



Predicting skin sensitisation using a decision tree integrated testing strategy with an *in silico* model and *in chemico/in vitro* assays



Donna S. Macmillan^{*}, Steven J. Canipa, Martyn L. Chilton, Richard V. Williams, Christopher G. Barber

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

ARTICLE INFO

Article history:

Received 18 December 2015

Received in revised form

13 January 2016

Accepted 14 January 2016

Available online 18 January 2016

Keywords:

Skin sensitisation

Derek Nexus

DPRA

KeratinoSens

LuSens

h-CLAT

U-SENS

Integrated testing strategy

In silico assessment

ABSTRACT

There is a pressing need for non-animal methods to predict skin sensitisation potential and a number of *in chemico* and *in vitro* assays have been designed with this in mind. However, some compounds can fall outside the applicability domain of these *in chemico/in vitro* assays and may not be predicted accurately.

Rule-based *in silico* models such as Derek Nexus are expert-derived from animal and/or human data and the mechanism-based alert domain can take a number of factors into account (e.g. abiotic/biotic activation). Therefore, Derek Nexus may be able to predict for compounds outside the applicability domain of *in chemico/in vitro* assays.

To this end, an integrated testing strategy (ITS) decision tree using Derek Nexus and a maximum of two assays (from DPRA, KeratinoSens, LuSens, h-CLAT and U-SENS) was developed. Generally, the decision tree improved upon other ITS evaluated in this study with positive and negative predictivity calculated as 86% and 81%, respectively. Our results demonstrate that an ITS using an *in silico* model such as Derek Nexus with a maximum of two *in chemico/in vitro* assays can predict the sensitising potential of a number of chemicals, including those outside the applicability domain of existing non-animal assays.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Skin sensitisation is an important toxicological endpoint which leads to allergic contact dermatitis (ACD). ACD develops in two stages, beginning with the induction stage where a chemical (known as a hapten) forms a conjugate with nucleophilic skin proteins. This initiates a cascade, resulting in proliferation of allergen specific T-cells. The second stage, elicitation, arises when the subject is re-challenged with the same allergen (hapten). The

hapten-protein complex is formed again, triggering the allergen specific T-cells to induce the release of inflammatory cytokines which leads to ACD (Kimber et al., 2002).

The prediction of skin sensitisation potential is an important requirement for a number of chemical safety assessments. This is outlined in the EURL ECVAM (European Union Reference Laboratory for alternatives to animal testing) Strategy for Replacement of Animal Testing for Skin Sensitisation Hazard Identification and Classification (European Commission, 2013a) and includes the European Union regulation Registration, Evaluation, Authorisation and restriction of Chemicals (REACH), Classification Labelling and Packaging of substances and mixtures (CLP) Regulation, and risk assessments associated with occupational exposure. Furthermore, there is considerable economic and social pressure to replace, reduce and refine the use of animal tests for the safety assessments of chemicals across a wide range of industries (Russell and Burch, 1959).

The cosmetics industry faces a particular challenge, driven by the implementation of EU Regulation 1223/2009 (replacing EU Directive 76/768/EEC) which came into effect in March 2009 (European Union, 2013). This regulation ensures that any cosmetic

Abbreviations: ACD, allergic contact dermatitis; AOP, adverse outcome pathway; CLP, classification labelling and packaging; DC, dendritic cells; DPRA, direct peptide reactivity assay; DX, Derek Nexus; EURL ECVAM, European Union Reference Laboratory for alternatives to animal testing; GPMT, guinea pig maximisation test; h-CLAT, human cell line activation test; IATA, integrated approach to testing and assessment; ITS, integrated testing strategy/strategies; KS, KeratinoSens; LC-MS, liquid chromatography–mass spectrometry; Lhasa DT, Lhasa decision tree; LLNA, murine local lymph node assay; MIE, molecular initiating event; U-SENS, myeloid U937 skin sensitisation test; OECD, Organisation for Economic Co-operation and Development.

^{*} Corresponding author.

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

An adverse outcome pathway (AOP) is the sequence of events leading from the molecular initiating event to an *in vivo* adverse outcome. The AOP for skin sensitisation caused by covalent binding of a test chemical to skin proteins has been published by the Organisation for Economic Co-operation and Development (OECD) and has been summarised as eleven steps of which four are considered key (OECD, 2012). There has been a concerted effort in recent years to develop new non-animal based assays which address these key events in the AOP such as the DPRA (Direct Peptide Reactivity Test), KeratinoSens, LuSens, h-CLAT (human Cell Line Activation Test) and U-SENS (myeloid U937 skin sensitisation test) (Fig. 1).

A diverse range of ITS have been investigated, for example, artificial neural networks and Bayesian networks (Hirota et al., 2015; Jaworska et al., 2013; Rorije et al., 2013; Tsujita-Inoue et al., 2015, 2014), weight of evidence (Ellison et al., 2010; Gubbels-van Hal et al., 2005; Natsch et al., 2013; Urbisch et al., 2015) or score-based test batteries (Jowsey et al., 2006; Natsch et al., 2008; Nukada et al., 2013; Takenouchi et al., 2015), decision trees (Bauch et al., 2012; van der Veen et al., 2014) and miscellaneous ITS such as gated logic algorithms, global regression analysis and machine learning (Luechtefeld et al., 2015; McKim et al., 2010; Natsch et al., 2014). Nevertheless there is still substantial interest in this

(1) DPRA – This assay simulates the molecular initiating event (MIE) of the skin sensitisation AOP by measuring the reactivity of the test substance towards synthetic model peptides containing either lysine or cysteine. The sensitisation potential of a test chemical is evaluated by measuring the mean cysteine and lysine peptide depletion and assigning a class based on percentage of this depletion. One described drawback when using model peptides is that the true skin conditions are not reproduced and test chemicals with affinities for several other amino acid residues may not be adequately evaluated by the DPRA ([Divkovic et al., 2005](#)). Furthermore, a number of false positives have been reported due to test chemicals which induce cysteine oxidation ([European Commission, 2013b](#); [OECD, 2015a](#)) resulting in a misleading quantity of peptide depletion. Similarly, LC-MS (liquid chromatography–mass spectrometry) analysis indicates that aldehydes do not form

