

Bayesian integrated testing strategy to assess skin sensitization potency: from theory to practice

Joanna Jaworska^{a*}, Yuri Dancik^a, Petra Kern^a, Frank Gerberick^b and Andreas Natsch^c

ABSTRACT: Frameworks to predict *in vivo* effects by integration of *in vitro*, *in silico* and *in chemico* information using mechanistic insight are needed to meet the challenges of 21st century toxicology. Expert-based approaches that qualitatively integrate multifaceted data are practiced under the term 'weight of evidence', whereas quantitative approaches remain rare. To address this gap we previously developed a methodology to design an Integrated Testing Strategy (ITS) in the form of a Bayesian Network (BN). This study follows up on our proof of concept work and presents an updated ITS to assess skin sensitization potency expressed as local lymph node assay (LLNA) potency classes. Modifications to the ITS structure were introduced to include better mechanistic information. The parameters of the updated ITS were calculated from an extended data set of 124 chemicals. A detailed validation analysis and a case study were carried out to demonstrate the utility of the ITS for practical application. The improved BN ITS predicted correctly 95% and 86% of chemicals in a test set ($n=21$) for hazard and LLNA potency classes, respectively. The practical value of using the BN ITS is far more than a prediction framework when all data are available. The BN ITS can develop a hypothesis using subsets of data as small as one data point and can be queried on the value of adding additional tests before testing is commenced. The ITS represents key steps of the skin sensitization process and a mechanistically interpretable testing strategy can be developed. These features are illustrated in the manuscript via practical examples. Copyright © 2013 John Wiley & Sons, Ltd.

Additional Supporting Information may be found in the online version of this article.

Keywords: integrated testing strategy; skin sensitization; Bayesian Network; LLNA potency class

Introduction

Chemical safety assessment methods are moving away from traditional animal testing. Ethical considerations, regulatory pressure, as well as rapid advances in the understanding of molecular mechanisms of adverse effects drive this paradigm shift. Frameworks to predict *in vivo* effects by integration of *in vitro*, *in silico* and *in chemico* information using mechanistic insight are needed and actively being developed. The amount of available data to inform about *in vivo* effects is increasing rapidly and becoming more heterogeneous in metrics and scales (cell-based assays with dose–response curves, reactivity assays with reactivity parameters and *in silico* readouts with class attributions). 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. A poorly informed decision could result in unnecessary testing. 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, that can address different levels of biological organization, can be incorporated into the existing body of knowledge about the candidate chemical.

Expert-based approaches to combine information and decide about the need for additional testing have been practiced under the term 'weight of evidence'. However, many authors expressed concern that, given the evolution in knowledge and

data, the decision frameworks have to be more objective, consistent and transparent (Bus and Becker, 2009; Jaworska *et al.*, 2010). In recognition of this need by the broad community of toxicologists, an evidence-based toxicology collaboration has been established (Hartung, 2010). Its mission is to foster the development of systematic and structured approaches to assess the performance of new tests and synthesize information provided by them to meet the needs of the 'Toxicity Testing in the 21st Century' paradigm.

Integrated Testing Strategies (ITS) are one of the tools to meet 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 information gain is maximized in a testing sequence leading to a risk decision (Jaworska and Hoffmann, 2010; Jaworska *et al.*, 2010; JRC 2005). In narrative terms, ITS has been described as combinations of tests in a battery covering relevant mechanistic steps organized in a logical,

*Correspondence to: Joanna Jaworska, Procter & Gamble NV, 100 Temselaan, 1853 Strombeek - Bever, Belgium.
Email: Jaworska.j@pg.com

^aProcter & Gamble NV, 100 Temselaan, 1853 Strombeek - Bever, Belgium

^bProcter & Gamble Company, Cincinnati, OH, USA

^cGivaudan Schweiz AG, Ueberlandstrasse 138, CH-8600, Dübendorf, Switzerland

hypothesis-driven decision scheme, which is required to make efficient use of generated data and to gain a comprehensive information basis for making decisions regarding hazard or risk (Jaworska and Hoffmann, 2010). In our previous work, we discussed the need for a quantitative approach as the way to assure consistency, transparency and objectivity in dealing with this complex information. To this end, we developed conceptual requirements for ITS and a Bayesian Network (BN)-based methodology to build an ITS (Jaworska et al., 2010; Jaworska and Hoffmann 2010). 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 (Vol) calculations. The structure of the BN ITS represents the underlying mechanistic processes leading to an *in vivo* effect.

