



Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Assessing skin sensitization hazard in mice and men using non-animal test methods

Daniel Urbisch^a, Annette Mehling^b, Katharina Guth^a, Tzutzy Ramirez^a, Naveed Honarvar^a, Susanne Kolle^a, Robert Landsiedel^{a,*}, Joanna Jaworska^c, Petra S. Kern^d, Frank Gerberick^e, Andreas Natsch^f, Roger Emter^f, Takao Ashikaga^g, Masaaki Miyazawa^h, Hitoshi Sakaguchi^h

^a BASF SE, Experimental Toxicology and Ecology, Ludwigshafen, Germany^b BASF Personal Care and Nutrition GmbH, Düsseldorf, Germany^c Procter & Gamble NV, 100 Temselaan, 1853 Strombeek-Bever, Belgium^d Procter & Gamble Technology (Beijing) Co., Ltd., China^e Procter & Gamble Company, Cincinnati, OH, USA^f Givaudan Schweiz AG, Ueberlandstraße 138, CH-8600 Dübendorf, Switzerland^g Shiseido Research Center, Shiseido Co., Ltd., 2-2-1 Hayabuchi, Tsuzuki-ku, Yokohama-shi, Kanagawa 224-8558, Japan^h Safety Science Research Laboratories, Kao Corporation, 2606 Akabane, Ichikai, Haga, Tochigi 321-349, Japan

ARTICLE INFO

Article history:

Received 22 September 2014

Available online xxxx

Keywords:

Skin sensitization

Allergic contact dermatitis (ACD)

Adverse outcome pathway (AOP)

DPRA

KeratoSense™

LuSense

h-CLAT

(modified) MUSST

Database

Integrated testing strategy (ITS)

ABSTRACT

Sensitization, the prerequisite event in the development of allergic contact dermatitis, is a key parameter in both hazard and risk assessments. The pathways involved have recently been formally described in the OECD adverse outcome pathway (AOP) for skin sensitization. One single non-animal test method will not be sufficient to fully address this AOP and in many cases the use of a battery of tests will be necessary. A number of methods are now fully developed and validated. In order to facilitate acceptance of these methods by both the regulatory and scientific communities, results of the single test methods (DPRA, KeratoSense™, LuSense, h-CLAT, (m)MUSST) as well as for the simple '2 out of 3' ITS for 213 substances have been compiled and qualitatively compared to both animal and human data. The dataset was also used to define different mechanistic domains by probable protein-binding mechanisms. In general, the non-animal test methods exhibited good predictivities when compared to local lymph node assay (LLNA) data and even better predictivities when compared to human data. The '2 out of 3' prediction model achieved accuracies of 90% or 79% when compared to human or LLNA data, respectively and thereby even slightly exceeded that of the LLNA.

© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Allergic contact dermatitis (ACD¹) is the clinically relevant outcome of skin sensitization and it is estimated that 15–20% of the general population will be sensitized at some point in the course of their

lives (Thyssen et al., 2007; Bruckner et al., 2000). ACD can be associated with morbidity, affect the quality of life, and sensitization – the prerequisite for the development of ACD – is usually a lifelong effect. Sensitization is considered to be one of the key human health endpoints of toxicological assessments of a substance both in the

* Corresponding author. Fax: +49 621 60 58134.

E-mail address: robert.landsiedel@basf.com (R. Landsiedel).

¹ Abbreviations: ACD, allergic contact dermatitis; Ac, acylating agent; AOP, adverse outcome pathway; ARE, antioxidant response element; CD, cluster of differentiation; CV, cell viability; DC, dendritic cell; DPRA, direct peptide reactivity assay; EC, effective concentration; ECHA, European Chemical Agency; ECVAM, European Centre for the Validation of Alternative Methods; ESAC, ECVAM Scientific Advisory Committee; GPTs, guinea-pig based tests; FN, false negative; FP, false positive; h-CLAT, human cell-line activation test; IATA, integrated testing approaches and assessments; IC, inhibitory concentration; ITS, integrated testing strategy; KC, keratinocyte; Keap-1, Kelch-like ECH-associated protein 1; LC-MS, liquid chromatography coupled with mass spectrometry; LLNA, local lymph node assay; LMW, low molecular weight; MA, Michael acceptor; MHC, major histocompatibility complex; MIT, minimal induction threshold; (m)MUSST, (modified) myeloid U937 skin sensitization test; MTT, 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazoliumbromide; n = number of substances; NC, not classified; NQO1 = NADPH:quinone oxidoreductase 1; Nrf-2, nuclear factor (erythroid-derived 2)-like 2; OECD, Organization for Economic Co-operation and Development; QP, quinone precursor; QSAR, quantitative structure activity relationship; REACH, Registration, Evaluation, Authorization and Restriction of Substances; RFI, relative fluorescence intensity; RIFM, Research Institute for Fragrance Materials; SB, Schiff base former; S_N1/2, substances reacting by nucleophilic substitutions of type 1 or 2; S_NAr, Aromatics reacting by nucleophilic substitutions; TN, true negative; TP, true positive; WoE, weight of evidence.

occupational and consumer setting. Sensitization and ACD are the result of a complex multifactorial sequence of events. The chemical and biological pathways involved are relatively well characterized and can be simplified and structured into an adverse outcome pathway (AOP). The key events involved have now been formally described by the OECD in a document titled “The Adverse Outcome Pathway for Skin Sensitization Initiated by Covalent Binding to Proteins” with the goal of facilitating the development of methods and approaches addressing the relevant events (ENV/JM/MONO(2012)10/PART 1 and 2).

Contact allergies develop in two stages: (1) the sensitization phase in which antigen/allergen specific T-cells are generated, and (2) the elicitation phase in which renewed contact with the allergen leads to the allergic response (Goebel et al., 2012). Following entry into the skin, the AOP described by the OECD identifies eleven events involved whereby four are considered to be key events. The initiating event of the sensitization process is the covalent binding of a hapten, a low molecular weight (LMW) and typically electrophilic substance, to the skin proteins. This is the essential event to transform an otherwise non-immunogenic molecule into a potential allergen. The protein reactivity of the substance may be inherent or can develop following metabolic or abiotic transformations of the pro- or pre-haptens, respectively (Gerberick et al., 2008; Jäckh et al., 2012). Key event 2 is the activation of keratinocytes. Keratinocytes are the main cell population of the epidermis and are the first cells to come into contact with the potential allergen. Keratinocytes may respond to the contact with the hapten and/or the complete allergen. Among these responses, the oxidative and electrophilic stress-driven expression of genes under the control of the antioxidant response element (ARE) as part of the Keap1/Nrf2 pathway is well described (Natsch and Emter, 2008). The third key event is the activation of dendritic cells. Dendritic cells take up and process antigens and present fragments in form of major histocompatibility complexes (MHC) on their surfaces. Dendritic cells mature during this process and migrate to the lymph nodes. Mature dendritic cells are characterized by the up-regulation of cell surface markers such as CD54 and CD86 in order to activate naïve T-cells. The final event of the sensitization phase is the proliferation of the antigen-specific T-cells and the generation of antigen-specific memory T-cells (Goebel et al., 2012).

In the regulatory context, the skin sensitization potential of a chemical has traditionally been evaluated using animal tests, in particular the guinea-pig based tests (GPTs) described in OECD 406 or the mouse-based local lymph node assay (LLNA) described in OECD 429 (or OECD 442 A + B). The 7th amendment of the Cosmetic Directive (Council Directive 76/768/EEC of 1976-07-27; now Cosmetics Regulation: REGULATION (EC) No. 1223/2009), implemented a phasing out of animal testing for the purposes of this legislation. This in turn has made the development and use of non-animal tests for new cosmetic substances indispensable. Other legislations have followed or are in the process of following suit, and explicitly permit animal testing only as a last resort, e.g. the current European substances legislation Nr. 1907/2006 [Registration, Evaluation, Authorization and Restriction of Chemicals (REACH)]. Of the human health endpoints to be assessed under REACH, skin sensitization must be evaluated for all substances to be registered. In REACH phase 1 more than 3700 substances were assessed for skin sensitization (Angers-Loustau et al., 2011) and several thousand more are expected for REACH phase 3 until 2018.

The current consensus among the scientific community is that one single non-animal test might not be sufficient as a stand-alone method to cover the endpoint skin sensitization and that the use of an integrated testing strategy (ITS; in this context ITS includes sequential testing strategies (STS), etc.) will be necessary (Mehling et al. 2012, Rovida et al. 2014; Basketter et al., 2013).

Currently, two tests, namely the direct peptide reactivity assay (DPRA) (Gerberick et al., 2004) and the ARE-Nrf2 luciferase test method KeratinoSens™ (Emter et al., 2010) have passed the validation process at ECVAM and a statement of the ECVAM Scientific Advisory Committee (ESAC) has been published regarding their use within integrated testing approaches and assessments (IATA) (EURL ECVAM, 2013; EURL ECVAM, 2014); the OECD draft guidelines have been drafted and are in the final stages of the commutation process. The ECVAM recommendation and ESAC statement as well as an OECD draft guideline are available for a third test, namely the human cell-line activation test (h-CLAT) (Ashikaga et al., 2006). In addition, validation studies are ongoing for the LuSens assay (Bauch et al., 2012; Ramirez et al., 2014), an ARE-Nrf2 luciferase test method similar to the KeratinoSens™. Data on over a hundred substances is available for the myeloid U937-based skin sensitization tests MUSST (myeloid U937 skin sensitization test) (Ade et al., 2006) and the modified MUSST (mMUSST) (Bauch et al., 2011, 2012). The MUSST is also in the validation at ECVAM (ECVAM Test Method Submissions 2008–2014).

Integrated approaches to testing and assessment (IATA) and in particular the integrated testing strategies (ITS) used within the IATA should consist of methods with a mechanistic relevance for the endpoint being assessed. Placed in the context of the AOP, the DPRA evaluates key event 1 – the protein/peptide reactivity of a substance, the KeratinoSens™ and LuSens assays represent key event 2 and give a measure keratinocyte activation, and the h-CLAT, MUSST and mMUSST describe key event 3 – dendritic cell activation. Thus, when used together, they cover the first three of the four key events of the sensitization process, thus being of mechanistic relevance and supporting the scientific rationale for using a combination of these methods in an AOP-based ITS. Indeed, these tests are already being successfully used in combination in a number of ITS which include, but are not limited to, the ‘2 out of 3’ approach (also sometimes termed majority vote or ‘2 out of 3’ weight of evidence (WoE) approach (Bauch et al., 2012; Natsch et al., 2013; van der Veen et al., 2014)), an ITS based on Bayesian Networks (Jaworska et al., 2013) and an ITS which includes an assessment of potency (Nukada et al., 2013; Tsujita-Inoue et al., 2014). Currently, the OECD in conjunction with EURL ECVAM is developing a document describing the key elements of an (AOP-based) IATA and designing templates for reporting an ITS. According to the roadmap proposed by EURL ECVAM submission of the document to the OECD should take place in 2016 (Kinsner-Ovaskainen et al., 2012). Whether this document will be accepted in time for the testing phase preceding the REACH phase 3 (substances with an annual production tonnage of 1–100 tons) deadline is currently unknown.

As risk assessments are based on hazard assessments, the first goal is to adequately address the skin sensitization hazard potential of a substance. A study conducted by ECVAM on all registered substances in the new substances data base, revealed that 2745 of the evaluated 3792 substances were not classified (NC; approx. 72%) for sensitization and that 1047 (about 28%) were classified as sensitizers (Angers-Loustau et al., 2011). This indicates that if a sensitization potential (hazard) can be excluded, further information on potency for risk assessments or classification according to the Globally Harmonized System (GHS) would only be needed for approximately 30% of the substances. Regulatory acceptance, e.g. for GHS classification and/or REACH registration, and the use of non-animal tests for safety assessments is critically dependent on the correct predictivities of a method or an ITS. This in turn is supported by the availability of data allowing the evaluation of the types of substances tested and the limitations/strengths of the methods or ITS. Only such data driven analysis will help to build confidence to use the new approaches. This was also a key message from a workshop at which regulators and industry dis-

cussed how to best facilitate acceptance of non-animal methods and ITS (Basketter et al., 2013).