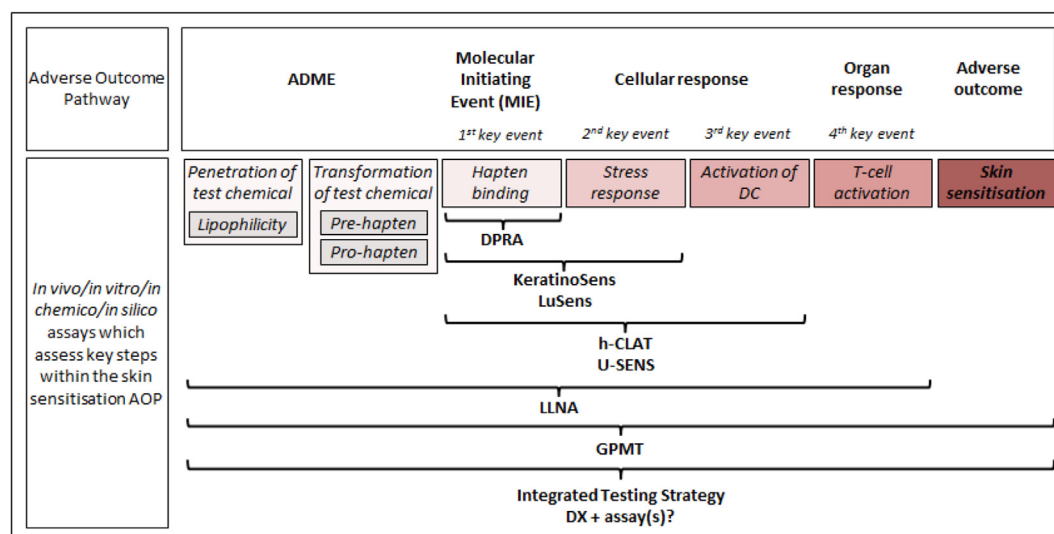


Fig. 1. Adverse outcome pathway for skin sensitisation and step/key events measured by the *in vivo/in chemico/in vitro/in silico* assay(s) shown. Adapted from OECD 2012.

covalent adducts with the model nucleophiles used in the DPRA, but instead cause peptide depletion through an unknown mechanism. Nevertheless, the assay result of sensitiser is generally correct for this class (Natsch et al., 2007).

- (2) KeratinoSens/LuSens - These assays measure the second key event in the AOP; the activation of keratinocytes by monitoring luciferase activity. The OECD guideline advises that negative results from the KeratinoSens may be unreliable as test chemicals selectively reactive to lysine residues may be predicted as non-sensitisers (OECD, 2015b). It has also been reported that non-sensitisers which can act as chemical stressors may lead to false positive results in the KeratinoSens assay (European Commission, 2014). Moreover, some chemical domains are predicted less well than others. Although the authors state there are few chemicals within this domain, it seems that anhydride-containing test chemicals are incorrectly predicted by the KeratinoSens assay (Natsch et al., 2014) – with phthalic anhydride mispredicted by both the KeratinoSens and the LuSens assay (Ramirez et al., 2014).
- (3) h-CLAT/U-SENS – These assays measure the third key event in the AOP; dendritic cell (DC) activation by measuring changes in the expression of cell surface markers (CD86 and CD54) (OECD, 2014). The h-CLAT has been described as a 'highly simplified' system compared to the skin sensitisation process and one proposed limitation is a lack of interaction with keratinocytes. There is evidence to suggest that the h-CLAT predicts weak sensitisers less well than other categories of sensitisers due to the relatively low sensitivity of the assay (Ashikaga et al., 2010) and that the flow cytometry analysis may be subject to interference by competing fluorescence and/or cytotoxicity caused by the test chemicals (European Commission, 2015). Again, as with all cell-based assays, it is suggested that negative results are regarded with caution (OECD, 2014).
- (4) Lipophilicity – A significant limitation of *in chemico/in vitro* assays occurs when testing lipophilic compounds (European Commission, 2013a,b; European Commission, 2014; European Commission, 2015; OECD – draft guideline 2014; OECD, 2015a; OECD, 2015b). The assays described above are performed in aqueous buffer or assay medium and therefore have associated intrinsic solubility limits. Several test chemicals have been observed to precipitate out of solution or remain as immiscible oils (Takenouchi et al., 2015, 2013) and a number of ITS have reported improved performance when logP/logK_{ow} limits have been applied (Natsch et al., 2008; Takenouchi et al., 2015, 2013).
- (5) Pre-/pro-haptens

A second potential limitation of *in chemico/in vitro* assays is when evaluating compounds which require abiotic (e.g. auto-oxidation) or biotic (e.g. P450 metabolism) activation before becoming allergens. Accordingly, numerous pre-/pro-haptens are mispredicted by the assays described above (European Commission, 2013a,b; European Commission, 2014; European Commission, 2015; OECD – draft guideline 2014; OECD, 2015a; OECD, 2015b). Research to combat this has been carried out using, for example, assay pre-treatments such as horseradish peroxidase and hydrogen peroxide (Gerberick et al., 2009) or rat liver S9 (Chipinda et al., 2011) to activate pro-haptens with some success.

1.2.3. Negative predictivity

Based on the limitations described above, whilst positive *in chemico/in vitro* results are generally associated with sensitisers, negative results, even as part of an ITS, are generally less reliable.

The negative predictivity of the DPRA, KeratinoSens, LuSens, h-CLAT and U-SENS range from 52%–59% for each single assay and an ITS based on a 2 out of 3 weight of evidence approach using DPRA, KeratinoSens and h-CLAT reported a negative predictivity of 59% for the chemicals assessed (Urbisch et al., 2015). Consequently, a number of sensitisers within the data set would be incorrectly predicted as non-sensitisers and presents a significant risk of inadvertent human exposure to skin sensitisers.

It is possible that the limitations associated with *in chemico/in vitro* ITS may be improved by incorporating an *in silico* model. Structural alerts for skin sensitisation in Derek Nexus, derived from animal and/or human data, model the end-to-end biological process of ACD and therefore the entire skin sensitisation AOP. Furthermore, as Derek Nexus is a rule-based system developed by experts, the alert domain can be adjusted to take abiotic/biotic activation into account (e.g. resorcinol or precursor alert). Therefore, Derek Nexus can theoretically predict for test chemicals outside the AOP and outside the applicability domain of *in chemico/in vitro* assays e.g. pre-/pro-haptens and lipophilic compounds. This may help plug the gap for test chemicals not sufficiently covered by existing *in chemico/in vitro* assays. Thus, the number of false negatives predicted may be reduced, resulting in the improvement of negative predictivity compared to single assays and/or ITS lacking an *in silico* component.

With this in mind, an *in silico* model may offer a valuable complement to the current battery of *in chemico/in vitro* tests. This paper will focus on using combinations of results from DPRA, KeratinoSens, LuSens, h-CLAT, U-SENS and Derek Nexus to optimise prediction of skin sensitisation based on LLNA and human data.

