

The LLNA dataset was taken from the Derek EC3 Model - 1.2.0 in Derek 6.0.1 and consisted of 1069 publicly available LLNA studies corresponding to 674 unique chemicals (Canipa et al., 2017). The EC3 values were then mapped to the equivalent human potency categories (Table 1) based on previously suggested EC3 ranges (Basketter, 2016) to give LLNA-derived potency categories.

2.2.3. Combining human and LLNA data

The curated human data ($n = 203$) were added to the LLNA data ($n = 674$) to give a dataset of 877 chemicals. These were grouped by InChi string to ensure that each structure was represented only once in the dataset ($n = 761$). Where a human potency category and an LLNA-derived potency category was available for a single chemical, the human potency category was given precedence. Next, chemicals activating more than one skin sensitisation alert ($n = 48$) were removed (to prevent uncertainty arising from competing toxicophores with differing potencies) and non-alerting chemicals ($n = 205$) were also removed (as the model only predicts a potency category for chemicals activating a Derek alert). Finally, two multicomponent chemicals were removed (Fig. 4).

Overall, the dataset used for the potency prediction model in the DA consisted of 506 unique chemicals, 35 with human potency category data, 390 with LLNA-derived potency category data, and 81 with data for both species. 55 chemicals were assigned as potency category 1, 103 as potency category 2, 185 chemicals as potency category 3, 76 chemicals as potency category 4, 84 chemicals as potency category 5, and 3 chemicals as potency category 6.

2.2.4. Methodology of potency prediction model

The model uses a k-Nearest Neighbours approach as described in Canipa et al. (2017) but was generalised to deal with discrete potency categories of 1–6 rather than continuous EC3 values. For chemicals that fired a Derek skin sensitisation alert, Nearest Neighbours in the same alert domain (i.e. firing the same Derek skin sensitisation alert) were selected as they are believed to act through the same mechanism.

Table 1

Mapping of human potency categories to GHS classifications and EC3 values.

Human potency category	Human potency category name	GHS classification	Equivalent EC3 value (%)
1	Extreme	1A	< 0.2
2	Strong		< 2
3	Moderate	1B	2–20
4	Weak		> 20
5	Very weak	No category	> 80
6	Non-sensitiser		negative

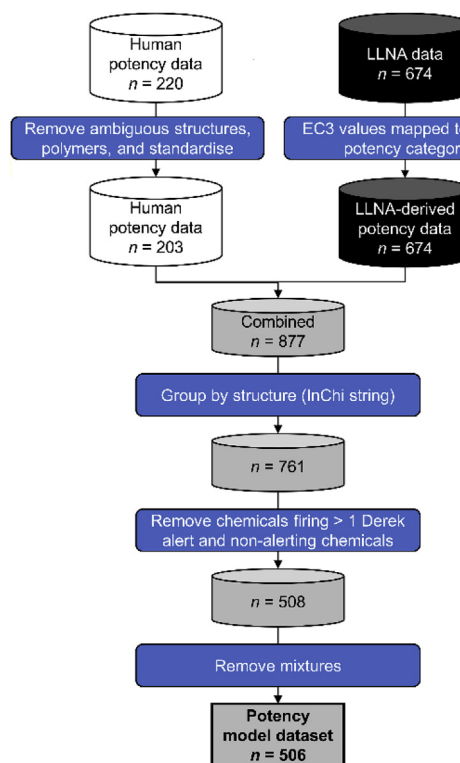


Fig. 4. Schematic showing combination of human data with mouse data to produce the combined dataset used in the potency prediction model of the DA.

When an alerting chemical was predicted to be a sensitiser by the DA, the potency categories of the most similar Nearest Neighbours in the potency prediction dataset (as judged by the Tanimoto index) within the same alert domain were used to predict the potency of the chemical used in the DA. A minimum of 3 and maximum of 10 of the most similar Nearest Neighbours were used to make a prediction (Fig. 5). A leave-one-out approach was used for all potency predictions to minimise any inflation of the predictivity from chemicals in the dataset used to evaluate the DA also being present in the potency prediction dataset.

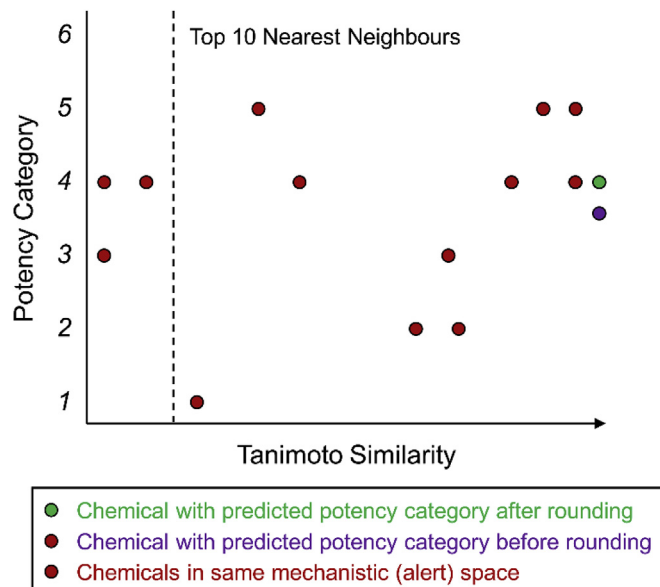


Fig. 5. Example of 10 Nearest Neighbours used in a typical potency prediction in the DA.

The harmonic mean of the potency categories of these Nearest Neighbours was then calculated, weighted by their similarity to the query chemical to provide a predicted potency category which is likely to be expressed as a decimal. To maximise protection of human health, this was rounded down to the closest potency category and then corresponding GHS classification (Equation (1)).

$$\text{Potency category}_{\text{predicted}} = \frac{\sum (\text{Tanimoto} * \text{Potency category}_{\text{NN}})}{\sum \text{Tanimoto}} \quad (1)$$

Equation (1). Potency prediction equation using the weighted average of Nearest Neighbours as calculated by Tanimoto scoring. NN = Nearest Neighbour.

2.3. Evaluation dataset

The DA was evaluated using a dataset of 213 previously published chemicals with human, LLNA, DPRA, KeratinoSens™, LuSens, and h-CLAT data (Urbisch et al., 2015). The corresponding U-SENS data ($n = 126$) was added, where available (Alépée et al., 2017; Bufalo et al., 2018; Hoffmann et al., 2018).

The dataset was standardised (e.g. by ensuring consistent aromaticity perception and removing any tetrahedral stereochemistry) and curated (e.g. by removing counterions from salts). The resulting chemical structures were then grouped by their InChI string to ensure that each structure was represented only once in the dataset. The DPRA data was used to assess KE1 (the MIE), and assays designed to measure the same KE (KeratinoSens™ & LuSens - KE2; h-CLAT & U-SENS™ - KE3) were combined into an overall call for each KE (Fig. 6).

2.3.1. Human

Urbisch and co-workers (Urbisch et al., 2015) used both HMT/HRIP assay data and expert-derived human potency categories to derive an overall human call of sensitiser or non-sensitiser for 111 chemicals. One chemical, glutaraldehyde was noted to contain a typographical error in the human potency category which was corrected - from category 1 to the correct category of 2, as reported by the original authors (Basketter et al., 2014). Ethyl acrylate was reported as a human potency category 4 chemical by Urbisch et al. (2015) based on the data provided in Basketter et al., (2014). This was later amended to category 3 in Api et al. (2017), and the result updated accordingly in the evaluation dataset. The human call and the human potency category were removed for 3 chemicals reported as borderline (amyl cinnamic aldehyde, hexyl cinnamic aldehyde, 6-methyl coumarin) and for 3 chemicals with ambiguous structures (benzalkonium chloride, Dextran, and Tween 80). Finally, where the binary human call assigned by Urbisch and co-workers and the human potency category conflicted (linalool, hexyl salicylate, pentachlorophenol, benzaldehyde), the binary human

call was taken (as this also considered additional human data). As a result, the human potency data were removed for the chemicals with a positive human call, and re-assigned as human potency category 5/6 for the two chemicals with a negative human call (Table 2). 108 chemicals with human sensitisation data remained.

2.3.2. LLNA

The LLNA data was obtained from Urbisch et al. (2015). Data for 210 chemicals were available after the removal of data for 3 chemicals with ambiguous structures (benzalkonium chloride, Dextran, and Tween 80). Where more than one EC3 value was reported the median value was calculated. Non-integer EC3 values (e.g. > 25) were removed from the dataset. The EC3 values were used to map each chemical to an LLNA-derived potency category and corresponding GHS classification (Table 1). Where the LLNA call conflicted with the LLNA-derived potency category, the LLNA category was amended (Table 3).