This manuscript compiles and evaluates the available data – both previously published and unpublished – obtained with the above mentioned *in vitro* methods to help facilitate the acceptance of non-animal testing approaches for the toxicological endpoint ‘skin sensitization’. The predictivities of the individual methods are evaluated as is their use in one ITS for skin sensitization hazard assessment, namely the ‘2 out of 3’ approach. The evaluation was limited to this example of an ITS as the design of an ITS needs to be adapted depending on the assessments being made, e.g. an ITS for risk assessments will usually differ from an ITS for hazard assessments. Additionally, flexibility is needed, e.g. if the ITS is part of a read across approach, if some methods may not be available at contract research institutes, or if an ITS has proprietary elements. Therefore, only the weight of evidence based ‘2 out of 3’ approach for hazard assessments which is based on the first three key events of the adverse outcome pathway was assessed. In this study, the applicability of the test methods and ITS for different substance properties and mechanisms of reactivity was also assessed. The compilation of non-animal test method results for over 200 substances now provides a very comprehensive data base which exceeds that used for the formal validation of the “benchmark” animal method – the LLNA.

2. Materials and methods

2.1. Data set

The data set consists of both previously published (Natsch et al., 2013; Nukada et al., 2012; Bauch et al., 2012) and additional unpublished data for 213 test substances. All substances, physico-chemical properties, types of use, test data, proposed reaction mechanisms as well as related references and data bases are listed in [Supplementary Table \(supporting information\)](#). Fig. 1 shows the use categories of these substances.

For 208 substances, high quality LLNA data were described in the literature or data bases; for 5 substances hitherto unpublished LLNA data were used. Of the 213 substances evaluated in this study, 151 (71%) are considered to be sensitizers and 62 (29%) to

be non-sensitizers according to available LLNA data. In addition, human data were available for 114 of these substances. For all substances within this data compilation, results were available for at least two of the investigated non-animal test methods. The set of non-animal test data comprises data originating from the DPRA (results for 199 substances), KeratinoSens™ (results for 195 substances), LuSens (results for 77 substances), h-CLAT (results for 166 substances), MUSST (results for 145 substances) and mMUSST data (results for 65 substances). In order to create mechanistic domains from a chemical perspective, the putative reaction mechanisms for peptide reactivity of the 213 substances was evaluated using the freely available OECD QSAR Toolbox v3.2 and searching the scientific literature.

2.2. Non-animal test methods

2.2.1. Direct peptide reactivity assay (DPRA)

Peptide reactivity data were generated using a method to measure reactivity of a test chemical with model hepta-peptides containing lysine (Ac-RFAAKAA-COOH) or cysteine (Ac-RFAACAACOOH) (Gerberick et al., 2004). Peptide reactivity was reported as percent depletion based on the decrease in non-reacted peptide concentration in the sample relative to the average concentration measured in the control.

Peptides were prepared and purified by the SynPep Corporation (Dublin, CA, USA) to >90% purity as measured by HPLC, and molecular weight confirmation was determined by flow injection positive-ion electrospray mass spectrometry. Briefly, 400 µL of a 1.25 mM peptide stock solution prepared in buffer and a 100 mM test chemical stock solution prepared in either acetonitrile or DMSO/acetonitrile were added to 100 mM ammonium acetate buffer (pH 10.2) for the lysine peptide or 100 mM sodium phosphate buffer (pH 7.5) for the cysteine peptide. The final reaction, containing 0.5 mM of the peptide and 5 or 25 mM of the test chemical, representing 1:10 and 1:50 M ratios, was mixed and incubated in the dark for 24 h at 25 °C. Control samples and standards used for defining the calibration curve for each analysis were prepared without test chemical for each peptide and ranged from 0.0156 to 1.0 mM. All samples were prepared in triplicate. Following incubation, the peptide was quantified by reverse-phase HPLC (Waters 2695 Alliance) on a Zorbax SB-C18 column (3.5 µm, 100 × 2.1 mm) with UV detection at 220 nm (Waters 996 PDA detector) using an external standard linear calibration curve. The UV spectrum was collected from 210 to 400 nm to permit verification of the peptide peak identity. Results for 199 substances were generated by P&G (referred to in [Supplementary Table](#) as DPRA I) or BASF (DPRA II) and thus available for this study. Five substances were not considered for further analyses due to discordant results in the two independent labs.

2.2.2. KeratinoSens™ assay

The standard operating procedure described (Natsch et al., 2011) and published online (ECVAM, 2014) was used to test additional substances in the KeratinoSens™ assay. Briefly, cells were grown for 24 h in 96-well plates. The medium was then replaced with medium containing the test substance and a final level of 1% of the solvent DMSO. Each test substance was subsequently tested at 12 twofold dilutions (0.98–2000 µM). In each repetition, three parallel replicate plates were run for luciferase determination and a fourth parallel plate was prepared for cytotoxicity determination. Cells were incubated for 48 h with the test substances, and then luciferase activity and cytotoxicity (with the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazoliumbromide (MTT) assay (Mosmann, 1983)) were determined. For each chemical the EC1.5, EC2 and EC3 values (concentration in µM for 1.5, 2 and 3-fold induction of the luciferase activity) were calculated along with

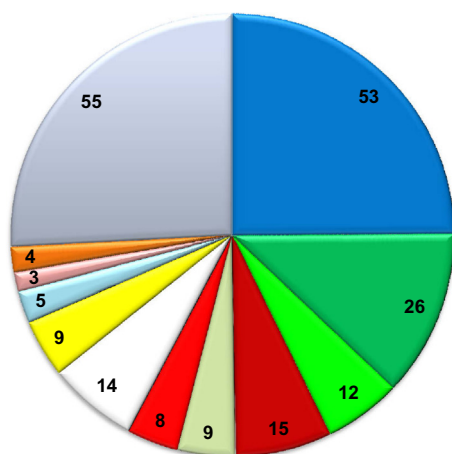


Fig. 1. Schematic representation of the uses of the investigated substances. The numbers of substances for each use are indicated in parentheses as well as in the pie chart itself. Blue – fragrances (n = 53), dark green – preservatives/disinfectants (n = 26), light green – dyes (n = 12), dark red – monomers (n = 15), mint green – pesticides (n = 9), red – solvents (n = 8), white – cosmetics (n = 14), yellow – pharmaceuticals (n = 9), light blue – surfactants (n = 5), pink – plasticizers (n = 3), orange – food/feed (n = 4), gray – other uses (n = 55).

IC50 values for the concentration yielding 50% reduction in cellular viability. Substances were tested in at least two independent experiments. A substance is considered to have a sensitizing potential if an induction equal to or exceeding 1.5-fold compared to the vehicle control is observed at a concentration below 1000 μ M and at which cells remain >70% viable. If the results of the two experiments were concordant, a prediction according to the prediction model was derived. Substances with discordant results or results close to the 1.5-fold threshold (borderline) were tested in additional independent experiments. The number of experiments and the number of positive results for each chemical is given in the database. Results for 195 substances were generated by Givaudan (referred to in [Supplementary Table](#) as KeratinoSens assay I) or BASF (KeratinoSens assay II) and thus available for this study. Eight substances were not considered for further analyses due to discordant results in the two independent labs.

2.2.3. LuSens assay

The LuSens assay is a keratinocyte-based assay which utilizes the luciferase gene under the control of the antioxidant response element (ARE) originating from the rat NQO1 gene as the reporter construct. The LuSens assay used in this paper is similar to that described in [Bauch et al., 2012](#) with some modifications ([Ramirez et al., 2014](#)). In brief, a cytotoxicity range finding experiment (consisting of twelve concentrations) was performed, to calculate the concentration in which cell viability corresponds to no less than 75% (CV75). Following the range finder experiment, a main experiment was set up using six concentrations of test substance (in triplicates), the highest tested concentration was $1.2 \times$ CV75 (or 2000 μ M if no cytotoxicity was observed). After 48 h treatment, luminescence and cytotoxicity were measured. A test substance is considered to have skin sensitization potential when the luciferase induction is above or equal to 1.5-fold compared to the vehicle control in two (or more than) consecutive non-cytotoxic tested concentrations whereby at least three tested concentrations must be non-cytotoxic (viability $\geq 70\%$). A test compound is considered not to have sensitizing potential if the above effects are not observed. The skin sensitization potential of a test substance is determined by the result of the majority of the repetitions of an experiment. If two of two or two of three repetitions are negative/positive, the substance is considered as negative/positive. In order to come to a conclusion on the skin sensitization hazard of a substance, one complete experiment needs to be conducted. A complete experiment consists of two valid independent repetitions ([Ramirez et al., 2014](#)). Results for 77 substances were obtained by BASF and considered for this study.

2.2.4. Human cell line activation test (h-CLAT)

In the h-CLAT assay, THP-1 cells (American Type Culture Collection, Manassas, VA, USA) were used as surrogate for dermal dendritic cells. For dose finding, cytotoxicity tests were conducted and the concentration resulting in 75% cell viability, termed CV75, was calculated based on the analysis of viable cells. THP-1 cells were treated with eight different concentrations, decided based on dose finding cytotoxicity test, for 24 h. After removing the test substance, the expression of CD86 and CD54 on the cell surface was measured by flow cytometry. The relative fluorescence intensity (RFI) was used as an indicator of CD86 and CD54 expression. If the RFI of CD86 or CD54 was greater than 150% or 200% at any dose in at least two out of three experiments, the substance was judged as a sensitizer. Otherwise, it was considered a non-sensitizer ([Ashikaga et al., 2006](#)). From the dose-dependency curves of three experiments, the median concentration inducing 150% of CD86 RFI and/or 200% of CD54 RFI (EC150 or EC200) was calculated like EC3 value determination in the LLNA. The lower EC value was defined as minimal induction threshold (MIT) ([Nukada et al.,](#)

[2013](#)). Results for 166 substances were available for this study and for the determination of Cooper statistics; data generated by the respective method developer was used.

2.2.5. Myeloid U937 skin sensitization test (MUSST) and modified MUSST

The MUSST uses the U937 cell-line purchased from the American Type Culture Collection (Rockville, MD, USA). Four to six concentrations are chosen based on preliminary propidium iodide cytotoxicity experiments and are applied in duplicate for 48 h. The highest tested concentration in the main experiment is twice the concentration causing a cytotoxicity of 25% (CV75) determined in a pretest. A test substance is predicted to have a dendritic cell activating potential indicative of being a sensitizer when CD86 induction (measured by flow cytometry) exceeds the threshold of 1.5-fold with respect to vehicle treated cells at any tested concentration showing sufficient cell viability ($\geq 70\%$) in at least two independent experiments ([Natsch et al., 2013](#)). Results for 145 substances were made available for this study by P&G (referred to in [Supplementary Table](#) as U-937 Test).

A modified version of the MUSST (mMUSST) uses the U937 cell line from German Resource Center for Biological Material DSMZ, Braunschweig, Germany. In the mMUSST, a test substance is predicted to have a dendritic cell activating potential when CD86 induction exceeds a threshold of 1.2-fold ([Bauch et al., 2012](#)). Data for 65 substances were generated in the mMUSST by BASF (referred to in [Supplementary Table](#) as mMUSST).

For the analyses within this study, the results from the MUSST and mMUSST were taken together to create a dataset of 161 substances. Data for 12 substances were not further considered for analyses due to discordant results being obtained in both tests.

2.2.6. '2 out of 3' prediction model

The least complicated way to assess the skin sensitization hazard potential of a substance is to use the results of single assays which reflect key steps of the AOP within a '2 out of 3' prediction model. For the assays addressing the three key events described in the OECD AOP on skin sensitization mentioned above, a '2 out of 3' assessment was introduced for the first time by [Bauch et al. \(2012\)](#). In the current study, this prediction model was applied using DPRA, KeratinoSens™ and h-CLAT data. Any two congruent results of the three tests rule the overall assessment: If at least two of the three assays were positive, the substance was rated to be a skin sensitizer. If at least two of the three assays were negative, the substance was rated to be a non-sensitizer. The classification as a sensitizer or non-sensitizer is therefore based on a weight of evidence pertaining to key events of the AOP. Cooper statistics for this classification were determined in comparison to LLNA or human data. Results for 180 or 101 substances, respectively, were obtained using this prediction model.