2. Materials and methods

2.1. Data set

The data set included a total of 213 previously published test substances which incorporated multiple data sets (Urbisch et al., 2015). For all 213 compounds, LLNA data were available. 111 of these compounds also had human data available. *In chemico/in vitro* test data were available for the majority of this data set; DPRA, 194 compounds; KeratinoSens, 187 compounds; LuSens, 78 compounds; h-CLAT, 166 compounds; U-SENS (referred to as MUSST in original data set), 149 compounds.

2.2. Derek Nexus

Derek Nexus 4.1.0 (using Knowledge Base Derek 2014 1.0), an expert knowledge base system for toxicity predictions, containing rules derived from both public and proprietary data, was used to make an *in silico* assessment of the skin sensitising potential of the test substances. If the predicted skin sensitisation outcome was defined as certain/probable/plausible/equivocal, the test chemical was considered as a sensitiser. If no skin sensitisation alert was fired with a likelihood of equivocal or above, the test chemical was considered as a non-sensitiser.

2.3. Data evaluation and statistics

Assessment of *in chemico/in vitro* sensitisation potential was based on the call assigned by the authors cited (Urbisch et al., 2015), however, conflicting and/or borderline results were removed from each data set. The data set was processed in Derek Nexus 4.1.0 against the endpoint of skin sensitisation. Tautomers and mixtures were turned off. The logP was used as calculated by Derek Nexus using the ClogP method from BioByte. Compounds were assigned as potential pre-/pro-haptens by assigning compounds which fire a

skin sensitisation alert in Derek Nexus and have only pre-hapten, pro-hapten or pre-/pro-hapten in the potential mechanism summary in the corresponding alert description.

All analysis was carried out using KNIME 2.12.0 or MS Excel 2013. The predictivity of single assays or combinations were calculated according to Cooper statistics (Cooper et al., 1979) to assess the following: sensitivity ($TP/[TP + FN]*100$); specificity ($TN/[TN + FP]*100$); positive predictivity ($TP/[TP + FP]*100$); negative predictivity ($TN/[TN + FN]*100$); accuracy: $(TP + TN/[TP + TN + FP + FN]*100)$; coverage: (number of compounds which can be predicted by assay or ITS/total number of compounds in data set).

2.4. Decision tree construction

A decision tree is a graphical model to predict the outcome of a specific variable based on discrete domains/variables. A Lhasa decision tree was designed by human experts, building on some previously published work (Macmillan et al., 2015), guided by the following principles:

- Derek Nexus covers the entire skin sensitisation AOP
- Reduce unnecessary *in chemico/in vitro* assays where possible
- Concordant Derek Nexus and *in chemico/in vitro* results – reliable result – no additional assays required
- Discordant Derek Nexus and *in chemico/in vitro* results – unclear result – run another assay and use majority call
- Consider the *in chemico/in vitro* applicability domains
- Maximise the negative predictivity

Firstly, a test chemical is processed through Derek, then assigned as in the domain or outside the domain of the *in chemico/in vitro* assays. Test chemicals within the domain are defined as chemicals with a ClogP below 3.5 and where no prior activation is required before it can behave as a skin sensitizer. Once separated by domain, the Lhasa decision tree indicates whether one or two *in chemico/in vitro* assays are required to assign the test chemical as a sensitizer or a non-sensitizer based on the Derek Nexus outcome (alert fired = skin sensitizer, no alert fired = non-sensitizer) and the first assay result. The Lhasa decision tree was constructed using a 213 test chemical data set and the mean predictivity of compounds

within each overall call is shown in Fig. 2. The number of compounds and predictivity at each branch in the tree is the mean value based on the 20 combinations of Derek Nexus, DPRA, KeratinoSens, LuSens, h-CLAT and U-SENS which can be input into the tree. The individual results for each combination of assays can be found in the Supporting Information.

3. Results

In this study, the performance of individual *in chemico* (DPRA), *in vitro* (KeratinoSens, LuSens, h-CLAT, U-SENS) and *in silico* (Derek Nexus) methods to predict skin sensitisation were evaluated alongside ITS using Derek Nexus with *in chemico/in vitro* assays against LLNA data ($n = 213$) and human data ($n = 111$) (data not shown).

The assays/ITS discussed in this paper are as follows:

- Single assays
 - Individual assays - DPRA, KeratinoSens, LuSens, h-CLAT, U-SENS and Derek Nexus
- Conservative Call
 - Combination of Derek Nexus prediction and one assay. If one result is positive, then overall call is positive. Both assays must be negative for a negative overall call.
- Weight of Evidence
 - Combination of Derek Nexus prediction and 2 assays. Overall call is assigned according to the majority result.
- Concordance Call
 - Combination of Derek Nexus prediction and two assays. Overall call is assigned when all three assay results are concordant. If the three assays are discordant then no call is assigned.
- Lhasa Decision Tree
 - Combination of Derek Nexus prediction and either one or two *in chemico/in vitro* assays.

When calculating the performance figures for each ITS, test chemicals without results for each assay required in the specific ITS were removed from the data set. This resulted in slightly different test data sets for different ITS combinations, according to the assays being evaluated. Full data including the individual results of each assay combination investigated are listed in the Supporting

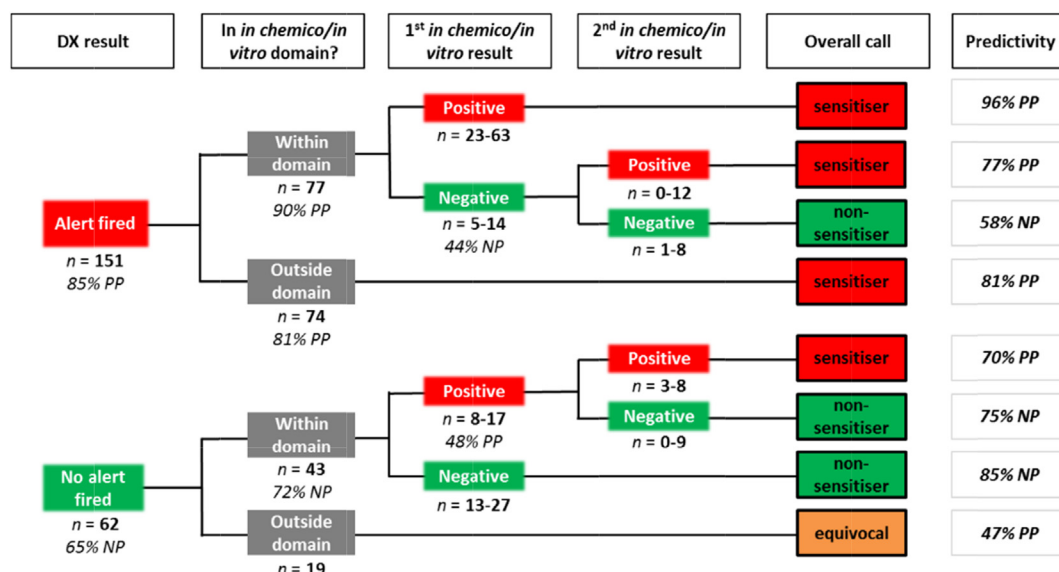


Fig. 2. Lhasa decision tree showing an ITS using Derek Nexus and one or more *in chemico/in vitro* assay (DPRA, KeratinoSens, LuSens, h-CLAT, U-SENS).