2.3.3. Direct peptide reactivity assay (DPRA)

The DPRA data was obtained from Urbisch et al. (2015). Data for 192 chemicals was available after the removal of data for 5 chemicals with borderline results (ethylene diamine, imidazolidinyl urea, salicylic acid, sodium lauryl sulfate, vanillin) and two chemicals with ambiguous structures (benzalkonium chloride and Tween 80).

2.3.4. KeratinoSens™

The KeratinoSens™ data was obtained from Urbisch et al. (2015). Data for 184 chemicals was available after the removal of data for 3 chemicals with borderline results (amyl cinnamic aldehyde, hexyl cinnamic aldehyde, zinc sulfate), and 8 chemicals with discordant data (4-allylanisole, aniline, eugenol, methyl methacrylate, n-hexane, nickel chloride, tartaric acid, vanillin), and 3 chemicals with ambiguous structures (benzalkonium chloride, Dextran, and Tween 80).

2.3.5. LuSens

The LuSens data was obtained from Urbisch et al. (2015). Data for 77 chemicals was available after the removal of data for 1 chemical with an inconclusive result (dioctyl carbonate) and 1 chemical with an ambiguous structure (Tween 80).

2.3.6. h-CLAT

The h-CLAT data was obtained from Urbisch et al. (2015). Data was available for 163 chemicals after the removal of data for 3 chemicals with ambiguous structures (benzalkonium chloride, Dextran, and Tween 80).

2.3.7. U-SENS™

The U-SENS™ data was obtained from Alépée et al., 2017; Bufalo et al. (2018); and Hoffmann et al., 2018. Data was available for 123 chemicals after the removal of data for 3 chemicals with ambiguous structures (benzalkonium chloride, Dextran, and Tween 80).

2.3.8. Grouping of *in vivo*/*in chemico*/*in vitro* data into AOP key events

Data from *in chemico* and *in vitro* assays designed to measure the same KE in the skin sensitisation AOP were combined. Discordant results were removed from the dataset (e.g. positive in h-CLAT, negative in U-SENS™) and not considered in further analyses.

2.3.9. Distribution and overlap of dataset

The final composition of the data set is described in Table 4. Data for assays measuring the same KE (KE2: KeratinoSens™ & LuSens; KE3: h-CLAT & U-SENS™) were combined and discordant data removed (KE2: 5 chemicals; KE3: 28 chemicals). LLNA data was available for all chemicals whereas human data was only available for 100, 103 and 90 chemicals with KE1, KE2, and KE3 data, respectively. The majority of chemicals (60%) had data covering all of the first 3 KE; 35% and 5% of chemicals had data for only 2 or 1 KE, respectively (Fig. 7). One

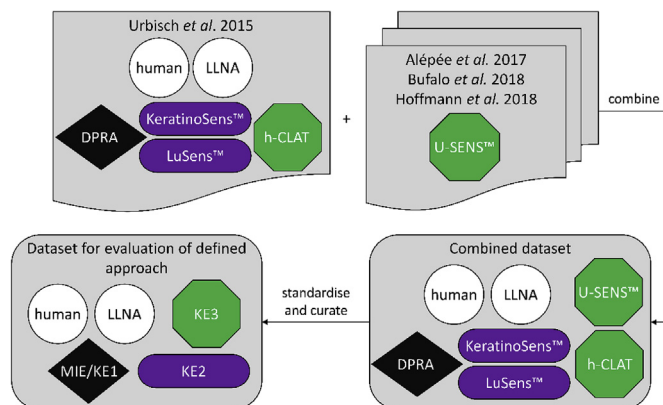


Fig. 6. Combination and curation of *in vivo*, *in chemico*, and *in vitro* data used to evaluate the DA.

Table 2

Amendments made to published human data when curating the evaluation dataset. For the human call 0 = negative, 1 = positive.

CAS	Name	Human call	Human potency category	Comment
122-40-7	Amyl cinnamic aldehyde	borderline	4	Human call and human potency category removed
101-86-0	Hexyl cinnamic aldehyde	borderline	5	Human call and human potency category removed
92-48-8	6-Methylcoumarin	borderline	No data	Human call removed
78-70-6	Linalool	0	4	Human potency category re-assigned as 5/6
6259-76-3	Hexyl salicylate	0	4	Human potency category re-assigned as 5/6
87-86-5	Pentachlorophenol	1	5	Human potency category removed
100-52-7	Benzaldehyde	1	5	Human potency category removed

Table 3

Amendments made to published LLNA data when curating the evaluation dataset. For the LLNA call 0 = negative, 1 = positive.

CAS	Name	LLNA call	EC3 (%)	LLNA-derived potency category	Comment
80-62-6	Methylmethacrylate	1	90.0	5	LLNA call amended to 0
62-53-3	Aniline	1	89.0	5	LLNA call amended to 0
1330-20-7	Xylene	1	95.8	5	LLNA call amended to 0

chemical, 12-bromo-1-dodecanol, had no data for KE1 or KE2 and discordant data for KE3 and could not be used to evaluate the DA.

2.4. Data evaluation and statistics

To assess the performance of the DA against skin sensitisation hazard predictions (sensitiser/non-sensitiser) the evaluation dataset was first processed in Derek Nexus 6.0.1 against the endpoint of skin sensitisation. Tautomer and mixture perception were unselected. All data analysis was carried out using KNIME 3.5.3 or MS Excel 2016 and Cooper statistics were calculated as follows (Cooper et al., 1979):

- Accuracy (Acc): $([TP + TN]/[TP + FP + TN + FN]) * 100$
- Sensitivity (Se): $(TP/[TP + FN]) * 100$
- Specificity (Sp): $(TN/[TN + FP]) * 100$
- Positive predictivity (PP): $(TP/[TP + FP]) * 100$
- Negative predictivity (NP): $(TN/[TN + FN]) * 100$
- Coverage (Cov): (number of compounds predicted/total number of chemicals in dataset) * 100

The performance of the DA potency prediction was assessed using the following performance metrics:

- % Correctly predicted: when the predicted and experimental classification were identical.
- % Over-predicted: when the predicted potency classification was more potent (i.e. a stronger sensitiser) than the experimental classification.
- % Under-predicted: when the predicted potency classification was less potent (i.e. a weaker sensitiser) than the experimental classification.

Table 4Dataset used to evaluate the DA consisting of *in vivo* data and *in chemico/in vitro* data for assays measuring KE1–4 and the AO in the skin sensitisation AOP.

Key Event (KE)	Assay	Positive	Negative	Assay total	Discordant data upon KE combination	KE total
1/MIE	DPRA	120	72	192	n/a	192
2	KeratinoSens™	122	62	184	5	195
	LuSens	45	32	77		
3	h-CLAT	113	50	163	28	147
	U-SENS™	98	25	123		
4	LLNA	151	59	210	n/a	210
AO	Human	75	33	108	n/a	108

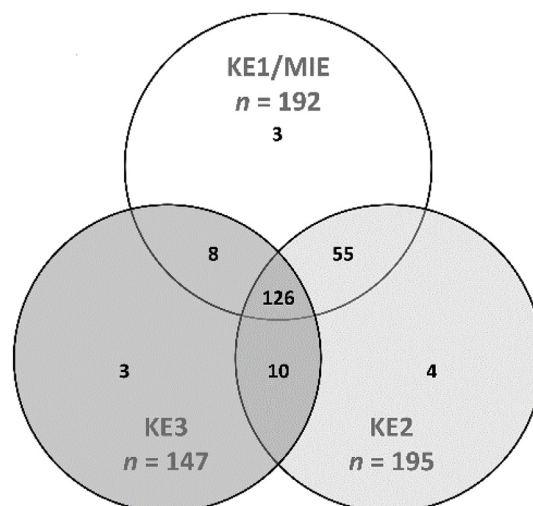


Fig. 7. Overlap of chemicals with *in chemico* and *in vitro* assay data measuring KE1 (MIE), KE2, and KE3 after removal of discordant data.