2.2.7. OECD QSAR Toolbox

The QSAR Toolbox developed by OECD in collaboration with the European Chemical Agency (ECHA) ([Raunio, 2011](#)) is a standalone software application for filling gaps of (eco)toxicity data that are needed to assess the potential hazards of substances. *In silico* analysis was performed with the OECD QSAR Toolbox in the version 3.2., which is freely available on the OECD website (<http://toolbox.oasis-lmc.org/?section=download&version=latest>). In order to support the identification of a chemical's toxicity, the OECD QSAR Toolbox contains numerous databases with results from experimental studies for over 55,000 substances and profilers for calculating specific properties (references are added at each profiler within this tool). Two such profilers are the protein-binding profilers based on OECD and OASIS algorithms ("Protein binding by OECD", "Protein binding by OASIS v1.2"). In order to also identify

substances, which require abiotic or metabolic activation, the “auto-oxidation profiler” and the “skin metabolism profiler” were used Teubner et al., 2013).

2.2.8. Cooper statistics

The Cooper statistics of predictivity were calculated for the single assays as well as for the ‘2 out of 3’ prediction model (Cooper et al., 1979). All parameters are based on a 2×2 contingency table counting the number of compounds that are “true positive” (TP), “false positive” (FP), “true negative” (TN) and “false negative” (FN):

$$\text{Sensitivity [\%]} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100$$

$$\text{Specificity [\%]} = \frac{\text{TN}}{\text{TN} + \text{FP}} * 100$$

$$\text{Positive predictive value [\%]} = \frac{\text{TP}}{\text{TP} + \text{FP}} * 100$$

$$\text{Negative predictive value [\%]} = \frac{\text{TN}}{\text{TN} + \text{FN}} * 100$$

$$\text{Accuracy [\%]} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} * 100$$

3. Results

3.1. Availability of in vivo data

Due to the fact that human data were available for only a limited number of substances, two different sets of substances were defined (Table 1). Murine LLNA data were available for all 213 substances (set A) whereas additional human data were available for a subset of 114 substances (subset B). No reliable human data were found in the literature or data bases for 99 substances (data were not further analyzed for this group). In subset B, direct comparisons of non-animal test results to both human and LLNA data were possible.

3.2. Predictivities compared to LLNA and human data

Detailed results (including concentration–response data, cytotoxicity information, peptide depletions) of the investigated assays are listed in Supplementary Table (supporting information). When calculating the Cooper statistics (Table 2), the assessments were reduced to yes/no ratings according to the prediction models of the individual assays. It should be noted that data for all 213 substances was not available for every assay. The number of tested substances being considered for further analyses hence ranges from $n = 77$ (LuSens) to $n = 194$ (DPRA). Accuracies of the single assays compared to LLNA data range from 73% to 76% (Table 2,

set A). The ‘2 out of 3’ approach (DPRA, KeratinoSens™ and h-CLAT data) provides an accuracy of 79% when compared to LLNA data.

Using the data from subset B, a direct comparison between the predictivities of the non-animal test methods and human or LLNA data is possible (Table 3). Interestingly, when comparing the Cooper statistics of almost all single non-animal test methods with human data, the percentages were higher than comparisons with LLNA data (Table 3). The ‘2 out of 3’ approach (DPRA, KeratinoSens™ and h-CLAT data) provides somewhat higher accuracies compared to the predictivities of the single assays, whereby the accuracies compared to human or LLNA data were 90% or 82%, respectively. These results indicate that the investigated non-animal methods generally predict human data more accurately than LLNA data.

To determine the performance of the LLNA to predict skin sensitization hazard of this data set, LLNA data were directly compared to the available human data for the substances within subset B. A total of 20 (18%) of the substances were incorrectly identified as FN or FP by the LLNA. Seven substances were assessed to be FN (benzaldehyde, benzyl alcohol, coumarin, streptomycin sulfate, kanamycin, benzocaine and nickel chloride) and 13 to be FP (pyridine, limonene, isopropyl myristate, citronellol, linalool, α -iso-methylionone, benzyl benzoate, SDS, tocopherol, benzyl salicylate, hexyl salicylate, xylene and phthalic anhydride). In total, comparison to human data revealed a sensitivity of 91%, a specificity of 64% and an overall accuracy of 82% of the LLNA (Table 3). For comparison, the ‘2 out of 3’ prediction model revealed a sensitivity of 90%, a specificity of 90% and an overall accuracy of 90% for the same set of substances suggesting predictions to be more relevant for humans.

3.3. Comparing results of single assays in different labs

3.3.1. Interlaboratory reproducibilities

For some of the 213 substances the DPRA, KeratinoSens™ and (m)MUSST were conducted in two labs enabling tentative analyses of interlaboratory reproducibilities. Within this study, 45 substances were tested in the DPRA in two labs (P&G and BASF). Among these, 40 substances gave congruent (89%) and five substances gave incongruent results. 52 Substances were tested in the KeratinoSens™ assay in two labs (Givaudan and BASF). Here, 44 substances gave congruent (85%) and eight substances gave incongruent qualitative outcomes. Comparing MUSST and mMUSST results obtained at P&G and BASF, the analyses of 45 substances resulted in 32 congruent (71%) and 13 incongruent qualitative outcomes (this comparison needs to be viewed in the light of slightly different protocols and use of the 1.5-fold and 1.2-fold CD86 induction cut-off, respectively). Substances with incongruent results were excluded from subsequent analyses.

3.3.2. Interchangeabilities

Since both the KeratinoSens™ and the LuSens address keratinocyte activation and the h-CLAT and (m)MUSST address dendritic cell activation, interchangeabilities of the assays representing the same AOP key event were analyzed based on predictivities of the assays and in the ‘2 out of 3’ approach.

A set of 69 substances was tested in the KeratinoSens™ (Givaudan) and the LuSens assay (BASF). The same overall outcome for 61 of the 69 substances resulted in an interchangeability of 88%. The discordant substances among the two ARE activation assays are listed in Table 11. The KeratinoSens™ offered an accuracy of 75% or 83% and the LuSens an accuracy of 71% or 79%, when comparing the results to LLNA or human data, respectively. The implementation of the KeratinoSens™ into a ‘2 out of 3’ approach with DPRA and h-CLAT data for the set of 69 substances resulted in an accuracy of 85% or 91%, when comparing the predictions to LLNA or

Table 1

Sensitizers and non-sensitizers among substances with LLNA data and with human data on skin sensitization.

Chemical set and reference data	Set A	Subset B	
	LLNA data	Human data	LLNA data
Sensitizer	151	75	83
Non-sensitizer	62	36	31
Borderline	0	3	0
Total	213	114	114

(A) All substances within the data set ($n = 213$) with available LLNA data. (B) Substances within the data set ($n = 114$) with available human data.

Table 2

Cooper statistics for the single non-animal test methods and the '2 out of 3' approach for the whole data set (set A).

Chemical set and reference data	Set A LLNA data			
	Sensitivity [%]	Specificity [%]	Accuracy [%]	n
'2 out of 3' approach	82	72	79	180
DPRA	76	72	75	194
KeratinoSens™	77	63	73	188
h-CLAT	81	64	76	166
LuSens	76	67	73	77
(m)MUSST	73	74	73	150

n = number of substances analyzed.

Table 3

Cooper statistics for the single non-animal test methods and the '2 out of 3' approach when compared to human and LLNA data (subset B).

Chemical set and reference data	Subset B Human data ^a				LLNA data			
	Se [%]	Sp [%]	Acc [%]	n	Se [%]	Sp [%]	Acc [%]	n
'2 out of 3' approach	90	90	90	101	81	83	82	103
DPRA	84	84	84	102	77	85	79	105
KeratinoSens™	82	84	82	102	74	73	74	103
h-CLAT	89	64	82	98	86	68	81	101
LuSens	78	79	79	60	73	70	71	62
(m)MUSST	74	88	78	85	71	83	75	87
LLNA	91	64	82	111	–	–	–	–

Se = sensitivity; Sp = specificity; Acc = accuracy; n = number of substances analyzed.

^a α -Amyl cinnamic aldehyde, α -hexyl cinnamic aldehyde and 6-methyl coumarin could not finally be evaluated as human sensitizers or non-sensitizers and were thus rated as "borderline".**Table 4**

Substances with discordant test results among KeratinoSens™ and LuSens.

Name	Cas #	KeratinoSens™			LuSens		
		Final	vs. LLNA	vs. Human	Final	vs. LLNA	Vs. Human
Farnesal	502-67-0	1	TP	–	0	FN	–
Ethylenediamine	107-15-3	1	TP	TP	0	FN	FN
Propyl gallate	121-79-9	1	TP	TP	0	FN	FN
Nickel chloride	7718-54-9	1	FP	TP	0	TN	FN
Methyl salicylate	119-36-8	0	TN	TN	1	FP	FP
Benzyl alcohol	100-51-6	0	TN	TN	1	FP	TP
Eugenol	97-53-0	0	FN	FN	1	TP	TP
α -Hexyl cinnamic aldehyde	101-86-0	Bord.	–	–	1	TP	–

1, positive overall outcome; 0, negative overall outcome; bord., borderline outcome; TN/TP, true negative/positive; FN/FP, false negative/positive.

human data, respectively. If the LuSens assay was used for this approach instead of the KeratinoSens™, Cooper statistics estimated an accuracy of 83% or 93%, when compared to LLNA or human data, respectively.

Regarding the h-CLAT (Kao and Shiseido) and the (m)MUSST (BASF and P&G), a common set of 105 substances was tested to investigate dendritic cell activation. The same overall outcome for 76 of 105 substances resulted in an interchangeability of 72%. The 29 substances with discordant results can be found in [Supplementary Table \(supporting information\)](#). The h-CLAT offered an accuracy of 80% or 86% and the (m)MUSST an accuracy of 75% or 78%, when comparing the predictions to LLNA or human data, respectively. The implementation of the h-CLAT into a '2 out of 3' approach with DPRA and KeratinoSens™ data for the set of 105 substances resulted in an accuracy of 85% or 91%, when comparing the predictions to LLNA or human data, respectively. If the (m)MUSST was used for this approach instead of the h-CLAT, Cooper statistics estimated an accuracy of 81% or 90%, when compared to LLNA or human data, respectively.

3.4. Predictivities for groups of substances

3.4.1. Mechanistic domains by protein-binding mechanisms

Several means for grouping of substances into specific chemical classes are possible, and the chosen approach may depend on the specific purpose (Aptula et al., 2005). The molecular initiating event (MIE) starting the skin sensitization process is the binding of haptens to proteins. In order to analyze if the non-animal test methods may detect classes of substances with different reaction mechanisms with similar performances, nine different mechanistic domains were defined (Table 4). The probable protein-binding mechanisms were assigned by the OECD QSAR Toolbox v3.2. Protein-binding mechanisms described in the scientific literature were added to build a mechanistic chemistry framework for the 213 substances (Roberts and Aptula, 2014; Karlberg et al., 2013; Aptula et al., 2009, 2007, 2005; Patlewicz et al., 2008; Roberts et al., 2007a) (Supplementary Table). For 11 substances, two possible protein-binding mechanisms were proposed to be probable. For example, maleic anhydride is supposed to react both as an acylat-

Table 5

Proposed mechanistic domains assigned by the OECD QSAR Toolbox, literature search and expert judgment as well as number and percentage of positive LLNA results within each mechanistic domain.

