Arch Toxicol (2015) 89:2355–2383
DOI 10.1007/s00204-015-1634-2

IN VITRO SYSTEMS

Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy

Joanna S. Jaworska[1] · Andreas Natsch[2] · Cindy Ryan[3] · Judy Strickland[4] · Takao Ashikaga[5] · Masaaki Miyazawa[6]

Received: 27 September 2015 / Accepted: 20 October 2015 / Published online: 26 November 2015

Abstract The presented Bayesian network Integrated Testing Strategy (ITS-3) for skin sensitization potency assessment is a decision support system for a risk assessor that provides quantitative weight of evidence, leading to a mechanistically interpretable potency hypothesis, and formulates adaptive testing strategy for a chemical. The system was constructed with an aim to improve precision and accuracy for predicting LLNA potency beyond ITS-2 (Jaworska et al., J Appl Toxicol 33(11):1353–1364, 2013) by improving representation of chemistry and biology. Among novel elements are corrections for bioavailability both in vivo and in vitro as well as consideration of the individual assays' applicability domains in the prediction process. In ITS-3 structure, three validated alternative assays, DPRA, KeratinoSens and h-CLAT, represent first three key events of the adverse outcome pathway for skin sensitization. The skin sensitization potency prediction is provided as a probability distribution over four potency classes. The probability distribution is converted to Bayes factors to: 1) remove prediction bias introduced by the training set potency distribution and 2) express uncertainty in a quantitative manner, allowing transparent and consistent criteria to accept a prediction. The novel ITS-3 database includes 207 chemicals with a full set of in vivo and in vitro data. The accuracy for predicting LLNA outcomes on the external test set ($n = 60$) was as follows: hazard (two classes)—100 %, GHS potency classification (three classes)—96 %, potency (four classes)—89 %. This work demonstrates that skin sensitization potency prediction based on data from three key events, and often less, is possible, reliable over broad chemical classes and ready for practical applications.

Keywords Integrated testing strategy · Skin sensitization · Bayesian network · LLNA potency class

Introduction

Chemical agents are the principal cause of occupational skin disease NIOSH (2012). Skin diseases comprise 17 % of all reported occupational diseases (Bureau of Labor Statistics 2014). The economic impact is significant; the estimated annual cost of occupational contact dermatitis is more than $1 billion (NIOSH 2012). Contact dermatitis is the most common type of occupational skin disease, and allergic contact dermatitis (ACD) is responsible for 20 % of the contact dermatitis cases (Sasseville 2008).

Disclaimers ILS staff provides technical support for NICEATM, but do not represent NIEHS, NTP, or the official positions of any Federal agency. The use of commercial product names is for comparative purposes only and does not constitute endorsement by any of the authors, organizations, or agencies.

Electronic supplementary material The online version of this article (doi:10.1007/s00204-015-1634-2) contains supplementary material, which is available to authorized users.

✉ Joanna S. Jaworska
Jaworska.j@pg.com

[1] Procter and Gamble Company, 1853 Strombeek-Bever, Belgium

[2] Givaudan Schweiz AG, 8600 Duebendorf, Switzerland

[3] Procter and Gamble Company, Mason, OH 45040, USA

[4] ILS/Contractor Supporting NICEATM, Research Triangle Park, NC 27709, USA

[5] Shisheido Company Limited, Tokyo, Japan

[6] Kao Corporation, R&D Safety Science Research, Tochigi 321-3497, Japan

ACD is also a public health problem, accounting for more than seven million outpatient visits every year (Middleton et al. 1998). Over 3700 substances are considered contact allergens (Beltrani et al. 2006; De Groot 1994). Consumers are exposed to contact allergens in skin care products, shampoos, and pesticides. To help consumers and workers avoid ACD, national and international regulatory authorities require chemicals and products to be tested to identify and label potential contact allergens (Boeniger and Ahlers 2003).

Although traditional skin sensitization tests are conducted in animals, legislative changes increasingly mandate that skin sensitization potential be assessed with non-animal methods. Since 2013, the European Union has banned the use of animals for testing cosmetic products and ingredients, as well as the marketing of finished products that have either been tested on animals or that contain ingredients that have been tested on animals following the ban (European Union 2009). Similar legislation may be proposed in the USA in 2015 (AP 2014). Additionally, many institutional animal care and use policies encourage animal use to be minimized.

Consequently, the skin sensitization field has been an area of very active research that resulted in great advances in mechanistic understanding of the processes leading to sensitization (Kimber et al. 2011; Mehling et al. 2012), many novel promising tests (Reisinger et al. 2015), and proposals for frameworks to make classification decisions with alternative data.

Expert approaches to classification, described as weight-of-evidence (WoE) approaches, have been used to integrate existing information and determine the need for additional testing. However, the amount of available alternative assay data is increasing rapidly and becoming more diverse. The new data streams are heterogeneous in metrics and scales (cell-based assays with dose–response curves for both cell marker induction and cytotoxicity, reactivity assays with reaction kinetics or peptide depletion, diverse in silico readouts) as are the biological events leading to skin sensitization that they address. Such heterogeneity makes traditional, subjective expert-based evaluations increasingly more challenging. To facilitate information exchange, increase shared knowledge, and encourage the development of mechanistic frameworks, the Organisation for Economic Co-operation and Development (OECD) coordinates the development of Adverse Outcome Pathways (AOP) for understanding the adverse effects of chemicals (OECD 2015a). The AOP framework codifies the mechanistic steps leading to an adverse effect and allows existing and novel tests to be mapped to the biological events of the AOP. The AOP framework construction practices are rapidly evolving (Villeneuve et al. 2014a, b). The AOP for skin sensitization produced by chemicals that bind covalently to proteins (OECD 2012) includes four key events that occur after a substance penetrates through the skin (Villeneuve et al. 2014a, b) and is potentially transformed to active metabolites:

- Key Event 1: covalent binding to skin proteins
- Key Event 2: activation of inflammatory cytokines and induction of cytoprotective genes in the keratinocyte
- Key Event 3: activation (induction of inflammatory cytokines and surface molecules) and mobilization of dendritic cells in the skin
- Key Event 4: activation and proliferation of antigen-specific T-cells

The final adverse outcome is the inflammatory response (e.g., erythema, edema, blisters, itching) that occurs in the skin of animals or humans upon re-challenge with an allergen.

There is a growing consensus that, given the evolution in knowledge and data, decision frameworks for risk must be more objective, consistent, and transparent (Bus and Becker 2009; Jaworska et al. 2010). Integrated testing strategies (ITS) are tools that can support the new paradigm of toxicity testing. They present a conceptual framework for the cumulative synthesis of information and for guiding testing in such a way that the information gain is maximized in a testing sequence that leads to a risk decision (Jaworska et al. 2010; Jaworska and Hoffmann 2010). As described by Jaworska and Hoffmann (2010) and further emphasized by Hartung et al. (2013) and Rovida et al. (2015), ITS are combinations of tests in a battery covering the relevant mechanistic steps organized in a logical, hypothesis-driven decision scheme, which is required to make efficient use of generated data and to provide a comprehensive information basis for making decisions regarding hazard or risk.

Recognition that the process of skin sensitization is too complex to be able to predict an adverse in vivo outcome using a single alternative test and the need for an ITS approach was very convincingly illustrated in the ITS conceptual model of Jowsey et al. (2006) and later reiterated by Basketter and Kimber (2009). The development of very diverse ITS approaches soon followed.

The majority of approaches focus on data integration and require that all data needed for a particular approach be available for a given chemical to make a prediction. The simplest approach is based on majority voting from the outcome of three in vitro tests (Bauch et al. 2012). Approaches based on machine learning algorithms are very popular. They use hybrid sets of inputs, most often combinations of physico-chemical properties, in silico predictions, and experimental data from one or more in vitro assays. Among them are linear regression-based methods (McKim Jr et al. 2010; Natsch et al. 2009; Nukada et al. 2012) and nonlinear methods such as neural networks (Tsujita-Inoue et al. 2014, 2015) or support vector machines (Strickland et al. manuscript in preparation) and random-forest models (Luechtefeld et al. 2015). The difference between linear and nonlinear lies in the assumption that the predicted variable is a linear or nonlinear combination of inputs. The common characteristic of these models is that the underlying model structure is dictated by the chosen machine learning algorithm, while parameters are data driven. As such, despite the fact that they use mechanistically relevant input data, these approaches do not have the ability to make mechanistically interpretable integrated predictions.

To overcome this shortcoming, Natsch et al. (2015b) and Patlewicz et al. (2014a) utilized mechanistic knowledge, from a chemistry perspective, to develop an ITS framework based on local reactivity domain models, physico-chemical properties, structural alerts, in silico simulators of skin metabolism, auto-oxidation, hydrolysis, and in vitro experimental data. While the approach in Natsch et al. (2015b) is quantitative and estimates the pEC3, a measure of potency, the approach taken by Patlewicz et al. (2014a) does not have a built-in algorithm to make quantitative estimates. It is a WoE tool with decisions driven by expert opinion. The WoE tool is limited to hazard characterization, and the authors suggest that it be used in read-across determinations. The work of Natsch et al. (2015b) and Patlewicz et al. (2014a) serves as inspiration to better integrate knowledge of reaction chemistry in quantitative ITS frameworks, which other approaches tend to lack.

The reason that there are so few ITS approaches in which model structure encodes the skin sensitization process, i.e., in a fully mechanistic framework, is the complexity and remaining uncertainty of the process. It is a very challenging task to formalize this process into equations and then populate the model with data for the parameters. Only two models of this kind are known to the authors: Maxwell and Mackay (2008) and Su et al. (2009). The latter is a model aimed to discover the fundamental principles of the immune response to antigens, while the objective of the first model is to eventually develop a tool for risk assessment for human health. The Maxwell and Mackay (2008) model is a classic pharmacodynamic-pharmacokinetic model represented by a set of ordinary differential equations to model the underlying chemical and biological dynamic processes of mass transport, reaction kinetics, cell population dynamics and receptor binding events. To date, the model has been parameterized only for 2,4-dinitrochlorobenzene and its readiness for routine risk assessment is compromised by the lack of parameters for other chemicals.

Expectations regarding risk decision frameworks reach beyond data integration and, for resource efficiency, should include methods to identify an optimal testing strategy. Among diverse ITS approaches, there is a class of sequential or tiered test batteries that require particular tests or information to be evaluated in a predetermined way (Bauch et al. 2012; Natsch 2014; Nukada et al. 2013; van der Veen et al. 2014). These approaches attempt to introduce efficiency to the strategy by stopping the testing whenever a chemical is predicted positive in the first tier, which usually consists of an assay. Sequential batteries of this kind have a tendency to yield a low number of false negatives and a higher number of false positives. This situation occurs because these strategies use prediction models initially developed as stand-alone prediction models for the individual assays, which usually have higher sensitivity and lower specificity. This effect is exacerbated in the sequential battery because sequential testing does not take into account information overlap between the assays in the battery (Jaworska et al. 2010, 2013; Natsch 2014). Jaworska et al. (2010) demonstrated that prescribed tiered strategies are not optimal and that mandating a single generic set of tests, either tiered or not, as a replacement strategy, is unlikely to be effective, and that ITS must be flexible and adaptive.

Despite high demand to predict sensitization potency, which is driven by the needs of classification and labeling (UN 2013) as well as quantitative risk assessment, the majority of ITS approaches address hazard only. Why is predicting potency so difficult? As measured by the murine local lymph node assay (LLNA), skin sensitization potency may span four orders of magnitude (Gerberick et al. 2005). Existing alternative test methods may be suitable to predict sensitization hazard, but are deemed as not appropriate for potency predictions (Adler et al. 2011; Basketter et al. 2012) partly due to insufficient dynamic range. Further, there is no agreement on what measurements, other than reactivity, are necessary to predict potency (Basketter and Kimber 2009). ITS potency assessment approaches developed to date predict four classes of the LLNA EC3 (Tsujita-Inoue et al. 2014, 2015), or the molar equivalent pEC3 (Natsch et al. 2015b). The authors of the former approach reported 66 % accuracy for the four classes, while the latter approach predicts the pEC3 within two- to fourfold of the experimental values for the training set; however, the performance of either approach on an external validation set is unknown.

Our approach to construct an ITS started with analyses of the needs and resulting conceptual requirements for ITS (Jaworska et al. 2010; Jaworska and Hoffmann 2010). Having these requirements, we identified Bayesian network (BN) approach as the best suited to meet these needs. In short, the BN ITS framework formulates a probabilistic hypothesis about the target variable (in our case, the induction of skin sensitization) based on cumulative evidence from initial data and guides subsequent testing by value of information (VoI) calculations. The rationale to use a

Table 1 Rationale for Bayesian network ITS approach

Feature	Function
The structure of the BN ITS represents the underlying mechanistic processes leading to an in vivo adverse effect while recognizing the uncertainty of the exact formalism The AOP sequence of key events (KE 1, 2, 3), bioavailability, and chemistry are encoded in the network structure	Allows interpretation in the biological context and is chemical specific
ITS framework uses only data as inputs	Eliminates potential inconsistency and uncertainty propagation due to use of the prediction models of multiple individual assays
Information overlap between individual assays regarding adverse effect is accounted for	Reduces false positives and false negative classifications
Can build a hypothesis with partial data in any sequence	Flexible and adaptive. Data outside the applicability domain of individual tests can be eliminated
Quantifies uncertainty for the hypothesis with any partial data	Facilitates consistent prediction acceptance criteria. Guides testing strategy using value of information

BN approach is summarized in Table 1. We have applied the BN ITS framework to skin sensitization hazard and potency prediction first in a proof of concept study (Jaworska et al. 2011) and then as a more mature approach demonstrating its readiness for practical applications (Jaworska et al. 2013).

Practical applications of ITS approaches are lagging behind because investigators often neglect evaluation of the approaches with an external validation (test) set (De Wever et al. 2012). Performance of the test batteries and testing strategies reviewed above cannot be adequately compared to one another because the investigators used different chemical sets to evaluate the methods. Evaluations based on an external dataset were lacking except for the approaches from the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) (Strickland et al. manuscript in preparation) and our work (Jaworska et al. 2011, 2013). The ITS-2 demonstrated a very good balanced accuracy of 88 % on the external test set ($n = 21$). The external dataset was fairly large for those that contain animal data in the toxicological literature; however from a statistical point of view, it had low power and therefore is not very robust. Frequently different accuracy is obtained on a different chemical sets, which leads to disappointments of the interested users. Other factors inhibiting the practical application of ITS is the lack of hands-on guidance in implementing ITS and the lack of regulatory guidance regarding the acceptability of ITS approaches (De Wever et al. 2012). Another hurdle is broad accessibility of the ITS approaches to interested users. A unique effort was undertaken by the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), which provides support to the ICCVAM, to facilitate data and information exchange among stakeholders and encourage the further development, evaluation, and acceptance of these types of non-animal approaches. NICEATM collaborated with us to reproduce and distribute the ITS-2 in an open source version (Pirone et al. 2014) at http://ntp.niehs.nih.gov/pubhealth/evalatm/integrated-testing-strategies/index.html.

The primary goal of the present work was to increase accuracy, precision, and robustness of the BN ITS predictions for the entire range of potency beyond the results achieved in ITS-2. We aimed to achieve this goal by a better integration of chemistry and biology as well as a refinement of the manner in which bioavailability is considered in the BN ITS. Further, large efforts in data generation were undertaken. As the result, the ITS-3 database includes 207 chemicals (training plus test sets), an almost 50 % increase over the ITS-2 database of 145 chemicals (Natsch et al. 2013).

The second goal was to refine the prediction process. Specifically, we wanted to exploit the fact that the BN ITS framework can build hypotheses with partial data. This feature can be used to apply the applicability domains of the individual assays in the process of gathering evidence by eliminating evidence if it was outside of the applicability domain of a particular assay. The intention was to feed to the ITS only relevant data to avoid mispredictions.