Skin sensitization is one of the key endpoints for which frameworks to make decisions using data from alternative test data are urgently needed. This is especially driven by the European cosmetic regulation which requires replacement of animal testing for skin sensitization by March 2013. Consequently, the field was the subject of very active research that resulted in many novel tests and great advances in mechanistic understanding of the processes leading to sensitization (Kimber et al., 2011; Mehling et al., 2012). Recognizing that the process of skin sensitization is too complex to be predicted by a single alternative test, several approaches to integration of non-animal tests have been published. They include conceptual frameworks (Basketter and Kimber, 2009; Jowsey et al., 2006), statistical regression-based methods (McKim et al., 2010; Natsch et al., 2009; Nukada et al., 2010), 'majority voting' from the outcome of three *in vitro* tests (Bauch et al., 2012) and mechanism-based approaches (Maxwell and Mackay, 2008). Existing alternative test methods are suitable to predict sensitization hazard, but are not yet appropriate for potency predictions (Adler et al., 2011; Basketter et al., 2012), and there is no agreement on what measurements, other than reactivity, are necessary to predict potency (Basketter and Kimber, 2009).

As explained above, expectations regarding decision frameworks reach beyond data integration and need to include methods to identify an optimal testing strategy. In the proof of concept work on BN ITS (further referred to as ITS-1) (Jaworska et al., 2011), we had shown how to combine these two needs in the context of skin sensitization potency assessment. On a small external test set ($n = 12$), ITS-1 achieved 94% accuracy for hazard. Potency prediction was promising – 50% of predictions were correct and mispredictions were at most one class off – but improvement was needed. Most importantly, however, we showed that mandating a single generic set of tests as a replacement strategy is unlikely to be effective, and that ITS must be flexible and adaptive. We also illustrated a situation in which generating additional data would not reduce uncertainty about the target information – i.e. local lymph node assay (LLNA) potency.

The goal of the present work was to improve on potency prediction and transform the proof of concept ITS-1 into a practical tool for skin sensitization potency assessment based on the assays under the European Centre for Validation of Alternative Methods (ECVAM) pre-validation. Specifically, the objective was to develop an ITS to characterize potency in the LLNA assay as a four-way classification scheme: non-sensitizers (NS), weak (W), moderate (M) and strong (S). To meet this objective, we developed the next generation BN ITS, referred to later as ITS-2. We refined the ITS structure to better reflect the skin

sensitization process by considering the ongoing evolution of the mechanistic understanding. Next, we replaced the AREC32 assay with the more standardized KeratinoSens™ assay, and introduced cytotoxicity and a direct Michael acceptor (MA) alert as *in vivo* potency modifying factors. In parallel, large efforts in data generation were undertaken. All data gaps in the ITS-1 database were filled in and data on additional chemicals were added, resulting in the ITS-2 database of 145 chemicals (Natsch et al., 2013). An extensive evaluation of the ITS-2 performance was carried out as this has been identified as the main hurdle before considering practical application of the ITS approach (De Wever et al., 2012). Finally, we demonstrated how the ITS-2 could be used in practice with a case study by developing hypotheses using partial evidence sets.

Materials and Methods

Data

The 145 chemicals were tested in the *in chemico* Direct Peptide Reactivity Assay (DPRA) assay and two cell-based assays: KeratinoSens™ and a CD86 activation assay based on the U937 cell line. An *in silico* hypothesis about sensitization potency was generated using the TIMES software (Dimitrov et al., 2005). Bioavailability-related variables were calculated from simulations using the skin penetration model developed by Kasting and coworkers (Dancik et al., 2013). An overview of all variables used in the ITS-2 is provided in Table 1. The dataset comprises a variety of chemical classes including fragrances, preservatives, dyes, dye precursors, halogenated alkanes and solvents, and covers a wide range of physicochemical properties. The details of the experimental part of the database are presented and discussed in the accompanying paper (Natsch et al., 2013). The whole database, including SMILES experimental data and *in silico* predictions, is available online (see Supplementary File A). There were 124 chemicals used to estimate the parameters of the BN ITS-2 (training set) and 21 chemicals to evaluate its performance (test set).