Protein-binding mechanism	n	n of positives in LLNA
No protein-binding mechanisms assigned (no alert)	65	17 (26%)
Acylation agents (Ac)	21	19 (91%)
Michael acceptors (MA)	45	41 (91%)
Quinone precursors (QP)	22	19 (86%)
Schiff base formers (SB)	30	25 (83%)
Substances reacting by nucleophilic substitutions of type 1 or 2 (S _N 1/2)	30	26 (87%)
Aromatics reacting by nucleophilic substitutions (S _N Ar)	6	6 (100%)
Substances reacting by a nucleophilic addition	1	1 (100%)
Metal ions forming coordination bonds	4	2 (50%)
Total (including two possible protein-binding mechanism for 11 substances)	224	–

n = number of substances analyzed.

ing agent and as a Michael acceptor. Furthermore, in some cases it was not possible to assess whether a chemical may react either in a S_N1 or S_N2 reaction. Thus, both organic reaction types were placed into one mechanistic domain. With exception of the one substance reacting in a nucleophilic addition and the four metal containing substances forming coordination bonds, all mechanistic domains are described in more detail in the following sections.

3.4.1.1. No reaction mechanism assigned (no alert). This substance group contains 65 substances whose molecular structures do not contain any obvious alerts for protein reactivity. Many substances in this group represent aliphatic alcohols, organic acids, nitriles, amides, α,β -unsaturated esters or simple dialkylketones. Accuracies for the non-animal tests calculated within this substance group were compared to LLNA and human data. One aspect which should be taken into account is that the predictivities are influenced by the number and type of substances being assessed and data sets may therefore vary in this study. DPRA data, which directly indicate peptide reactivity, were available for 61 substances in this group. The accuracy of the DPRA was 67% when compared to LLNA data or 76% when compared to human data. Regarding keratinocyte activation, the KeratinoSens™ offered an accuracy of 65% or 82% and the LuSens an accuracy of 60% or 80%, when comparing the predictions to LLNA or human data, respectively. Regarding dendritic cell activation, the accuracy of the h-CLAT was 67% or 59% and that of the (m)MUSST was 68% or 82% when compared to LLNA or human data, respectively. Using the '2 out of 3' approach, 73% or 80% of the substances were correctly identified in this domain when compared to LLNA or human data, respectively. The number of tested substances and predictions is given in Table 5.

3.4.1.2. Acylation agents (Ac). In an acylation, acyl moieties can be transferred to the nucleophilic centers of proteins. This mechanistic domain contains 21 acylation agents like esters of acidic alcohols (e.g. phenol esters) or carboxylic anhydrides. Accuracies for the non-animal tests calculated for the substances within this domain were compared to LLNA and human data. The accuracy of the DPRA was 100% or 82% when compared to LLNA or human, respectively. The ARE-based assays did not allow a reliable prediction of the sensitization potential of substances in this domain. The KeratinoSens™ offered an accuracy of 56% or 58% and the LuSens an accuracy of 44% or 50%, when comparing the predictions to LLNA or human data, respectively. Regarding dendritic cell activation test methods, Cooper statistics estimated an accuracy of 88% or 83% for the h-CLAT and an accuracy of 69% or 56% for the (m)MUSST, when compared to LLNA or human data, respectively. Using the '2 out of 3' approach, 88% or 83% of the substances were correctly identified in this domain when compared to LLNA or

Table 6

Total number of analyzed substances, number of TP, FN, TN, FP and accuracies for each non-animal test method and the '2 out of 3' approach for substances whose molecular structures do not contain any obvious alerts for protein reactivity.

No reaction mechanism assigned (n = 65)		Set A	Subset B	
		LLNA data	Human data	LLNA data
DPRA	n total	61	34	34
	TP/FN	5/11	2/4	3/8
	TN/FP	36/9	24/4	20/3
	Accuracy [%]	67	76	68
KeratinoSens™	n total	55	33	33
	TP/FN	5/9	3/3	2/8
	TN/FP	31/10	24/3	19/4
	Accuracy [%]	65	82	64
LuSens	n total	30	25	25
	TP/FN	1/5	2/1	1/4
	TN/FP	17/7	18/4	15/5
	Accuracy [%]	60	80	64
h-CLAT	n total	48	29	29
	TP/FN	10/4	3/3	6/3
	TN/FP	22/12	14/9	14/6
	Accuracy [%]	67	59	69
(m)MUSST	n total	47	28	28
	TP/FN	4/6	3/2	3/5
	TN/FP	28/9	20/3	17/3
	Accuracy [%]	68	82	71
'2 out of 3' approach	n total	51	30	30
	TP/FN	7/8	2/3	3/7
	TN/FP	30/6	22/3	18/2
	Accuracy [%]	73	80	70

n, number of substances analyzed; TN/TP, true negative/positive; FN/FP, false negative/positive.

human data, respectively. The number of substances, accuracies and number of correct or incorrect predictions can be found in Table 6.

3.4.1.3. Michael acceptors (MA). MAs contain α,β -unsaturated ester, ketone or aldehyde functions. α,β -unsaturated alcohols can also react as MAs after the alcohol group is oxidized to an aldehyde (Karlberg et al., 2013). Due to their high probability to react as MAs, quinone precursors could also be placed into this domain (Aptula et al., 2009), but are handled as a separate substance domain on account of their specific structural characteristics. This mechanistic domain contains 45 Michael acceptors. Accuracies for the non-animal tests calculated within this mechanistic domain were compared to LLNA and human data. In this domain, the DPRA provided an accuracy of 81% when compared to LLNA data or 86% when compared to human data. The KeratinoSens™ offered an accuracy of 85% or 100% and the LuSens an accuracy of 88% or

100%, when comparing the predictions to LLNA or human data, respectively. Regarding dendritic cell activation, Cooper statistics estimated an accuracy of 92% or 90% for the h-CLAT and an accuracy of 86% or 80% for the (m)MUSST, when compared to LLNA or human data, respectively. Using the '2 out of 3' approach, 88% or 95% of the substances were correctly identified in this domain when compared to LLNA or human data, respectively. The number of substances, accuracies and number of correct or incorrect predictions can be found in Table 7.

3.4.1.4. Quinone precursors. This mechanistic domain contains di- or poly-substituted aromatic compounds with alkoxy, hydroxyl or amine residues. These compounds are pre- or pro-haptens and thus require abiotic or metabolic oxidation to form quinones, quinone imines or quinone methides. These oxidation products are most likely to react as Michael acceptors (free radical binding via Wuerster-type radical may also be possible) (Aptula et al., 2009). 22 substances were identified as possible quinone precursors. Accuracies for the non-animal test methods calculated within this mechanistic domain were compared to LLNA and human data. In this domain, the DPRA provided an accuracy of 71% when compared to LLNA data or 91% when compared to human data. Regarding keratinocyte activation, the KeratinoSens™ offered an accuracy of 79% or 90% and the LuSens an accuracy of 71% or 71%, when comparing the predictions to LLNA or human data, respectively. However, only 7 substances (set A and B) in this domain have LuSens data. Regarding dendritic cell activation, Cooper statistics estimated an accuracy of 94% or 91% for the h-CLAT and an accuracy of 83% or 80% for the (m)MUSST, when compared to LLNA or human data, respectively. Using the '2 out of 3' approach, 83% or 91% of the substances were correctly identified in this domain when compared to LLNA or human data, respectively. The number of substances, accuracies and number of correct or incorrect predictions can be found in Table 8.

Table 7

Total number of analyzed substances, number of TP, FN, TN, FP and accuracies for each non-animal test method and the '2 out of 3' approach for possible acylating agents.

Acylating agents (n = 21)		Set A	Subset B	
		LLNA data	Human data	LLNA data
DPRA	n total	19	11	11
	TP/FN	17/0	9/1	10/0
	TN/FP	2/0	0/1	1/0
	Accuracy [%]	100	82	100
KeratinoSens™	n total	18	12	12
	TP/FN	10/6	6/5	5/6
	TN/FP	0/2	1/0	0/1
	Accuracy [%]	56	58	42
LuSens	n total	9	6	6
	TP/FN	4/4	2/3	2/4
	TN/FP	0/1	1/0	0/0
	Accuracy [%]	44	50	33
h-CLAT	n total	16	12	12
	TP/FN	12/2	9/2	9/2
	TN/FP	2/0	1/0	1/0
	Accuracy [%]	88	83	83
(m)MUSST	n total	16	9	9
	TP/FN	9/5	5/4	5/3
	TN/FP	2/0	0/0	1/0
	Accuracy [%]	69	56	67
'2 out of 3' approach	n total	16	12	12
	TP/FN	12/2	9/2	9/2
	TN/FP	2/0	1/0	1/0
	Accuracy [%]	88	83	83

n, number of substances analyzed; TN/TP, true negative/positive; FN/FP, false negative/positive.

3.4.1.5. Schiff 'base formers' (SB). Aldehydes and activated ketones predominantly react with hard nucleophiles to form imines (Schiff 'bases'). Pre/pro-SB formers like primary amines were also included in this domain (Foussereau et al., 1983) as well as 1,3 dicarbonyl compounds due to their tendency to enolize (Roberts et al., 2007b). Aromatic aldehydes of the general formula ArCHO were assigned to this domain, although a non-sensitizing effect of this alert is discussed in the literature (Patlewicz et al., 2001). In total, 30 of the 213 substances were supposed to react as Schiff 'base formers. Regarding peptide reactivity, the DPRA provided an accuracy of 65% when compared to LLNA data or 77% when compared to human data. The KeratinoSens™ offered an accuracy of 66% or 79% and the LuSens an accuracy of 75% or 86%, when comparing the predictions to LLNA or human data, respectively. However, only 8 (set A) or 7 substances (subset B) in this domain have LuSens data. Regarding dendritic cell activation, Cooper statistics estimated an accuracy of 75% or 93% for the h-CLAT and an accuracy of 64% or 75% for the (m)MUSST, when compared to LLNA or human data, respectively. Using the '2 out of 3' approach, 69% or 92% of the substances were correctly identified in this domain when compared to LLNA or human data, respectively. The number of substances, accuracies and number of correct or incorrect predictions can be found in Table 9.

3.4.1.6. Nucleophilic substitutions ($S_N1/2$). Nucleophilic substitutions exchange a moiety of a molecule (leaving group) for an attacking nucleophilic group. This can be a concerted reaction (S_N2) or stepwise with the intermediate formation of a carbocation (S_N1). Since a clear distinction of the reaction order is not always possible and the reaction products are identical (except for stereochemistry), both reaction types were placed into one mechanistic domain. The accuracy of the DPRA was 73% when compared to LLNA data or 100% when compared to human data. The KeratinoSens™ offered an accuracy of 88% or 83% and the LuSens an accu-

Table 8

Total number of analyzed substances, number of TP, FN, TN, FP and accuracies for each non-animal test method and the '2 out of 3' approach for possible Michael acceptors.

Michael acceptors (n = 45)		Set A	Subset B	
		LLNA data	Human data	LLNA data
DPRA	n total	42	22	22
	TP/FN	32/6	18/3	18/3
	TN/FP	2/2	1/0	1/0
	Accuracy [%]	81	86	86
KeratinoSens™	n total	41	21	21
	TP/FN	35/2	20/0	19/1
	TN/FP	0/4	1/0	0/1
	Accuracy [%]	85	100	90
LuSens	n total	17	10	10
	TP/FN	15/1	10/0	10/0
	TN/FP	0/1	0/0	0/0
	Accuracy [%]	88	100	100
h-CLAT	n total	37	20	20
	TP/FN	30/3	18/1	19/0
	TN/FP	4/0	0/1	1/0
	Accuracy [%]	92	90	100
(m)MUSST	n total	29	15	15
	TP/FN	22/3	12/3	12/2
	TN/FP	2/1	0/0	1/0
	Accuracy [%]	83	80	87
'2 out of 3' approach	n total	43	22	22
	TP/FN	36/3	20/1	20/1
	TN/FP	2/2	1/0	1/0
	Accuracy [%]	88	95	95

n, number of substances analyzed; TN/TP, true negative/positive; FN/FP, false negative/positive.

Table 9

Total number of analyzed substances, number of TP, FN, TN, FP and accuracies for each non-animal test method and the '2 out of 3' approach for possible quinone precursors.