3. Results and discussion

3.1. Interchangeabilities

3.1.1. Between assays covering the same key event

The KeratinoSens™ and LuSens assays address the second KE in the skin sensitisation AOP, activation of keratinocytes, and are described in detail in Key Event Based Test Guideline 442D (OECD, 2018a). The h-CLAT and U-SENS™ assays measure the third KE, activation of dendritic cells, and are described in Key Event Based Test Guidelines 442E (OECD, 2018b). Both Test Guidelines state that results from the assays within can be used interchangeably to fulfil an information requirement

for each KE, even though slightly different protocols and readouts are carried out and obtained from each. Consequently, it was decided to combine the data from assays designed to measure the same KE to increase the size of the respective data pools and to simplify the DA.

Prior to grouping the assay data by KE for the DA, the interchangeability of the assays was evaluated. The KeratinoSens™ and LuSens assays have been reported as functionally equal (Ramirez et al., 2016) and a comparison of the data used in the evaluation dataset has already been reported (Urbisch et al., 2015). However, as the dataset described in the present paper has had some additional curation the equivalence between the assays was re-assessed. Although there is over twice the amount of data for KeratinoSens™ than for LuSens the assays show nearly identical performance, correctly predicting the skin sensitisation potential for 74% (136/184) and 75% (58/77) of chemicals with corresponding LLNA data, and 83% (82/99) and 82% (49/60) with corresponding human data in the evaluation dataset. A subset of 61 chemicals were tested in both the KeratinoSens™ and LuSens assays, with a concordance of 92% (56/61 with the same outcome). Of the 5 chemicals with discordant data, 5/5 were correctly predicted by the KeratinoSens™ and 0/5 by the LuSens assay when compared against LLNA data. Only four of the chemicals with discordant KE2 data had corresponding human data, where the KeratinoSens™ assay was more often correct (3/4). The KeratinoSens™ and LuSens data were combined, and after removal of data with discordant results, data from assays measuring KE2 correctly predicted 74% of chemicals with LLNA data and 83% with human data (Fig. 8).

The h-CLAT and U-SENS™ assays correctly predict the skin sensitisation potential for 74% (121/163) and 83% (102/123) of chemicals with corresponding LLNA data, and 81% (77/95) and 82% (75/91) with corresponding human data in the evaluation dataset. The concordance between a subset of 111 chemicals with data for both assays was 75% (83/111 with the same outcome) (Fig. 8). Of the 28 chemicals with discordant data, 36% (10/28) were correctly predicted by the h-CLAT and 64% (18/28) by the U-SENS™ assay when compared against LLNA data. Only thirteen of the chemicals with discordant KE3 data had corresponding human data, and each assay was correct approximately half of the time (h-CLAT (6/13) and U-SENS™ (7/13)). The h-CLAT and U-SENS™ data were combined, and after removal of data with discordant results, data from assays measuring KE3 correctly predicted

82% of chemicals with LLNA data and 88% with human data.

Overall, the intra-KE concordance is reasonable considering the assays have individual protocols - and are reported to have similar limitations and applicability domains. This indicates that a KE-based approach (i.e. where an assay measuring a single KE can be exchanged for another assay measuring the same KE) can be used instead of a specific assay-based approach. Therefore, there should be no restriction in the future for the inclusion other 'me-too' assays within their related KE in the DA, assuming they measure the KE in a similar manner.

3.2. Establishing exclusion criteria

The DA consists of a decision tree, built on previously published work (Macmillan et al., 2016). The original DA used a Derek prediction in combination with up to two additional *in chemico/in vitro* assays, using a global applicability domain (where pre/prohaptens and/or chemicals with a logP > 3.5 were considered to be outside the domain of every *in chemico/in vitro* assay). Now, five specific rule-based exclusion criteria have been hypothesised using the recognised limitations and applicability domain restrictions associated with the individual information sources:

- Derek reasoning level
- Derek negative prediction outcome
- Metabolic activation
- Lipophilicity
- Lysine-reactivity

When a chemical matches an exclusion criterion, an assay or prediction outcome is de-prioritised, and thus not used in the DA.

3.2.1. Derek reasoning level

For alerting chemicals, the reasoning level provided by Derek is generally associated with the proportion of compounds thought to be active within a given alert. This measure of confidence is expressed using the terms certain, probable, plausible, equivocal, doubted, improbable, and impossible. An alert firing with a reasoning level of equivocal has approximately an equal weight of evidence for and against skin sensitisation potential whereas other reasoning levels are

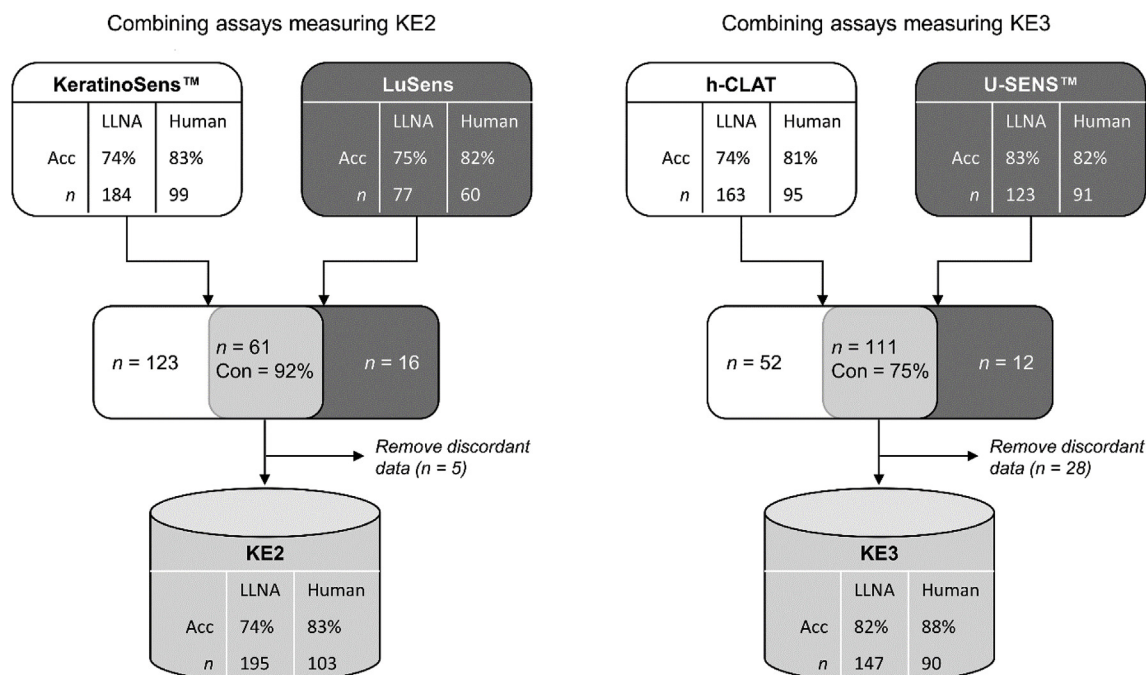


Fig. 8. Combination of assays measuring KE2 and KE3. Acc = accuracy, Con = concordance.

Table 5
Summary of exclusion criteria for the DA.

Exclusion criteria used in DA		Information source used in DA?			
		Derek	MIE/KE1	KE2	KE3
Derek reasoning	Certain, probable, plausible, doubted, improbable, impossible	✓			
	Equivocal	✗			
Derek negative prediction outcome	Non-sensitiser	✓			
	Non-sensitiser containing misclassified feature(s)	✗			
	Non-sensitiser containing unclassified feature(s)	✗			
Metabolic activation	Prohaptens	✓	✗	✓	✓
logP	>3.5	✓	✓	✓	✗
	>5	✓	✓	✗	✗
Lysine-reactivity		✓	✓	✗	✓

associated with more confidence (Judson et al., 2013). As a result of the reduced confidence in a Derek prediction of equivocal, chemicals firing an alert with this reasoning level were de-prioritised and thus not used in the DA.

3.2.2. Derek negative prediction outcome

For non-alerting chemicals, Derek identifies features that are associated with mispredicted sensitiser (misclassified features) and novel chemistry (unclassified features), which can modulate the confidence placed in a negative prediction. It is beyond the scope of this paper to go into the methodology used in detail but these features, although rarely observed, may represent sources of uncertainty for the model (Chilton et al., 2018). Therefore, a negative prediction of non-sensitiser containing misclassified and/or unclassified features was de-prioritised.