The murine local lymph node assay (LLNA)

The LLNA data (OECD 2002) were compiled from multiple sources, which included published literature and previously unpublished data from several laboratories. The chemicals were chosen based on the quality of the LLNA studies. The training set includes 36 non-sensitizers (NS), 28 weak (W), 35 moderate (M) and 25 strong or extreme (S) sensitizers. The test set contains 6 NS, 6 W, 4 M and 5 S chemicals.

Direct peptide reactivity assay (DPRA)

Peptide reactivity data were generated using a method measuring the reactivity of a test chemical with model hepta-peptides containing lysine or cysteine (Gerberick et al., 2004, 2007). 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™

Data were generated using the HaCaT keratinocyte-based cell line KeratinoSens™. The average concentrations (in μM) inducing a 1.5- or a 3-fold enhanced luciferase activity (KEC1.5 or KEC3.0, respectively) and the concentration leading to 50% cytotoxicity

Table 1. *In vitro*, *in chemico* and *in silico* tests data used for the development of the Integrated Testing Strategy (ITS)-2

Variable name	Description	Metric
LLNA	Skin sensitization potency in the LLNA	Potency class (NS, W, M, S)
CD86	150% induction of the cell surface activation marker CD86 in the U937 assay	$\mu\text{M} = 10^{-3} \mu\text{mol cm}^{-3}$
DPRACys, DPRALys	Reaction with the cysteine- (Cys), and lysine- (Lys) peptide in the DPRA-assay	% remaining peptide
KEC1.5, KEC3	1.5-fold (KEC1.5) or a 3-fold (KEC3) Induction of Nrf2-dependent luciferase activity in the KeratinoSens™ assay	μM
IC50	50% reduction in cell viability in the KeratinoSens™ assay	μM
TIMES	Skin sensitization potential prediction by the TIMES-M	Classes (NS, W, S)
AUC120	Area under the total flux curve in the epidermis after 120 h of exposure as in LLNA assay (approximated amount systemically absorbed after 120h)	$\mu\text{mol cm}^{-2}$
C_{free}	Maximum free (unbound) concentration in the mid-epidermis	$\mu\text{mol cm}^{-3}$
Kow	Octanol/water partition coefficient	Log (dimensionless)

after 24 h (IC50, in μM) 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 Emter *et al.* (2010) or generated at Givaudan laboratories.

Dendritic cells activation

The data were generated using a modified U937 cell activation test (Python *et al.*, 2007), as the 50% activation of cell surface marker CD86 is measured and expressed as EC150.

TIMES

The TIMES software (V.2.25.7) (Dimitrov *et al.*, 2005) is an *in silico* expert system that classifies chemicals into non-sensitizers, weak sensitizers and combines others into one class called Strong. As we differentiate between moderate and strong, class Strong is mapped onto the moderate (M) and strong/extreme (S) classes as distributed in the ITS-2 training set. In the ITS-1 study, predictions considering metabolism were more robust and we continue to use them in the ITS-2.

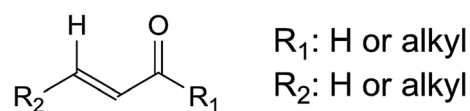
Epidermal bioavailability

The epidermal bioavailability variables come from simulations of *in vivo* penetration using the skin permeation model from G. Kasting (Dancik *et al.*, 2013; Ibrahim *et al.*, 2012). The exposure scenario mimics the low dose used in the LLNA assay: three $0.5 \mu\text{mol cm}^{-2}$ applications of permeant in ethanol at times $t=0$, 24, and 48 h with no further applications until the end of the simulation at 120 h. The following variables were chosen to characterize epidermal bioavailability: the maximum mid-epidermal free (unbound) concentration, C_{free} [$\mu\text{mol /cm}^3$], the area under the total flux curve in the epidermis after 120 h of penetration, AUC120 [$\mu\text{mol /cm}^2$] and the octanol-water partitioning coefficient, log Kow.

Michael acceptor alert

While evaluating the test set predictions, an alert for directly acting Michael acceptors (MA) was considered as additional evidence. The general structure is shown below, i.e. α,β -unsaturated ketones and aldehydes with an unbranched β -position (Fig. 1).