The third goal was to increase the standardization of inputs and use only validated assays to increase the practical utility of ITS-3. To this end, we replaced the in vitro U937 test related to Key Event 3, activation of dendritic cells, with the human cell line activation test (h-CLAT), which has been validated by the European Union Reference Laboratory for alternatives to animal testing (EURL ECVAM) (Joint Research Centre of the European Union 2015). In addition, we simplified bioavailability inputs to just physico-chemical properties and eliminated the need to run Kasting's skin penetration model (Dancik et al. 2013).

The fourth goal was to carry out an extensive evaluation of the ITS-3 performance. For this purpose, tremendous effort was spent in preparing a large database that allowed us to use a 60-chemical external validation (test) set. In addition to evaluating the model with a large external dataset, we demonstrate how the ITS-3 could be used in practice with several case studies. First, we demonstrate a prediction when all evidence is in agreement. Second, we illustrate use of ITS-3 for a chemical that is exclusively reactive with lysine and not with cysteine. The third example deals with a potency prediction where alternative data appear to be in conflict due to testing beyond applicability domain limits. In the fourth example, we demonstrate a post-processing step to correct the potency prediction for a direct Michael acceptor (MA).

The ITS-3 developed here provides potency information which can be used for:

1. Hazard identification and classification and labeling under the Globally Harmonized System of Classification and Labelling of Chemicals (GHS) scheme (UN 2013)
2. Quantitative risk assessment (QRA) especially when combined with in vivo evidence on analogs.
3. Development of an efficient testing strategy, thus it is a decision strategy. There is no one best, predefined, testing strategy for all chemicals, but the optimal sequence of tests depends on the information at hand, and is chemical specific (Jaworska et al. 2011). The ITS guides testing by VoI, expressed as mutual information (MI), and measures progress by uncertainty reduction in the probability distribution. VoI informs on whether the prediction class will change once the new information is added.

Materials and methods

The target variable: LLNA potency classes

Although LLNA potency is typically expressed as a percent weight per volume of the concentration required to produce a threshold positive response (e.g., a stimulation index = 3; EC3), in this work we express in vivo potency data in molar units. The driving force for toxic effects is a function of the number of molecules present at the target sites, not the mass of these molecules. For the same reasons, we express both in vitro assay results in molar concentrations. Also, from the potency modeling perspective, mixing inputs expressed in molar and weight units lead to compromised results. To this end, we converted all the data into mol/L concentrations.

The ITS-3 estimates skin sensitization potency in the LLNA, TG 429, (OECD 2010), expressed as probability distribution of LLNA pEC3 spread among 4 potency classes (C1–C4), where $\mathrm{pEC3} = \mathrm{Log}\left(\frac{\mathrm{MW}}{250*\mathrm{EC3}\,\%}\right)$. For non-sensitizing chemicals, for which EC3 was not determined, EC3 % was set to 101 % to provide a corresponding pEC3 value. Next, the pEC3 cutoffs to obtain C1–C4 classes were set to −1.9, −1.1, −0.35. These cutoffs were chosen to follow closely the weight-based classification representing non-sensitizer (NS), weak sensitizer (W), moderate sensitizer (M), and strong or extreme sensitizer (S) classes based on EC3 % (Kimber et al. 2003) (NS, 100-10, 10-1, <1 %). Molecules for which this relationship is not maintained are the ones with very low/high MW or those for which the EC3 % value was close to a cutoff, i.e., EC3 % = 90 (see training and test set).

The LLNA data were compiled from published literature and from previously unpublished data from several laboratories. The chemicals were chosen based on the quality of the LLNA studies. The dataset is comprised of 207 chemicals including fragrances, preservatives, dyes, dye precursors, halogenated alkanes, and solvents and covers a wide range of physico-chemical properties. The training set ($n = 147$) includes 36 NS, 28 W, 35 M, and 25 S sensitizers. The test set ($n = 60$) contains 12 NS, 21 W, 13 M, and 14 S chemicals. Distribution in the pEC3 space, i.e., C1–C4 classes, is 39, 39, 40, 29 for the training set and 14, 19, 12, 15 for the test set. To facilitate reading, equivalence of C1 and NS, C2 and W, C3 and M, and C4 and S is suggested as it does not compromise the interpretation.

Further, after prediction of the pEC3 class distribution is made, it is always possible to convert it to EC3 % distribution, and eventually to specific EC3 % percentiles (Sheet 2 Supplementary file). Usually the most representative summary for the distribution is the 50th percentile, albeit other percentiles can be chosen for a given application. The conversion is provided in the Appendix 2. We discuss the utility of different percentiles later in the manuscript.

Data inputs

The ITS-3 uses the following data sources as inputs (Table 2): (1) bioavailability-related variables (physico-chemical properties: distribution coefficient at pH = 7 log-$\mathrm{D}_{\mathrm{pH}=7}$, water solubility $\mathrm{Ws}_{\mathrm{pH}=7}$, fraction ionized at pH = 7, % plasma protein binding (PB)—ACDlabs Percepta 2014); (2) in silico potency prediction which considers metabolism and potential for auto-oxidation (TIMES); (3) Key Event 1: peptide reactivity [OECD 442 C: Direct peptide reactivity test (DPRA) (OECD 2015c)]; (4) Key Event 2: keratinocyte activation [OECD 442 D: ARE-Nrf2 luciferase test method (KeratinoSens™) (OECD 2015d)]; (5) Key Event 3: dendritic cell activation [human cell line activation test (h-CLAT) (OECD 2015b)]. The whole database, including SMILES experimental data and in silico predictions, is available online (see Supplementary file A).

Table 2 In vitro, in chemico, and in silico data used in the ITS-3

Input type	Endpoint	Unit
Bioavailability	Ws—Water solubility at pH = 7 Log D—Distribution coefficient at pH = 7 PB—Plasma protein binding fraction Fraction ionized at pH = 7	M [−] [−] [−]
In silico prediction of potency in vivo: TIMES	1. Mechanistic alert for direct reactivity (including direct Michael acceptor) and auto-oxidation 2. Prediction of 3 classes (non-sensitizer, weak, or moderate/strong) based on the most potent among parent and metabolites	Classes (NS, W, S)
Key Event 1: DPRACys, DPRALys	% of the cysteine-(Cys), and lysine-(Lys) peptide remaining in the DPRA assay	% remaining peptide
Key Event 2: KeratinoSens™ KEC1.5, KEC3, IC50	Concentration yielding 1.5-fold (KEC1.5); threefold (KEC3) induction of Nrf2-dependent luciferase activity in the KeratinoSens™ assay; 50 % reduction in cell viability in the KeratinoSens™ assay	μM
Key Event 3: h-CLAT EC150, EC200, CV75	Concentrations yielding 150 % induction of the cell surface activation marker CD86 in the h-CLAT; 200 % induction of the cell surface activation marker CD54 in the h-CLAT; 25 % reduction in cell viability in the h-CLAT	μM

Data sources and relevance

ACD/labs Percepta Platform 2014 (ACD Labs, Toronto, Canada) is used to calculate physico-chemical properties related to bioavailability.

DPRA (OECD 2015c) addresses Key Event 1, protein binding. Haptens applied to the skin covalently bind to the nucleophilic residues (i.e., cysteine [Cys] and lysine [Lys]) of proteins in the skin. Binding of chemicals to protein in the skin is an essential step for a sensitizer to produce allergenicity (OECD 2012). Because reactivity is important for the molecular initiating event (MIE), intrinsic or metabolically triggered reactivity has key biological relevance. Binding to the cysteine and lysine peptides provides two measures of the MIE. DPRA data were generated by measuring the reactivity of a test chemical with model heptapeptides containing lysine or cysteine (Gerberick et al. 2007, 2004). Peptide reactivity is reported as percent of free peptide remaining in the sample, which is opposite to the original method reporting percent depletion. The data were generated at Procter & Gamble laboratories.

KeratinoSens™ ARE-Nrf2 Luciferase Test Method [*OECD TG 442d* (OECD 2015d)] addresses Key Event 2, activation of the keratinocyte. Based on recent data (El Ali et al. 2013; van der Veen et al. 2013), the Nrf2 pathway is a key pathway of defense triggered by sensitizers in vivo. According to current knowledge, it is the key common molecular pathway which triggers gene expression in response to electrophilic chemicals at sub-toxic concentrations (Dinkova-Kostova et al. 2005; Natsch 2010). Sensitizers with an exclusive reactivity toward lysine might be negative in the KeratinoSens™ assay. Data were generated using the transfected HaCaT keratinocyte cell line KeratinoSens™. The average concentrations (in μM) inducing a 1.5-fold or a threefold enhanced luciferase activity (KEC1.5 or KEC3.0, respectively) and the concentration leading to 50 % cytotoxicity after 24 h (IC50) are determined. KEC3 data are used in addition to KEC1.5 data, because KEC1.5 may be too low a threshold for some reactive chemicals (Emter et al. 2010). KeratinoSens™ data were obtained from Natsch et al. (2013) or generated at Givaudan laboratories.

h-CLAT (Ashikaga et al. 2006; OECD 2015b) addresses Key Event 3, dendritic cell (DC) activation. When a hapten is applied to the skin, surface molecules (i.e., CD54, CD86) on skin DCs are upregulated through the activation process. Since CD54 is involved in DC migration to draining lymph nodes and CD86 stimulates T cell activation during antigen presentation by DC, both molecules are essential in the induction of skin sensitization. The h-CLAT data were obtained using the THP-1 cell line. The average test chemical concentrations (in μM) inducing 150 % of vehicle control CD86 cell surface marker expression or 200 % of control cell surface CD54 expression (EC150 or EC200, respectively) and the concentration leading to 25 % cytotoxicity after 24 h (CV75) are determined. Data were mainly generated in Kao and Shiseido laboratories.

Tissue Metabolism Simulator for predicting skin sensitization (TIMES) Software V.2.27.13 (Dimitrov et al. 2005) is an in silico hybrid expert system that (1) generates a reactivity alert, (2) assesses potential auto-oxidation, (3) assesses metabolic transformation potential, and (4) semi-quantitatively classifies chemicals into the three categories: non-sensitizers, weak, and strong sensitizers (Dimitrov et al. 2005; Patlewicz et al. 2014b). Intrinsic or metabolically triggered reactivity has a key biological relevance.

Since we differentiate between moderate and strong, the strong class from TIMES is mapped onto the C3 and C4 classes in the ITS-3 training set. The most potent molecule among the parent and metabolites is used for the quantitative prediction.

Bayesian network construction

We continue to pursue a BN-based ITS approach which we identified as the most suitable ITS framework that allows us to capture all of the biology and chemistry, with the ability to combine multiple, heterogenous data streams and use advanced decision-making tools (Jaworska et al. 2010). A BN is a probabilistic graphical model of a problem domain. It is uniquely suited to represent uncertain knowledge when one knows which variables, not necessarily all, are important in the process of interest, but where the relationships between the variables are not well characterized, or complex, or both. In a BN, each node represents one of the features of the problem domain and the arcs between the nodes represent the direct dependencies between the corresponding variables.

In BN parlance, the variable for which we develop a hypothesis (in this study, LLNA potency) is the target variable, while the variables providing evidence (in this study, all listed test results and data in Table 2) are referred to as the manifest variables. In addition to manifest variables, we use latent variables in the network structure. The latent variables are not directly observable—they combine information from similar tests and allow communicating summary results obtained from the parent nodes of the latent variable. From the computational perspective, they simplify the structure of the network by reducing the number of arcs between conditionally dependent variables, and they simplify the numerical computations for the joint probabilities. We divide learning a BN into two iterative sub-tasks: First we learn the structure and then learn the parameters for that structure.

The structure of the ITS-3 model was developed manually from mechanistic knowledge of the endpoint following the approach outlined in Lucas et al. (2004). The AOP structure (i.e., sequence of events, MIEs) as well as data related to AOP Key Events 1, 2, and 3 is encoded in the ITS-3. Aligning ITS-3 model structure with the AOP structure is a critical element of our approach. It makes the ITS-3 gain mechanistic model characteristics: (1) Interpretation of the results is possible in the biological context, i.e., the hypothesis generated by the ITS-3 model can be explained based on known mechanisms; and (2) mechanistic models are more robust and extrapolate better beyond data used to develop the model.

Both the construction method and the resulting structure of ITS-3 are similar to ITS-2 (Jaworska et al. 2013), but there are several refinements. First, as in ITS-2, the mechanistic scheme of the skin sensitization induction process (Basketter and Kimber 2009) with the AOP events of stratum corneum penetration, protein binding, keratinocyte activation and DC activation (Basketter and Kimber 2010) was translated into a Naïve Bayes network structure. Naïve Bayes structure assumes that these events are independent. In the network the Bioavailability latent node relates to stratum corneum penetration potential as well as free concentration in vitro. The Cys latent node and Lys nodes relate to AOP Key Events 1, peptide binding, and 2, keratinocyte activation (for Cys only). The h-CLAT latent node relates to Key Event 3, DC activation, and combines information from all h-CLAT readouts. Second, the tests used to observe the above process were mapped onto the initial network as manifest variables. There are tests that clearly measure different key events, and there are also tests that measure the same key event or part of the process but in different ways. Capturing this information is critical to the proper mapping of tests onto the initial network structure and is described below.

There are two possible MIEs: reaction with cysteine (Cys) and reaction with lysine (Lys), which are represented by two independent nodes. This allows identification of chemicals that act via both MIEs as well as only through one MIE. The Cys latent variable represents the event of cysteine haptenation that can be observed via the DPRA-Cys measurement and/or the KeratinoSens™ assay [a bias toward cysteine-reactive chemicals in Nrf2-dependent assays has been discussed previously (Natsch 2010)]. Reactivity toward cysteine is also measured indirectly in TIMES as electrophilicity molecular descriptors. Further, it has been postulated that the molecular basis of DC stimulation by electrophilic chemicals is a reflection of their ability to bind to sensor proteins (such as Keap1 or others). Therefore, it was even argued that DC-based assays might be a complicated measure of cysteine reactivity (Kimber et al. 2011). The fact that we observed a high shared information between CD86 and Cys nodes in ITS-2 seemed to support this postulate. To reflect this, arcs connecting Cys latent with h-CLAT, as well as Cys latent and TIMES, were introduced exactly as in ITS-2. The only difference is that h-CLAT is a latent variable as it combines information from two separate readouts (CD86 and CD54 surface marker induction).

The new elements of the ITS-3 relate to bioavailability and cytotoxicity. Despite the obvious fact that a chemical must pass through the skin's stratum corneum barrier, most authors did not find bioavailability, usually expressed as log Kow, to be a significant contributor to explain skin sensitization hazard (Alves et al. 2015) or even potency (Roberts and Aptula 2008). Our own efforts to express bioavailability using absorbed dose, as well as maximum epidermal concentration, from the skin penetration simulation of the LLNA dosing scenario using the model developed by Kasting and coworkers (Dancik et al. 2013) showed a somewhat stronger relationship, especially for maximum epidermal concentration, but the effect was still small except for weak sensitizers (Jaworska et al. 2013).

While the role of skin penetration kinetics in in vivo skin sensitization potency remains to be further elucidated, another kinetic component, kinetics in vitro, should also be considered in the ITS framework. Kinetics in vitro aims to assess the free concentration of a tested chemical in an in vitro test. The need for consideration of in vitro kinetics and the importance of using free instead of nominal concentration in the interpretation of the in vitro result has been demonstrated (Groothuis et al. 2015; Kramer et al. 2012) but remains to be routinely used. To this end, we decided to generalize the bioavailability latent variable to consider both skin penetration in vivo and kinetics in vitro in the ITS-3 framework structure. The bioavailability latent variable is constructed from the following physico-chemical properties: water solubility at pH = 7, distribution coefficient, log D at pH = 7, fraction ionized at pH = 7, and % plasma protein binding (PB). These variables are relevant determinants of skin penetration, cell membrane penetration, and free concentration. The bioavailability latent variable is connected by arcs to LLNA pEC3, Cys, Lys, and h-CLAT nodes. The pEC3-bioavailability arc represents bioavailability in vivo, while the arcs with Cys, Lys and h-CLAT represent the respective bioavailabilities in vitro and in chemico.