3.2.3. Metabolic activation

Some chemicals do not act directly as sensitiser and may require prior activation, known as prehaptens (chemicals requiring non-enzymatic activation, e.g. autoxidation) and prohaptens (chemicals requiring metabolic activation). However, which process occurs may not be known or cannot always be differentiated, which can make it difficult to assess whether *in chemico* and *in vitro* assays adequately predict these chemicals. The *in chemico* DPRA assay is generally thought to provide reliable predictions for prehaptens but less so for prohaptens, as described by the KE1 OECD Test Guideline (OECD, 2015) and reported elsewhere (Casati et al., 2016; Patlewicz et al., 2016; Urbisch et al., 2016), due to a lack of metabolic capacity in the assay. Conversely, cell-based assays measuring KE2 and KE3 have some intrinsic metabolic capacity (Fabian et al., 2013) and would be expected to correctly predict both prehaptens and prohaptens (Casati et al., 2016; OECD, 2018a, 2018b; Patlewicz et al., 2016; Urbisch et al., 2015). Consequently, for chemicals thought to exclusively undergo a prohaptens mechanism exclusively, the KE1 assay was de-prioritised in the DA.

3.2.4. Lipophilicity

Cell-based assays are often thought to be less predictive for lipophilic chemicals due to their aqueous test system. This has been documented in the corresponding OECD Test Guidelines and updated as more data has become available (OECD, 2018a, 2018b). In general, if a chemical is soluble or can form a stable dispersion then testing can be conducted. Nevertheless, in the absence of solubility information, exclusion criteria based on logP were assigned. The present OECD Test Guideline for KE2 state that for the KeratinoSens™ and LuSens assays, chemicals with a logP greater than 7 may be insoluble in the cell media (OECD, 2018a). At the time of DA development, an earlier version of the OECD Test Guideline stated that results for chemicals with a logP between 5 and 7 should also be treated with caution. Therefore, to be conservative, assays measuring KE2 were de-prioritised for chemicals with a logP greater than 5. For KE3, poor predictivity has been documented in the h-CLAT for chemicals with a logP greater than 3.5,

although positive results are usually more robust (OECD, 2018b; Takenouchi et al., 2013). However, the U-SENS™ assay reportedly predicts lipophilic chemicals well (Alépée et al., 2017; Piroird et al., 2015) and no logP restrictions are given for IL8-Luc. Regardless, to be conservative, and minimise false negatives, assays measuring KE3 were de-prioritised for chemicals with a logP greater than 3.5. No logP limit was set for the assay measuring KE1 (DPRA) as current understanding is that *in chemico* assays have fewer problems with solubility (OECD, 2015).

3.2.5. Lysine-reactivity

As the activation of the Nrf2-ARE pathway during KE2 is associated with skin proteins reacting with cysteine residues, chemicals thought to preferentially react with lysine residues such as acyl-transfer agents (Aptula et al., 2005) may not be reliably predicted by the KeratinoSens™ assay (Emter et al., 2013; Natsch, 2017; OECD, 2018a) and the LuSens assay (OECD, 2018a; Ramirez et al., 2014). Consequently, for chemicals with a preference for lysine residues, assays measuring KE2 were de-prioritised.

3.3. Summary of exclusion criteria

The exclusion criteria outlined above for the DA can be summarised together as follows (Table 5):

3.4. Using the DA decision tree

Derek is the first information source to be considered in the DA decision tree. Derek also provides the information required to assess a chemical against the exclusion criteria. After any relevant exclusion criteria is documented, the Derek prediction is considered. If the reasoning level is plausible/probable/certain then the Derek prediction is considered evidence that the chemical is a sensitiser. If the reasoning level is improbable/doubted or a negative prediction outcome of non-sensitiser is given, the Derek prediction is considered evidence that the chemical is a non-sensitiser. If the reasoning level is equivocal or a negative prediction outcome is given which states that misclassified or unclassified features are present, the Derek prediction is not further used in the DA. This step dictates which branch of the decision tree is to be followed. The prioritised assays are then run in order of KE in the AOP (KE1 > KE2 > KE3), unless assays measuring certain KE are precluded by the exclusion criteria assessment in the previous step, until two concordant results are obtained - effectively a '2 out of 3' approach (Fig. 9).

When the chemical is predicted by the DA to be a non-sensitiser, it was assigned as GHS no category. Deciding which potency category to assign was more difficult as both category 5 (very weak sensitiser) and category 6 (non-sensitiser) would be assigned as non-sensitisers if aligning to the GHS classification. Consequently, these were collectively assigned as 5/6. Alerting sensitiser, however, are assessed by the potency prediction model and assigned a potency category (1–4), and corresponding GHS classification (1A/1B).

3.5. Predictivity

3.5.1. Hazard prediction

The DA accurately predicted skin sensitisation hazard when compared against both LLNA data and human data (85% DA vs LLNA; 86% DA vs human) indicating that it can predict both species equally well. Furthermore, the DA is more predictive of human skin sensitisation (86% accuracy - DA vs human) than the LLNA (81% accuracy - LLNA vs human) for the dataset analysed as well as being more predictive than any other assay alone (Table 6). On average, for each chemical in the evaluation dataset, 1.27 assays were necessary to make a prediction of sensitiser/non-sensitiser. For most chemicals, a Derek prediction and a single result from a KE was adequate (76%, 147/194), occasionally an

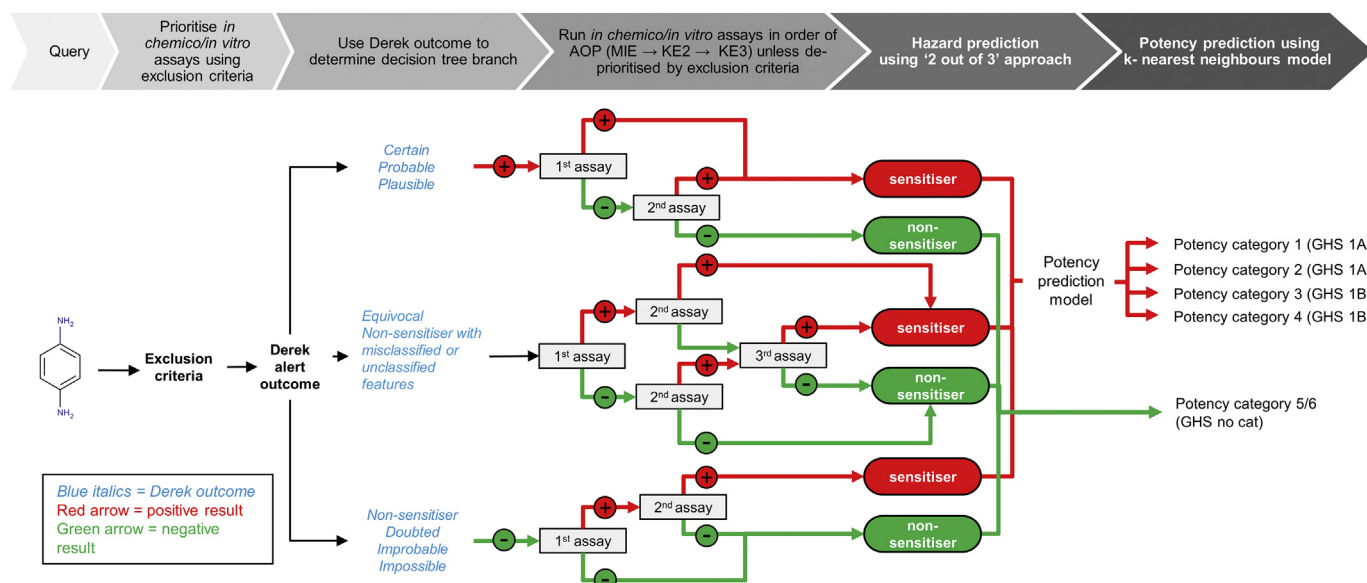


Fig. 9. The DA decision tree uses exclusion criteria to de-prioritise *in chemico/in vitro* assays, then uses the Derek outcome to determine which branch of the tree to follow. Between 1 and 3 *in chemico/in vitro* assays are then run in order of the AOP (MIE → KE2 → KE3) unless de-prioritised in the previous step and the outcome(s) used to assign a hazard classification (sensitiser/non-sensitiser) and predict the potency category (1–4, 5/6) and GHS classification.

additional assay measuring another KE was used (21%, 41/194), and only rarely were assays covering all three KE utilised in the DA (3%, 6/194). The DA is not able to make a prediction in every case as some chemicals lack sufficient KE data to assign a call.

3.5.2. Potency prediction

The DA correctly predicted the potency category (1–4, 5/6) for 59% and the GHS classification (1A, 1B, no category) for 73% of the evaluation dataset when compared to LLNA data. When compared against human data the DA correctly predicted the potency category for 68% and the GHS classification for 76% of the dataset, respectively (Table 7). The DA was also shown to predict human sensitisation more accurately than the LLNA (potency category 68% (DA) vs 54% (LLNA); GHS classification 76% (DA) vs 65% (LLNA)).