Quinone precursors (n = 22)		Set A	Subset B	
		LLNA data	Human data	LLNA data
DPRA	n total	21	11	11
	TP/FN	14/5	10/1	10/1
	TN/FP	1/1	0/0	0/0
	Accuracy [%]	71	91	91
KeratinoSens™	n total	19	10	10
	TP/FN	14/3	9/1	9/1
	TN/FP	1/1	0/0	0/0
	Accuracy [%]	79	90	90
LuSens	n total	7	7	7
	TP/FN	4/2	4/2	4/2
	TN/FP	1/0	1/0	1/0
	Accuracy [%]	71	71	71
h-CLAT	n total	16	11	11
	TP/FN	13/1	9/1	9/1
	TN/FP	2/0	1/0	1/0
	Accuracy [%]	94	91	91
(m)MUSST	n total	18	10	10
	TP/FN	14/2	8/2	8/2
	TN/FP	1/1	0/0	0/0
	Accuracy [%]	83	80	80
'2 out of 3' approach	n total	18	11	11
	TP/FN	14/3	10/1	10/1
	TN/FP	1/0	0/0	0/0
	Accuracy [%]	83	91	91

n, number of substances analyzed; TN/TP, true negative/positive; FN/FP, false negative/positive.

racy of 100% or 100%, when comparing the predictions to LLNA or human data, respectively. However, only 6 (set A) or 5 substances (subset B) in this domain have LuSens data. Regarding dendritic cell activation, Cooper statistics estimated an accuracy of 59% or 92% for the h-CLAT and an accuracy of 64% or 75% for the (m)MUSST, when compared to LLNA or human data, respectively. Using the '2 out of 3' approach, 76% or 100% of the substances were correctly identified in this domain when compared to LLNA or human data, respectively. The number of substances, accuracies and number of correct or incorrect predictions can be found in Table 10.

3.4.1.7. Nucleophilic substitutions in aromatic compounds (S_NAr). A specific case of nucleophilic substitution takes place in aromatic compounds with electron-withdrawing groups. Within this study, only six substances were assigned to this domain whereby only two had human data. Therefore, Cooper statistics were only calculated against LLNA data (set A). The accuracies of the single assays ranged from 75% to 100% (DPRA: 100%, KeratinoSens™: 83%, LuSens: 100%, h-CLAT: 100%, (m)MUSST: 75%) and the accuracy of the '2 out of 3' approach achieved 100% within this mechanistic domain.

3.4.1.8. Pre- and pro-haptens. The class of pre- and pro-haptens represents substances which require abiotic (e.g. via autoxidation) or biotic (e.g. via metabolic pathways) activation, respectively, in order to become electrophilic. Opposite to pro-haptens, the activation of pre-haptens could be prevented by precautionary measures in the handling and storage of the substances (Gerberick et al., 2008). Nevertheless, pre- and pro-haptens are not always distinct since autoxidation and metabolic oxidation can result in the same product, although the underlying mechanisms may differ (Karlberg et al., 2013). Since pre-haptens may also be pro-haptens and vice versa no distinction between substances requiring biotic and

Table 10

Total number of analyzed substances, number of TP, FN, TN, FP and accuracies for each non-animal test method and the '2 out of 3' approach for Schiff 'base formers.

Schiff 'base formers (n = 30)		Set A	Subset B	
		LLNA data	Human data	LLNA data
DPRA	n total	26	13	13
	TP/FN	15/7	12/1	11/1
	TN/FP	2/2	0/0	0/0
	Accuracy [%]	65	77	85
KeratinoSens™	n total	29	14	14
	TP/FN	18/7	11/3	10/3
	TN/FP	1/3	0/0	0/1
	Accuracy [%]	66	79	71
LuSens	n total	8	7	7
	TP/FN	6/1	5/1	5/1
	TN/FP	0/1	1/0	1/0
	Accuracy [%]	75	86	86
h-CLAT	n total	24	14	14
	TP/FN	15/5	12/1	11/1
	TN/FP	3/1	1/0	1/1
	Accuracy [%]	75	93	86
(m)MUSST	n total	22	12	12
	TP/FN	13/6	9/3	8/3
	TN/FP	1/2	0/0	0/1
	Accuracy [%]	64	75	67
'2 out of 3' approach	n total	26	13	13
	TP/FN	17/6	12/1	11/1
	TN/FP	1/2	0/0	0/1
	Accuracy [%]	69	92	85

n, number of substances analyzed; TN/TP, true negative/positive; FN/FP, false negative/positive.

abiotic activation was made. The data set contains at least 30 pre/pro-haptens (compare Supplementary Table). The '2 out of 3' approach could be performed for 25 substances within this domain and resulted in 21 correct predictions, when compared to LLNA data (accuracy = 84%).

4. Discussion

With the compilation of currently available non-animal test method results for the distinction of skin sensitizers and non-sensitizers the database could be enlarged to a set of 213 substances. Included in the present dataset are 54 and 145 substances published previously by Bauch et al. (2012) and Natsch et al. (2013), respectively, whereby 41 substances were reported in both studies.

4.1. Predictivities

Compared to the predictivity described by Bauch et al. (2012) and Natsch et al. (2013), the performance of the investigated non-animal test methods is generally slightly lower when analyzing the data set of 213 substances (Table 12). For example, the DPRA provided an accuracy of 79% for 52 substances in Bauch et al. (2012) and 75% for 194 substances in the current study (both compared to LLNA data). This difference may simply reflect the variations occurring with various datasets on the one hand, but on the other hand newly developed methods are often first compared against 'gold-standard' lists of chemicals with the most consistent *in vivo* evidence from multiple tests and test methods, while upon expansion of the database chemicals are included for which, e.g. only a single LLNA study is available. In this regard, the degree of variability of the *in vivo* method needs to be kept in mind. In the light of hazard prediction 15.7% of the investigated non-sensitizers were falsely predicted as sensitizers in repeated experiments and 3.1% of sensitizers were falsely predicted as

Table 11

Total number of analyzed substances, number of TP, FN, TN, FP and accuracies for each non-animal test method and the '2 out of 3' approach for substance reacting in nucleophilic substitutions of type 1 or 2.

Nucleophilic substitutions ($S_N1/2$) ($n = 30$)		Set A	Subset B	
		LLNA data	Human data	LLNA data
DPRA	<i>n</i> total	26	11	11
	TP/FN	18/4	9/0	9/0
	TN/FP	1/3	2/0	0/2
	Accuracy [%]	73	100	82
KeratoSens™	<i>n</i> total	26	12	12
	TP/FN	22/0	10/0	10/0
	TN/FP	1/3	0/2	2/0
	Accuracy [%]	88	83	100
LuSens	<i>n</i> total	6	5	5
	TP/FN	6/0	5/0	5/0
	TN/FP	0/0	0/0	0/0
	Accuracy [%]	100	100	100
h-CLAT	<i>n</i> total	27	12	12
	TP/FN	14/9	9/1	9/2
	TN/FP	2/2	2/0	0/1
	Accuracy [%]	59	92	75
(m)MUSST	<i>n</i> total	17	9	9
	TP/FN	10/5	7/1	7/1
	TN/FP	2/0	1/0	0/1
	Accuracy [%]	64	75	78
'2 out of 3' approach	<i>n</i> total	25	12	12
	TP/FN	18/3	10/0	10/0
	TN/FP	1/3	2/0	0/2
	Accuracy [%]	76	100	83

n, number of substances analyzed; TN/TP, true negative/positive; FN/FP, false negative/positive.

non-sensitizers (Hoffmann, 2014). Similarly, by retesting 22 LLNA performance standards in the standard LLNA protocol, a reproducibility of only 77% was found for the investigated substances (Kolbe et al., 2013).

In addition to evaluating the results of the single test methods, the overall yes/no-ratings of the '2 out of 3' prediction models from the previously published studies and the current study was compared (Table 12). The '2 out of 3' prediction model in Bauch et al. (2012) consisted of the DPRA, LuSens and mMUSST. An overall accuracy of 83% and 94% compared to LLNA or human data, respectively, was achieved. Natsch et al. (2013) used the DPRA, KeratoSens™ and MUSST in their '2 out of 3' prediction model and

reported a similar accuracy for 145 substances (81% compared to LLNA). In the current study, the results of the DPRA, KeratoSens™ and h-CLAT were used for applying this prediction model. Table 12 illustrates, that all three '2 out of 3' prediction models provide very similar and high overall accuracies. Even though the dataset was expanded from 54 to 180 substances the accuracy is still similar when compared to the LLNA; 79% compared to 83% in the Bauch et al. study and 81% in the Natsch et al. study. Compared to human data, the accuracies reported in Bauch et al. (2012) and in the current study are very similar (94% and 90%, respectively), although the current analysis is based on an almost twofold larger dataset. The higher accuracy when comparing to human instead of LLNA data indicates that the investigated non-animal test methods predict human data more accurately than LLNA data.

In addition, the direct comparison of LLNA data to human data revealed an overall accuracy of 82% of the LLNA. The '2 out of 3' prediction model however revealed an accuracy of 90% when compared to human data for the same set of substances. This indicates that the '2 out of 3' predictions might be more relevant for humans than the LLNA predictions. However, some of the substances with different *in vivo* outcomes are pre-haptens (e.g. limonene, citronellol or linalool). This could be one reason for discordant LLNA and human results, since pre-hapten activation could be prevented by precautionary measures in handling or storage (Gerberick et al., 2008). But this does not imply that pre-haptens are not possibly a risk for humans. Substances leading to different predictions in the LLNA, human tests and the '2 out of 3' are discussed in more detail in the following part.

4.2. False negative and false positives

The discussion of false predictions provided by the single non-animal test methods would be beyond the scope of this study. Hence, the analysis was restricted to the detailed investigation of false positive and false negative predictions by the '2 out of 3' approach, when compared to LLNA and human data. The 28 substances leading to false negative results are divided into 4 groups and possible explanations for the discrepancy are discussed individually (Table 13). In brief, in group FN-1 negative human data confirm the negative overall results of the non-animal test methods for 11 substances (the LLNA is over-predictive for this set of substances). For the three substances in group FN-2 (diethylenetriamine, resorcinol and benzoyl peroxide) concordant human and

Table 12

Performances of the investigated non-animal test methods and the '2 out of 3' approach in different datasets.

		Bauch et al. (2012)		Natsch et al. (2013)		This paper	
		Acc [%]	<i>n</i>	Acc [%]	<i>n</i>	Acc [%]	<i>n</i>
<i>Compared to LLNA data</i>							
Peptide reactivity	DPRA	79	54	80	145	75	194
KC activation	KeratinoSens™	81	54	77	145	73	188
	LuSens	77	54	–	–	73	77
DC activation	(m)MUSST	74	54	71	141	73	149
	h-CLAT	–	–	–	–	76	166
<i>Compared to human data</i>							
Peptide reactivity	DPRA	86	51	–	–	84	102
KC activation	KeratinoSens™	80	51	–	–	82	102
	LuSens	84	51	–	–	79	61
DC activation	(m)MUSST	86	51	–	–	78	85
	h-CLAT	–	–	–	–	82	98
<i>Prediction model</i>							
'2 out of 3' approach ^a (vs. LLNA data)		83	54	81	145	79	180
'2 out of 3' approach ^a (vs. human data)		94	51	–	–	90	101

Acc, accuracy; *n*, number of analyzed substances; KC, keratinocyte; DC, dendritic cell; "–", no data available or data not considered in this study.

^a '2 out of 3' prediction model in Bauch et al. (2012) (DPRA, LuSens, mMUSST) and Natsch et al. (2013) (DPRA, KeratoSens™, MUSST) were slightly different compared to the prediction model underlying this paper (DPRA, KeratoSens™, h-CLAT).

Table 13

False negative predictions within the '2 out of 3' approach.