In order to trigger the sensitization response in vivo there is, after hapten formation, the need for a danger signal in the form of local trauma triggering the emigration of DC. This danger signal appears to involve the formation of extracellular ATP and breakdown products of hyaluronic acid generated by sensitizers (Esser et al. 2012; Weber et al. 2010). The release of ATP from cells is, at least under certain circumstances, triggered by cytotoxicity. For example, cytotoxic surfactants have the ability to provide this local trauma. In the LLNA, which we model in our analysis, no such adjuvant is given. Thus, in the LLNA, a chemical must provide both the hapten and the danger signal in order to trigger the response. Therefore, the LLNA measures both the haptenic potential and the danger signal provided by the chemical, and a chemical with stronger danger signal potential in principle will generate a stronger LLNA response. To account for the presence of the danger signal in the network, we connect the cytotoxicity and pEC3 nodes. The cytotoxicity latent variable is constructed from cytotoxicity measured in h-CLAT assay (CV75) and cytotoxicity measured in the KeratinoSens™ assay (IC50). The arcs connecting IC50 with KEC1.5, KEC3, as well as CV75 with EC150 and EC200, inform about cell viability in relation to the sensitization-specific response. Cytotoxicity in cell-based assays to a certain extent may mimic the 'danger' signals elicited by skin sensitizers in vivo, which might explain why cytotoxicity can partly explain LLNA potency for some chemicals. However, it is important to keep in mind that this reasoning specifically applies to the experimental situation of the LLNA test which is modeled in this work.

Discretization

All input data were discretized using k-means algorithm weighed by MI with the target node, i.e., LLNA pEC3. The number of bins per variable was aimed to be at least four for the latent variables and up to six for the manifest variables. The process of establishing the final number of bins was iterative with the objective to optimize performance of the network on the training set across all potency classes.

Learning parameters of the network

Once the network structure was set, the parameters of the network, i.e., conditional probability tables (CPT), and the resulting joint distributions, were learned from data (the details are described in the Supplementary file B available online). The final network was constructed using the Taboo algorithm in BayesiaLab 5.4 software (Bayesia SAS, Laval Cedex, France).

Learning the network involves calculating a joint probability distribution over all of the variables in the network. As a result, every node of the network has its own joint probability distribution conditioned on the other network variables. Parameters of the network that characterize the arcs, the CPTs, are derived directly from data. CPT are matrices in general form p(A|B)—the probability of data A occurring, given that data B occurs. Associated with each node is a CPT that gives the probability of the node being in a particular state, given the values of the parent nodes.

Value of information (VoI) analysis

VoI, expressed as mutual information, MI (X, Y) between variable X and Y was used to characterize the relationship between variables. MI measures the amount of uncertainty in Y (equal to entropy), which is removed by knowing X. We expressed fraction of entropy of the parent node Y, H (Y), reduced by knowledge of X, i.e., MI (X,Y)/H(Y), and expressed it in percent. The one-step look-ahead hypothesis is used as the methodology to guide testing (Kjaerulff and Madsen 2013). The one-step look-ahead hypothesis calculates the VoI from all possible individual information sources and chooses the one for which the information gain about the target variable is maximized. The foundation of this reasoning is the analysis of the changes in the probability distribution of the information target, given a set of existing data versus a set of existing and new data.

Assessment of applicability domain for in vitro and in chemico assays

Consideration of applicability domain is recognized as very important in interpretation of the assay results. However, this is most often done a posteriori usually in situations when there is a conflict in results. We introduced the consideration of applicability domain in pretreatment of the data based on (1) biological domains; (2) physico-chemical properties such a water solubility at pH = 7 and fraction ionized at pH = 7. Based on the guidance from the developer, all TIMES predictions were accepted whether or not the structure of interest was considered in domain of the model.

Biological domains

We consider the potential for metabolic activation (pro-hapten) and auto-oxidation (pre-hapten) by examination of TIMES predicted sensitization potential for a parent molecule and metabolites. The following information from the TIMES model is taken into consideration: the predicted skin sensitization potency for parent and predicted metabolites, comments on the metabolite prediction (e.g., biotic or abiotic (auto-oxidation) activation), the nature of the metabolic transformation (chemical functionality for protein interaction), and the protein binding alert. Currently, the sensitivity and predictivity of the abiotic transformations in TIMES is 88 and 85 %, respectively (Patlewicz et al. (2014a). We flag the potential pre- and pro-hapten chemicals for more careful examination during the process of hypothesis building.

Water solubility cutoffs

A unifying limitation of the cell-based assays and, to a lesser degree, the in chemico assay, is water solubility (Joint Research Centre of the European Union 2013, 2014, 2015). Traditionally the solubility limitation of an assay has been expressed as a function of log Kow. The published cutoff for h-CLAT is log Kow up to 3.5 for negative h-CLAT results (Takenouchi et al. 2013), and for KeratinoSens™ it is log Kow up to 5 (OECD 2015d). However, log Kow is a good surrogate of water solubility only for neutral chemicals. Partially or fully ionized chemicals are much more soluble in water than their neutral counterparts. In general, water solubility is pH dependent. In order to generalize the solubility cutoff to chemicals that are ionized at physiological pH, we calculated water solubility at pH = 7 and express the minimum cutoff based on this variable. Water solubility at pH = 7 was calculated using ACD labs Percepta software. It is worth noting that the majority of software solutions calculate only solubility of the neutral molecule without clearly explaining this fact. As expected, chemicals with log Kow up to 3.5 for h-CLAT and log Kow up to 5 for KeratinoSens™ revealed a wide range of solubility values (see supplementary material) as these chemicals have diverse degrees of ionization. The cutoffs were chosen as the highest solubility value among the chemicals with log Kow >3.5 and log Kow >5 for h-CLAT and KeratinoSens™, respectively. This resulted in very close cutoff values for h-CLAT and KeratinoSens™ (Table 3). Our result of very similar solubility cutoffs makes sense as both assays require similar medium composition but have slightly different buffer capacity. Till now, the DPRA assay was considered not to have a solubility cutoff because for poorly soluble chemicals the protocol allows to add DMSO up to 10 % of the volume in contrast to 1 % for KS and 0.2 % for h-CLAT. In this work, the choice for the DPRA cutoff was established based on the intersection of the most soluble and potent chemical in the database that had 0 % depletion values for both lysine and cysteine—7,12-dimethylbenz[α]anthracene.

Table 3 Water solubility at pH = 7 cutoffs for DPRA, KeratinoSens™, and h-CLAT

Ws@pH7 [M]	DPRA	KeratinoSens™	h-CLAT
<2.5e−08	x	x	x
2.5e−08 to 1.7e−04	Ok	x	x
1.7e−04 to 2.1e−04	Ok	Ok	x
>2.1e−04	Ok	Ok	Ok

Only data records with solubility greater than the solubility cutoff were considered in the analysis of the test set. Solubility cutoffs were not considered in the training set as this would require removal of all the chemicals with solubility less than 2.1e−4M and would result in a loss of valuable information. Instead, we chose to keep all the chemicals and retain all records below solubility limits that may introduce noise to the data. We say ‘may’ because some of these records do not influence the model parameters. For chemicals with water solubility <2.5e−08M, only the TIMES input was used with the physico-chemical inputs (i.e., no assay data were used).

Fraction ionized

Chemicals that were 100 % ionized at pH = 7 were deemed not to be suitable for cell-based assays due to poor bioavailability, i.e., due to their inability to cross the cell membrane. For partially ionized chemicals, we assumed that while bioavailability is impaired in terms of the rate of crossing cell membranes, the testing period is sufficient for the chemical to cross and reach the target. Among test set chemicals there are two that are 100 % ionized: squaric acid and tartaric acid. The KeratinoSens™ and h-CLAT data records were removed from the test set file for these chemicals. It is worth noting that these chemicals do not exert any reactivity in either KeratinoSens™ or h-CLAT. Fraction ionized was calculated with the formula

$$f\_\text{ionized} = \left| 1 - \frac{10^{\log D}}{10^{\log \text{Kow}}} \right|$$

where || means an absolute value.

Post-processing steps after making pEC3 class probability distribution prediction

Michael acceptor (MA) alert

An alert for directly acting MA triggers an additional post-processing step. The general structure of the alert is shown below, i.e., α,β-unsaturated ketones and aldehydes with an unbranched β-position (Fig. 1).

The direct MA alert is identified by TIMES as chemicals with the following transformation or active alert:

- alpha, beta-carbonyl compounds with polarized double bonds
- alpha, beta-aldehydes
- conjugated systems with electron withdrawing groups
- alpha, beta-carbonyl compounds with polarized triple bond
- conjugated alkenyl pyridines, pyrazines, pyrimidines or triazines
- di-substituted alpha, beta-unsaturated aldehydes

Natsch et al. (2011) noted that chemicals with this substructure are less sensitizing in vivo than would be inferred from chemical reactivity data, due to the anti-inflammatory action of MA. Further, Natsch et al. (2011) showed that the anti-inflammatory activity increases with chemical reactivity for this class of molecules. Since this MA alert does not translate to potency a priori, we have not included it in the structure of the network. However, we use information about this alert in the predictions by manually modifying the hypothesis toward a weaker class in the following manner:

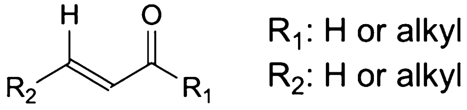

Fig. 1 A structural alert for directly acting Michael acceptors

$$P(Cn_{\text{new}}) = \sum_{i=1,2,3,4} A_i \cdot P(C_i)$$

where $n = 1, 2, 3, 4$.

This modification corresponds to the direct MA effect presented in Natsch et al. (2011) in graphical abstract and was applied only to the MA test set chemicals. The conversion is available on sheet 1 in Supplementary file (Table 4).

Bayes factors

The predicted probability distribution is converted to Bayes factors to: 1) remove prediction bias introduced by the training set class distribution, and 2) express prediction uncertainty, which allows transparent and consistent criteria for accepting the prediction (see Table 5). The conversion is done using the following formula:

$$B = \frac{P(H = x|e)/P(H = \text{not\_}x|e)}{P(H = x)/P(H = \text{not\_}x)} = \frac{\text{posterior odds}}{\text{prior odds}}$$

where: Prior distribution (distribution of the chemicals in the training set):$P(H = x)$ probability that a chemical is in class x (x = C1–C4) in the training set, $P(H = \text{not } x)$ probability that a chemical is not in class x; Posterior distribution (chemical and evidence provided specific prediction distribution): $P(H = x|e)$ probability that a chemical is in class x (x = C1–C4) given the evidence provided to ITS-3, $P(H = \text{not } x)$ probability that a chemical is not in class x given the evidence provided to ITS-3.

Success criteria

Previously established success criteria (Jaworska et al. 2013) were reapplied here. First, an ITS framework should be transparent, consistent, and objective in terms of the decision process, as well as mechanistically interpretable for every prediction made. These are conceptual requirements and have been discussed previously (Jaworska and Hoffmann 2010). Second, the ITS-3 should produce higher accuracy than individual tests on an external test set. Third,

Table 4 Coefficients A of probability transformation for Michael acceptors for the post-processing step

Old\new	$C1_{\text{new}}$	$C2_{\text{new}}$	$C3_{\text{new}}$	$C4_{\text{new}}$
C1	1	0	0	0
C2	0.6	0.4	0	0
C3	0	0.6	0.4	0
C4	0	0.2	0.5	0.3

Table 5 Interpretation of Bayes factors in terms of strength of evidence (Goodman 1999)

Bayes factor	Strength of evidence for acceptance of prediction
<1	Negative (supports alternative)
1–3	Barely worth mentioning (weak)
3–10	Substantial
>30	Strong

the ITS-3 should predict equally well, or better, on both the training and the external tests set. Fourth, if no prediction can be made with the available data, the BN ITS should be able to determine whether additional testing will reduce uncertainty about the prediction.

Process applied to derive the prediction for a new chemical

The process of deriving a prediction for a new chemical consists of two steps: gathering evidence and developing a quantitative hypothesis. This process was applied to all chemicals in the test set and in the case studies and it is summarized below:

1. Gathering evidence

   (a) Calculation of physico-chemical properties of chemicals

   (b) Prediction of sensitization potency category using TIMES:

(i) Potency is based on the highest potency among parent molecule and predicted metabolites;

(ii) Assessment of potential of metabolic activation (pro-hapten) and auto-oxidation (pre-hapten) to facilitate interpretation of DPRA, KeratinoSens™, and h-CLAT assay results;

(iii) Determine whether a chemical is a direct MA based on reactivity alerts.

   (c) Evaluation of the completeness of the evidence for MIEs: Does the dataset have evidence for both cysteine and lysine reactivity?

   (d) Assessment of applicability domains:

(i) If the chemical is deemed a potential pre- or pro-hapten via TIMES prediction, then DPRA, KeratinoSens™, and h-CLAT data are examined with caution, against potential conflict with other data. A hypothesis without these data is considered.

(ii) Solubility domain. Only data records not exceeding solubility cutoffs are considered in the analysis (Table 3). For chemicals with water solubility <2.5e−08M, only TIMES and physico-chemical inputs characterizing bioavailability are used.

(iii) Ionization: chemicals that are completely ionized were not considered suitable for the in vitro assays.

2. Integration of all relevant in-domain evidence via ITS-3 and prediction of the pEC3 probability distribution

   (a) Post-processing correction of the probability distribution for MA, if applicable. See supplementary information.

   (b) Analysis of the hypothesis based on cumulative evidence from combinations of relevant assays.

   (c) Conversion of probability distribution to Bayes factors for final interpretation and acceptance of prediction.

Results

Network structure

Figure 2 shows the structure of the network in ITS-3 (2b) in comparison with the previous version of the network (2a). The biggest change is integration of the h-CLAT assay and the new latent variable Cytotox. The key differences between the networks are presented in Table 6.

Assessing value of a single test using mutual information

MI is a useful measure of interdependence between two variables. Using MI we quantified the values of the individual tests to assess the entire range of LLNA potency (Table 7 first column, MI potency) as well as their values for predicting individual potency classes (Table 7 columns 2–5, MI for NS, W, M, and S, respectively). Ranking of the assays depends on the prediction target. This has an important implication for ITS/WoE. It demonstrates that the assays have very different contributions or 'weights' when predicting different potency classes. Models that use only one weight per assay, such as regressions and decision trees, are not able to capture this robust 'weight' representation.