The confusion matrices comparing the DA predicted GHS classification against LLNA and human reference data are shown in Fig. 10.

3.5.3. Performance of exclusion criteria

3.5.3.1. Prohaptens. Only three chemicals (2,2-azobis phenol, benzo(a) pyrene, and 7,12-dimethylbenz[a]anthracene) in the evaluation dataset were thought to be exclusive prohaptens (as identified by Derek), all of which were positive in the LLNA. The DA suggests to de-prioritise the DPRA (KE1) for prohaptens (due to the lack of metabolic capacity in the assay) and as expected, the DPRA mispredicted both prohaptens tested

in this assay (2,2-azobis phenol and 7,12-dimethylbenz[a]anthracene) as non-sensitisers. Assays measuring KE2 and KE3 predicted all chemicals correctly, where data were available: benzo(a)pyrene and 7,12-dimethylbenz[a]anthracene have been tested in assays covering KE2 and benzo(a)pyrene and 2,2-azobis phenol in assays covering KE3 (Table 8). No human data were available for these assays.

The DA was unable to make a prediction for any of these prohaptens as, although data for KE2 and KE3 were available, the relatively high lipophilicity of these chemicals resulted in assays measuring either KE being de-prioritised according to the DA exclusion criteria.

3.5.3.2. Lipophilic chemicals. The exclusion criteria used in the DA de-prioritises assays measuring KE2 and/or KE3 when chemicals have a logP over 5, and/or 3.5, respectively.

For chemicals with a logP greater than 3.5, no significant difference was observed in predictivity when using assays measuring KE3 (LLNA - 71% acc, human - 80% acc) in comparison to assays measuring KE1 (LLNA - 65% acc, human - 84% acc) or KE2 (LLNA - 70% acc, human - 74% acc). Similarly, for chemicals with a logP greater than 5, assays measuring KE2 were generally as predictive as assays measuring KE1 or KE3 when compared against both LLNA (Acc - KE1, 73%; KE2, 77%; KE3, 56%) or human data (Acc - KE1, 100%; KE2, 100%; KE3, 100%). This suggests that cell-based assays are more predictive of lipophilic chemicals than previously thought, as indicated by the recent update to

Table 6

The skin sensitisation hazard predictivity of individual test methods and the DA compared to both LLNA and human reference data.

Test method	LLNA				Human			
	n	Accuracy	Sensitivity	Specificity	n	Accuracy	Sensitivity	Specificity
LLNA	–	–	–	–	108	81%	88%	64%
DPRA	192	76%	77%	71%	100	81%	93%	52%
KeratinoSens™	184	74%	78%	64%	99	84%	84%	83%
LuSens	77	75%	77%	72%	60	83%	82%	86%
h-CLAT	163	74%	80%	59%	95	82%	78%	87%
U-SENS™	123	83%	93%	56%	91	81%	89%	60%
Derek	210	83%	93%	61%	108	81%	88%	64%
DA	194 ^a	85%	88%	77%	102 ^b	86%	89%	81%

^a 16 chemicals not predicted to be a sensitiser/non-sensitiser by the DA.

^b 6 chemicals not predicted to be a sensitiser/non-sensitiser by the DA.

Table 7

The skin sensitisation potency predictivity of the DA (and LLNA) compared to both LLNA and human reference data.

Benchmark data	Potency category				GHS classification			
	<i>n</i>	Over-predicted	Correctly predicted	Under-predicted	<i>n</i>	Over-predicted	Correctly predicted	Under-predicted
DA vs LLNA	174 ^a	20%	59%	21%	174 ^a	12%	73%	15%
DA vs Human	79 ^b	14%	68%	18%	79 ^b	10%	76%	14%
LLNA vs Human	89	25%	54%	21%	79	20%	65%	15%

^a 20 chemicals with a hazard prediction not given a potency prediction by the DA.^b 23 chemicals with a hazard prediction not given a potency prediction by the DA.

the OECD Test Guideline for KeratinoSens™ (OECD, 2018a) and results reported for U-SENS™ (Alépée et al., 2017). Regardless of the minor difference in predictivity between the assays, the DA, using the exclusion criteria, correctly predicted the hazard (sensitiser/non-sensitiser) for 78% of chemicals with a logP greater than 3.5 with LLNA data and 77% with human data (Table 9). For potency, the potency category was correctly predicted for 59%, and GHS classification correctly predicted for 77% of lipophilic chemicals when compared against LLNA data. When compared against human data, the accuracy improved to 91% for both potency category and GHS classification prediction (Table 10). Similar trends were observed for chemicals with a logP greater than 5 with the DA correctly predicting the hazard for 70% and 100% of these chemicals when compared to LLNA and human reference data, respectively (Table 9). The potency category was correctly identified for 60% and GHS classification for 70% of chemicals with LLNA data and both classification types predicted correctly for 100% of chemicals with human data (Table 10).

3.5.3.3. Lysine-reactive chemicals. Although only 14 chemicals in the evaluation dataset were assigned by Derek as preferentially lysine-reactive (14 with LLNA data, 6 with human data) it was clear that assays measuring KE2 were much less predictive (LLNA - 39% acc, human - 50% acc) than assays measuring KE1 (LLNA - 100% acc, human - 83% acc) or KE3 (LLNA - 86% acc, human - 100% acc) (Table 8).

The DA correctly predicted the skin sensitisation hazard of 100% of lysine-reactive chemicals with LLNA data and 83% with human data (Table 9). The potency category was correctly predicted for 54% and GHS classification correctly predicted for 92% of lysine-reactive chemicals with LLNA data. When compared against the human data, 67% were predicted correctly when assessed against both classification types (Table 10).

3.6. Mispredictions

The majority of mispredictions made by the DA fall into the following three categories:

(1) Inconsistency between *in vivo* outcomes

102 chemicals with both LLNA and human data were given a hazard prediction by the DA. When compared against the *in vivo* reference data the DA predicted both species well, however, 20 chemicals had a species-dependant outcome due to discordant LLNA and human calls. This underlines the importance of considering the species when evaluating the results from the DA as it is not uncommon for LLNA and human data to conflict. For example, nickel chloride is a human sensitiser but is outside the applicability domain of the LLNA (ICCVAM, 2010), although often reported as a non-sensitiser. Nevertheless, the DA predicted it to be a strong sensitiser (GHS 1A, potency category 2) due to the potency prediction model containing positive human data for both nickel chloride (human potency category 2) and nickel sulfate (human potency category 1), as well as containing positive human and LLNA data for other metals in the same Derek alert domain. This underlines

the importance of considering data from more than one species as the LLNA does not always correspond to the human outcome.

Similarly, phthalic anhydride is not a human skin sensitiser (Basketter and Kimber, 2016) but is a strong sensitiser in the LLNA (Gerberick et al., 2007) as the LLNA does not differentiate between skin and respiratory sensitisers (Dearman et al., 2013). When compared against human reference data, phthalic anhydride was incorrectly predicted by the DA to be a sensitiser due to the firing of a Derek alert and a positive result in the DPRA. It was predicted to be a strong sensitiser (potency category 2, GHS 1A) as the potency prediction model contains LLNA data for this chemical (EC3 = 0.16%) as well as several similar LLNA-positive chemicals within the same alert domain.