Chemical name	CAS #	LLNA EC3	Discussion
<i>Group FN-1</i>	'2 out of 3' approach (–), LLNA (+), human (–)		
Phthalic anhydride	85-44-9	0.16	Respiratory sensitizer, but no skin sensitizer (Dearman et al., 2013); LLNA probably FP
α -iso-Methylionone	127-51-5	21.8	In HRIPT ^a no induction at 70,866 $\mu\text{g}/\text{cm}^2$; LLNA probably FP
Xylene	1330-20-7	95.8	No evidence for skin sensitization in humans (Basketter et al., 2014); LLNA probably FP
Pyridine	110-86-1	71.2	No evidence for skin sensitization in humans (Schneider and Akkan, 2004); LLNA probably FP
Isopropyl myristate	110-27-0	44	No evidence for skin sensitization in humans (Basketter et al., 2014); LLNA probably FP
Linalool	78-70-6	30	No or weak evidence in humans with no induction in HRIPT ^a at 15,000 $\mu\text{g}/\text{cm}^2$; LLNA possibly FP
Sodium dodecyl sulfate (SDS)	151-21-3	14	Well characterized false positive response in LLNA, no evidence for skin sensitization in humans (Basketter et al., 2014); LLNA probably FP
Hexyl salicylate	6259-76-3	0.18	In HRIPT ^a no induction at 35,433 $\mu\text{g}/\text{cm}^2$ and in HMT ^a no induction at 20,654 $\mu\text{g}/\text{cm}^2$; LLNA probably FP
D,L- α -Tocopherol	10191-41-0	7.4	No evidence for skin sensitization in humans (Basketter et al., 2014); LLNA probably FP
Benzyl benzoate	120-51-4	17	No or only weak evidence in humans (Basketter et al., 2014) with no induction in HRIPT ^a at 59000 $\mu\text{g}/\text{cm}^2$; LLNA probably FP
Benzyl salicylate	118-58-1	2.9	No or only weak evidence in humans (Basketter et al., 2014); in HRIPT ^a no induction at 17,717 $\mu\text{g}/\text{cm}^2$ and in HMT ^a no induction at 20,690 $\mu\text{g}/\text{cm}^2$; LLNA probably FP
<i>Group FN-2</i>	'2 out of 3' approach (–), LLNA (+), human (+)		
Benzoyl peroxide	94-36-0	0.004	Well-known contact allergen in humans (Basketter et al., 2014), clearly positive in DPRA
Resorcinol	108-46-3	5.5	Known contact allergen in humans after considerable exposure (Basketter et al., 2014); positive in h-CLAT and MUSST ^b
Diethylenetriamine	111-40-0	5.8	Positive in HMT (Kligman, 1966b), all non-animal tests negative, putative pro-hapten
<i>Group FN-3</i>	'2 out of 3' approach (–), LLNA (+), human (no or conflicting data)		
α -Hexyl cinnamic aldehyde	101-86-0	11.97	Inconclusive human data; borderline in KeratinoSens TM , positive in LuSens ² and MUSST ^b
N,N-Dibutylaniline	613-29-6	19.6	Putative pro-hapten, all non-animal tests negative
2-Methoxy-4-methylphenol	93-51-6	5.8	Putative pro-hapten, positive in h-CLAT and MUSST ^b
3-Aminophenol	591-27-5	3.2	Putative pro-hapten, positive in h-CLAT and MUSST
2,2,6,6-Tetramethyl-3,5-heptanedione	1118-71-4	27	All non-animal tests negative
3-Methyl-1-phenylpyrazolone	89-25-8	8.5	All non-animal tests negative
Undec-10-enal	112-45-8	6.8	Positive in KeratinoSens TM
Squaric acid diethyl ester	5231-87-8	0.9	All non-animal tests negative
Methyl pyruvate	600-22-6	2.4	All non-animal tests negative
Benzyl cinnamate	103-41-3	18.4	Positive in KeratinoSens TM
<i>Group FN-4</i>	'2 out of 3' approach (–), LLNA (–), human (+)		
Coumarin	91-64-5	NC	Well-known contact allergen in humans (Basketter 2014-62), positive in KeratinoSens TM
Streptomycin sulfate	3810-74-0	NC	Positive in HMT (Kligman, 1966a)
Kanamycin	8063-07-8	NC	Known contact allergen in humans after considerable exposure (Schneider and Akkan, 2004)
Benzyl alcohol	100-51-6	NC	Positive in HMT ^a at 8858 $\mu\text{g}/\text{cm}^2$

^a HRIPT and HMT data were found in the data base of the Research Institute for Fragrance Materials (RIFM).^b DPRA, KeratinoSensTM and h-CLAT results were considered for '2-out-of-3' approach.

LLNA data are available. In this case, *in vivo* evidence gives a strong hint that the negative outcome in the '2 out of 3' approach is false. Diethylenetriamine and resorcinol are putative pro-haptens. They are not detectable in the cell-free DPRA and may not be activated in the cellular assays due to a limited metabolic capacity (Fabian et al., 2013; Oesch et al., 2014). Benzoyl peroxide was strongly positive in the DPRA but negative in the KeratinoSensTM and h-CLAT. The ten substances in group FN-3 were tested positively by LLNA, but no human data are available. Group FN-4 contains four substances which are true negatives when compared to LLNA data but false negatives, when compared to human data. Among these are the two water soluble antibiotics streptomycin and kanamycin. Both are negative in all investigated non-animal tests and also in the LLNA, but decades of human use indicate an incidence for an ACD after considerable exposure (Kligman, 1966b; Schneider and Akkan, 2004). These oligoaminoglucosides are structurally

significantly different from all other known haptens and pre-/pro-haptens and it is still not clear whether they sensitize by the classical hapten-based mechanism.

Similar to the discussion of the false negative results, also the overall 15 false positive substances could be divided into 4 groups (Table 14). In group FP-1 positive human data confirm the positive overall results of the non-animal test methods and identify the LLNA to be under-predictive for the three substances benzocaine, benzaldehyde and nickel chloride in this group. Benzocaine and benzaldehyde are known contact allergens in humans but only after considerable exposure (Griem et al., 2003) (RIFM database). Hence, the possibility of a false negative response in the LLNA has to be considered. Nickel chloride directly activates the human Toll-like receptor 4 (TLR4). As the mouse cannot mount this TLR4 based response (Schmidt et al., 2010), this indicates a true and well recognized false negative LLNA result. Group FP-2 with concordant

negative human and negative LLNA data just contains propyl parabens, which was tested positively in the KeratinoSens™ and h-CLAT. Although some positive human patch test results are described in the literature, the frequent use of parabens in general corroborates this class to be non-sensitizers from a regulatory point of view (Schnuch et al., 2011; Basketter et al., 2006). Group FP-3 contains nine substances which were predicted to be sensitizers within the '2 out of 3' approach but are non-sensitizers according to LLNA data. For these substances no human data were found in the literature. Peptide adduct formation was detected in a modified peptide reactivity test using LC–MS detection (Natsch et al., submitted for publication) for six substances within this group (i.e. 2-hydroxypropyl methacrylate, 2-acetyl-cyclohexanone, furil, 1-bromobutane, 1-iodohexane and methyl-3-bromopropionate) which indicates that the positive outcome of the '2 out of 3' prediction model might possibly be correct as these are clearly protein-modifying haptens. Group FP-4 contains R(+)-limonene and D,L-citronellol. For both, negative human data are described in the literature (Basketter et al., 2014). Thus, human data are supposed to overrule the positive overall outcome of the non-animal test methods as well as the positive LLNA results. For citronellol, also a recent LLNA on highly pure material was negative (Rudback et al., 2014). However, under the EU Cosmetics Regulation both limonene and citronellol are considered to be allergens although only oxidative metabolites may be reactive suggesting probable pre-/pro-haptens.

4.3. Integrated and sequential testing strategies

From the 144 chemicals with clear-cut results in all three tests (DPRA, KeratinoSens™ and h-CLAT), a congruent result in all three tests was obtained in 76 cases ('3 out of 3'), while for 68 cases a '2 out of 3' assessment was made. For an additional set of 36 chemicals the prediction model is based on a '2 out of 2' assessment from either the combination of DPRA and KeratinoSens™ ($n = 24$), KeratinoSens™ and h-CLAT ($n = 11$) or DPRA and h-CLAT ($n = 1$).

The high overall accuracy of the '2 out of 3' approach indicates that in many cases positives or negatives in single assays are actually FP or FN, respectively, what underlines the importance of making an majority voting. False-positives in the different assays might be due to different mechanisms; thus in KeratinoSens™ unspecific activation of the antioxidant response due to other mechanisms than covalent modification of Keap1 is possible, while false-positives in DPRA may be generated by unspecific peptide oxidation. In the h-CLAT, non-sensitizing irritants such as octanoic acid may also lead to surface marker expression. False negatives results might occur due to solubility issue or limited metabolic or oxidative activation. Among the 68 substances relying on two concordant of three results ('2 out of 3'), the DPRA and KeratinoSens™ rule the '2 out of 3' overall prediction in 24 cases, the DPRA and h-CLAT in 29 cases and the KeratinoSens™ and h-CLAT in 15 cases.

At the same time this analysis indicates that to arrive at the final conclusions often not all three tests are needed, as with two congruent tests the third assay can be waived as it would not change the assessment. This was actually applied in 36 cases with a congruent result in 2 out of 2 tests.

4.4. Interchangeabilities

Within the current dataset both the KeratinoSens™ and the LuSens assay cover the AOP key event 'activation of keratinocytes'. For a common subset of 69 substances an interchangeability of 88% could be calculated. Only 8 test substances did not provide concordant data among both assays (Table 11). Reasons for the different results might be due to differences in the test procedures (e.g. cytotoxicity range finders), the nature of the used ARE sequence in the

keratinocyte cell line (e.g. in KeratinoSens™ and LuSens derived from human or rat, respectively), putative differential metabolic capacities of the cell lines (four substances were pro-haptens) or incidental (borderline read-out in one case). In addition, some of these differences could be also related to the different prediction models used, for instance, the KeratinoSens™ only requires one single concentration of the test substance yielding an induction higher than 1.5-fold whereas the LuSens requires at least two consecutive concentrations. A detailed discussion of the similarity of both assays can be found in Ramirez et al. (2014). Compared to human and LLNA data, both assays provide comparable Cooper statistics. And also the use of the LuSens assay instead of the KeratinoSens™ within a '2 out of 3' approach leads to similar accuracies, although the available dataset of the LuSens is smaller compared to the KeratinoSens™.

An interchangeability of only 72% was calculated for the h-CLAT and the (m)MUSST that both cover the AOP key event 'activation of dendritic cells'. The MUSST and the mMUSST use slightly different protocols, prediction models and cell lines. For the analyses within this study, the results from the MUSST and mMUSST were taken together and discordant results were excluded. When comparing the results of a common subset to *in vivo* data, the (m)MUSST provides a slightly increased specificity whereas the h-CLAT provides an increased sensitivity and overall accuracy. Reasons for the higher sensitivity of the h-CLAT might be the additional marker CD54 and the suitability for this specific common subset of substances. A detailed analysis of the CD54 and CD86 induction might provide further valuable information.

4.5. Mechanistic domains

In order to analyze if the non-animal test methods may detect classes of substances with different reaction mechanisms with similar performances, nine different mechanistic domains were defined by probable protein-binding mechanisms of the 213 substances underlying this study (Table 4). One aspect which should be taken into account is that the predictivities are influenced by the number and type of substances being assessed and data sets may therefore vary in this study. No reaction mechanism was assigned for a group of substances ($n = 65$) with a lack of obvious structural characteristics associated with skin sensitization. However, some of the substances in this group are sensitizing in the LLNA or in humans (e.g. hexyl salicylate). *In vivo* evidence for hexyl salicylate resulting from a single LLNA test provides the lowest EC3 value within this substance group (EC3 = 0.18, compare RIFM data base). Like the structurally similar methyl salicylate, which was tested negatively in the LLNA, the chemical structure of hexyl salicylate reveals no obvious alert for peptide reactivity. Irritation is a confounding factor in the LLNA since it leads to overestimations of sensitization potentials (Ball et al., 2011). In addition, hexyl salicylate was negative even at high concentrations in HRIPT (NOEL $\approx 35,400 \mu\text{g}/\text{cm}^2$) and HMT (NOEL $\approx 20,600 \mu\text{g}/\text{cm}^2$) (compare RIFM database). Therefore, the very low EC3 may well be due to its irritating properties or possibly also due to sensitizing impurities. Another substance which was sensitizing in both human clinical trials and in the murine LLNA is abietic acid. Nevertheless, abietic acid itself is considered to be a non-sensitizer, but depending on storage, sensitization via hydro peroxides derived from autoxidation is probable (Roberts et al., 2007a). A further chemical with a positive LLNA outcome in this group is the well-characterized irritant SDS (EC3 = 14%) (Gerberick et al., 2005; Ball et al., 2011). SDS is considered to be the classic example of a substance yielding a false positive response in the LLNA, what is also confirmed by negative human patch tests (Basketter et al., 2014). Likewise, anhydrous oxalic acid was tested positively in the LLNA (EC3 = 15). This substance was further analyzed for its capacity to

Table 14

False positive predictions within the '2 out of 3' approach.