Similar to earlier findings with ITS-2, TIMES appears to be the most dominant variable in the ITS-3 model. We know that this is inflated because the training set of TIMES partially overlaps with the training set of the ITS-3 model. However, in our work we always use model predictions and not data. TIMES has the highest MI in global ranking as well as for the C1, C2 and C3 categories. The fact that TIMES has a lower MI for the C4 class is the result of two factors. When mapping the TIMES class 3 onto the C3 and C4 classes, there are fewer strong sensitizers than moderate

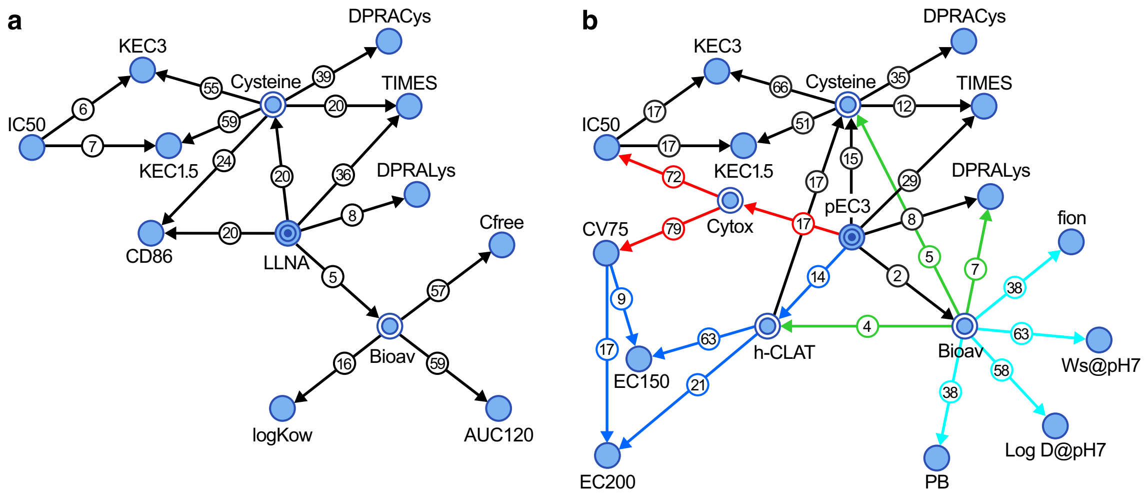

Fig. 2 Comparison of ITS-2 (**a**) and ITS-3 (**b**) network structures. *AUC120* area under the total flux curve for epidermal concentration over 120 h of exposure as in LLNA, *Bioav* bioavailability, *CD86* concentration producing 150 % induction of the CD86 cell surface marker in the U937 assay, *Cfree* maximum free concentration in the mid-epidermis, *DPRACys* remaining cysteine peptide from the direct peptide reactivity assay, *DPRALys* remaining lysine peptide from the direct peptide reactivity assay, *CV75* concentration that reduces cell viability by 25 % in the h-CLAT, *EC150* concentration that produces 150 % induction of the CD86 cell surface activation marker in the h-CLAT, *EC200* concentration that produces 200 % induction of the CD54 cell surface activation marker in the h-CLAT, *fion* fraction ionized at pH = 7, *IC50* concentration that reduces cell viability by 50 % in the KeratinoSens™ assay, *KEC1.5* concentration that produces 1.5-fold induction of Nrf2-dependent luciferase activity in the KeratinoSens™ assay, *KEC3* concentration that produces threefold induction of Nrf2-dependent luciferase activity in the KeratinoSens™ assay, *log Kow* log octanol/water partition coefficient, *log D@pH7* distribution coefficient at pH = 7, *TIMES* Tissue Metabolism Simulator software for predicting skin sensitization potency, *WspH7* water solubility (M) at pH = 7

Table 6 Summary of ITS-3 refinements over ITS-2 structure (color figure and table online)

ITS-2	ITS-3
Key Event 3:U937	Key Event 3: h-CLAT
Bioavailability considered only for neutral chemicals: logKow, Cfree, AUC120 [a]	Generalized Bioavailability inputs applicable for both neutral and ionized chemicals (WspH7, logDpH7, fion, PB)
Bioavailability in vivo	Bioavailability in vivo and in vitro considered in the ITS structure
Cytotoxicity impacts KeratinoSens$^{\text{TM}}$	Cytotoxicity impacts KeratinoSens$^{\text{TM}}$, h-CLAT, and LLNA

Color-coded new elements match Fig. 2

[a] *AUC120* area under the total flux curve for epidermal concentration over 120 h of exposure as in LLNA, *Cfree* maximum free concentration in the mid-epidermis

sensitizers (40/60 %). In addition, S is the smallest class (19 % of the 147 chemicals) in comparison with the other classes (ca 27 % each of the 147 chemicals).

The Cytox node representing combined information on cytotoxicity from KeratinoSens™ and h-CLAT is second in the global ranking. In the ITS-3, cytotoxicity provides more information about potency than the in chemico and biological assays. This result requires a careful explanation. There are three potential reasons for this correlation: the necessity for a chemical to provide both the hapten and the danger signal in order to trigger the response in the LLNA, an intrinsic link between cytotoxicity and reactivity, and a potential bias in the database toward an exaggerated high number of non-cytotoxic non-sensitizers. This will be discussed in more detail below.

Among assays quantification of Cys-reactivity has the highest overall MI, closely followed by h-CLAT. Further we confirm a high degree in information overlap in

Table 7 Mutual information of individual assays used in ITS-3 based on the training set

MI potency overall		MI for C1		MI for C2		MI for C3		MI for C4	
TIMES	28	TIMES	58	TIMES	16	TIMES	18	Cys	21
Cytox	17	Cytox	35	Cys	5.7	h-CLAT	9.6	KEC3	16
Cys	15	CV75	29	Cytox	5.4	EC150	7.4	KEC1.5	15
CV75	14	IC50	28	h-CLAT	4.6	EC200	3.4	h-CLAT	13
IC50	13	Cys	21	KEC1.5	4.5	KEC1.5	1.8	Cytox	12
h-CLAT	13	KEC1.5	20	CV75	3.9	Cytox	1.7	DPRALys	12
KEC1.5	12	KEC3	20	IC50	3.8	Cys	1.5	DPRACys	11
KEC3	12	EC200	17	KEC3	3.5	CV75	1.5	CV75	10
EC150	10	h-CLAT	17	DPRALys	3.0	IC50	1.3	IC50	10
EC200	9.1	EC150	17	EC150	2.3	DPRALys	1.0	TIMES	10
DPRALys	7.5	DPRACys	10	Bioav	2.1	Bioav	0.8	EC150	7.7
DPRACys	7.0	DPRALys	9.0	WspH7	1.8	KEC3	0.5	EC200	7.5
Bioav	2.4	Bioav	3.8	Log DpH7	1.5	Prot Bind	0.3	Bioav	1.5
fion	1.4	fion	2.9	DPRACys	1.1	WspH7	0.3	fion	1.1
Log DpH7	1.3	Log DpH7	2.1	PB	1.1	DPRACys	0.3	Log DpH7	0.5
WspH7	1.3	WspH7	1.7	EC200	1.0	Log DpH7	0.2	WspH7	0.3
PB	0.5	PB	0.2	fion	0.5	fion	0.2	PB	0.0

Table 8 Prior and posterior distribution probabilities and Bayes factors for benzo(α)pyrene, CAS# 50-32-8

Prior distribution as in the training set				Posterior distribution predicted by ITS-3				Bayes factors			
pEC3 C1	pEC3 C2	pEC3 C3	pEC3 C4	pEC3 C1	pEC3 C2	pEC3 C3	pEC3 C4	B (C1)	B (C2)	B (C3)	B (C4)
0.27	0.27	0.27	0.19	0.04	0.29	0.37	0.30	0.11	1.11	1.60	1.75

Cys-reactivity and DC activation measurement as observed in ITS-2. MI between h-CLAT and Cys-reactivity is the same as U937 and Cys-reactivity in ITS-2–20. Taking a closer look h-CLAT is clearly more valuable than Cys-reactivity in identifying C3 class, while it provides very similar information for C1, C2 and C4 classes.

Quantification of uncertainty for decision-making-converting probability-based predictions to Bayes factors

Since many results below are expressed as Bayes factors, the impact of changing from probability to Bayes factor needs to be explained first. Use of Bayes factors corrects for the distribution of the chemicals in the training set and therefore provides a more objective prediction than a posterior probability distribution. In the ITS-3 training set, Class S chemicals are slightly underrepresented (compare the probability of the prior distribution of pEC3 Class 4 with those of the other three classes in Table 8). This results in the deflated posterior probabilities for this class. In the example of benzo(α) pyrene, this leads to the conclusion that this is a C3 chemical based on probabilities [Pr(C3) = 0.37]. However, when predictions are based on the highest Bayes factor (*B*), benzo(α) pyrene is predicted as C4, which is concordant with experimental data. Since B(C4) is only 1.75, we conclude that the evidence for this chemical to be C4 is weak. Similarly, for chemical classes that are overrepresented, the prediction probabilities are inflated.

Corrections for MA chemicals

There were seven direct MA chemicals in the test set for which the probability distributions were corrected (Table 9). The correction resulted in class changes to the class that was experimentally observed for four chemicals. Class assignments for ethyl and methyl acrylates moved from C4 to C3. As discussed later, in vivo data for these chemicals are considered not reliable due to the high volatility of low molecular weight acrylates. The safranal class prediction did not change; however, the Bayes factor distribution shifted it toward the less potent classes (see Example 4 in the case studies).

Table 9 Bayes factors before and after corrections for Michael acceptor chemicals

Chemical	Before MA correction				After MA correction			
	B (C1)	B (C2)	B (C3)	B (C4)	B (C1)	B (C2)	B (C3)	B (C4)
Methylmethoxy acrylate	***156.5***	0.0	0.0	0.1	***156.6***	0.0	0.0	0.0
Ethyl acrylate	0.0	**0.3**	0.1	*33.8*	0.2	**0.8**	*2.2*	1.5
Methyl acrylate	0.0	**0.0**	0.1	*92.5*	0.0	**0.8**	*2.6*	1.6
Farnesol	0.0	**0.4**	*6.3*	0.9	0.2	***2.8***	1.6	0.2
Safranal	0.0	0.1	**3.6**	2.7	0.06	1.7	***2.3***	0.6
α-Damascone	0.0	0.2	**0.6**	*11.6*	0.1	1.1	***2.2***	1.2
5-Methyl-2-phenyl-2-hexenal	0.0	0.3	**1.7**	*4.21*	0.2	1.7	***1.8***	0.7

Bolded numbers denote class attributed based on in vivo data, italicized numbers show the class predicted by ITS-3

Table 10 Predictive capacity of the approach given as a contingency matrix based on the highest Bayes factor

GHS category	Observed									
	Training set (147)					Test set (60)				
	Class	C1(39)	C2(39)	C3(40)	C4(29)	Class	C1(14)	C2(19)	C3(12)	C4(15)
None	C1	36	2	1	0	C1	14	0	0	0
1B	C2	2	32	3	3	C2	0	17	3	0
	C3[a]	0	3	38	5	C3	0	2	9	2
1A[a]	C4	1	2	8	21	C4	0	0	0	13

Numbers in parentheses indicate number of chemicals

[a] Since the GHS classification cutoff for 1A is ≤ 2 %, the table provides a more conservative classification. Further differences are to be expected due to the conversion from weight to molar units

Predictive capacity

Accuracy and precision with full and partial evidence

The strength of the BN ITS is its ability to reason with either all or partial evidence. Therefore only results that are within the applicability domains of the individual assays are recommended to be used when predicting potency of new molecules. When all evidence is entered to the system the hazard prediction accuracy for the test set expressed as a balanced accuracy (bac) is 100 % (Table 10). Bac accounts for uneven distribution of positive and negative chemicals in a dataset. For a binary classification, the formula is:

$$\text{bac} = \frac{\text{Se} + \text{Sp}}{2}$$

where Se = sensitivity and Sp = specificity.

Extending bac to multiple classes, one gets bac for GHS C&L = 96 %; bac for four-class potency = 89 %. The four-class potency accuracy of 89 % is in excellent agreement with accuracy for the training set (85 %). It demonstrates that the ITS-3 model is not overfit. The higher accuracy for the test set reflects the fact that we have a pre-processing step of selecting data only from their physico-chemical applicability domains and a post-processing step of MA correction. The ITS-3 achieved 100 % accuracy for 14 C1 chemicals. It also reliably predicted the weak and strong classes. When the two problematic acrylates are removed (Table 9), it predicted C2 chemicals 100 % correctly. However, the model has a drop in accuracy for C3, correctly predicting 75 % of the chemicals. Previously, in ITS-2 we observed this drop in accuracy for W and M (Jaworska et al. 2013) and attributed it to insufficient dynamic range of in vitro assays.

In Table 10 predictions with varying degree of uncertainty are accepted as long Bayes factor (B) > 1. One can easily modify criteria for acceptance. For example a non-sensitizer prediction can be accepted only when $B > 3$ (strong evidence), while $B > 1$ can be deemed sufficient to accept chemical to be a sensitizer.

When using all information, the following seven chemicals in the test set were mispredicted (Table 11). One should look at the mispredictions from the side of in vivo data and alternative data inputs. Among the seven chemicals mispredicted, there are four data points where the in vivo data are not reliable: 2 acrylates, tocopherol and anhydride. Of the remaining three, two chemicals are out of the in vitro assay domains due to poor water solubility.

Table 11 List of chemicals mispredicted by ITS-3 using full datasets

Chemical	EC3 %	B (C1)	B (C2)	B (C3)	B (C4)	Explanation
Ethyl acrylate	28.0	0.2	**0.8**	2.2	1.5	High vapor pressure, in vivo results likely underpredicted due to evaporation
Methyl acrylate	20.0	0.0	**0.8**	2.6	1.6	High vapor pressure, in vivo results likely underpredicted due to evaporation
Dihydroeugenol(2-methoxy-4-propyl-phenol)	6.8	0.0	5.4	**0.7**	0.6	Pro-hapten, however removal of DPRA yields class S
Farnesol	4.1	0.3	2.0	**1.6**	0.4	Predicted by TIMES as pre-hapten, however removal of DPRA yields class S. KeratinoSens™ and h-CLAT out of solubility domain
Tocopherol	7.4	0.4	5.1	**0.4**	0.5	logP = 10.6, result based on TIMES only, DPRA, KeratinoSens™ and h-CLAT out of solubility domain. Tocopherol/vitamin E is not a human sensitizer and LLNA may be false positive
1,2-Cyclohexane dicarboxylic anhydride (hexahydrophthalic anhydride)	0.8	0.1	0.3	4.6	**1.3**	This chemical quickly hydrolyzes in water. However, in DPRA reactivity is so fast that it is even faster than hydrolysis (if peptide added first). KeratinoSens™ and h-CLAT out of solubility domain for the parent molecule however it is more likely that an acid is tested. Because the acid is very strong it will fall out from applicability domain based on fion[a]
Squaric acid diethyl ester	0.9	0.4	1.1	3.9	**0.1**	This chemical quickly hydrolyzes in water, in vitro assays test not the parent chemical but acid and alcohol (Cohen and Cohen 1966)

[a] Similar chemical (phthalic anhydride [CAS# 85-44-9], a known misclassified extreme sensitizer, hydrolyzes in water at pH 6.8–7.24 with half-lives of 0.5–1 min at 25 °C, forming phthalic acid, and is therefore not within the applicability domain of the in vitro assays (UNEP 2005). Phthalic acid [CAS# 88-99-3] is classified as a non-sensitizer by a modification of the Maguire method and the LLNA (ECHA database on registered substances, searched on 25.07.2014). Bolded numbers denote class based on experimental in vivo data

Table 12 Accuracy of potency predictions for the test set in % for either full data input or with omission of one of the key event assays

GHS C&L	Potency class	All data	w/o DPRA	w/o Ksens	w/o h-CLAT	w/o TIMES	*n*
	All	89	82	77	75	74	60
None	C1	100	100	93	93	87	14
1B	C2	90	83	83	83	89	19
	C3	75	58	50	45	58	12
1A	C4	87	87	80	80	60	15

Predicting potency without one of the AOP key event assays

From the mechanistic point of view, the three assays have fairly large information overlap. Many models were published, using information on subsets of key events with good results. We (Jaworska et al. 2013) and others (Natsch et al. 2015a) have shown previously that a correct prediction of potency does not always require entering information from all possible assays. Our results in Table 12 further confirm this observation.

However, the loss of accuracy when information from one of the assays is not provided is more prominent for C3 and C4 classes than for C1 and C2. In other words, our results indicate that the highest information overlap is in predicting NS. This may not be too surprising given that all three in vitro assays (and the LLNA) were initially developed to discriminate sensitizers from non-sensitizers—so with the same primary goal to provide this specific information. Omitting TIMES has the biggest effect overall driven by the loss of accuracy for C1, C3, and C4. This result is a combination of the high TIMES predictivity for C1 but also the fact that six of 15 chemicals in the C4 class have very poor solubility that makes the in vitro data out of the applicability domain. Thus these predictions without TIMES use only bioavailability. Omitting h-CLAT and KeratinoSens™ has the same effect on accuracy for all the classes while DPRA has the smallest effect.