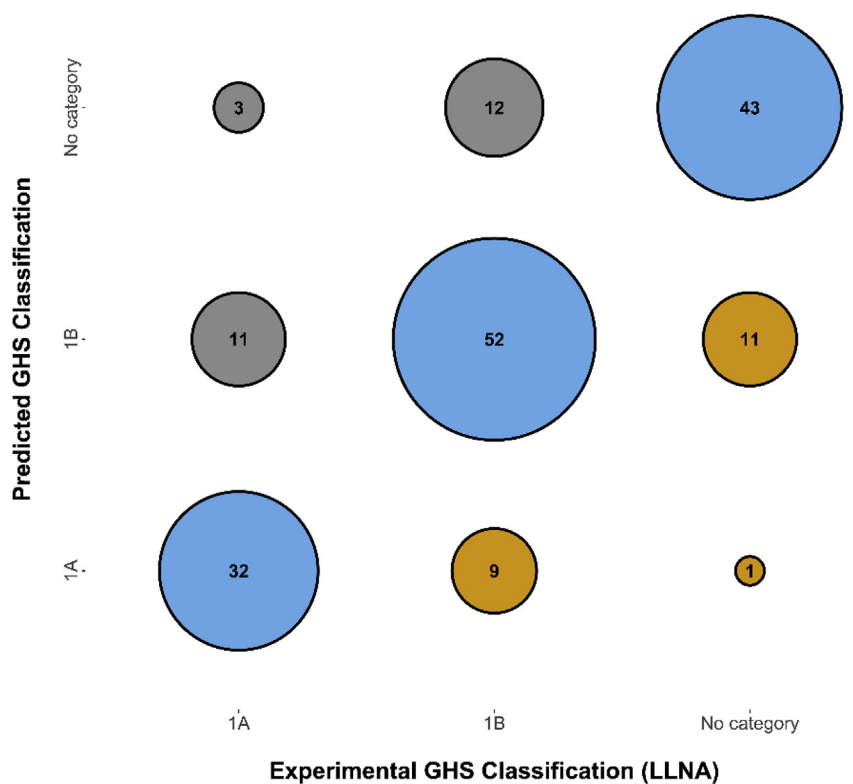
There are also several prehaptns with differing human and LLNA outcomes which may lead to a misprediction by the DA depending on the *in vivo* reference data used. As the oxidation of prehaptns such as limonene and citronellol will differ based on the storage and batch of the chemical tested this may explain the incongruity between the two species.

(2) Derek outcome being overturned by *in chemico/in vitro* assay results

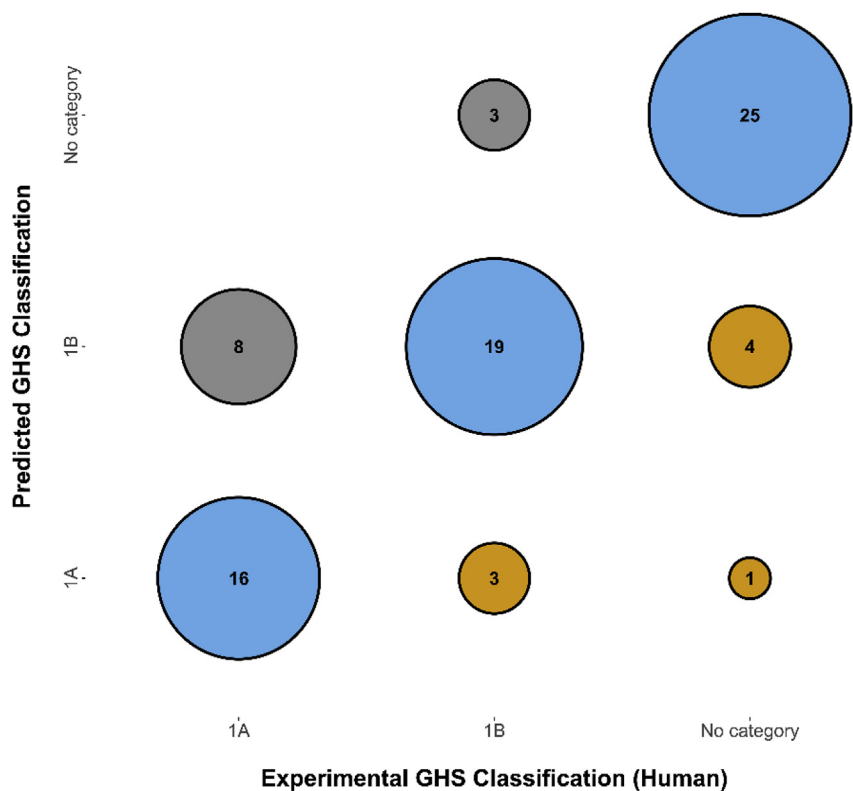
It is very concerning from a human health perspective when sensitising chemicals are predicted by the DA as non-sensitisers as this could lead to accidental human sensitisation. Three chemicals predicted to be non-sensitisers by the DA were reported as strong sensitisers in the LLNA (3-aminophenol, hexyl salicylate, and squaric acid diethyl ester). All three fired a Derek alert but the *in silico* prediction was overturned by mainly negative *in chemico/in vitro* assay data. Likewise, fourteen weak sensitisers (GHS 1B) were mispredicted as non-sensitisers when compared against either LLNA or human reference data. Derek fired an alert for 9/14 of these chemicals, but again due to generally negative results from the *in chemico/in vitro* assays, the Derek result was overturned in the '2 out of 3' overall prediction.

(3) Chemicals close to potency category/GHS classification boundaries

When the model predicts the potency category, it does so by taking the harmonic mean of the Nearest Neighbours' potency categories, weighted by their Tanimoto similarities, and rounding half down to give an integer used as the potency category (e.g. 2.5 is rounded down to category 2). By rounding half down, the predictions are more protective of human health. However, if a prediction lies relatively close to the boundary between two categories this can have a detrimental impact on the accuracy of the prediction, as has been investigated recently by Leontaridou and co-workers (Leontaridou et al., 2017). Using the human *in vivo* data as reference it was found that predictions closer to category borders were indeed less accurate. For chemicals predicted to be GHS 1A sensitisers (e.g. the predicted potency category was below 2.5) the correct GHS classification was predicted for 88% (*n* = 16) of chemicals with a predicted potency category below 2.0 whereas the accuracy fell to 50% (*n* = 4) when evaluating chemicals with a predicted category closer to the border (between 2.01 and 2.5). Similarly, chemicals predicted to be GHS 1B sensitisers (e.g. the predicted potency category falls between 2.51 and 4.5), the accuracy was slightly higher,



Category under-predicted
 Category correctly predicted
 Category over-predicted



Category under-predicted
 Category correctly predicted
 Category over-predicted

Fig. 10. GHS classification confusion matrices comparing the predicted GHS classification against LLNA and human reference data. The x-axis shows the experimentally-derived GHS classification and the y-axis the predicted GHS classification from the DA. Correctly predicted categories are illustrated by the blue circles with the black trim forming a diagonal line, in increasing size according to number of chemicals predicted. Dark grey circles (shown on the left of the correctly predicted diagonal) represent chemicals which have been predicted as less potent than observed experimentally (under-prediction). Orange circles (shown on the right of the correctly predicted diagonal) represent chemicals which have been predicted as more potent than observed experimentally (over-prediction). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 8

The skin sensitisation hazard predictivity of the DA for chemicals matching the exclusion criteria compared to both LLNA and human reference data.

Hazard predictivity of KE assays												
Chemical subset	LLNA						Human					
	KE1		KE2		KE3		KE1		KE2		KE3	
	<i>n</i>	Acc	<i>n</i>	Acc	<i>n</i>	Acc	<i>n</i>	Acc	<i>n</i>	Acc	<i>n</i>	Acc
Prohaptens	2	0%	2	100%	2	100%	0	n/a	0	n/a	0	n/a
logP > 3.5	46	65%	43	70%	31	71%	19	84%	19	74%	15	80%
logP > 5	15	73%	13	77%	9	56%	7	100%	6	100%	6	100%
Lysine-reactive	13	100%	13	39%	7	86%	6	83%	6	50%	5	100%

Table 9

The skin sensitisation hazard predictivity of the DA for chemicals matching the exclusion criteria compared to both LLNA and human reference data.

Hazard predictivity of DA									
Chemical subset	LLNA				Human				
	<i>n</i>	Accuracy	Sensitivity	Specificity	<i>N</i>	Accuracy	Sensitivity	Specificity	
Prohaptens	0	n/a	n/a	n/a	0	n/a	n/a	n/a	n/a
logP > 3.5	36 ^a	78%	75%	88%	13 ^c	77%	80%	75%	
logP > 5	10 ^b	70%	57%	100%	3 ^d	100%	100%	100%	
Lysine-reactive	14	100%	100%	n/a	6	83%	100%	n/a	

^a Prediction could not be made for 14 chemicals.^b Prediction could not be made for 5 chemicals.^c Prediction could not be made for 5 chemicals.^d Prediction could not be made for 2 chemicals.**Table 10**

The skin sensitisation potency predictivity of the DA for chemicals matching the exclusion criteria compared to both LLNA and human reference data.