Information. The graphs below show the mean and range of performance metrics (PP, NP, coverage and accuracy) of each assay or ITS evaluated in this study (Fig. 3–6).

3.1. Predictive performance of ITS

The individual and mean positive predictivity for each assay combination for each ITS is shown in Fig. 3. The mean positive predictivity ranges from 80% to 96%. As might be expected, the highest mean positive predictivity was observed for the concordant call ITS (96%). The lowest mean positive predictivity was observed for the conservative call ITS (80%) - likely to be caused by the increased number of false positives introduced when using this ITS (Table 1). The mean positive predictivity for single assays and the mean Lhasa decision tree (Lhasa DT) ITS were very similar at 85% and 86%, respectively and the weight of evidence ITS was slightly higher at 90%.

The mean negative predictivity for single assays was relatively low at 57% but by using Derek Nexus with one other assay the negative predictivity improved to 75% (Fig. 4). The mean negative predictivity for the 2 out of 3 weight of evidence ITS is slightly lower at 69% but the 3 out of 3 concordance call ITS and Lhasa decision tree surpass both of these with a mean negative predictivity of 81% for each. This is attributed to a similar number of false negatives being generated for both these ITS within this data set (Table 1).

The coverage differs markedly depending on which ITS is evaluated (Fig. 5). There were significantly fewer test chemicals with LuSens results ($n = 77$) which was responsible for the low coverage outliers in each ITS. Overall, the mean coverage for each ITS was similar with the exception of the 3 out of 3 concordance call. The coverage was 31% which was significantly lower than the others (73%–80%) and was attributed to the reduced number of test chemicals, with 3 concordant results, used in this ITS ($n = 66$, see Table 1). Conversely, the accuracy for the 3 out of 3 concordance ITS was higher than the other ITS evaluated (Fig. 6). However, the reduction in number of compounds within the coverage of this ITS

must be taken into consideration; the 2 out of 3 weight of evidence and the Lhasa decision tree ITS have an accuracy of 84% and 85%, respectively, but have significantly higher coverage than the 3 out of 3 concordance ITS which has an accuracy of 92%. Remarkably the Lhasa decision tree has a comparable number of false negatives ($n = 5$) generated compared to the 3 out of 3 concordance ITS ($n = 3$) but almost 100 more test chemicals evaluated in the ITS (Table 1).

Similar trends were observed for the ITS against human data (data not shown) and results can be found in the Supporting Information.

4. Discussion

This study used a data set of 213 compounds to develop an ITS incorporating an *in silico* model, Derek Nexus, and at least one *in chemico/in vitro* assay from the DPRA, KeratinoSens, LuSens, h-CLAT and U-SENS. To improve the predictive performance of our ITS we designed a decision tree which after firing or not firing a skin sensitisation alert in Derek Nexus, the test chemical was assessed as in domain or out of domain of the *in chemico/in vitro* assays. We defined the *in chemico/in vitro* domain as test chemicals with a ClogP under 3.5 and/or test chemicals known to elicit skin sensitisation through a hapten mechanism with no prior activation required (as generated by Derek Nexus). Once divided by domain, results from *in chemico/in vitro* assays were used to evaluate the sensitisation potential of a test chemical. When all potential assay combinations were taken into account ($n = 20$), the mean accuracy was 85%, positive predictivity 86% and negative predictivity 81%.

4.1. Improvement in negative predictivity between ITS

Although the positive predictivity was not quite as high as for the 2 out of 3 weight of evidence or the concordance call (86% compared with 90% and 96%, respectively), the main advantage to using the Lhasa decision tree over the other ITS was the improvement in negative predictivity. The Lhasa decision tree and the 3 out

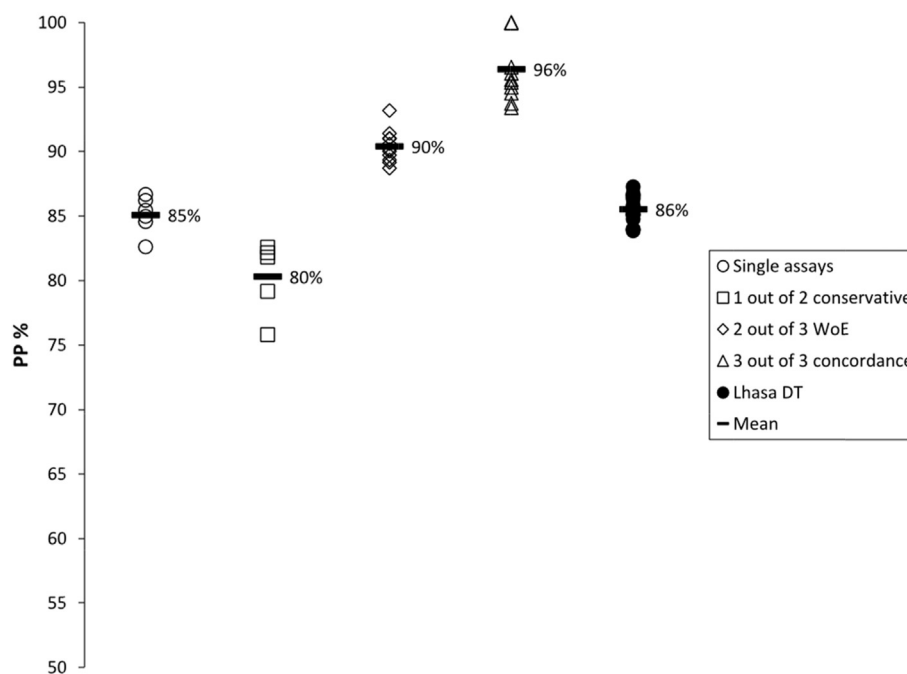


Fig. 3. Positive predictivity for single assays and integrated testing strategies against LLNA data.