Accuracy presents only one aspect of the predictive value of the system. Accuracy refers to the agreement between

Table 13 Precision of the predictions for the test set expressed as the median Bayes factor in a given class

Class	All	w/o DPRA	w/o KS	w/o h-CLAT	w/o TIMES	*n*
C1	133.1	95.4	64.5	107.2	6.9	14
C2	4.0	4.1	4.3	4.2	2.7	18
C3	2.2	2.1	2.3	1.7	2.0	12
C4	7.7	2.1	3.6	5.9	7.1	15

measured and predicted value. The other aspect of the predictive value, independent from accuracy, is provided by precision. It tells us about uncertainty of the prediction. Bayes factors are expressions of precision. To this end, we provide prediction precision information in Table 13.

Assessment of precision varies greatly with a class and evidence. When all data are used, the system indicates that NS are predicted with highly decisive strength of evidence ($B > 100$). For other classes, there is a remarkable, over 20-fold drop in the precision. When using all of the data, the system concludes that strength of evidence for W and S is strong ($B > 3$). The ITS-3 system has the smallest precision predicting class M where the median strength of evidence is weak ($1 < B < 3$).

When the system makes predictions without DPRA, there is a drop in the strength of evidence by about 30 % in predicting C1. There is no drop in the strength in predicting C2 and C3, while the strength of evidence in predicting S is reduced by 70 %. Not entering KeratinoSens™ data halves the strength of evidence for predicting NS and S, has no practical effect on predicting W, and slightly improves prediction reliability for M. Lack of h-CLAT had the weakest effect on the reliability of predictions of NS, with reduction of only 20 %. There is no significant reduction for W, and 25 % for M and S. Strength of evidence without TIMES for NS is reduced by 95 % but nevertheless remains strong, in fact well above the Bayes factor threshold for strong. Exclusion of TIMES halves strength of evidence for W class and had little effect on M and S.

Omitting TIMES has the largest effect on prediction precision of the system. We know that this is somewhat biased because the training set of TIMES partially overlaps with the training set of the ITS-3 model. Interestingly, leaving out TIMES has a marginal effect on the prediction precision for class S, while it had the highest impact on accuracy. To determine C1, leaving out KeratinoSens™ results in the largest loss in precision, followed by DPRA, and eventually h-CLAT. Leaving out DPRA results in the highest loss of precision for the C4 class.

Only while analyzing accuracy and precision together one can make a choice about the evidence needed to make a decision. For example one needs all three assays and TIMES to conclude in a highly decisive manner ($B > 100$) whether a chemical is a C1. With incomplete records, one can still make a correct prediction but with a lower precision. Analysis of changes in accuracy and precision when providing partial evidence can be also explained by MI values of individual tests (Table 7).

Case studies

Four chemicals were selected from the test set and are presented here as case studies. These case studies illustrate the steps described in the Methods section under the heading ‘Process applied to derive the prediction for a new chemical’ and they indicate how considerations regarding applicability domain, MA correction and conflicting information are handled.

Example 1: octanenitrile CAS# 124-12-9; LLNA EC3 not determined, non-sensitizer

1. *Prediction of physico-chemical properties of chemicals* Calculated (see Table 14).

2. *Prediction of TIMES:*

   (a) TIMES prediction results—parent NS, metabolite NS
   (b) Not identified to be a pre/pro-hapten
   (c) No direct MA alert

3. *Completeness of MIE evidence check: Does the dataset have evidence on both: cysteine and lysine?*

   Data are available for Cys and Lys MIEs. Complete dataset.

4. *Assessment of applicability domains*

   (a) Not identified to be a pre/pro-hapten by TIMES
   (b) Water solubility within acceptable range for all assays
   (c) Chemical mostly in a non-ionized form (f_ ion = 0.06)

Table 14 Input data overview for octanenitrile

EC150 µM	EC200 µM	CV75 µM	DPRA-Cys % rem	DPRA-Lys % rem	KEC1.5 µM	KEC3 µM	IC50 µM	TIMES	fion	Log D@pH7	PB	Ws@pH7
10000	10000	3430	100	96.4	2000	1512	2000	1	0.06	2.72	79.2	0.013

Table 15 pEC3 probability distribution and Bayes factors for individual and combinations of inputs for octanenitrile

Evidence	pEC3 C1	pEC3 C2	pEC3 C3	pEC3 C4	B(C1)	B(C2)	B(C3)	B(C4)
TIMES	0.93	0.03	0.00	0.03	**38.3**	0.1	0.0	0.1
DPRA (Cys + Lys)	0.52	0.24	0.21	0.03	**3.0**	0.9	0.7	0.1
KS (KEC1.5, KEC3, IC50)	0.71	0.11	0.16	0.02	**6.8**	0.4	0.5	0.1
h-CLAT (EC150, EC200, CV75)	0.69	0.14	0.12	0.05	**6.1**	0.5	0.4	0.2
DPRA + KS	0.79	0.09	0.11	0.00	**10.5**	0.3	0.3	0.0
h-CLAT + KS	0.76	0.12	0.09	0.02	**9.0**	0.4	0.3	0.1
h-CLAT + KS + DPRA(Cys + Lys)	0.83	0.10	0.06	0.00	**13.8**	0.3	0.2	0.0
h-CLAT + KS + DPRA(C + L) + bioav.	0.71	0.18	0.09	0.01	**6.9**	0.6	0.3	0.0
h-CLAT + KS + DPRA(Cys + Lys) +bioav. + TIMES	0.98	0.02	0.00	0.00	**129.1**	0.1	0.0	0.0

Bold values indicate Bayes factor that drives the decision

5. *Integration of all the in-domain evidence via ITS-3 and prediction of the pEC3 probability distribution*(Table 15).

   (a) All individual assays predict this chemical a non-sensitizer with a substantial strength of evidence ($B \geq 3$)
   (b) Bioavailability has a negative but weak effect on the hypothesis that the chemical is a non-sensitizer because of relatively high protein binding and low water solubility.
   (c) All combinations of assays predict the chemical to be a non-sensitizer with increasing strength of evidence as data is added in a stepwise manner. This means that there no data conflicts.

6. *Post-processing step of probability distribution correction for MA, if applicable.*

Not needed. This chemical is not a direct MA.

7. *Conversion of probability distribution to Bayes factors for final interpretation and decision.*

Taking into account all input parameters according to the AOP, this chemical is deemed to be a non-sensitizer with very strong strength of evidence ($B = 130$). All data are in agreement (Table 15). Bolded number indicates Bayes factor that drives the decision.

Example 2: 2-methyl-4H-3,1-benzoxazin-4-one CAS# 525-76-8, LLNA EC3 = 0.7 %, strong sensitizer illustrating the need for checking both Lys and Cys MIEs

1. *Prediction of physico-chemical properties of chemicals* Calculated (see Table 16)

2. *Prediction of TIMES:*

   (a) Predicted to be strong sensitizer based on parent structure
   (b) Not identified to be a pre/pro-hapten
   (c) No direct MA alert

3. *Completeness of MIE evidence check: Does the dataset have evidence on both: cysteine and lysine?*

DPRA data are available only for Lys, due to technical problem with Cys-reactivity measurement (e.g., co-elution of the Cys-peptide with the chemical). MIE of cysteine binding is reflected by KeratinoSens™ (KS) and probably h-CLAT.

4. *Assessment of applicability domains*

   (a) Not considered to be a pre/pro-hapten
   (b) Water solubility within acceptable range for all assays. Chemical is neutral.

Table 16 Input data overview for 2-methyl-4H-3,1-benzoxazin-4-one

EC150 µM	EC200 µM	CV75 µM	DPRA-Cys % rem	DPRA-Lys % rem	KEC1.5 µM	KEC3 µM	IC50 µM	TIMES	f_ion	Log D@pH7	PB	Ws@pH7
10000	1980	3530	Co-elution	65.7	135	688	2000	3	0	1.5	69.8	0.01

Table 17 pEC3 probability distribution and Bayes factors for individual and combinations of inputs for 2-methyl-4H-3,1-benzoxazin-4-one

Evidence	pEC3 C1	pEC3 C2	pEC3 C3	pEC3 C4	B(C1)	B(C2)	B(C3)	B(C4)
TIMES	0.05	0.24	0.42	0.29	0.1	0.9	**1.9**	1.7
DPRA (Lys)	0.10	0.10	0.40	0.40	0.3	0.3	1.8	**2.7**
KS (KEC1.5, KEC3, IC50)	0.43	0.33	0.19	0.05	**2.1**	1.4	0.7	0.2
h-CLAT (EC150, EC200, CV75)	0.68	0.14	0.12	0.05	**6.0**	0.5	0.4	0.2
h-CLAT + KS	0.66	0.17	0.13	0.04	**5.4**	0.6	0.4	0.2
TIMES + DPRA(Lys)	0.01	0.07	0.46	0.45	0.0	0.2	2.3	**3.4**
h-CLAT + KS + DPRA(Lys) + bioav. + TIMES	0.28	0.15	0.01	0.56	1.1	0.5	0.0	**5.1**
h-CLAT + KS + bioav. + TIMES	0.44	0.34	0.01	0.21	2.2	1.4	0.0	1.1

Bold values indicates Bayes factor that drives the decision

5. *Integration of all the in-domain evidence and prediction of the pEC3 probability distribution* (Table 17)

   (a) The chemical is predicted as a moderate or strong sensitizer by TIMES.
   (b) Based on Lys-peptide depletion values, the chemical is predicted to be a strong allergen.
   (c) The KS or h-CLAT data, however, are in conflict with Lys-reactivity and TIMES data. KS points toward a NS or maybe weak sensitizer. The results from the h-CLAT activation assay are clearly indicating a NS.
   (d) Combination of KS with h-CLAT data strongly supports the NS class ($B = 5.4$), while the combination of TIMES and DPRA-Lys supports hypothesis that the chemical is a strong sensitizer ($B = 3.4$). The latter hypothesis is a bit weaker. This is an example where statistics alone would be misleading; the chemical is acting via lysine as the only MIE, which is rare. KS and h-CLAT which are preferential toward detecting chemicals acting via the cysteine MIE did not detect this chemical in the context of ITS-3. The KS stand-alone prediction model, which does not use cytotoxicity as an input, would detect it, but rate it rather weak (data not shown).
   (e) Combination of all data (including bioavailability) results in a clear hypothesis that the chemical is a strong sensitizer ($B = 5.1$).

6. *Post-processing step of probability distribution correction for MA, if applicable.*

   Not needed, this is not a MA chemical.

7. *Conversion of probability distribution to Bayes factors for final interpretation and decision.*

   Taking into account all input parameters according to the AOP, this chemical is considered to be a strong sensitizer with substantial strength of evidence ($B = 5.1 > 3$). The strong sensitization potency is driven by strong reactivity toward lysine because this chemical is not reactive with cysteine (Table 17).

Example 3: (ethoxymethoxy)cyclododecane CAS# 58567-11-6; LLNA EC3 = 25.1 %, weak sensitizer, illustrating importance of considering cytotoxicity and applicability domains

1. *Prediction of physico-chemical properties of chemicals*
   Calculated (see Table 18).

2. *Prediction of TIMES:*

   (a) Predicted to be weak sensitizer based on metabolite, while parent was predicted as NS
   (b) Identified to be a pre-hapten due to auto-oxidation
   (c) No direct MA alert

Table 18 Input data overview for (ethoxymethoxy)cyclododecane

EC150 µM	EC200 µM	CV75 µM	DPRA-Cys % rem	DPRA-Lys % rem	KEC1.5 µM	KEC3 µM	IC50 µM	TIMES	Log D@pH7	PB	Ws@pH7	fion
10000	38.99	88.70	96.1	99.4	2000	2000	24.34	2	4.71	89.73	9.77e-04	0.9

Table 19 pEC3 probability distribution and Bayes factors for individual and combinations of inputs for (ethoxymethoxy)cyclododecane

Evidence	pEC3 C1	pEC3 C2	pEC3 C3	pEC3 C4	B(C1)	B(C2)	B(C3)	B(C4)
TIMES	0.14	0.60	0.14	0.11	0.5	**4.2**	0.5	0.5
DPRA (Cys + Lys)	0.52	0.24	0.21	0.03	**3.0**	0.9	0.7	0.1
KEC1.5, KEC3	0.50	0.20	0.24	0.06	**2.7**	0.7	0.9	0.3
KEC1.5, KEC3, IC50	0.01	0.32	0.49	0.19	0.0	1.3	**2.5**	0.9
h-CLAT EC150, EC200	0.01	0.29	0.34	0.36	0.0	1.1	1.4	**2.3**
EC150, EC200, CV75	0.01	0.33	0.29	0.37	0.0	1.4	1.1	**2.4**
DPRA +KS	0.01	0.40	0.51	0.08	0.0	1.8	**2.8**	0.3
h-CLAT + KS	0.01	0.39	0.37	0.23	0.0	1.8	1.6	1.2
DPRA +TIMES	0.20	0.56	0.21	0.02	0.71	**3.5**	0.7	0.1
h-CLAT + KEC + DPRA(Cys + Lys)	0.01	0.53	0.38	0.08	0.03	**3.1**	1.7	0.4
h-CLAT + KEC + DPRA(Cys + Lys) + TIMES	0.00	0.69	0.26	0.05	0.00	**6.2**	0.9	0.2
h-CLAT + KEC+ DPRA(Cys + Lys) + bioav.	0.01	0.55	0.34	0.10	0.01	**3.8**	1.4	0.5
h-CLAT + KEC + DPRA(Cys + Lys) +bioav. + TIMES	0.00	0.71	0.22	0.06	0.00	**6.8**	0.8	0.3

Bold values indicate Bayes factor that drives the decision

3. *Completeness of MIE evidence check: Does the dataset have evidence on both: cysteine and lysine?*

Data are available for Cys and Lys MIEs. Complete dataset.

4. *Assessment of applicability domains*

   (a) Identified to be a pre-hapten by TIMES.
   (b) Water solubility within acceptable range for all assays.
   (c) Chemical mostly in ionized form (*f*_ion = 0.9) but within acceptable range

5. *Integration of all the in-domain evidence and prediction of the pEC3 probability distribution* (Table 19)

   (a) DPRA: based on Cys- and Lys-peptide remaining values, the chemical is predicted to be a NS; however, this result needs to be taken with caution because it is also predicted a pre-hapten.
   (b) Similarly in the KeratinoSens™, based on KEC1.5 and KEC3, the chemical is predicted to be a NS; however, this result needs to be taken with caution because it is predicted a pre-hapten. In addition, this chemical is cytotoxic at a level below reactivity making the reactivity readouts less reliable. Taking into account KEC1.5, KEC3 and IC50, the chemical is predicted to be most likely a moderate sensitizer ($B = 2.5$). However, the chance that the chemical is a weak sensitizer is about half of that ($1.3/2.5 = 0.\ 5$).
   (c) h-CLAT data show only CD54 and no CD86 induction, indicating a weak-to-strong sensitizer based on the whole h-CLAT evidence.
   (d) Combination of DPRA, KeratinoSens™, and h-CLAT data supports a W class ($B = 3.1$). By addition of TIMES, the combined data strongly supports the hypothesis that this chemical is a weak sensitizer ($B = 6.2$).
   (e) Combination of all data, including bioavailability, further confirms hypothesis this chemical is a weak sensitizer ($B = 6.8$). If less reliable data are removed due to cytotoxicity (i.e., KS) or violation of the applicability domain (i.e., DPRA) due to the pre-hapten feature, the hypothesis is based only on cytotoxicity, TIMES and bioavailability, which suggests the chemical is a weak sensitizer ($B = 4.6$)

6. *Post-processing step of probability distribution correction for MA, if applicable.*

Not needed. This chemical is not a direct MA.

7. *Conversion of probability distribution to Bayes factors for final interpretation and decision.*

Taking into account all input parameters according to AOP, this chemical is deemed to be a weak sensitizer with substantial strength of evidence ($B = 6.8 > 3$). Estimated EC3 % 50th percentile is 18 %.