Potency predictivity of DA								
Chemical subset	Potency category				GHS classification			
	<i>n</i>	Over-predicted	Correctly predicted	Under-predicted	<i>n</i>	Over-predicted	Correctly predicted	Under-predicted
LLNA								
Prohaptens	0	n/a	n/a	n/a	0	n/a	n/a	n/a
logP > 3.5	34 ^a	15%	59%	27%	34 ^a	3%	77%	21%
logP > 5	10 ^b	10%	60%	30%	10 ^b	0%	70%	30%
Lysine-reactive	13 ^c	15%	54%	31%	13 ^c	8%	92%	0%
Human								
Prohaptens	0	n/a	n/a	n/a	0	n/a	n/a	n/a
logP > 3.5	9 ^d	9%	91%	0.0%	9 ^d	9%	91%	0%
logP > 5	3	0%	100%	0.0%	3	0%	100%	0%
Lysine-reactive	3 ^e	33%	67%	0.0%	3 ^e	33%	67%	0%

^a Prediction could not be made for 2 chemicals.^b Prediction could not be made for 5 chemicals.^c Prediction could not be made for 1 chemical.^d Prediction could not be made for 4 chemicals.^e Prediction could not be made for 3 chemicals.

64% ($n = 14$) for chemicals with a predicted potency category between 3.0 and 4.5 in contrast to chemicals with a predicted category between 2.51 and 2.99, 59% ($n = 17$). Further work will focus on developing a method to assign a confidence level associated with the potency prediction. See Supporting Data for the full prediction list and corresponding *in vivo/in silico/in chemico/in vitro* data.

3.7. Limitations

The limitations of the DA can be separated into (1) limitations of the *in vivo* reference data, (2) limitations inherent in the *in silico/in chemico/*

in vitro information sources used, and (3) limitations associated with the DA itself.

- (1) The predictive capacity of defined approaches is typically assessed by comparing against LLNA and/or human data, however, both have drawbacks. The EC3 values determined from the LLNA can vary significantly between studies on the same chemical, and potency can differ when different vehicles are used (Dumont et al., 2016; Hoffmann, 2015; Roberts et al., 2016). As a result, where more than one EC3 value was available for a chemical in the evaluation dataset, the median value was used to try and limit the

impact of high variance. The human data categorised by Basketter et al. is commonly used as reference data as the categories were assigned after careful consideration of all available human data, taking exposure and population into account. However, it should be kept in mind that the No Observed Effect Level (NOEL) which may be used to support categorisation may not actually be the Lowest Observed Effect Level (LOEL) (Api et al., 2017; Basketter et al., 2014). Finally, although the correlation between LLNA and human outcomes is generally quite high, there are several outliers which may have an impact on the predictivity of the DA.

- (2) The applicability domains of the *in chemico/in vitro* assays have not yet been fully explored indicating that there may be some ambiguity over assay data for specific chemical classes, especially for negative assay outcomes. Other issues such as intra- and inter-laboratory reproducibility, protocol issues such as difficulties in predicting fluorescent and/or highly cytotoxic chemicals, and interchangeability between assays measuring the same KE could have an impact when using data from these assays in the DA (OECD, 2018a, 2018b, 2015). One limitation of the *in silico* component, Derek, is that chemicals with an ambiguous structure (e.g. Tween 80) and/or undefined mixtures (e.g. fragrances such as oak moss) cannot be used in the tool. Another potential limitation is that it is possible that the scope of a skin sensitisation alert could be modified upon the release of new versions of Derek. As alert changes are documented in the Derek release notes this potential change could be anticipated in advance and chemicals firing affected alerts re-evaluated. Another limitation is that with Derek being a proprietary software tool, access to the alerts, and by extension, the information used in the exclusion criteria is restricted to Lhasa Limited members. Alternatively, information from other *in silico* models and/or expert judgement could be used to assess chemicals against the exclusion criteria used in the DA without the use of Derek, although other models have not yet been evaluated.
- (3) For the DA, one limitation is related to the potency prediction model. For example, no potency prediction is provided when fewer than 3 Nearest Neighbours are present in the same alert domain. However, Lhasa Limited is pursuing data sharing opportunities with members to add historical data to these alerts to address this. Furthermore, some relatively small alert domains (due to scarcity of public data) in the potency prediction model may lead to more uncertainty in the prediction. More uncertainty may be associated with predictions which are closer to category boundaries (as discussed in Section 3.6). Other ways to evaluate potency using *in silico* results, possibly in combination with *in chemico* and *in vitro* data, are being investigated. Another limitation is related to the availability of public datasets containing *in chemico/in vitro* assay data which can be used to evaluate the performance of the DA. The dataset used to evaluate the DA consisted of only 210 chemicals and represents one of the largest available in the public domain. The DA also does not use a discrete information source that measures KE4 or the adverse outcome, although this information is present in Derek as the alerts are based on *in vivo* mouse, guinea pig, and human data. An additional limiting factor is that the DA is more useful for mono constituent substances and defined mixtures - as Derek requires a structure as input, UVCBs (Unknown or Variable Composition, Complex Reaction Products and Biological Materials) cannot currently be predicted by the DA. Finally, the DA has not yet been validated using a test set of new *in chemico/in vitro* data. Lhasa Limited is actively exploring opportunities and collaborations to ensure that this is achieved.

4. Conclusions

A DA decision tree has been designed using exclusion criteria based on known limitations of *in silico* outcomes and *in chemico/in vitro* assays. The hypothesis proposed is that excluding less reliable assays and/or *in*

silico outcomes for specific chemicals will produce a rational and transparent DA for the prediction of skin sensitisation potential.

When evaluated against a curated dataset ($n = 210$) comprising of previously published human, LLNA, DPRA, KeratinoSens™, LuSens, h-CLAT, and U-SENS™ data the DA correctly predicts the hazard (sensitiser/non-sensitiser) for 85% and 86% of chemicals with LLNA and human reference data, respectively. In terms of potency, the DA can accurately predict the potency category (1–4, 5/6) for 59%, and the GHS classification (1A/1B/no category) for 73% of the dataset with LLNA reference data and 68% and 76%, respectively, with human reference data. For both hazard and potency predictions, the DA is more predictive of human skin sensitisation than the LLNA as the LLNA predicts the hazard with 81% accuracy (c.f. DA 86%) and the GHS classification with 65% accuracy (c.f. DA 76%). The concordance of assays addressing the same KE suggests that these can be used interchangeably (as well as future ‘me-too’ assays, assuming they measure the KE in a similar manner). The majority of chemicals in the evaluation dataset were accurately predicted using only Derek and one other (concordant) assay which makes the DA a quick and easy way to predict the skin sensitisation potential of a given chemical. The DA is applicable to a wide variety of chemicals with varying functional groups and physicochemical properties and chemicals with poor predictivity in *in chemico* and *in vitro* assays can be correctly predicted by the DA e.g. pre- and prohaptenes, lipophilic chemicals, metals. Due to the high predictivity of the DA, it is suitable to be used in place of animal tests for regulatory purposes as it can discriminate between sensitisers and non-sensitisers as well as provide a potency prediction (potency category and/or GHS classification).

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.yrtph.2018.11.001>.

Transparency document

Transparency document related to this article can be found online at <https://doi.org/10.1016/j.yrtph.2018.11.001>.

References

- Alépée, N., Piroird, C., Nardelli, L., 2017. U-SENS™: a U937 cell line activation test for skin sensitization. In: *Alternatives for Dermal Toxicity Testing*. Springer International Publishing, pp. 311–330.
- Api, A.M., Parakhia, R., O'Brien, D., Basketter, D.A., 2017. Fragrances categorized according to relative human skin sensitization potency. *Dermatitis* 28, 299–307.
- Aptula, A.O., Patlewicz, G., Roberts, D.W., 2005. Skin sensitization: reaction mechanistic applicability domains for structure-activity relationships. *Chem. Res. Toxicol.* 18, 1420–1426.
- Aptula, A.O., Roberts, D.W., 2006. Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: general principles and application to reactive toxicity. *Chem. Res. Toxicol.* 19, 1097–1105.
- Basketter, D., 2016. Skin sensitisation, adverse outcome pathways and alternatives. *Altern. Lab. Anim.* 44, 431–436.
- Basketter, D.A., Alépée, N., Ashikaga, T., Barroso, J., Gilmour, N., Goebel, C., Hibatallah, J., Hoffmann, S., Kern, P., Martinozzi-Teissier, S., Maxwell, G., Reisinger, K., Sakaguchi, H., Schepky, A., Tailhardat, M., Templier, M., 2014. Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis* 25, 11–21.
- Basketter, D.A., Kimber, I., 2016. Phthalic anhydride: illustrating a conundrum in chemical allergy. *J. Immunot.* 13, 767–769.
- Bufalo, A., Del, Pauloin, T., Alepee, N., Clouzeau, J., Detroyer, A., Eilstein, J., Gomes, C., Nocairi, H., Cile Piroird, C., Rousset, F., Tourneix, F., Basketter, D., Teissier, S.M., 2018. Alternative integrated testing for skin sensitization: assuring consumer safety. *Appl. In Vitro Toxicol.* 4, 30–43.
- Canipa, S.J., Chilton, M.L., Hemingway, R., Macmillan, D.S., Myden, A., Plante, J.P.,