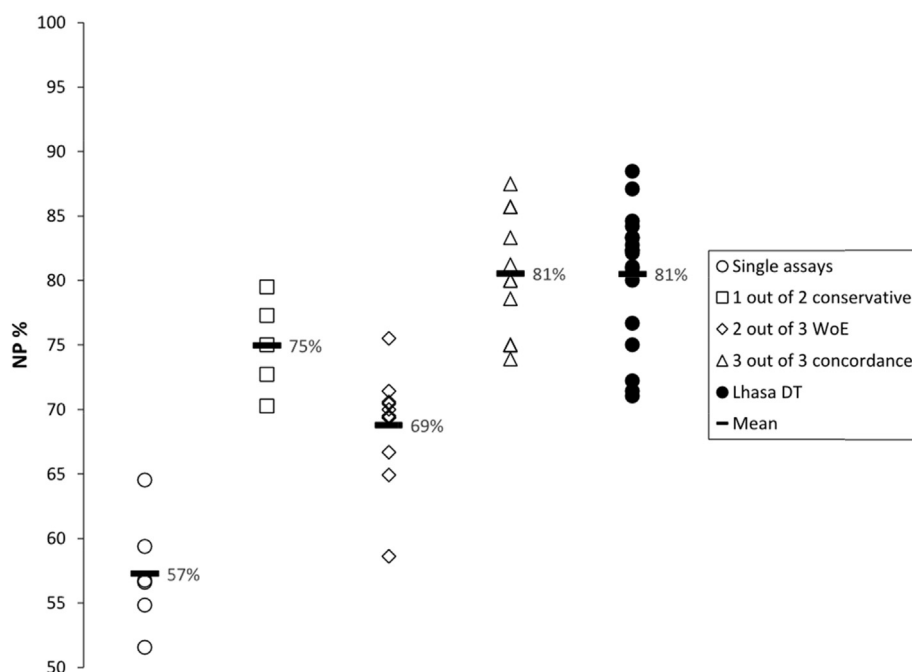


Fig. 4. Negative predictivity for single assays and integrated testing strategies against LLNA data.

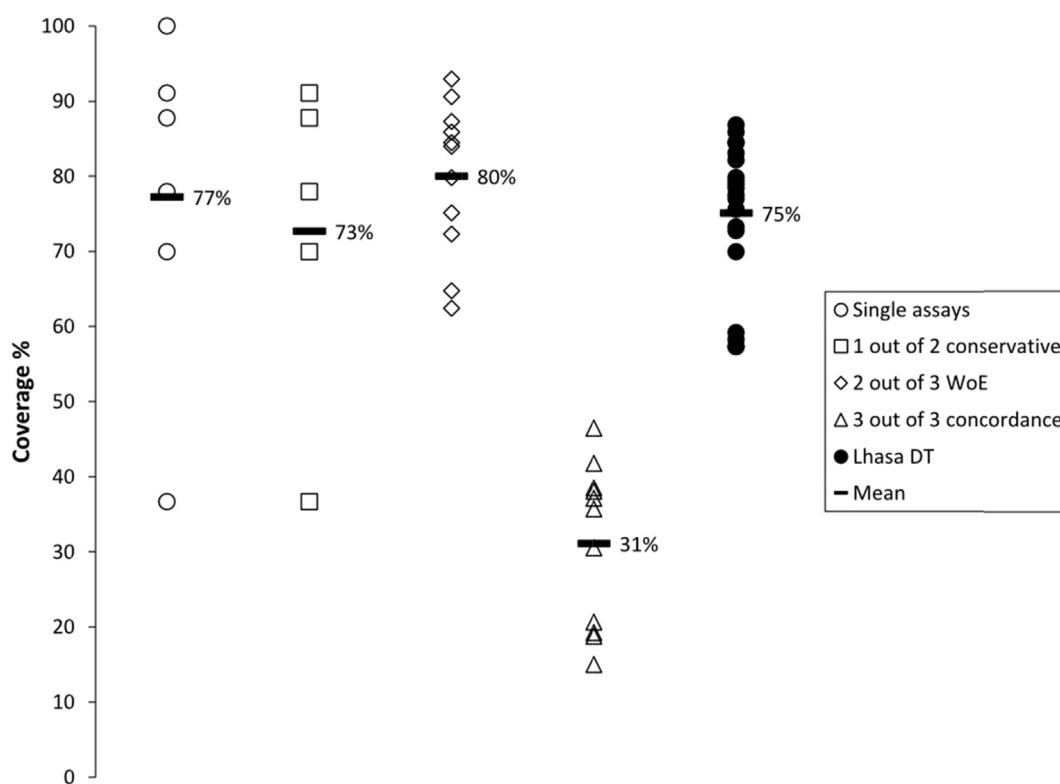


Fig. 5. Coverage of single assays and integrated testing strategies against LLNA data.

of 3 concordance ITS both demonstrated a mean negative predictivity of 81%, compared with 70% for the 2 out of 3 weight of evidence. The coverage of each ITS was similar (between 73% and 80%) with the exception of the 3 out of 3 concordance ITS. This was considerably lower than the other ITS evaluated, at 31% coverage, ascribed to the large number of test chemicals which were out of

domain of the ITS (as the results were discordant).

4.2. Reducing the number of assays required – comparison of Lhasa decision tree and 2 out of 3 weight of evidence

The Lhasa decision tree also has the potential to reduce the

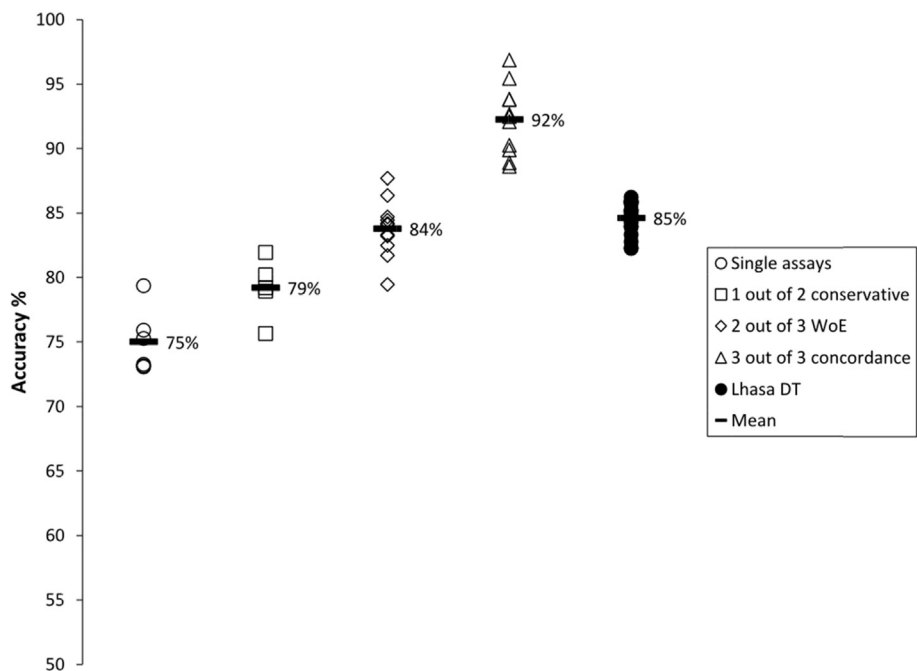


Fig. 6. Accuracy of single assays and integrated testing strategies against LLNA data.