There are two caveats here. TIMES predicts hydroperoxide formation. However, this is a transformation happening upon forced oxidation of chemicals during many months, and not in the timescale of the LLNA or product application (Bodin et al. 2003; Skold et al. 2002), and thus it is not relevant to the chemical itself. In addition, we need to keep in mind that the LLNA sometimes generates false positives due to irritation, especially for ethoxylated surfactants (Ball et al. 2011), the irritation being triggered by the strong cytotoxicity of such chemicals. The ITS-3 analysis actually hints at the possibility that a similar mechanism applies to the chemical investigated here, as the prediction is strongly driven by cytotoxicity. When evaluating such a result, we need to keep in mind that we have trained the model specifically to predict the LLNA response.

Example 4: 2,6,6-trimethylcyclohexa-1,3-dienyl methanol (safranal) CAS# 116-26-7; LLNA EC3 = 7.5 %, moderate sensitizer illustrating MA correction

1. *Prediction of physico-chemical properties of chemicals* Calculated (see Table 20)

2. *Prediction of TIMES:*

   (a) Predicted to be strong sensitizer based on parent structure; metabolite by auto-oxidation predicted to be weak sensitizer
   (b) Possible pre-hapten predicted by TIMES
   (c) Direct MA active alert, di-substituted α,β-unsaturated aldehyde

3. *Completeness of MIE evidence check: Does the dataset have evidence on both: cysteine and lysine?*

Data are available for Cys and Lys MIEs. Complete dataset.

4. *Assessment of applicability domains*

   (a) Considered to be a possible pre-hapten. Despite this, all assays indicate reactivity, and thus all data can be used as evidence.

Table 20 Input data overview for safranal

EC150 µM	EC200 µM	CV75 µM	DPRA-Cys % rem	DPRA-Lys % rem	KEC1.5 µM	KEC3 µM	IC50 µM	TIMES	f_ion	Log D@pH7	PB	Ws@pH7
176.2	256.2	456.8	8.2	100	5.4	33.5	337.3	3	0	2.8	40	0.008

Table 21 EC3 probability distribution and Bayes factors for individual and combinations of inputs for safranal

Evidence	pEC3 C1	pEC3 C2	pEC3 C3	pEC3 C4	B(C1)	B(C2)	B(C3)	B(C4)
TIMES	0.05	0.24	0.42	0.29	0.1	0.9	**1.9**	1.7
DPRA (Cys + Lys)	0.12	0.3	0.36	0.22	0.4	1.2	**1.5**	1.2
KS (KEC1.5, KEC3, IC50)	0.17	0.13	0.23	0.64	0.5	0.4	0.7	**4.9**
h-CLAT (EC150, EC200, CV75)	0.13	0.25	0.56	0.18	0.4	0.8	**2.7**	0.8
h-CLAT + KS	0	0.04	0.42	0.53	0.0	0.1	2.0	**4.7**
TIMES + DPRA(Cys + Lys)	0.02	0.23	0.48	0.27	0.06	0.8	**2.5**	1.50
h-CLAT + KS + bioav. + TIMES	0	0.02	0.49	0.49	0.0	0.1	2.6	**3.9**
h-CLAT + KS + DPRA(Cys + Lys) + bioav. + TIMES	0	0.03	0.57	0.4	0.0	0.1	**3.6**	2.7
h-CLAT + KEC + DPRA(Cys + Lys) + bioav. + TIMES + MA correction	0.02	0.35	0.43	0.12	0.06	1.7	**2.3**	0.6

Bolded number indicates Bayes factor that drives the decision

(b) Water solubility within acceptable range for all assays. Chemical is neutral.

5. *Integration of all the in-domain evidence and prediction of the pEC3 probability distribution* (Table 21)

   (a) The chemical is predicted as a strong or moderate sensitizer by TIMES.
   (b) Based on Cys and Lys-peptide depletion values, the chemical is predicted to be a moderate allergen.
   (c) The KS data, however, are in conflict with Cys- and Lys-reactivity and TIMES data. KS points toward a strong sensitizer. The results from the h-CLAT activation assay are indicating a moderate sensitizer.
   (d) Combination of KS with h-CLAT data strongly supports a strong potency class ($B = 4.7$) without MA correction and moderate after MA correction (data not shown).
   (e) Combination of all data (including bioavailability) results in a clear hypothesis that a chemical is a moderate sensitizer ($B = 3.6$).

6. *Post-processing step of probability distribution correction for MA, if applicable.*

In this case, the MA correction does not change the predicted potency class—Bayes factor is still the largest for class C3 (Table 21). However, before the correction the ITS-3 predicts the chemical to be a moderate sensitizer, but there is still a large probability associated with class C4. The prediction changes after MA correction. Based on Bayes factors, safranal is 1.3 times more likely to be a moderate than a strong sensitizer ($3.6/2.7 = 1.3$). The MA correction shifts the probability mass toward the weak sensitizer class; safranal is 3.8 (2.3/0.6) times more likely to be a moderate sensitizer than a strong one. Expressed as an estimated EC3 % 50th percentile, predicted EC3 is 2 % without and 4.9 % with MA correction.Bolded number indicates Bayes factor that drives the decision.

7. *Conversion of probability distribution to Bayes factors for final interpretation and decision.*

Taking into account all input parameters according to the AOP, this chemical is considered to be a moderate sensitizer with a weak strength of evidence ($B < 3$). While application of the MA correction factor reduced the strength of evidence for the moderate class, it shifted the remaining probability away from the strong sensitizer class and toward the weak sensitizer class, which is in a better agreement with the EC3 value of 7.5 %.

Discussion

The presented ITS for skin sensitization potency assessment—ITS-3—builds upon previously published work, ITS-1 (Jaworska et al. 2011) and ITS-2 (Jaworska et al. 2013), in which we use a Bayesian network as the underlying framework of the ITS.

The main goal of the present work, to increase accuracy, precision, and robustness of the predictions for the entire range of potency beyond the results achieved in ITS-2, was met with success. This result was possible by (1) refinement of the skin sensitization process representation in the network structure, (2) conversion of potency from weight to molar units, (3) generation of a large dataset that increased size and diversity of the underlying database, (4) establishment of a structured prediction process that considers the assays' applicability domains, and (5) consideration of bioavailability in vivo and in in vitro assays. This work demonstrates that skin sensitization potency prediction based on data from 3 key events, and often less, is possible, reliable over broad chemical classes, and ready for practical applications.

First, we improved the integration of biology knowledge into our quantitative BN structure. We believe that our approach is uniquely suited to represent a complex biological process by coding it to the structure of the ITS. Proper representation of the process is very central in the system analysis field and regarded critical to achieve robust predictions (Brase and Brown 2009). BNs in conjunction with Bayesian statistical techniques facilitate the combination of domain knowledge and data. Importance of prior or domain knowledge is critical when data are scarce or expensive as is the case with toxicity testing in general. BNs are based on causal semantics that makes the encoding of causal prior knowledge particularly straightforward. In addition, BNs encode the strength of causal relationships with probabilities.

In the BN ITS, the individual information sources are not used as stand-alone assays, but the outcomes are used to derive interim conclusions and to select, on the basis of VoI, which assays are needed next. The BN ITS uses quantitative WoE based on Bayesian statistics to update the hypothesis about LLNA potency every time after new information is provided. Most importantly, in the BN, existing prediction models for individual assays previously published are not used, and thus a potential bias from a prediction model which was trained for a stand-alone use of an assay is not integrated into the ITS-3. This is a key difference from most other ITS approaches, which aim at using a combination of stand-alone prediction models to arrive at improved predictions. The weighing of AOP events during model development is accomplished by CPTs. CPTs are matrices representing dependency relationships between discretized variables connected by arcs. The CPTs are populated by data from the database that serves as the training set for the ITS-3. This is a much more robust representation of the dependency between variables than, for example, a constant value.

Aligning the ITS-3 network structure with the AOP structure allows interpretation of results in the biological context and is chemical specific. This issue is currently underappreciated in the toxicology community where the efforts have been more focused on the development of assays representing a single element of the adverse outcome process or statistical models using combinations of inputs, including biological assays. It is our opinion that further development of AOP frameworks will facilitate transition from statistical to mechanistic ITS frameworks.

In the ITS-3 the AOP structure (i.e., sequence of events, MIEs), as well as data related to AOP Key Events 1, 2 and 3, is encoded in the ITS-3 network structure. Cysteine and lysine reactivity are treated as two separate, independent MIEs. The need for such a structure is driven by the molecular targets of the in vitro assays. The KeratinoSens™ assay is a luciferase reporter cell line assay which targets activation of a single signaling pathway, the Nrf2 (nuclear factor-erythroid 2-related factor 2)-antioxidant response element (ARE) pathway. The specific molecular events in this pathway are well characterized. The sensor protein Keap1 (Kelch-like ECH-associated protein 1) contains highly reactive Cys-residues. In an uninduced state, Keap1 is bound to the Nrf2 transcription factor. When electrophilic molecules covalently bind to the Cys-residues, Keap1 dissociates from Nrf2 which then is available to activate ARE-dependent genes such as those coding for phase II detoxifying enzymes (Dinkova-Kostova et al. 2005; Natsch 2010). Due to this mode of action, the KeratinoSens™ assay is highly responsive to Cys-reactive chemicals and not to Lys-only reactive chemicals. Specific molecular targets reacting specifically to Cys or Lys have not been fully described for the h-CLAT. The h-CLAT measures the upregulation of two cell surface proteins, CD86 and CD54, as markers of DC activation in THP-1 cells, a human monocytic leukemia cell line. The increased expression of these markers can be attributed to many different cellular pathways which have been partially revealed. One key pathway for CD86 activation is upregulation of p38 kinase activity (Miyazawa et al. 2008). Interestingly, it was shown that p38 is upregulated in presence of sensitizers due to modifications of Cys-residues on the cell surface (Kagatani et al. 2010). Based on this proposed mechanism, there is, at least for CD86, probably also a Cys-dependent pathway at work. This is maybe reflected in our network by the fact that we found a relatively high MI between the Cys and the h-CLAT nodes. Only DPRA assay ability to identify Lys selective chemicals.

Another novel aspect for ITS-3 is handling of bioavailability. Bioavailability is applied to both in vivo and in vitro data. The former estimates the potential of a chemical to penetrate the stratum corneum, while the latter aims to correct the nominal concentration to the free concentration in in vitro assays. Despite the fact that contribution of bioavailability is relatively small, we found it still a significant contributor for predicting potency.

Next, we further integrated chemistry into the ITS-3 framework by formulating applicability domains for DPRA, KeratinoSens™, and h-CLAT assays and then considering them in the prediction process. The ITS-3 process considers metabolic transformation and auto-oxidation using TIMES predictions as a way to assess a chemical's potential for being a pre- or pro-hapten. Moreover, physico-chemical properties (water solubility and fraction ionized) are used to define chemistry-oriented applicability domains. We exploit the fact that a BN allows building a hypothesis based on partial information to introduce the notion of applicability domain in the prediction process, i.e., only relevant data are used for predictions. Data outside applicability domains are not included in the integrated prediction or are treated with caution according to the prediction process.

The next foundational element for refined potency prediction was a switch from weight to molar units for potency. The prediction is provided as a pEC3 probability distribution for four potency classes. They closely represent traditional classes based on EC3 in %. One may ask—was it necessary? We strongly believe it was. Predictive models, especially for potency, need to be grounded in chemistry and biology over the entire dynamic range of responses, and expression of potency in molar units is necessary to achieve this. Till now, many predictive models were constructed with the goal to directly predict the in vivo result expressed on a weight basis in order to directly meet existing regulatory needs. While this seems pragmatic, and we did this in the past as well (ITS-1, ITS-2), our opinion evolved as we are pushing the boundaries of what is possible in terms of refinement and robustness of potency prediction. In the animal testing paradigm, decisions have been based on external dose. In the toxicology of the twenty-first century paradigm, we expect an increasing number of studies to be performed in molar units and to use internal dose as the dose metric. This can always be converted to the traditional weight-based units to inform chemical management decisions.

To increase size and diversity of the underlying database large, efforts in data generation were undertaken to create the largest dataset to date with full records on DPRA, KeratinoSens™ and h-CLAT. The ITS-3 database includes 207 chemicals, an almost 50 % increase over the ITS-2 database of 145 chemicals (Natsch et al. 2013), and it integrates

Table 22 Balanced accuracy (bac) for four potency classes for ITS-2 (Jaworska et al. 2013) and ITS-3

	All			C3 and C4 (similar to $M + S$)		
	n	Bac %	95 % CI	n	Bac %	95 % CI
ITS-2	21	85	70–100	10	80	55–100
ITS-3	60	89	81–97	27	82	68–96

all three assays validated by EURL ECVAM. This allowed robust estimation of the network parameters, i.e., CPTs, as well as to carry out an extensive evaluation of the ITS-3 performance with an external validation set comprising 60 chemicals. Test sets in computational toxicology are notoriously small, and usually only absolute accuracies are reported and characterized by high uncertainty. Yet this uncertainty is not reported. This leads to surprises when models start to be used in practice because much lower accuracies are observed. We analyzed bac values along with their respective confidence intervals that take into account size of the test sets that provides a more comprehensive assessment of predictive capacity of a model for ITS-3 and ITS-2 (Jaworska et al. 2013) (Table 22).

As an overall conclusion, incorporation of all these novel components summarized above reduces the uncertainty of predictions of all potency classes compared with ITS-2.

The most impactful variables

Times

The most impactful input variable is TIMES. TIMES represents collective expert knowledge, quantitative reactivity-based relationships and metabolism. Despite the 60 % overlap between the ITS-3 and the TIMES training sets, there is no inflation of the weight of TIMES for two reasons. First, TIMES is an expert system that makes a prediction based on alerts, reactivity parameters, in vivo data and expert mechanistic knowledge. In that sense it is very different from a machine learning approach where the overlap would be an issue. Second, the weight of TIMES in the ITS-3 is based on the accuracies achieved for the training set. These are 93 % for C1, 60 % for C2, and 70 % for C3 and C4. These numbers correspond quite well with TIMES external validation using 40 chemicals (Roberts et al. 2007) where accuracies were 88 % for NS and 56 % for S. The lower performance for S was attributed to the poor prediction of metabolism at that time. Since 2007, there has been a lot of development to address this topic. A recent paper by Patlewicz et al. (2014b) describes these developments. Therefore the current 'weight' of TIMES is not strongly inflated by the overlap between training sets of TIMES and ITS-3 but is a reflection of reality.

Cytotoxicity

We observe a strong contribution of cytotoxicity to the prediction of potency classes, even stronger than the Cys- and Lys-reactivity nodes. Does this indicate that cytotoxicity is a stronger contributor to sensitization potency as compared to reactivity, contradicting common wisdom? There are a number of points to consider here. First there are two mechanistic explanations. The first one we used as motivation to add cytotoxicity as explanatory variable in the network structure.

To recall, in order to trigger the sensitization response in vivo there is, next to hapten formation, the need for a danger signal in the form of local trauma triggering the emigration of DC. This danger signal appears to involve formation of extracellular ATP and breakdown products of hyaluronic acid (Esser et al. 2012; Weber et al. 2010). Release of ATP from cells in particular is triggered by cytotoxicity, and cytotoxic surfactants therefore do have the ability to provide this local trauma. Classical maximization tests used this ability by amplifying the sensitization reaction to chemicals by simultaneous or pretreatment with irritants. In the LLNA, which we model in our analysis, no such adjuvant is given. Thus in the LLNA a chemical must provide both the hapten and the danger signal in order to trigger the response. Therefore the LLNA measures both the haptenic potential and the danger signal provided by the chemical, and a chemical with stronger danger signal potential will be rated stronger in the LLNA. Based on this reasoning, we would expect a lower contribution of cytotoxicity when modeling maximization tests, in which the danger signal is provided by (co-)treatment with irritating adjuvantia. Unfortunately this hypothesis is difficult to test in detail as the guinea pig maximization test yields mainly qualitative and only limited quantitative information. Still it is important to keep this simple fact in mind—when we model LLNA, we model a situation in which the chemical is applied without any concomitant danger or adjuvants, which is different from a human exposure situation in which the danger signal may come from other chemical or physical insults or from preexisting inflammation. Thus, there is a certain risk that the strong weight for cytotoxicity identified here is specific to the LLNA, which we model here, and not to sensitization in general. There is also the risk that cytotoxicity of a chemical exaggerates predicted sensitization risk in ITS-3, as may be the situation in Example 3 of the case studies.