Natsch *et al.* (2011) noted that chemicals with this substructure are less sensitizing *in vivo* than would be inferred from chemical reactivity data, owing to the anti-inflammatory action

**Figure 1.** Structural alert for directly acting Michael acceptors.

of MA. Further, Natsch *et al.* (2011) showed that the anti-inflammatory activity increases with chemical reactivity for this class of molecules. As this MA alert does not translate to potency *a priori*, we have not included it in the structure of the network. However, we decided to use information about this alert in the predictions by manually modifying the hypothesis toward a weaker class in the following manner: if a chemical is predicted S (strong sensitizer), 40% probability mass was moved from S to M (moderate); if predicted M, 30% probability mass was moved from M to W class; if predicted W, 15% probability mass was moved from W to NS; if predicted NS, there was no change in the probability distribution. This modification corresponds to the direct MA effect presented in Natsch *et al.* (2011) graphic for Table of Contents and was applied only to the test set in Table 3.

Bayesian Network Construction

A Bayesian network (BN) is a graphical high-level model of a problem domain. Intuitively speaking, a BN model is constructed by explicitly determining all the direct dependencies between the variables of the domain. 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. Learning the network amounts to calculating a joint probability distribution over a set of all variables in the network. As a result, every node of the network has its own joint probability distribution conditioned on other network variables.

In BN parlance, the variable for which we develop a hypothesis (in this study, LLNA potency) is the target variable, whereas the variables providing evidence (in this study, all listed tests (Table 1) are referred to as manifest variables. In addition to manifest variables, we introduced the latent variables, bioavailability and Cysteine, 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, as well as simplify numerical computations for the joint probabilities.

First, the mechanistic scheme of the skin sensitization induction process as described in Basketter and Kimber, (2009) was translated into a corresponding network structure. Second, the tests used to observe the process were mapped onto the network. There are tests that clearly measure different parts of the process leading to an effect *in vivo*: bioavailability, protein binding and dendritic cells activation. There are also tests that measure the same part of the process but in different ways. A good example of this is protein binding, in particular cysteine binding. This creates complex dependencies between the nodes related to cysteine reactivity. The Cysteine latent variable represents the event of cysteine haptenation that can be observed via the DPRA-Cys assay and/or the KeratinoSens™ assay [a bias toward cysteine-reactive chemicals in Nrf2-dependent assays had been discussed previously (Natsch, 2010)]. Reactivity towards cysteine was also measured indirectly in TIMES as electrophilicity molecular descriptors. Further, it has been postulated that the molecular basis of dendritic cell 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 dendritic cell-based assays might be a complicated measure of cysteine reactivity (Kimber *et al.*, 2011). To reflect this, arcs connecting cysteine latent with CD86, as well as cysteine-latent and TIMES were introduced. Once the network structure was set, the parameters and joint distributions were learned from data (the details are described in the Supplementary file B available online). The network was constructed using the BayesiaLab software (Bayesia).

Value of Information (Vol) expressed as mutual information (MI) (X, Y) between variable X and Y was used to characterize the relationship between variables. Mutual information 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 (Kjærulff and Madsen, 2012). The one-step look-ahead hypothesis calculates the Vol 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 set of existing and new data.

Success Criteria

Prior to evaluation, a set of success criteria was developed. First, an ITS framework should be transparent, consistent, objective and mechanistically interpretable. These are conceptual requirements and have been discussed previously (Jaworska and Hoffmann, 2010). Second, the ITS-2 should be able to predict better than individual tests on an external test set. Third, the ITS-2 should predict equally well on the training and on external tests set. Fourth, if no prediction can be made with available data, the BN ITS should be able

to determine whether additional testing will reduce uncertainty about the prediction.

Results

Structure of the network

The ITS-2 structure is shown in Fig. 2. The refinements of the structure compared with ITS-1 represent improvements in mechanistic understanding and we describe the rationale for these modifications, as follows.

Bioavailability

To generate variables characterizing bioavailability in the ITS-1, we used high-single, high-dose exposure and 24-h simulations. In the ITS-2, we changed the simulations to mimic repeated exposure conditions by a lower test dose in the LLNA assay. In addition to the parameters C_{free} and AUC120, the skin penetration model allows for the calculation of many kinetic parameters. We have investigated their predictive value through the process of constructing several networks and evaluating MIs (results not shown). The C_{free} , AUC120 (equivalent to DoseAbs in ITS-1) and C_{total} (total concentration in the epidermis, equivalent to EpiConc in ITS-1) were identified as the most informative parameters. The K_{ow} was kept as the general physicochemical descriptor of bioavailability and skin penetration and C_{total} was removed, because C_{total} and C_{free} are functionally dependent via K_{ow} and including both would be redundant. Variables Kp and MW, used in ITS-1 were removed because of their low MIs.