Table 1
Mean false positives, false negatives and number of compounds used in each ITS.

ITS	Mean FP	Mean FN	Mean n in ITS
Single assays	16	26	165
Conservative call	24	8	155
Weight of evidence	11	15	170
Concordance call	2	3	66
Lhasa DT	19	5	160

number of assays run. Based on our data analysis, the number of *in chemico/in vitro* assays required for a specific test chemical is contingent on the Derek Nexus prediction and whether the chemical is likely to be in the *in chemico/in vitro* assay domain.

We compared the number of assays required for the Lhasa decision tree ITS against a 2 out of 3 weight of evidence ITS using DPRA, KeratinoSens (KS) and h-CLAT (Fig. 7). To illustrate, Derek Nexus (DX), DPRA and KeratinoSens were used in the Lhasa decision tree. 74 test chemicals fired an alert but were outside the assay domain – these were assigned as sensitisers without any additional assays performed. 90 test chemicals were within the *in chemico/in vitro* assay domain and were assigned an overall call based on the concordant Derek Nexus and one assay result. Only 19 chemicals within this data set had discordant Derek Nexus and first *in chemico/in vitro* results, and required another assay result to assign an overall call. Finally, 13 compounds were non-alerting in Derek Nexus and were outside the assay domain – these were assigned as equivocal. 17 compounds could not be evaluated as they lacked data for one or more of the *in chemico/in vitro* assays being assessed. The categorisation of each other ITS assay combination using Derek Nexus, DPRA, KeratinoSens and h-CLAT is shown in Fig. 7.

In terms of assay quantities, based on the analysis above for this data set ($n = 213$), between 426 and 639 assays would be required to assess a 2 out of 3 weight of evidence using only *in chemico/in vitro* assays, whereas the Lhasa decision tree would use a maximum of 128 assays for this data set. Consequently, using the Lhasa decision tree with Derek Nexus and two assays instead of a 2 out of 3

weight of evidence ITS would have a substantial impact on time-scale and cost.

4.3. Improvement in negative predictivity between Derek Nexus and Lhasa decision tree

The negative predictivity calculated for this data set was 65% using Derek Nexus alone and 81% when using the Lhasa decision tree ITS. 22 false negatives were generated when using Derek Nexus alone but 14 of these were correctly predicted by at least one Lhasa decision tree ITS, which improved the negative predictivity between the two models. The ITS only overturns a Derek Nexus prediction when two *in chemico/in vitro* assays are concordant thus test chemicals in this category could indicate an area of chemical space requiring refinement within Derek Nexus. The 14 false negatives mispredicted by Derek Nexus have been examined and 7 compounds have now been incorporated into existing alerts in the Derek Nexus knowledge base (released January 2016). The remaining seven false negatives are being investigated further.

4.4. Removing test chemicals used in Derek Nexus training set

As the Lhasa decision tree ITS was built on a publicly available data set, it is likely that some of the data have previously been used to train Derek Nexus. We found 43 test chemicals used as alert examples in Derek Nexus, removed them from the data set, and compared the performance to the full data set [data not shown].

We found that the negative predictivity improved from 81% to 84% and there was a slight decrease in positive predictivity and accuracy from 86% to 80% and from 85% to 81%, respectively.

4.5. Limitations of the Lhasa decision tree

The Lhasa decision tree we have designed takes the applicability domain of the *in chemico/in vitro* assays into account before assigning an overall call. The positive predictive performance of the Lhasa decision tree is high (86%) and has superior negative

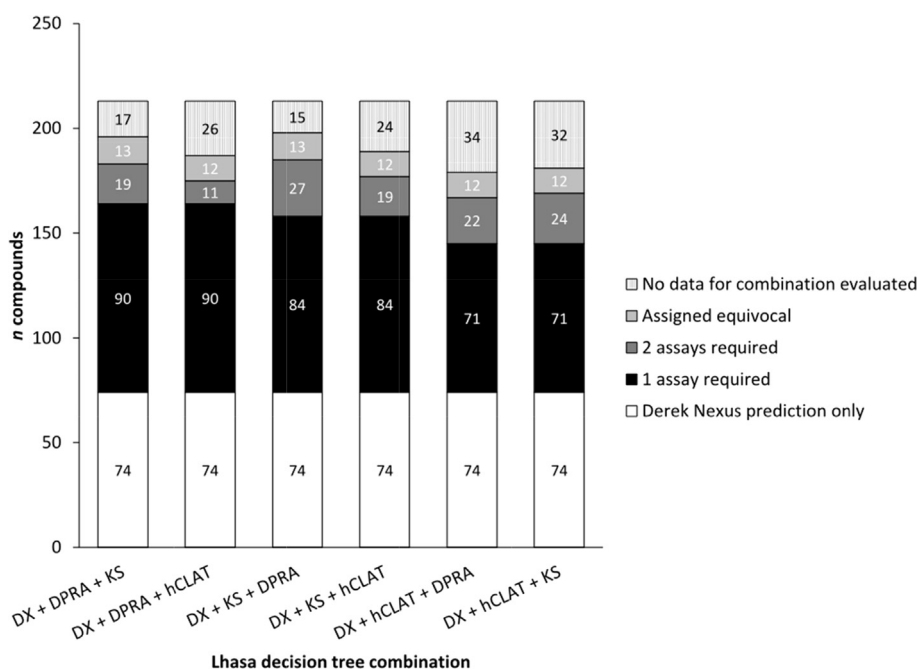


Fig. 7. Number of compounds in each category and how many *in chemico/in vitro* assays are required when applying the Lhasa decision tree ITS.

predictivity (81%) when compared to the other ITS evaluated in this study. Nevertheless, the Lhasa decision tree approach has some limitations:

- The data set ($n = 213$) on which the Lhasa decision tree ITS was based was relatively small (c.f. > 1000s compounds used to train Ames models), and the predictivities of some categories of overall call are based on a very small subset of the data. There is a distinct lack of *in chemico/in vitro* assay data available in the public domain and as more data become available, the Lhasa decision tree may require refinement or adjustment.
- Furthermore, the Lhasa decision tree has not yet been validated using an external data set due to the aforementioned scarcity of data.
- This is significant as a portion (20%) of the data set used to design the Lhasa decision tree has been used in the training set for Derek Nexus. This could potentially skew the Lhasa decision tree ITS results. However, all the compounds used in the data set have also been used to train, design and evaluate *in chemico/in vitro* assays, therefore, there is a chance that an unseen external data set could have a sub-optimal correlation to the non-animal assays. Our analysis above indicates that removing the training set data does not have a substantial detrimental effect on the Lhasa decision tree performance and actually improves the negative predictivity of the Lhasa decision tree.
- Finally, at the moment, there are 19 test chemicals which do not fire a Derek Nexus alert and are assessed as outside the *in chemico/in vitro* assay domain. These are currently assigned an overall call of equivocal. Notably, the predictivity of compounds which fall within this category was not significantly improved by using additional *in chemico/in vitro* assays. This may be a facet of this specific data set or could be an intrinsic feature of this Lhasa decision tree category – more data are required to resolve this. Importantly, it may signal that more and more *in chemico/in vitro* testing does not necessarily lead to greater clarity of skin sensitisation predictions.

We are eager to validate this Lhasa decision tree ITS with an

external data set, and have plans to do so in the near future. Moreover, as additional assay data become available in the public domain, it may become possible to expand the Lhasa decision tree for non-alerting compounds outside the assay applicability domain – assigning an overall call to test chemicals currently assigned as equivocal. Further work could also include incorporating the mechanistic domain into the Lhasa decision tree ITS with a view to indicate which assay combinations would be suitable for specific chemical classes. Finally, to facilitate risk assessment, future work will also focus on developing a prediction model which as well as assigning hazard, also assigns risk e.g. potency classification (EC3).

5. Conclusion

This study analysed 213 test compounds in a combination of different non-animal assays (DPRA, KeratinoSens, LuSens, h-CLAT and U-SENS). The results indicate that our Lhasa decision tree using Derek Nexus to complement two of the aforementioned *in chemico/in vitro* assays displays high concordance with the LLNA and is a superior method to assess skin sensitisation potential compared to the other ITS evaluated in this study. This is attributed to *in silico* models being able to take into account some of the intrinsic limitations of *in chemico/in vitro* assays: poor prediction of metals, insolubility of some test chemicals, lack of metabolic capacity within assays etc. Furthermore, by using Derek Nexus as well as *in chemico/in vitro* assays in a ITS, the user has access to all public data, expert commentary, mechanistic information and references assessed by Lhasa Limited experts to derive the alert scope.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.yrtph.2016.01.009>.

Transparency document

Transparency document related to this article can be found online at <http://dx.doi.org/10.1016/j.yrtph.2016.01.009>.