Secondly, there is an intrinsic link between cytotoxicity and reactivity: chemicals with strong cysteine reactivity are in general cytotoxic, as modification of cysteine residues in enzymes blocks key vital processes and directly leads to cytotoxicity (referred to as 'excess toxicity' in environmental toxicology as compared to 'narcotic baseline toxicity' (Bohme et al. 2009). Thus, part of the good correlation between cytotoxicity and sensitization potency must not be a causal effect but rather a correlation effect, due to the simple fact that strongly cysteine-reactive chemicals are always both sensitizing and cytotoxic. Hence, cytotoxicity may be a predictor of potency even if it is not cytotoxicity, but rather the underlying reactivity, which leads to the sensitization potency.

There is also a caveat that the result may partly be driven by the database composition. The publicly available LLNA database contains a relatively high number of non-sensitizers with low molecular weight and very low cytotoxicity (such as butanol, propylene glycol, glycerol, benzoic acid, lactic acid). The database may not fully represent the distribution of non-sensitizers in the chemical universe, and only due to this database bias, has cytotoxicity attained the highest weight to predict non-sensitizers in our analysis.

Finally, one needs to keep in mind that the TIMES node is also mainly reflecting reactivity, as TIMES predictions are based on reactive alerts in the molecules. Thus, taken together the cysteine, lysine and TIMES nodes contribute more to the predictions as compared to cytotoxicity (which itself is partly driven by reactivity)—confirming the common wisdom.

Practical use

Standardization of inputs

To increase the practical utility of ITS-3, only validated assays are used. To this end, we replaced the in vitro U937 test related to Key Event 3, DC activation, with the h-CLAT that has been validated by EURL ECVAM (Joint Research Centre of the European Union 2015). Thus, this is the first ITS incorporating data of the three assays validated by EURL ECVAM (DPRA, KeratinoSens™, and h-CLAT) in a quantitative manner. In addition, we simplified bioavailability inputs to just physico-chemical properties and eliminated the need to run Kasting's skin penetration model (Dancik et al. 2013). This new bioavailability representation allows for easier yet more robust characterization of bioavailability that handles both neutral and ionized chemicals.

Improved process facilitating optimized testing strategy

We refined the prediction process. We made it a very structured, step-by-step process and demonstrated how to use it with four diverse examples. We exploited the fact that the BN ITS framework can build a hypothesis with partial data. We introduced consideration of applicability domains of individual assays into the process of gathering evidence, and elimination of evidence if it was deemed outside the applicability domain of a particular assay due to physico-chemical property limits or biological domain. By feeding to the ITS-3 only relevant data, we reduce mispredictions and frequently occurring input data conflicts. Usually when there is no conflict among data, a correct hypothesis is formed. In contrast, predictions based on input data which are in conflict with each other always result in a flat probability distribution and compromised precision/higher uncertainty. Being able to explain data conflict is therefore critical to the successful use of ITS. Eliminating data on the basis of being outside of the applicability domains is a key tool for removing data conflict.

In case of missing evidence on one MIE (i.e., Lys or Cys-reactivity), we recommend applying additional caution to the prediction or collecting data on the lacking MIE. Based on a simple statistical analysis, there are many test combinations that may lead to a prediction that will be deemed acceptable for the purpose of a decision [see e.g., Bauch et al. (2012); van der Veen et al. (2014); Urbisch et al. (2015)]. We propose to separately check the completeness of the evidence in a biological sense. Since most assays are cysteine reactivity oriented, this check is critical for Lys selectively reactive chemicals. In cases where evidence on MIE Lys-reactivity is missing, one may be misled by an apparent high certainty of a prediction, i.e., a high Bayes factor for a particular class (see Example 2 in the case studies).

As discussed earlier, in BN ITS, the weight of the individual information source is context specific. This adaptive nature of Bayesian frameworks has important implications for decision-making. It treats the decision-making process as a dynamic process. Bayesian frameworks account for dynamically changing interrelationships between tests based on evidence provided, and thus there is no constant weight of for an AOP event. Since the process of adding evidence to the BN ITS can be sequential (and not all at once), interim predictions and decisions can be generated. As a result, BN ITS can also be used to guide and optimize a testing strategy before testing is commenced.

Furthermore, there are circumstances when more information on a particular chemical is necessary. For those situations, the assessor must decide which study—or studies—would yield the most relevant information for the risk decision while being mindful of the applicability domain for the input assays. A poorly informed decision could result in unnecessary testing. For example, when the outcome of TIMES indicates that the chemical of interest is a potential pro-hapten (i.e., the parent molecule is predicted to be non-sensitizing and a metabolite is predicted to be reactive after metabolic transformation), conduct of a DPRA would be of limited value as pro-haptens are out of its applicability domain. Similarly, results from one of the cell-based assays, KeratinoSens™ or h-CLAT, for chemicals with low water solubility (e.g., between 2.5e−08 and 1.7e−04M) may not be reliable and therefore have little relevance. In addition to the decision on the relevance of potential experiments, there is the equally complex problem of deciding how data from a new experiment, which can address different levels of biological organization, can be incorporated into the existing body of knowledge about the candidate chemical. Using the BN ITS, an assessor may explore the impact of the additional experimental data on the reliability of the BN prediction by observing its effect on the Bayes factor.

Utility of the ITS-3 output for quantitative risk assessment (QRA)

The ultimate goal of the sensitizer potency assessment is to use the results for QRA for skin sensitization of chemicals (Api et al. 2008; Gerberick et al. 2001). Currently, after a No Expected Sensitization Induction Level (NESIL) is established, sensitization assessment factors (SAF) are used to transform the NESIL into an acceptable exposure level (AEL), which is a finite maximal dose considered safe for human exposure. When predictions are made for a sensitization potency class, one approach to apply this output in a QRA is to assign a conservative default value to each class for use as the NESIL (e.g., one may assume a 'worst case' of NESIL of 100 µg/cm^2 for all chemicals predicted as moderate by the ITS-3) (Gerberick et al. 2001).

The output of the ITS-3 is a potency probability distribution. In the current framework of QRA, the probabilistic readout would need to be transformed to a deterministic value, which then could be used as a NESIL. The question arises—what would be the most appropriate finite value representing the predicted distribution? To this end, we explored different pEC3 values derived from 50th, 60th, 70th, 80th and 90th percentiles of the pEC3 probability distribution for the test set chemicals (Sheet 3 Supplementary file).

To guide us in determining what percentile would be acceptable, we examined which chemicals would be underpredicted by a given percentile. First we considered 'most likely predicted LLNA EC3 values.' This most likely value by definition is the 50-percentile, i.e., the dose at which the likelihoods for a lower or higher EC3 values are balanced. In two-thirds of the cases, the in vivo LLNA EC3 value is not more than a factor of two below this predicted value. As a factor of two is considered a typical variability of the LLNA EC3 value itself, such a result may be used directly—but for one-third of the chemicals there remains a more significant underestimation of the sensitization potential as revealed by the LLNA. So the assessment may need to be somewhat more conservative. One option is to move from a 50 % percentile to, e.g., the 90 % percentile, i.e., the concentration at which the chance of a lower real EC3 value is only 10 %.

Clearly the 90th percentile is overly conservative, attributing very low EC3 values to a number of chemicals (Table 1 in the supplementary file). This happens especially for chemicals with a relatively flat probability distribution or chemicals with a low maximal Bayes factor for any of the classes. These include the three polycyclic compounds 7,12-dimethylbenz[α]anthracene, 1-chloromethylpyrene and benzo(α)pyrene. Very poor solubility and a need for metabolic activation indicated these specific chemicals are out of the applicability domain of all three experimental assays. Thus, their prediction was based only on TIMES and bioavailability-related physico-chemical variables. As such, the evidence for these chemicals to be strong sensitizers is rather weak, i.e., characterized by large uncertainty (Table 8, $B(S) = 1.75$). In addition, we do not know the exact EC3 values for these chemicals, as the reported values are all extrapolations from experiments done at (much) higher concentration. Finally, by design, extreme sensitizers (four of the nine chemicals that are strongly underpredicted) are not precisely predicted with ITS-3 model, which bins strong and extreme chemicals into one class. Due to these limitations, we conclude that for extreme and to some degree strong class we cannot reliably estimate EC3.

Our analysis indicates that reasonable predictions for EC3 values for moderate and weak sensitizers can be made. To identify a protective percentile for these classes, we need to consider LLNA experimental variability, which is considered to vary by a factor of two in both directions. To this end, we halved the EC3 predicted by a given percentile to represent a conservative prediction of EC3, 0.5* EC3. 70th percentile corresponded best to 0.5*EC3 (closest yet conservative) for the all C2 and C3 chemicals in the test set (sheet 3 in Supplementary File). Overall, this discussion indicates that a probability distribution for classes predicted by the ITS-3 may be transformed into a NESIL for use in a skin sensitization QRA, but continued learning for this process and comparison to other approaches will be needed.

Uncertainty in the ITS-3 approach prediction

From a policy perspective, the value of a model-based analysis lies not simply in its ability to generate a precise point estimate for a specific outcome, but also in the systematic examination and reporting of uncertainty surrounding the prediction and the ultimate decision for which it is applied. Below we discuss sources of uncertainty in the ITS-3. Next, we discuss how these sources of uncertainty translate into prediction uncertainty. Finally, we explain how assessment of uncertainty is quantified in prediction for a new chemical.

The uncertainty in the ITS-3 framework comes from two main sources: uncertainty in the model structure and uncertainty in the experimental data. The former is related to uncertainty in knowledge and limitation in the coverage of AOP key events by the current tests. The latter is associated with the inherent variability of biological data. The ITS-3 model structure aims to correctly represent the mechanisms of the skin sensitization process. It is developed purely based on mechanistic knowledge with the aim to follow the sequence of the mechanistic events in the existing AOP. The uncertainties associated with the existing AOP are reflected in the ITS-3 model structure.

Inherent variability of data is an important aspect of biological data. Variability of skin sensitization data is relatively well documented. Regulatory acceptance and low cost, combined with the lack of a central repository for sharing LLNA data, resulted in several LLNA assays per chemical. Due to the freedom to choose a vehicle, as well as the inherent biological variability of the response in vivo, EC3 values from multiple tests generated in different laboratories may vary by an order of magnitude (ICCVAM 2011). Expressing potency as a four-class distribution, as in our work, partially compensates for EC3 variability. Using the traditional five-category system for potency (Kimber et al. 2003), Hoffmann (2015) reported that 29 % of such repeated LLNA tests may change potency category compared with the median value of the repeat tests when different vehicles are used. Thus, potency categories were consistent with the median for 71 % of the tests. The target data variability has the largest impact for the moderate sensitizer predictions by the ITS-3 since this class, within the chemicals with positive evidence, is the only class flanked on both sides by another class with positives. It is followed by the weak sensitizer class representing one order of potency (but since many chemicals have maximal test concentration of 25–50 % it is, in practice, less than one-order-potency class). Predictions of non-sensitizers and strong sensitizers are least impacted. Such impacts are consistent with that shown by Hoffmann (2015).

Variability in vitro is expected to be smaller than variability in vivo. In vitro assay systems are more biologically simple and more standardized than the in vivo systems. The between-laboratory reproducibilities for non-sensitizer/sensitizer outcomes for DPRA, KeratinoSens™, and h-CLAT were: 75, 86, and 80 %, respectively (Joint Research Centre of the European Union 2013, 2014, 2015). Reproducibility is defined as the proportion of chemicals tested that had concordant results among the laboratories that participated in the validation study, i.e., all laboratories had the same sensitizer/non-sensitizer classification for this proportion of substances tested. While it is not possible to assess reproducibility of these assays for four-class potency without assuming a prediction model, it is nevertheless safe to say that it will be lower than the one reported for sensitizer/non-sensitizer classification. Thus it appears that variability in vivo and the variability of the three considered assays is probably comparable.

Deterministic models have very limited scope for correctly handling intrinsic data uncertainty, while probabilistic models have a naturally built-in capability to handle it. Specific input data variability can be explicitly considered in the ITS-3 prediction. First, the variability of the input data from the information source considered must be quantified. Next, every piece of evidence is represented as a range. There are two possibilities regarding how the evidence fits into bins. One possibility is that the evidence range is within one bin of the discretized distribution. In such situations, the prediction is not impacted by the variability of this particular information source. The other possibility is that the evidence range straddles across two discretization bins (theoretically, it could be more, but the principle is the same). In this case, the evidence can be entered as two bins so as to distribute the probability mass according over the ranges of the two bins (i.e., 30 % for $x < 10$ and 70 % for $x > 10$). Consideration of the uncertainty of the data related to data quality can be examined in a similar way.

The ITS-3 prediction for a new chemical, being probabilistic, inherently includes assessment of uncertainty associated with this prediction. Further, conversion to Bayes factors allows for a consistent acceptance of uncertainty in predictions based on fit for purpose criteria. This uncertainty reflects the combined uncertainty associated with ITS-3 structure and, in part, uncertainty due to the variability of input information sources as well as the target, i.e., LLNA pEC3.

Restricting predictions to potency classes and inputs to the intervals of the discretized distributions partially compensates for the inherent variability described above. Given the accuracy of the test set predictions achieved in this study, there is a reason to believe that much more precision in the potency prediction is not possible with existing skin sensitization data. In our opinion, further improvement in the predictive capability of new models hinges on the reduction of variability of experimental methods used both as inputs as well as the benchmark and less in generating more data.

ITS-3 accessibility

To access ITS-3 for a web-based application please contact directly Joanna Jaworska.

Acknowledgments Roger Emter generated new KeratinoSens data. Leslie Foertsch generated new DPRA data. Morihiko Hirota and Yuuki Ootsubo generated new h-CLAT data. Data generation was partially funded by Research Institute of Fragrance Materials. Yuri Dancik was involved in the early stages of data compilation. Judy Strickland was supported by NIEHS contract HHSN273201500010C in support of NICEATM.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical statement This article does not contain any studies with human participants or animals performed by any of the authors.