Protein binding

In the ITS-1, all reactivity-related variables: DPRACys, DPRALys and ARE were grouped together under the latent variable 'reactivity'. In the ITS-2, we differentiate reactivity towards cysteine (as a cysteine-latent variable) and lysine (as the lysine-manifest variable), as evidence is accumulating that cysteine- and lysine-reactive chemicals can trigger distinctive biological pathways (Lalko *et al.*, 2012; Natsch, 2010). As the KeratinoSens™ assay is a cell-based assay, cytotoxicity was added as an additional variable that modifies KeratinoSens™ readouts. Owing to direction of the arcs, IC50 values modify only KEC3 and KEC1.5. A substantial correlation of cytotoxicity to potency of skin sensitizers has been described by several previous authors (dos Santos *et al.*, 2011; Lambrechts *et al.*, 2010). Similar to these studies, we found a strong dependence between LLNA and IC50 (MI = 16%). A network structure with IC50 to LLNA arc, instead of arcs to

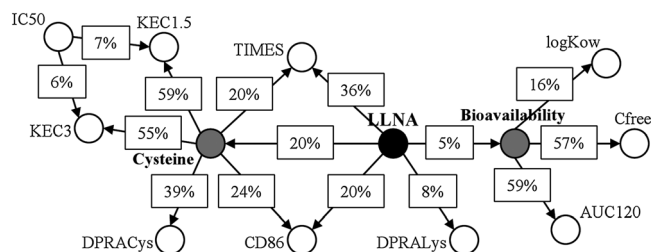


Figure 2. The structure of the Integrated Testing Strategy (ITS)-2. Each arc is annotated by mutual information (MI) [%] between the child node and the parent node. Open nodes denote manifest variables, gray nodes denote latent variables, and the black node is the target variable.

KEC variables, was explored. However, its performance was worse than that of the ITS-2, suggesting an overfit, and therefore stopped.

The second node related to protein haptentation is lysine. Currently, it can be observed only by the DPRA-Lys assay.

Dependence between Cysteine–Latent and CD86 and TIMES

The TIMES-M to cysteine-latent arc is very strong [MI (Cys, TIMES) = 20%], demonstrating large overlap. To note, an arc lysine-TIMES was also explored but eventually not used owing to a low MI. The cysteine-CD86 arc is even stronger [MI(Cys, CD86) = 24%], indicating strong dependence. The association seems to be strong for highly and non-reactive chemicals and complex for those with intermediate potency (Fig. 3). In order not to over predict potency class when Cys and CD86 (or TIMES) results agree and under-predict when they are in (partial) conflict, it is important to have these arcs. Further elucidation of these relationships is beyond the scope of this study.

ITS-2 Predictive Performance

Full evidence

The ITS-2 calculates probabilities for each potency class for the training and the test set chemicals. The confusion matrices based on assigning a chemical to a class with the highest probability are shown in Table 2. During the process of training, the ITS-2 predicted 13 moderate sensitizers as strong. This indicates the ITS-2's tendency to over-predict moderate sensitizers. Among the 21 test set chemicals, eight are also in the TIMES training set and three are direct MA (see Supporting Information). After applying MA correction,

the ITS-2 misclassified three chemicals as one class lower. We analyze the reasons for the misclassifications in the Discussion.

An Area Under the Curve of Receiver-Operating Curve (AUC ROC) removes bias owing to an uneven distribution of the number of chemicals in different classes and, therefore, are a useful way of evaluating performance of a classification model (Pepe, 2003). ROC AUC-based performance of the ITS-2 on the training and the test sets are very similar, which demonstrates a robust model and meets our 3rd success criterion (Table 3).

Analyzing the results shown in Tables 2 and 3 provides useful summary information but excludes much information generated by the ITS-2. The strength of the BN ITS is the fact that it predicts the probability distribution over all classes, which is a far richer information representation (Table 4). It allows discerning situations in which the hypothesis has a high uncertainty (i.e. flat distribution, for example 2-mercapto-benzothiazole) or low uncertainty (i.e. chlorobenzene).