- Tennant, R.E., Vessey, J.D., Steger-Hartmann, T., Gould, J., Hillegass, J., Etter, S., Smith, B.P.C., White, A., Sterchele, P., De Smedt, A., O'Brien, D., Parakhia, R., 2017. A quantitative *in silico* model for predicting skin sensitization using a nearest neighbours approach within expert-derived structure-activity alert spaces. *J. Appl. Toxicol.* 37, 985–995.
- Casati, S., Aschberger, K., Asturiol, D., Basketter, D., Dimitrov, S., Dumont, C., Karlberg, A.-T., Lepoittevin, J.-P., Patlewicz, G., Roberts, D.W., Worth, A., 2016. Ability of Non-animal Methods for Skin Sensitisation to Detect pre- and pro-haptens: Report and Recommendations of an EURL ECVAM Expert Meeting. EUR 27752 EN.
- Chilton, M.L., Macmillan, D.S., Steger-Hartmann, T., Hillegass, J., Bellion, P., Vuorinen, A., Etter, S., Smith, B.P.C., White, A., Sterchele, P., De Smedt, A., Glogovac, M., Glowienke, S., O'Brien, D., Parakhia, R., 2018. Making reliable negative predictions of human skin sensitisation using an *in silico* fragmentation approach. *Regul. Toxicol. Pharmacol.* 95, 227–235.
- Cooper, J.A., Saracci, R., Cole, P., 1979. Describing the validity of carcinogen screening tests. *Br. J. Canc.* 39, 87–89.
- Dearman, R.J., Basketter, D.A., Kimber, I., 2013. Inter-relationships between different classes of chemical allergens. *J. Appl. Toxicol.* 33, 558–565.
- 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., van der Veen, J.W., Adamson, G., Ezendam, J., van Loveren, H., Natsch, A., 2013. Gene expression changes induced by skin sensitizers in the KeratinoSens™ cell line: discriminating Nrf2-dependent and Nrf2-independent events. *Toxicol. In Vitro* 27, 2225–2232.
- Environmental Protection Agency, 2018. Interim Science Policy: Use of Alternative Approaches for Skin Sensitization as a Replacement for Laboratory Animal Testing.
- European Union, 2009. Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on Cosmetic Products. O.J., vol 52. pp. 342.
- European Union, 2006. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, Amending Directive 1999/4. O.J., vol 49. pp. 396.
- Fabian, E., Vogel, D., Blatz, V., Ramirez, T., Kolle, S., Eltze, T., van Ravenzwaay, B., Oesch, F., Landsiedel, R., 2013. Xenobiotic metabolizing enzyme activities in cells used for testing skin sensitization *in vitro*. *Arch. Toxicol.* 87, 1683–1696.
- Gerberick, G.F., Vassallo, J.D., Foertsch, L.M., Price, B.B., Chaney, J.G., Lepoittevin, J.-P., 2007. Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. *Toxicol. Sci.* 97, 417–427.
- 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., Alépée, N., Allen, D., Api, A.M., Ashikaga, T., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Goebel, C., Kern, P.S., Klaric, M., Kühn, J., Lalko, J.F., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Parakhia, R., van Vliet, E., Zang, Q., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database. *Crit. Rev. Toxicol.* 48, 344–358.
- ICCVAM, 2010. ICCVAM Test Method Evaluation Report on Using the LLNA for Testing Pesticide Formulations, Metals, Substances in Aqueous Solutions, and Other Products, NIH Publication No. 10-7512. National Institute of Environmental Health Sciences, Research Triangle Park, NC.
- Judson, P.N., Stalford, S.A., Vessey, J., 2013. Assessing confidence in predictions made by knowledge-based systems. *Toxicol. Res.* 2, 70–79.
- Leontaridou, M., Urbisch, D., Kolle, S.N., Ott, K., Mulliner, D.S., Gabbert, S., Landsiedel, R., 2017. The borderline range of toxicological methods: quantification and implications for evaluating precision. *ALTEX* 34, 525–538.
- Macmillan, D.S., Canipa, S.J., Chilton, M.L., Williams, R.V., Barber, C.G., 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.
- Natsch, A., 2017. The KeratinoSens™ assay for skin sensitization screening. In: Alternatives for Dermal Toxicity Testing. Springer International Publishing, pp. 235–248.
- OECD, 2018a. Key Event Based Test Guideline 442D: *In Vitro* Skin Sensitisation Assays Addressing the Key Event on Keratinocyte Activation on the Adverse Outcome Pathway for Skin Sensitisation, OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris.
- OECD, 2018b. Key Event Based Test Guideline 442E: *In Vitro* Skin Sensitisation Assays Addressing the Key Event on Activation of Dendritic Cells on the Adverse Outcome Pathway for Skin Sensitisation. OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris.
- OECD, 2016a. Guidance Document on the Reporting of Defined Approaches to Be Used within Integrated Approaches to Testing and Assessment, OECD Series on Testing and Assessment. No. 255. OECD Publishing, Paris.
- OECD, 2016b. Annex I: Case Studies to the Guidance Document on the Reporting of Defined Approaches and Individual Information Sources to Be Used within Integrated Approaches to Testing and Assessment (IATA) for Skin Sensitisation. OECD Series on Testing and Assessment. No.256. OECD Publishing, Paris.
- OECD, 2015. Test No. 442C: *In Chemico* Skin Sensitisation: Direct Peptide Reactivity Assay (DPRA). OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris.
- OECD, 2012. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins Part 1: Scientific Evidence. OECD Series on Testing and Assessment. OECD Publishing, Paris.
- OECD, 2010. Test No. 429: Skin Sensitisation: Local Lymph Node Assay. OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris.
- OECD, 1992. Test No. 406: Skin Sensitisation. OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris.
- Patlewicz, G., Casati, S., Basketter, D.A., Asturiol, D., Roberts, D.W., Lepoittevin, J.-P., Worth, A.P., Aschberger, K., 2016. Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? *Regul. Toxicol. Pharmacol.* 82, 147–155.
- Piroid, C., Ovigne, J.-M., Rousset, F., Martinozzi-Teissier, S., Gomes, C., Cotovio, J., Alépée, N., 2015. The Myeloid U937 Skin Sensitization Test (U-SENS) addresses the activation of dendritic cell event in the adverse outcome pathway for skin sensitization. *Toxicol. In Vitro* 29, 901–916.
- Ramirez, T., Mehling, A., Kolle, S.N., Wruck, C.J., 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.
- Ramirez, T., Stein, N., Aumann, A., Remus, T., Edwards, A., Norman, K.G., Ryan, C., Bader, J.E., Fehr, M., Burleson, F., Foertsch, L., Wang, X., Gerberick, F., Beilstein, P., Hoffmann, S., Mehling, A., van Ravenzwaay, B., Landsiedel, R., 2016. Intra- and inter-laboratory reproducibility and accuracy of the LuSens assay: a reporter gene-cell line to detect keratinocyte activation by skin sensitizers. *Toxicol. In Vitro* 32, 278–286.
- Roberts, D.W., Api, A.M., Aptula, A.O., 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.
- Takenouchi, O., Miyazawa, M., Saito, K., Ashikaga, T., Sakaguchi, H., 2013. Predictive performance of the human Cell Line Activation Test (h-CLAT) for lipophilic chemicals with high octanol-water partition coefficients. *J. Toxicol. Sci.* 38, 599–609.
- Urbisch, D., Becker, M., Honarvar, N., Kolle, S.N., Mehling, A., Teubner, W., Wareing, B., Landsiedel, R., 2016. Assessment of pre- and pro-haptens using nonanimal test methods for skin sensitization. *Chem. Res. Toxicol.* 29, 901–913.
- Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., Landsiedel, R., Jaworska, J., Kern, P.S., Gerberick, F., Natsch, A., Emter, R., Ashikaga, T., Miyazawa, M., Sakaguchi, H., 2015. Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul. Toxicol. Pharmacol.* 71, 337–351.