References

Adler S, Basketter D, Creton S, Pelkonen O, van Benthem J, Zuang V, Andersen KE et al (2011) Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. Arch Toxicol 85(5):367–485

Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A (2015) Predicting chemically-induced skin reactions. Part II: QSAR models of skin permeability and the relationships between skin permeability and skin sensitization. Toxicol Appl Pharmacol 284(2):273–280

AP (2014) federal ban sought for animal testing on cosmetics USA Today. http://www.usatoday.com/story/news/politics/2014/11/15/federal-ban-animal-testing-cosmetics/19090873/. Accessed 26 Oct 2015

Api AM, Basketter DA, Cadby PA, Cano MF, Ellis G, Gerberick GF, Griem P et al (2008) Dermal sensitization quantitative risk assessment (QRA) for fragrance ingredients. Regul Toxicol Pharmacol 52(1):3–23

Ashikaga T, Yoshida Y, Hirota M, Yoneyama K, Itagaki H, Sakaguchi H, Miyazawa M et al (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(5):767–773

Ball N, Cagen S, Carrillo JC, Certa H, Eigler D, Emter R, Faulhammer F et al (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 Pharmacol 60:389–400

Basketter DA, Kimber I (2009) Updating the skin sensitization in vitro data assessment paradigm in 2009. J Appl Toxicol 29(6):545–550

Basketter D, Kimber I (2010) Re: updating the skin sensitization in vitro data assessment paradigm in 2009—a chemistry and QSAR perspective. J Appl Toxicol 30(3):289

Basketter DA, Clewell H, Kimber I, Rossi A, Blaauboer B, Burrier R, Daneshian M et al (2012) A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing. ALTEX 29:3–89

Bauch C, Kolle SN, Ramirez T, Eltze T, Fabian E, Mehling A, Teubner W et al (2012) Putting the parts together: combining in vitro methods to test for skin sensitizing potentials. Regul Toxicol Pharmacol 63(3):489–504

Beltrani VS, Bernstein IL, Cohen DE, Fonacier L (2006) Contact dermatitis: a practice parameter. Ann Allergy Asthma Immunol 97(SUPPL. 2):S1–S38

Bodin A, Linnerborg M, Nilsson JL, Karlberg AT (2003) Structure elucidation, synthesis, and contact allergenic activity of a major hydroperoxide formed at autoxidation of the ethoxylated surfactant C12E5. Chem Res Toxicol 16(5):575–582

Boeniger MF, Ahlers HW (2003) Federal government regulation of occupational skin exposure in the USA. Int Arch Occup Environ Health 76(5):387–399

Bohme A, Thaens D, Paschke A, Schuurmann G (2009) Kinetic glutathione chemoassay to quantify thiol reactivity of organic electrophiles–application to alpha, beta-unsaturated ketones, acrylates, and propiolates. Chem Res Toxicol 22(4):742–750

Brase JM, Brown DL (2009) Modeling, simulation and analysis of complex networked systems. A program plan. U.S. Department of Energy

Bureau of Labor Statistics (2014) Employer-reported workplace injuries and illnesses–2013. Supplemental News Release Tables. Table SNR10. Numbers of nonfatal occupational illnesses by industry and category of illness. http://www.bls.gov/iif/oshwc/osh/os/ostb3971.pdf. Accessed 26 Oct 2015

Bus JS, Becker RA (2009) Toxicity testing in the 21st century: a view from the chemical industry. Toxicol Sci 112(2):297–302

Cohen S, Cohen S (1966) Preparation and reactions of derivatives of squaric acid. Alkoxy-, hydroxy-, and aminocyclobutenediones. J Am Chem Soc 88(7):1533–1536

Dancik Y, Miller MA, Jaworska J, Kasting GB (2013) Design and performance of a spreadsheet-based model for estimating bioavailability of chemicals from dermal exposure. Adv Drug Deliv Rev 65(2):221–236

De Groot AC (1994) Patch testing: test concentrations and vehicles for 3700 chemicals, 2nd edn. Elsevier, New York

De Wever B, Fuchs HW, Gaca M, Krul C, Mikulowski S, Poth A, Roggen EL et al (2012) Implementation challenges for designing integrated in vitro testing strategies (ITS) aiming at reducing and replacing animal experimentation. Toxicol In Vitro 26:526–534

Dimitrov SD, Low LK, Patlewicz GY, Kern PS, Dimitrova GD, Comber MH, Phillips RD et al (2005) Skin sensitization: modeling based on skin metabolism simulation and formation of protein conjugates. Int J Toxicol 24(4):189–204

Dinkova-Kostova AT, Holtzclaw WD, Kensler TW (2005) The role of Keap1 in cellular protective responses. Chem Res Toxicol 18(12):1779–1791

El Ali Z, Gerbeix C, Hemon P, Esser PR, Martin SF, Pallardy M, Kerdine-Romer S (2013) Allergic skin inflammation induced by chemical sensitizers is controlled by the transcription factor Nrf2. Toxicol Sci 134(1):39–48

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(3):281–290

Esser PR, Wolfle U, Durr C, von Loewenich FD, Schempp CM, Freudenberg MA, Jakob T et al (2012) Contact sensitizers induce skin inflammation via ROS production and hyaluronic acid degradation. PLoS One 7(7):e41340

European Union (2009) Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. OJL 342(59):59–209

Gerberick GF, Robinson MK, Felter SP, White IR, Basketter DA (2001) Understanding fragrance allergy using an exposure-based risk assessment approach. Contact Dermat 45(6):333–340

Gerberick GF, Vassallo JD, Bailey RE, Chaney JG, Morrall SW, Lepoittevin JP (2004) Development of a peptide reactivity assay for screening contact allergens. Toxicol Sci 81(2):332–343

Gerberick GF, Ryan CA, Kern PS, Schlatter H, Dearman RJ, Kimber I, Patlewicz GY et al (2005) Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods. Dermatitis 16(4):157–202

Gerberick GF, Vassallo JD, Foertsch LM, Price BB, Chaney JG, Lepoittevin JP (2007) Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. Toxicol Sci 97(2):417–427

Goodman SN (1999) Toward evidence-based medical statistics. 2: the Bayes factor. Ann Intern Med 130(12):1005–1013

Groothuis FA, Heringa MB, Nicol B, Hermens JL, Blaauboer BJ, Kramer NI (2015) Dose metric considerations in in vitro assays to improve quantitative in vitro-in vivo dose extrapolations. Toxicology 332:30–40

Hartung T, Luechtefeld T, Maertens A, Kleensang A (2013) Integrated testing strategies for safety assessments. ALTEX 30(1):3–18

Hoffmann S (2015) LLNA variability: an essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. ALTEX. http://dx.doi.org/10.14573/altex.1505051. Accessed 26 Oct 2015

ICCVAM (2011) ICCVAM test method evaluation report: usefulness and limitations of the murine local lymph node assay for potency categorization of chemicals causing allergic contact dermatitis in humans. National Institute of Environmental Health Sciences, Research Triangle Park

Jaworska J, Hoffmann S (2010) Integrated Testing Strategy (ITS)—opportunities to better use existing data and guide future testing in toxicology. Altex 27(4):231–242

Jaworska J, Gabbert S, Aldenberg T (2010) Towards optimization of chemical testing under REACH: a Bayesian network approach to Integrated Testing Strategies. Regul Toxicol Pharmacol 57(2–3):157–167

Jaworska J, Harol A, Kern PS, Frank Gerberick G (2011) Integrating non-animal test information into an adaptive testing strategy—skin sensitization proof of concept case. ALTEX 28(3):211–225

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

Joint Research Centre of the European Union (2013) EURL ECVAM recommendation on the Direct Peptide Reactivity Assay (DPRA) for skin sensitisation testing. Publications Office of the European Union, Luxembourg

Joint Research Centre of the European Union (2014) EURL ECVAM recommendation on the KeratinoSens™ assay for skin sensitisation testing. Publications Office of the European Union, Lusembourg

Joint Research Centre of the European Union (2015) EURL ECVAM recommendation on the human cell line activation test (h-CLAT) for skin sensitisation testing. Publications Office of the European Union, Lusembourg

Jowsey IR, Basketter DA, Westmoreland C, Kimber I (2006) A future approach to measuring relative skin sensitising potency: a proposal. J Appl Toxicol 26(4):341–350

Kagatani S, Sasaki Y, Hirota M, Mizuashi M, Suzuki M, Ohtani T, Itagaki H et al (2010) Oxidation of cell surface thiol groups by contact sensitizers triggers the maturation of dendritic cells. J Invest Dermatol 130:175–183

Kimber I, Basketter DA, Butler M, Gamer A, Garrigue JL, Gerberick GF, Newsome C et al (2003) Classification of contact allergens according to potency: proposals. Food Chem Toxicol 41(12):1799–1809

Kimber I, Basketter DA, Gerberick GF, Ryan CA, Dearman RJ (2011) Chemical allergy: translating biology into hazard characterization. Toxicol Sci 120(SUPPL.1):S238–S268

Kjaerulff UB, Madsen AL (2013) Bayesian networks and influence diagrams: a guide to construction and analysis, 2nd edn. Springer, New York

Kramer NI, Krismartina M, Rico-Rico A, Blaauboer BJ, Hermens JL (2012) Quantifying processes determining the free concentration of phenanthrene in Basal cytotoxicity assays. Chem Res Toxicol 25(2):436–445

Lucas PJ, van der Gaag LC, Abu-Hanna A (2004) Bayesian networks in biomedicine and health-care. Artif Intell Med 30(3):201–214

Luechtefeld T, Maertens A, McKim JM, Hartung T, Kleensang A, Sa-Rocha V (2015) Probabilistic hazard assessment for skin sensitization potency by dose-response modeling using feature elimination instead of quantitative structure-activity relationships. J Appl Toxicol 35(11):1361–1371

Maxwell G, Mackay C (2008) Application of a systems biology approach to skin allergy risk assessment. Altern Lab Anim 36(5):521–556

McKim JM Jr, Keller DJ 3rd, Gorski JR (2010) A new in vitro method for identifying chemical sensitizers combining peptide binding with ARE/EpRE-mediated gene expression in human skin cells. Cutan Ocul Toxicol 29(3):171–192

Mehling A, Eriksson T, Eltze T, Kolle S, Ramirez T, Teubner W, van Ravenzwaay B et al (2012) Non-animal test methods for predicting skin sensitization potentials. Arch Toxicol 86(8):1273–1295

Middleton E, Reed CE, Ellis EF, Adkinson NF, Yunginger JW, Busse WW (eds) (1998) Allergy principles and practice. Mosby, St. Louis

Miyazawa M, Ito Y, Kosaka N, Nukada Y, Sakaguchi H, Suzuki H, Nishiyama N (2008) Role of MAPK signaling pathway in the activation of dendritic type cell line, THP-1, induced by DNCB and $NiSO_4$. J Toxicol Sci 33(1):51–59

Natsch A (2010) The Nrf2-Keap1-ARE toxicity pathway as a cellular sensor for skin sensitizers–functional relevance and a hypothesis on innate reactions to skin sensitizers. Toxicol Sci 113(2):284–292

Natsch A (2014) Integrated approaches to safety testing: general principles and skin sensitization as a test case. In: Reducing, refining and replacing the use of animals in toxicity testing. Issues in toxicology, vol 19. Royal Society of Chemistry, Cambridge, UK, pp 364–288

Natsch A, Emter R, Ellis G (2009) Filling the concept with data: integrating data from different in vitro and in silico assays on skin sensitizers to explore the battery approach for animal-free skin sensitization testing. Toxicol Sci 107(1):106–121

Natsch A, Haupt T, Laue H (2011) Relating skin sensitizing potency to chemical reactivity: reactive Michael acceptors inhibit NF-κB signaling and are less sensitizing than S NAr- and S N2-reactive chemicals. Chem Res Toxicol 24(11):2018–2027

Natsch A, Ryan CA, 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, Emter R, Gfeller H, Haupt T, Ellis G (2015a) Predicting skin sensitizer potency based on in vitro data from keratinosens and kinetic peptide binding: global versus domain-based assessment. Toxicol Sci 143(2):319–332

Natsch A, Emter R, Gfeller H, Haupt T, Ellis G (2015b) Predicting skin sensitizer potency based on *In Vitro* data from KeratinoSens and kinetic peptide binding: global versus domain-based assessment. Toxicol Sci 143(2):319–332

NIOSH (2012) Skin exposures and effects. Workplace safety and health. In: Centers for disease control and prevention. http://www.cdc.gov/niosh/topics/skin/ Accessed 3 Apr 2013

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(7):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(2):609–618

OECD (2010) Test No. 429. Skin sensitisation: local lymph node assay OECD guidelines for the testing of chemicals, section 4: health effects. OECD Publishing, Paris

OECD (2012) OECD series on testing and assessment no. 168. The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins. Part 1: scientific assessment. OECD Publishing, Paris

OECD (2015a) Adverse outcome pathways, molecular screening and toxicogenomics. In: OECD Publishing. http://www.oecd.org/chemicalsafety/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm Accessed 22 Jul 2015

OECD (2015b) Draft proposal for a new test guideline. *In Vitro* skin sensitisation: human Cell Line Activation Test (h-CLAT). In:

OECD Publishing. http://www.oecd.org/env/ehs/testing/Draft-Proposal-for-a-new-Test-Guideline-on-invitro-skin-sensitisation-h-CLAT.pdf. Accessed 12 Aug 2015

OECD (2015c) Test No. 442C. *In chemico* skin sensitization: direct peptide reactivity assay (DPRA) OECD guidelines for the testing of chemicals, section 4: health effects. OECD Publishing, Paris

OECD (2015d) Test No. 442D. *In vitro* skin sensitisation: ARE-Nrf2 Luciferase Test Method OECD guidelines for the testing of chemicals, section 4: health effects. OECD Publishing, Paris

Patlewicz G, Kuseva C, Kesova A, Popova I, Zhechev T, Pavlov T, Roberts DW et al (2014a) Towards AOP application–implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization. Regul Toxicol Pharmacol 69(3):529–545

Patlewicz G, Kuseva C, Mehmed A, Popova Y, Dimitrova G, Ellis G, Hunziker R et al (2014b) TIMES-SS–recent refinements resulting from an industrial skin sensitisation consortium. SAR QSAR Environ Res 25(5):367–391

Pirone JR, Smith M, Kleinstreuer NC, Burns TA, Strickland J, Dancik Y, Morris R et al (2014) Open source software implementation of an integrated testing strategy for skin sensitization potency based on a Bayesian network. ALTEX 31(3):336–340

Reisinger K, Hoffmann S, Alepee N, Ashikaga T, Barroso J, Elcombe C, Gellatly N et al (2015) Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. Toxicol In Vitro 29(1):259–270

Roberts DW, Aptula AO (2008) Determinants of skin sensitisation potential. J Appl Toxicol 28(3):377–387

Roberts DW, Patlewicz G, Dimitrov SD, Low LK, Aptula AO, Kern PS, Dimitrova GD et al (2007) TIMES-SS—a mechanistic evaluation of an external validation study using reaction chemistry principles. Chem Res Toxicol 20(9):1321–1330

Rovida C, Alepee N, Api AM, Basketter DA, Bois FY, Caloni F, Corsini E et al (2015) Integrated testing strategies (ITS) for safety assessment. ALTEX 32(1):25–40

Sasseville D (2008) Occupational contact dermatitis. Allergy Asthma Clin Immunol 4(2):59–65

Skold M, Borje A, Matura M, Karlberg AT (2002) Studies on the autoxidation and sensitizing capacity of the fragrance chemical linalool, identifying a linalool hydroperoxide. Contact Dermatitis 46(5):267–272

Su B, Zhou W, Dorman KS, Jones DE (2009) Mathematical modelling of immune response in tissues. Comput Math Method M 10(1):9–38

Takenouchi O, Miyazawa M, Saito K, Ashikaga T, Sakaguchi H (2013) Predictive performance of the human Cell Line Activation Test (h-CLAT) for lipophilic chemicals with high octanol-water partition coefficients. J Toxicol Sci 38(4):599–609

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(4):626–639

Tsujita-Inoue K, Atobe T, Hirota M, Ashikaga T, Kouzuki H (2015) In silico risk assessment for skin sensitization using artificial neural network analysis. J Toxicol Sci 40(2):193–209

UN (2013) Globally harmonised system of classification and labelling of chemicals (GHS), Fifth revised edition. United Nations, New York

UNEP (2005) OECD SIDS. Phthalic anhydride. CAS No: 85-44-9. UNEP Publishing. http://www.inchem.org/documents/sids/sids/85449.pdf. Accessed 26 Oct 2015

Urbisch D, Mehling A, Guth K, Ramirez T, Honarvar N, Kolle S, Landsiedel R et al (2015) Assessing skin sensitization hazard in mice and men using non-animal test methods. Regul Toxicol Pharmacol 71(2):337–351

van der Veen JW, Gremmer ER, Vermeulen JP, van Loveren H, Ezendam J (2013) Induction of skin sensitization is augmented in Nrf2-deficient mice. Arch Toxicol 87(4):763–766

van der Veen JW, Rorije E, Emter R, Natsch A, van Loveren H, Ezendam J (2014) Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals. Regul Toxicol Pharmacol 69(3):371–379

Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, Landesmann B et al (2014a) Adverse outcome pathway (AOP) development I: strategies and principles. Toxicol Sci 142(2):312–320

Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, Landesmann B et al (2014b) Adverse outcome pathway development II: best practices. Toxicol Sci 142(2):321–330

Weber FC, Esser PR, Muller T, Ganesan J, Pellegatti P, Simon MM, Zeiser R et al (2010) Lack of the purinergic receptor P2X(7) results in resistance to contact hypersensitivity. J Exp Med 207(12):2609–2619