The overall accuracy of the predictions for the test set without MA alert correction is 81% for a four-way potency classification (17/21) and 95% (20/21) for hazard classification. Among the incorrectly predicted chemicals, there was 2, 4-heptadienal, which contains a direct MA alert. The MA correction shifted peak of its probability distribution from the Strong to the Moderate potency class observed experimentally and improved potency overall prediction to 86% (18/21). A prohaptent, cinnamyl alcohol and a prehaptent 1, 4-phenylenediamine were both under predicted by the BN ITS-2: cinnamyl alcohol was predicted NS but observed *in vivo* as W, 1,4-phenylenediamine was predicted M but observed *in vivo* as S. Prohaptens are expected to be under-predicted in chemico and, to a lesser extent, in cell-based assays compared with an *in vivo* response, because the

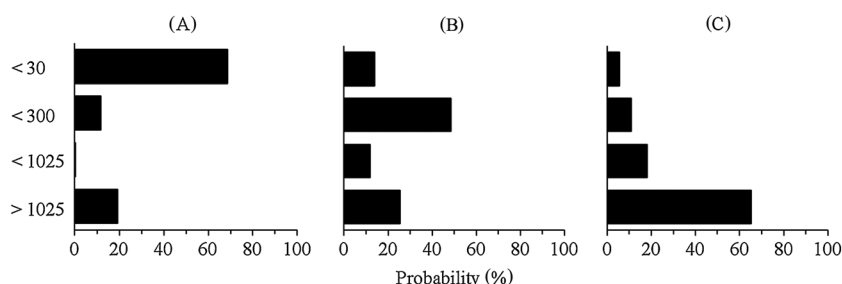


Figure 3. CD86 Probability distribution given cysteine-latent variable has (A) high, (B) medium and (C) low value corresponding to high, medium and low reactivity.

Table 2. The Integrated Testing Strategy (ITS-2) confusion matrices for the training and test sets

		Training set				Test set			
		Observed				Observed			
		NS (36)	W (28)	M (35)	S (25)	NS (6)	W (5)	M (5)	S (5)
Predicted	NS (37)	31	2	1	3	NS (7)	6	1	0
	W (27)	2	22	2	1	W (5)	0	4	0
	M (27)	2	3	19	3	M (4)	0	0	3 (4) ^a
	S (33)	1	1	13	18	S (5)	0	0	1 (0) ^a

^awith MA correction

Table 3. Area under the curve (AUC) receiver-operating characteristic (ROC) values (%) for the Integrated Testing Strategy (ITS)-2 training and test set

LLNA class	Training set	Test set
NS	93 ^a	100
W	89	93
M	75	83
S	81	73

^aROC values for a given local lymph node assay (LLNA) class are computed from that class's ROC curve (True Positive Rate (TPR) on the y-axis vs. False Positive Rate (FPR) on the x-axis). Points on the ROC curve correspond to (TPR, FPR) pairs obtained by considering whether the target (the LLNA class) is met for various discrimination thresholds (probabilities). The percentages are the ratio of the area under the ROC curve for the actual prediction to the total area, which represents a perfectly accurate prediction.

assays miss metabolic activation *in vivo* completely or may have reduced metabolic competency (Aeby *et al.*, 2009). Prehaptens are also often under-predicted in the *in chemico* or cell-based assays because spontaneous oxidation needed to activate prehaptens may be slow, thus reducing the *in vitro* or *in chemico* response. Finally, 2-mercapto-benzothiazole was predicted as W in the ITS-2, whereas it is M in the LLNA assay. Although W class had the highest probability, it is a very weak inference and we observed a fairly flat distribution over the W, M, and S classes, indicating data conflict for this chemical.

When comparing the ITS-2 results with *in vivo* data, it is important to consider experimental variability. Experimental variability in EC3 owing to biological variability and/or use of different vehicles may lead to different class assignments (Basketter and

Cadby, 2004). The EC3 LLNA values for many chemicals span across two potency classes. The coefficients of variation for citral, eugenol and isoeugenol are 42%, 38% and 48%, respectively (ICCVAM 2008), which can be converted to the general rule of an EC3 variability range of $0.5 \times \text{EC3}$ to $2 \times \text{EC3}$. Therefore, some ITS-2 mispredictions by one class might be attributed to variability of the LLNA result.