Chemical name	CAS #	LLNA EC3	Discussion
<i>Group FP-1</i>	'2 out of 3' approach (+), LLNA (–), human (+)		
Nickel chloride	7718-54-9	NC	Well characterized false negative response in murine LLNA, most common contact allergen in humans (Griem et al., 2003); LLNA probably FN
Benzocaine	94-09-7	NC	Known contact allergen in humans after considerable exposure (Griem et al., 2003), all non-animal tests positive; LLNA probably FN
Benzaldehyde	100-52-7	NC	Positive HRIPT ^a at 2760 µg/cm ² , positive in KeratinoSens™, h-CLAT and MUSST; LLNA probably FN
<i>Group FP-2</i>	'2 out of 3' approach (+), LLNA (–), human (–)		
Propyl paraben	94-13-3	NC	No or only weak evidence in humans (Basketter et al., 2014)
<i>Group FP-3</i>	'2 out of 3' approach (+), LLNA (–), human (no or conflicting data)		
2-Hydroxypropyl methacrylate	923-26-2	NC	Only h-CLAT negative; peptide adduct formation in LC–MS test (Natsch et al., submitted for publication)
3-Phenoxypropionitrile	3055-86-5	NC	Only KeratinoSens™ negative
2-Acetyl-cyclohexanone	874-23-7	NC	All non-animal tests positive; peptide adduct formation in LC–MS test (Natsch et al., submitted for publication)
4-Methyl-2-nitroanisole	119-10-8	NC	Only DPRA negative
2-Fluoro-5-nitroaniline	369-36-8	NC	Only h-CLAT negative
Furil	492-94-4	NC	Only h-CLAT negative; peptide adduct formation in LC–MS test (Natsch et al., submitted for publication)
1-Bromobutane	109-65-9	NC	Only KeratinoSens™ and MUSST ^b negative; highly volatile; peptide adduct formation in LC–MS test (Natsch et al., submitted for publication)
1-Iodohehexane	638-45-9	NC	Only MUSST negative; peptide adduct formation in LC–MS test (Natsch et al., submitted for publication)
Methyl-3-bromopropionate	3395-91-3	NC	Only h-CLAT negative; peptide adduct formation in LC–MS test (Natsch et al., submitted for publication)
<i>Group FP-4</i>	'2 out of 3' approach (+), LLNA (+), human (–)		
R(+)-Limonene	5989-27-5	69	No or only weak evidence in humans (Basketter et al., 2014); putative pro-hapten
D,L-Citronellol	106-22-9	43.5	No or only weak evidence in humans (Basketter et al., 2014); putative pro-hapten, recent negative LLNA (Rudback et al., 2014)

^a HRIPT and HMT data were found in the data base of the Research Institute for Fragrance Materials (RIFM).^b DPRA, KeratinoSens™ and h-CLAT results were considered for '2-out-of-3' approach.

covalently bind to reactive side chains of model peptides but did not form any adducts in an LC–MS test (Natsch et al., submitted for publication) and also provided negative results in the DPRA.

For this mechanistic group, accuracies of the single non-animal test methods are in a range of 65–68% when compared to LLNA data and 56–84% when compared to human data. The accuracies provided by the '2 out of 3' approach are 73% and 80% when compared to LLNA or human data, respectively. If such substances as exemplified above with known *in vivo* sensitization potential and a theoretical non-binding capacity were excluded from this group, cooper statistics would have been higher. However, these data cannot be disregarded. Thus, the aim should be extension of the parameters used for assessment of the binding capacities and include factors such as abiotic or enzymatic activation processes.

In contrast to the above discussed group of substances with a lack of obvious structural characteristics being associated with peptide reactivity, the following mechanistic domains containing Michael acceptors, acylating agents, Schiff 'base formers, quinone precursors and substances reacting in nucleophilic substitutions are discussed. The substances within the domain of Michael acceptors were predicted with a generally high accuracy of at least 80% by all of the non-animal test methods. In this domain, human and LLNA data are also concordant for most of the substances.

Somewhat lower accuracies were calculated for acylating agents. The low accuracies of the KeratinoSens™ (56%, compared to LLNA data) and LuSens (44%, compared to LLNA data) are related to a lack of activation of the Keap1/Nrf-2 pathway. Molecular pathway activation triggered in KeratinoSens™ and LuSens is linked to cysteine reactivity with the Keap 1 sensor protein. However, acylating agents like anhydrides transfer their acyl moiety predominantly to lysine residues (Emter et al., 2013; Aptula et al., 2005). Seven substances of the false negatives in the keratinocyte-based assays show considerably increased lysine reactivity in the DPRA.

In this mechanistic domain more weight should be given to DPRA results, since the accuracy in this domain was 100% or 82%, when compared to LLNA or human data, respectively. Similar to acylating agents, short chain aldehydes and longer chain saturated alkanals in the domain of Schiff 'base formers represent hard electrophiles preferring to react with hard nucleophiles like lysine residues instead of cysteine residues (LoPachin and Gavin, 2014). In general, the battery of non-animal test methods appears to be more sensitive to cysteine-reactive substances, as the DPRA has a readout depending on cysteine reactivity, the KeratinoSens™ and LuSens are also dependent on cysteine binding to a significant degree, but also the CD86 expression may be associated with cysteine reactivity (Natsch et al., 2013). This might explain the slightly lower accuracies of most of the non-animal test methods within the domain of Schiff 'base formers. Quinone precursors act as pro-Michael acceptors and must first be activated in order to become electrophilic, but other protein-binding mechanisms cannot be ruled out for some of the substances (Roberts et al., 2007a). Compared to the other non-animal test methods in this substance domain, the DPRA provided the lowest accuracy of 71%, when compared to LLNA data. This can be explained by the fact, that some members of this group might require enzymatic activation, which is absent in the *in chemico* assay. The cell-based test methods also have only limited metabolic capacities (Oesch et al., 2014; Fabian et al., 2013) and thus have limitations in detecting putative pro-haptens.

Regarding the domain of nucleophilic substitutions (S_N1 and 2), the DPRA yields false negative results for one pro-hapten (i.e. dimethylbenz[α]anthracene) and four benzylic esters with common structural alerts (i.e. benzyl benzoate, benzyl salicylate, butylbenzyl phthalate and benzyl cinnamate). Concerning the latter, the benzylic sp³ carbon atom is supposed to react as electrophile. The resulting benzylic cation is resonance stabilized what indicates a

S_N1 mechanism to be probable. Resulting highly reactive S_N1 electrophiles might spontaneously react with water or other rival nucleophilic reaction partners of the model peptides such as hydroxyl ions or solvents what would explain the negative outcome and the slightly decreased accuracy of the DPRA in this domain. This would also be an explanation for the slightly decreased accuracies found in the dendritic cell-based assays when compared to LLNA data, although DC activation is not necessarily be associated with protein binding potential of a compound. The domain of aromatics that react by nucleophilic substitutions (S_NAr , $n = 6$) is quite small in this data set, although further data are described in the literature (Roberts and Aptula, 2014). Also the number of metal ions forming coordination bonds is small ($n = 4$). In contrast to the DC and KC based assays, the applicability domain of DPRA is not defined for compounds containing metal ions.

5. Conclusions

The compilation of non-animal test results provides a comprehensive reference dataset with additional information such as physicochemical properties, types of use, proposed organic reaction mechanisms as well as related *in vivo* reference data for 213 substances.

This study confirms the utility of the five investigated non-animal test methods, i.e. DPRA, KeratinoSens™, LuSens, h-CLAT and (m)MUSST, to predict the respective AOP key event with a high accuracy. When implemented into a '2 out of 3' test strategy, skin sensitizers can be discriminated from non-sensitizers with a high reliability be the use of these alternative methods. For several substances the '2 out of 3' approach does not predict the outcome of the LLNA. For those, human and LLNA data only show a limited concordance. The direct comparison of both *in vivo* references demonstrates that the non-animal test methods predict human data more accurately (Accuracy = 90%) than LLNA data (Accuracy = 82%).

The expanded dataset was further used to define different mechanistic domains by probable protein-binding mechanisms. This approach shows that Michael acceptors, substances reacting in nucleophilic substitutions and quinone precursors were predicted with the highest accuracies. In the domain of Schiff 'base' formers as well as in the group of substances with a lack of obvious alerts for peptide reactivity, accuracies were slightly decreased. In the domain of acylating agents, the keratinocyte based assays show mechanistically justifiable decreased predictivities. If a chemical is supposed to react by acylation, more weight should be given to the DC-based assays and especially the Lys reactivity in the DPRA. The number of substances tested in some specific groups is still low (only six aromatics reacting by nucleophilic substitutions (S_NAr) and four metal-containing complexes within this dataset). Overall, assigning a test substance to a domain according to its protein binding mechanisms offers a way to obtain a more accurate estimate of the predictive performance of the individual non-animal test methods as well as the overall '2 out of 3' prediction.

6. Outlook

In consideration of the obtained data, the presented strategy can be integrated in the regulatory assessment of the skin sensitization hazard potential. For this purpose the "2 out of 3" results should be interpreted under the consideration of the impact the mechanistic domain of the pertinent compound has on the outcome of the individual test method. Thus, the obtained experimental results of the non-animal test methods together with the reliability of the data based on the mechanistic domain provide a weight of evidence for predicting the hazard potential for the induction of skin sensitization.

Beside information for hazard assessment the reference standard dataset also contains concentration–response data on the non-animal test methods and potency information based on LLNA and in part on human data. Therefore, this reference database may further be used to develop prediction models for skin sensitization potential and in particular it may be used to analyze how to arrive at skin sensitization potency predictions based on current data.

Conflict of interest statement

The authors are not aware of any conflicts of interest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.yrtph.2014.12.008>.