References

- Ashikaga, T., Sakaguchi, H., Sono, S., Kosaka, N., Ishikawa, M., Nukada, Y., Miyazawa, M., Ito, Y., Nishiyama, N., Itagaki, H., 2010. A comparative evaluation of in vitro skin sensitisation tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). *Altern. Lab. Anim.* 38, 275–284.
- Bauch, C., Kolle, S.N., 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. <http://dx.doi.org/10.1016/j.yrtph.2012.05.013>.
- Chipinda, I., Ruwona, T.B., Templeton, S.P., Siegel, P.D., 2011. Use of the human monocytic leukemia THP-1 cell line and co-incubation with microsomes to identify and differentiate hapten and prohaptens sensitizers. *Toxicology* 280, 135–143. <http://dx.doi.org/10.1016/j.tox.2010.12.004>.
- Cooper, J.A., Saracci, R., Cole, P., 1979. Describing the validity of carcinogen screening tests. *Br. J. Cancer* 39, 87–89. <http://dx.doi.org/10.1038/bjc.1979.10>.
- Divkovic, M., Pease, C.K., Gerberick, G.F., Basketter, D.A., 2005. Hapten-protein binding: from theory to practical application in the in vitro prediction of skin sensitization. *Contact Dermat.* 53, 189–200. <http://dx.doi.org/10.1111/j.0105-1873.2005.00683.x>.
- Ellison, C.M., Madden, J.C., Judson, P., Cronin, M.T.D., 2010. Using in silico tools in a weight of evidence approach to aid toxicological assessment. *Mol. Inf.* 29, 97–110. <http://dx.doi.org/10.1002/minf.200900006>.
- European Commission, 2015. EURL ECVAM Recommendation on the Human Cell Line Activation Test (H-CLAT) for Skin Sensitisation Testing. <http://dx.doi.org/10.2788/29986>.
- European Commission, 2014. EURL ECVAM Recommendation on the KeratinoSens Assay for Skin Sensitisation Testing. <http://dx.doi.org/10.2788/52914>.
- European Commission, 2013a. EURL ECVAM Strategy for Replacement of Animal Testing for Skin Sensitisation Hazard Identification and Classification. <http://dx.doi.org/10.2788/84214>.
- European Commission, 2013b. EURL ECVAM Recommendation on the Direct Peptide Reactivity Assay (DPRA) for Skin Sensitisation Testing. <http://dx.doi.org/10.2788/48229>.
- European Union, 2013. Commission regulation (EU) no 1197/2013 of 25 november 2013 amending annex III to regulation (EC) no 1223/2009 of the european parliament and of the Council on cosmetic products. *Off. J. Eur. Union* 56, 34–66. <http://dx.doi.org/10.3000/19770677.L.2013.315.eng>.
- Gerberick, G.F., Troutman, J.A., Foertsch, L.M., Vassallo, J.D., Quijano, M., Dobson, R.L.M., Goebel, C., Lepoittevin, J.-P., 2009. Investigation of peptide reactivity of pro-haptens skin sensitizers using a peroxidase-peroxide oxidation system. *Toxicol. Sci.* 112, 164–174. <http://dx.doi.org/10.1093/toxsci/kfp192>.
- Gubbels-van Hal, W.M.L.G., Blaauboer, B.J., Barentsen, H.M., Hoitink, M.A., Meerts, I.A.T.M., van der Hoeven, J.C.M., 2005. An alternative approach for the safety evaluation of new and existing chemicals, an exercise in integrated testing. *Regul. Toxicol. Pharmacol.* 42, 284–295. <http://dx.doi.org/10.1016/j.yrtph.2005.05.002>.
- Hirota, M., Fukui, S., Okamoto, K., Kurotani, S., Imai, N., Fujishiro, M., Kyotani, D., Kato, Y., Kasahara, T., Fujita, M., Toyoda, A., Sekiya, D., Watanabe, S., Seto, H., Takenouchi, O., Ashikaga, T., Miyazawa, M., 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. <http://dx.doi.org/10.1002/jat.3105>.
- 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. <http://dx.doi.org/10.1002/jat.2869>.
- Jowsey, I.R., Basketter, D.A., Westmoreland, C., Kimber, I., 2006. A future approach to measuring relative skin sensitising potency: a proposal. *J. Appl. Toxicol.* 26, 341–350. <http://dx.doi.org/10.1002/jat.1146>.
- Kimber, I., Basketter, D.A., Gerberick, G.F., Dearman, R.J., 2002. Allergic contact dermatitis. *Int. Immunopharmacol.* 2, 201–211. [http://dx.doi.org/10.1016/S1567-5769\(01\)00173-4](http://dx.doi.org/10.1016/S1567-5769(01)00173-4).
- Luechtefeld, T., Maertens, A., McKim, J.M., Hartung, T., Kleensang, A., Sá-Rocha, V., 2015. Probabilistic hazard assessment for skin sensitization potency by dose-response modeling using feature elimination instead of quantitative structure-activity relationships. *J. Appl. Toxicol.* 35, 1361–1371. <http://dx.doi.org/10.1002/jat.3172>.
- Macmillan, D.S., Canipa, S.J., Williams, R.V., 2015. Examining the role of in silico assessment in skin sensitisation integrated testing strategies. In: *Poster Presented at: Society of Toxicology 54th Annual Meeting March 22–26-San Diego, CA, USA*.
- McKim, J.M., Keller, D.J., Gorski, J.R., 2010. A new in vitro method for identifying chemical sensitizers combining peptide binding with ARE/EpRE-mediated gene expression in human skin cells. *Cutan. Ocul. Toxicol.* 29, 171–192. <http://dx.doi.org/10.3109/15569527.2010.483869>.
- Natsch, A., Emter, R., Ellis, G., 2008. Filling the concept with data: integrating data from different in vitro and in silico assays on skin sensitizers to explore the battery approach for animal-free skin sensitization testing. *Toxicol. Sci.* 107, 106–121. <http://dx.doi.org/10.1093/toxsci/kfn204>.
- Natsch, A., Emter, R., Gfeller, H., Haupt, T., Ellis, G., 2014. 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. <http://dx.doi.org/10.1093/toxsci/kfu229>.
- Natsch, A., Gfeller, H., Rothaupt, M., Ellis, G., 2007. Utility and limitations of a peptide reactivity assay to predict fragrance allergens in vitro. *Toxicol. Vitro* 21, 1220–1226. <http://dx.doi.org/10.1016/j.tiv.2007.03.016>.
- Natsch, A., Ryan, C.A., Foertsch, L., Emter, R., Jaworska, J., Gerberick, F., Kern, P., 2013. A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. *J. Appl. Toxicol.* 33, 1337–1352. <http://dx.doi.org/10.1002/jat.2868>.
- 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. Vitro* 27, 609–618. <http://dx.doi.org/10.1016/j.tiv.2012.11.006>.
- OECD, 2015a. Test No. 442C: In Chemico Skin Sensitisation: Direct Peptide Reactivity Assay (DPRA). OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing. <http://dx.doi.org/10.1787/20745788>.
- OECD, 2015b. Test No. 442D: In Vitro Skin Sensitisation: ARE-nrf2 Luciferase Test Method. OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing. <http://dx.doi.org/10.1787/9789264229822-en>.
- OECD, 2014. Draft Guideline: in Vitro Skin Sensitisation: Human Cell Line Activation Test (H-CLAT) (OECD Guidelines for the Testing of Chemicals, Draft Proposal).
- OECD, 2012. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins Part 1: Scientific Evidence. <http://dx.doi.org/10.1787/9789264221444-en>. OECD Environment, Health and Safety Publications Series on Testing and Assessment No. 168.
- 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. Vitro* 28, 1482–1497. <http://dx.doi.org/10.1016/j.tiv.2014.08.002>.
- Rorije, E., Aldenberg, T., Buist, H., Kroese, D., Schuurmann, G., 2013. The OSIRIS weight of evidence approach: ITS for skin sensitisation. *Regul. Toxicol. Pharmacol.* 67, 146–156. <http://dx.doi.org/10.1016/j.yrtph.2013.06.003>.
- Russell, W.M.S., Burch, R.L., 1959. *The Principles of Humane Experimental Technique*. Methuen, London.
- Takenouchi, O., Fukui, S., Okamoto, K., Kurotani, S., Imai, N., Fujishiro, M., Kyotani, D., Kato, Y., Kasahara, T., Fujita, M., Toyoda, A., Sekiya, D., Watanabe, S., Seto, H., Hirota, M., Ashikaga, T., Miyazawa, M., 2015. Test battery with the human cell line activation test, direct peptide reactivity assay and DEREK based on a 139 chemical data set for predicting skin sensitizing potential and potency of chemicals. *J. Appl. Toxicol.* 35, 1318–1332. <http://dx.doi.org/10.1002/jat.3127>.
- 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. <http://dx.doi.org/10.2131/jts.38.599>.
- Tsujita-Inoue, K., Atobe, T., Hirota, M., Ashikaga, T., Kouzuki, H., 2015. In silico risk assessment for skin sensitization using artificial neural network analysis. *J. Toxicol. Sci.* 40, 193–209. <http://dx.doi.org/10.2131/jts.40.193>.
- Tsujita-Inoue, K., Hirota, M., Ashikaga, T., Atobe, T., Kouzuki, H., Aiba, S., 2014. Skin sensitization risk assessment model using artificial neural network analysis of data from multiple in vitro assays. *Toxicol. Vitro* 28, 626–639. <http://dx.doi.org/10.1016/j.tiv.2014.01.003>.
- 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. <http://dx.doi.org/10.1016/j.yrtph.2014.12.008>.
- van der Veen, J.W., Rorije, E., Emter, R., Natsch, A., van Loveren, H., Ezenam, J., 2014. Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals. *Regul. Toxicol. Pharmacol.* 69, 371–379. <http://dx.doi.org/10.1016/j.yrtph.2014.04.018>.