Partial evidence

In practice, the assessor often must work with limited data and different data gaps for different chemicals. Bayesian Network ITS can be used to develop a hypothesis based on partial evidence – even on one data point. To demonstrate that the ITS-2 can be used with missing data or simply using subsets of full data, we documented its performance on several incomplete records (Table 5). In all but one instance, the performance on partial evidence is lower compared with the full evidence. 2,4-heptadienal was always over predicted and MA correction always improved the number of correct predictions. Separation between NS and sensitizers yielded good results for all the subsets considered in the range 81% to 100%. In addition to full data, there were four subsets of data that predicted at least 95% correct. Remarkably, prediction of hazard without TIMES is perfect. Correct predictions of potency were, as expected, lower and were in the 59% to 86% range. Interestingly, the combinations of evidence without cysteine but with CD86 both had the worst potency prediction results. For two- and four-way classifications, there are different combinations of tests having the same discriminating capability. This illustrates that there is redundancy or information overlap in the input tests in terms of predicting hazard and potency in the LLNA. One can choose certain subsets of evidence without compromising strongly on the quality of the predictions. A very efficient battery of tests appears to be cysteine- and lysine-

Table 4. The Integrated Testing Strategy (ITS)-2 probability predictions for each potency class of the test set chemicals. Shaded cells denote the experimental local lymph node assay (LLNA) potency class. Bolded values mark the assigned LLNA potency class

Chemical	MA	P (LLNA=)				P (LLNA=) with MA correction			
		N	W	M	S	N	W	M	S
Chlorobenzene	0	93	4	0	3	93	4	0	3
Isopropanol	0	98	1	0	1	98	1	0	2
Lactic acid	0	98	1	0	1	98	1	0	2
Methyl salicylate	0	97	1	0	2	97	1	0	2
Salicylic acid	0	84	3	0	13	84	3	0	13
2-Hydroxypropyl methacrylate	1	95	3	2	0	95	3	2	0
Hexyl cinnamic aldehyde	1	0	99	1	0	15	84	0	0
Imidazolidinyl urea	0	0	95	5	0	0	95	5	0
Eugenol	0	4	69	27	0	4	69	27	0
1-Iodohexane	0	11	87	2	0	11	87	2	0
Cinnamyl Alcohol	0	78	8	13	0	78	8	13	0
Citral	0	0	3	93	4	0	3	93	4
Dihydroeugenol	0	4	9	62	25	4	9	62	25
Isoeugenol	0	0	16	67	17	0	16	67	17
2-Mercapto-benzothiazole	0	0	46	37	17	0	46	37	17
2,4-Heptadienal	1	0	1	39	60	0	31	49	20
Fluorescein-5-isothiocyanate	0	0	2	23	75	0	2	23	75
1-Chloro-2,4-dinitrobenzene	0	0	0	45	55	0	0	45	55
1,4-Phenylenediamine	0	0	0	99	1	0	0	99	1
Tetrachloro-salicylanilide	0	0	0	16	84	0	0	16	84
5-Chloro-2-methyl-4-isothiazolin-3-one	0	0	0	20	80	0	0	20	80

Table 5. The Integrated Testing Strategy (ITS)-2 performance on the test set with full and partial evidence. X denotes that the test was not and $\checkmark$ indicates that the test was used in the particular evidence set. The subsets are ordered according to performance to predict potency

TIMES	B	CD86 ^a	Cys	Lys	Correct Potency (%)		Correct Hazard (%)	
					w/o MA	With MA	w/o MA	With MA
$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	81	86	95	95
X	$\checkmark$	X	$\checkmark$	$\checkmark$	81	81	90	95
$\checkmark$	X	X	X	X	71	76	95	95
X	X	X	$\checkmark$	$\checkmark$	71	76	95	95
X	$\checkmark$	$\checkmark$	$\checkmark$	X	67	72	86	86
X	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	71	71	100	100
X	X	$\checkmark$	$\checkmark$	$\checkmark$	70	71	88	88
X	X	X	$\checkmark$	X	62	67	81	81
X	$\checkmark$	$\checkmark$	X	$\checkmark$	59	59	88	88
X	X	$\checkmark$	X	X	59	59	82	82

^aAs only full records were used for the analysis, the statistics for records with CD86 is based on $n = 16$. See also supporting information.

reactivity combined with bioavailability for both hazard and potency prediction, albeit more data are needed to confirm it. Given that biological data – both *in vivo* and alternative – always have inherent variability, further improvements of hazard prediction are not feasible. In order to increase accuracy in predicting chemical potency in LLNA, information other than reactivity or dendritic cells activation is likely needed.