References

- Ade, N., Martinozzi-Teissier, S., Pallardy, M., Rousset, F., 2006. Activation of U937 cells by contact sensitizers: CD86 expression is independent of apoptosis. *J. Immunotoxicol.* 3, 189–197.
- Angers-Loustau, A., Tosti, L., Casati, S., 2011. The regulatory use of the local lymph node assay for the notification of new chemicals in Europe. *Regul. Toxicol. Pharmacol.* 60, 300–307.
- Aptula, A.O., Enoch, S.J., Roberts, D.W., 2009. Chemical mechanisms for skin sensitization by aromatic compounds with hydroxy and amino groups. *Chem. Res. Toxicol.* 22, 1541–1547.
- 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., Pease, C.K., 2007. Haptens, prohaptens and prehaptens, or electrophiles and proelectrophiles. *Contact Dermatitis* 56, 54–56.
- Ashikaga, T., Yoshida, Y., Hirota, M., Yoneyama, K., Itagaki, H., Sakaguchi, H., Miyazawa, M., Ito, Y., Suzuki, H., Toyoda, H., 2006. Development of an *in vitro* skin sensitization test using human cell lines: the human cell line activation test (h-CLAT) I. Optimization of the h-CLAT protocol. *Toxicol. In Vitro* 20, 767–773.
- Ball, N., Cagen, S., Carrillo, J.C., Certa, H., Eigler, D., Emter, R., Faulhammer, F., Garcia, C., Graham, C., Haux, C., Kolle, S.N., Kreiling, R., Natsch, A., Mehling, A., 2011. Evaluating the sensitization potential of surfactants: integrating data from the local lymph node assay, guinea pig maximization test, and *in vitro* methods in a weight-of-evidence approach. *Regul. Toxicol. Pharm.* 60 (3), 389–400.
- Basketter, D., Alepee, N., Casati, S., Crozier, J., Eigler, D., Griem, P., Hubesch, B., de, K.J., Landsiedel, R., Louekari, K., Manou, I., Maxwell, G., Mehling, A., Netzeva, T., Petry, T., Rossi, L.H., 2013. Skin sensitisation – moving forward with non-animal testing strategies for regulatory purposes in the EU. *Regul. Toxicol. Pharmacol.* 67, 531–535.
- Basketter, D.A., Alepee, 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., McFadden, J., Evans, P., Andersen, K.E., Jowsey, I., 2006. Identification and classification of skin sensitizers: identifying false positives and false negatives. *Contact Dermatitis* 55, 268–273.
- Bauch, C., Kolle, S.N., Fabian, E., Pachel, C., Ramirez, T., Wiench, B., et al., 2011. Intralaboratory validation of four *in vitro* assays for the prediction of the skin sensitizing potential of chemicals. *Toxicol. In Vitro* 25 (6), 1162–1168.
- 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.
- Bruckner, A.L., Weston, W.L., Morelli, J.G., 2000. Does sensitization to contact allergens begin in infancy? *Pediatrics* 105, e3.
- Cooper, J.A., Saracci, R., Cole, P., 1979. Describing the validity of carcinogen screening-tests. *Br. J. Cancer* 39, 87–89.
- Dearman, R., Basketter, D., Kimber, I., 2013. Inter-relationships between different classes of chemical allergens. *J. Appl. Toxicol.* 33, 558–565.
- Emter, R., Ellis, G., Natsch, A., 2010. Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers *in vitro*. *Toxicol. Appl. Pharmacol.* 245, 281–290.
- 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.
- Fabian, E., Vogel, D., Blatz, V., Ramirez, T., Kolle, S., Eltze, T., van, R.B., Oesch, F., Landsiedel, R., 2013. Xenobiotic metabolizing enzyme activities in cells used for testing skin sensitization *in vitro*. *Arch. Toxicol.* 87, 1683–1696.

- Foussereau, J., Samsen, M., Hecht, M.T., 1983. Occupational dermatitis to Ampholyt G in hospital personnel. *Contact Dermatitis* 9, 233–234.
- Gerberick, F., Aleksic, M., Basketter, D., Casati, S., Karlberg, A.T., Kern, P., Kimber, I., Lepoittevin, J.P., Natsch, A., Ovine, J.M., Rovida, C., Sakaguchi, H., Schultz, T., 2008. Chemical reactivity measurement and the predictive identification of skin sensitizers. *Altern. Lab. Anim.* 36, 215–242.
- Gerberick, G.F., Ryan, C.A., Kern, P., Schlatter, H., Dearman, R., Kimber, I., Patlewicz, G., Basketter, D., 2005. Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods. *Dermatitis* 16, 157–202.
- Gerberick, G.F., Vassallo, J.D., Bailey, R.E., Chaney, J.G., Morrall, S.W., Lepoittevin, J.P., 2004. Development of a peptide reactivity assay for screening contact allergens. *Toxicol. Sci.* 81, 332–343.
- Goebel, C., Aeby, P., Ade, N., Alepee, N., Aptula, A., Araki, D., Dufour, E., Gilmour, N., Hibatallah, J., Keller, D., Kern, P., Kirst, A., Marrec-Fairley, M., Maxwell, G., Rowland, J., Safford, B., Schellau, F., Schepky, A., Seaman, C., Teichert, T., Tessier, N., Teissier, S., Weltzien, H.U., Winkler, P., Scheel, J., 2012. Guiding principles for the implementation of non-animal safety assessment approaches for cosmetics: skin sensitisation. *Regul. Toxicol. Pharmacol.* 63, 40–52.
- Griem, P., Goebel, C., Scheffler, H., 2003. Proposal for a risk assessment methodology for skin sensitization based on sensitization potency data. *Regul. Toxicol. Pharmacol.* 38, 269–290.
- Hoffmann, S., 2014. LLNA variability: an essential ingredient for a comprehensive assessment of non-animal skin sensitisation test methods and strategies. *Altex Proceedings* 3, 1/14, Prague 2014, 11–12–110.
- Jäckh, C., Fabian, E., van Ravenzwaay, B., Landsiedel, R., 2012. Relevance of xenobiotic enzymes in human skin in vitro models to activate pro-sensitizers. *J. Immunotoxicol.* 9 (4), 426–438.
- 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 (11), 1353–1364.
- Karlberg, A.T., Borje, A., Duus, J.J., Liden, C., Rastogi, S., Roberts, D., Uter, W., White, I.R., 2013. Activation of non-sensitizing or low-sensitizing fragrance substances into potent sensitizers – prehapten and prohapten. *Contact Dermatitis* 69, 323–334.
- Kinsner-Ovaskainen, A., Maxwell, G., Kreysa, J., Barroso, J., Adriaens, E., Alepee, N., Berg, N., Bremer, S., Coecke, S., Comenges, J.Z., Corvi, R., Casati, S., Dal, N.G., Marrec-Fairley, M., Griesinger, C., Halder, M., Heisler, E., Hirmann, D., Kleensang, A., Kopp-Schneider, A., Lapenna, S., Munn, S., Prieto, P., Schechtman, L., Schultz, T., Vidal, J.M., Worth, A., Zuang, V., 2012. Report of the EPAA-ECVAM workshop on the validation of integrated testing strategies (ITS). *Altern. Lab. Anim.* 40, 175–181.
- Kligman, A.M., 1966a. Identification of contact allergens by human assay. 3. Maximization test – a procedure for screening and rating contact sensitizers. *J. Invest. Dermatol.* 47, 393.
- Kligman, A.M., 1966b. The identification of contact allergens by human assay. II. Factors influencing the induction and measurement of allergic contact dermatitis. *J. Invest. Dermatol.* 47, 375–392.
- Kolle, S.N., Basketter, D.A., Casati, S., Stokes, W.S., Strickland, J., van, R.B., Vohr, H.W., Landsiedel, R., 2013. Performance standards and alternative assays: practical insights from skin sensitization. *Regul. Toxicol. Pharmacol.* 65, 278–285.
- LoPachin, R.M., Gavin, T., 2014. Molecular mechanisms of aldehyde toxicity: a chemical perspective. *Chem. Res. Toxicol.* 27, 1081–1091.
- Mehling, A., Eriksson, T., Eltze, T., Kolle, S., Ramirez, T., Teubner, W., et al., 2012. Non-animal test methods for predicting skin sensitization potentials. *Arch. Toxicol.* 86 (8), 1273–1295.
- Mosmann, T., 1983. Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *J. Immunol. Methods* 65, 55–63.
- Natsch, A., Bauch, C., Foertsch, L.M., Gerberick, G.F., Norman, K., Hilberer, H., Inglis, H., Landsiedel, R., Onken, S., Reuter, R., Schepky, A., Emter, R., 2011. The intra- and inter-laboratory reproducibility and predictivity of the KeratinoSens assay to predict skin sensitizers in vitro: results of a ring-study in five laboratories. *Toxicol. In Vitro* 25, 733–744.
- Natsch, A., Emter, R., 2008. Skin sensitizers induce antioxidant response element dependent genes: application to the in vitro testing of the sensitization potential of chemicals. *Toxicol. Sci.* 102, 110–119.
- Natsch, A., 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 (11), 1337–1352.
- Natsch, A. et al., submitted for publication. Predicting skin sensitizer potency based on in vitro data from KeratinoSens and kinetic peptide binding: global versus domain-based assessment. *Toxicol. Sci.*
- Nukada, Y., Ashikaga, T., Miyazawa, M., Hirota, M., Sakaguchi, H., Sasa, H., Nishiyama, N., 2012. Prediction of skin sensitization potency of chemicals by human cell line activation test (h-CLAT) and an attempt at classifying skin sensitization potency. *Toxicol. In Vitro* 26, 1150–1160.
- Nukada, Y., Miyazawa, M., Kazutoshi, S., Sakaguchi, H., Nishiyama, N., 2013. Data integration of non-animal tests for the development of a test battery to predict the skin sensitizing potential and potency of chemicals. *Toxicol. In Vitro* 27, 609–618.
- Oesch, F., Fabian, E., Guth, K., Landsiedel, R., 2014. Xenobiotic-metabolizing enzymes in the skin of rat, mouse, pig, guinea pig, man, and in human skin models. *Arch. toxicol.* 88 (12), 2135–2190.
- Patlewicz, G., Basketter, D.A., Smith, C.K., Hotchkiss, S.A., Roberts, D.W., 2001. Skin-sensitization structure–activity relationships for aldehydes. *Contact Dermatitis* 44, 331–336.
- Patlewicz, G., Roberts, D.W., Uriarte, E., 2008. A comparison of reactivity schemes for the prediction skin sensitization potential. *Chem. Res. Toxicol.* 21, 521–541.
- Ramirez, T., Mehling, A., Kolle, S.N., Wruck, C.J., Teubner, W., Eltze, T., Aumann, A., Urbisch, D., Ravenzwaay, B.V., 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 (8), 1482–1497.
- Raunio, H., 2011. In silico toxicology – non-testing methods. *Front. Pharmacol.* 2, 33.
- Roberts, D.W., Aptula, A.O., 2014. Electrophilic reactivity and skin sensitization potency of SNAr electrophiles. *Chem. Res. Toxicol.* 27, 240–246.
- Roberts, D.W., Aptula, A.O., Patlewicz, G., 2007a. Electrophilic chemistry related to skin sensitization. Reaction mechanistic applicability domain classification for a published dataset of 106 chemicals tested in the mouse local lymph node assay. *Chem. Res. Toxicol.* 20, 44–60.
- Roberts, D.W., Patlewicz, G., Kern, P.S., Gerberick, F., Kimber, I., Dearman, R.J., Ryan, C.A., Basketter, D.A., Aptula, A.O., 2007b. Mechanistic applicability domain classification of a local lymph node assay dataset for skin sensitization. *Chem. Res. Toxicol.* 20, 1019–1030.
- Rovida, C., Alépée, N., Api, A.M., Basketter, D.A., Bois, F.Y., Caloni, F., et al., 2014. Integrated Testing Strategies (ITS) for safety assessment. *ALTEx*.
- Rudback, J., Hagvall, L., Borje, A., Nilsson, U., Karlberg, A.T., 2014. Characterization of skin sensitizers from autoxidized citronellol – impact of the terpene structure on the autoxidation process. *Contact Dermatitis* 70, 329–339.
- Schmidt, M., Raghavan, B., Muller, V., Vogl, T., Fejer, G., Tchaptchet, S., Keck, S., Kalis, C., Nielsen, P.J., Galanos, C., Roth, J., Skerra, A., Martin, S.F., Freudenberg, M.A., Goebeler, M., 2010. Crucial role for human Toll-like receptor 4 in the development of contact allergy to nickel. *Nat. Immunol.* 11, 814–819.
- Schneider, K., Akkan, Z., 2004. Quantitative relationship between the local lymph node assay and human skin sensitization assays. *Regul. Toxicol. Pharmacol.* 39, 245–255.
- Schnuch, A., Mildau, G., Kratz, E.M., Uter, W., 2011. Risk of sensitization to preservatives estimated on the basis of patch test data and exposure, according to a sample of 3541 leave-on products. *Contact Dermatitis* 65, 167–174.
- Teubner, W., Mehling, A., Schuster, P.X., Guth, K., Worth, A., Burton, J., et al., 2013. Computer models versus reality: How well do *in silico* models currently predict the sensitization potential of a substance. *Regul. Toxicol. Pharm.* 67 (3), 468–485.
- Thyssen, J.P., Linneberg, A., Johansen, J.D., 2007. The epidemiology of contact allergy in the general population – prevalence and main findings. *Contact Dermatitis* 57, 287–299.
- 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. In Vitro* 28, 626–639.
- van der Veen, J.W., Rorije, E., Emter, R., Natsch, A., van, L.H., Ezendam, J., 2014. Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals. *Regul. Toxicol. Pharmacol.* 69 (3), 371–379.