Mutual information analysis

Mutual information between individual tests and LLNA was analyzed per potency class and globally. Global MI analysis gives information about average contributions of individual tests (Cys: 20 %, CD86: 20 %, DPRALys: 8 %, bioavailability: 5 %). However, there are substantial differences between classes that cannot be discovered in the global analysis (Fig. 4) that lead to the following observations. Among the tests considered, the NS class can be predicted with the highest accuracy and the M class with the lowest accuracy. Bioavailability MI peaks for the W class, suggesting that there is a unique epidermal disposition of weak sensitizers. Lysine is important in identifying NS and, to some

extent, S. Finally, Cys and CD86 have very similar MI values for W, M, and S potency classes, implying that they provide similar information about these classes. For NS, Cys carries more information than CD86. These observations are relevant for the development of an adaptive testing strategy.

Potency class signatures

Mutual information and modal values (outcomes of the tests most likely associated with particular classes) were calculated per potency class (Table 6). Mutual information analysis provides information about which different tests predict potency. Modal values can be interpreted as fingerprints for specific classes and used without running the network.

It is helpful to analyze Table 6 per LLNA potency class and per latent class. The same tests carry different information for different potency classes. Within the latent class, the tests are usually synergistic, although the degree of synergism can vary with potency class and variables. TIMES has highest MI across all potency classes. However, its MI values are inflated artificially, because 89 of 124 chemicals in the ITS-2 training set overlap with TIMES training set (see Supporting Information). Thus, our results do not indicate that TIMES is more informative than in chemico or biological assays. We explored the predictability of the ITS-2 without TIMES in the partial evidence section. Table 6 helps to identify equivalent tests by comparing MIs within potency class. For the non-sensitizers, we noted that the latent variable cysteine, which combines information from DPRA-Cys and KeratinoSens™, carries more information than any of the individual tests. This means that these tests are not exactly equivalent in spite of their similar MI values. Results associated with the cysteine latent variable may reflect the fact that the KeratinoSens™ assay has a larger dynamic range as compared to DPRA and, therefore, is better positioned to separate different potency classes. The KeratinoSens™ assay appears to have some intrinsic metabolic capacity, although it is not able to detect all prohaptenes. Especially α,β -unsaturated alcohols, such as cinnamic alcohol, geraniol, or some sensitizing amines such as ethylenediamine are correctly predicted and, therefore, the applicability domain of this assay may be a bit broader as compared to DPRA-Cys. On the other hand, although the

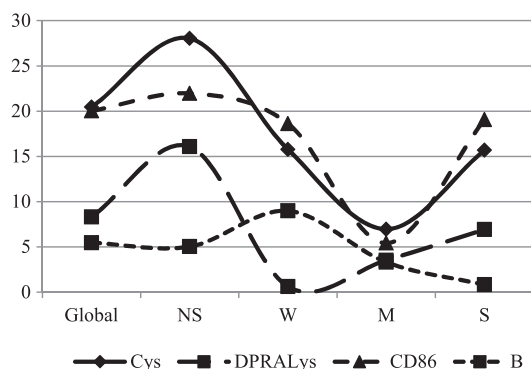


Figure 4. Mutual information (MI) between Cys, DPRALys, CD86, B nodes and local lymph node assay (LLNA) potency classes. For perspective Times MI values on the training set were: MI (LLNA = NS|Times) = 46, MI (LLNA = W|Times) = 42, MI (LLNA = M or S|Times) = 42. TIMES MI values are inflated because 89 chemicals out of 124 are common to the TIMES and the ITS-2 training sets.

Table 6. The mutual information (MI) and modal (most likely)^a values for each local lymph node assay (LLNA) potency class. Latent variables are marked in *italics*

NS			W			M			S		
Node	MI	Modal Value	Node	MI	Modal Value	Node	MI	Modal Value	Node	MI	Modal Value
TIMES	45	1	TIMES	42	2	TIMES	17	3	TIMES	25	3
Cysteine	28	Low	CD86	19	<295	Cysteine	7	Medium	CD86	19	<27
CD86	22	>1025	Cysteine	16	Medium	CD86	5.4	<27	Cysteine	16	High
KEC3	20	>945	Bioav	9.0	Low	KEC3	4.6	<476	KEC3	7.7	>945
DPRALys	16	>95	KEC1.5	8.8	<238	DPRACys	4.1	<15	KEC1.5	7.0	<10
KEC1.5	16	>1098	KEC3	7.3	>945	KEC1.5	4.0	<238	DPRALys	6.9	<70
DPRACys	15	>90	Cfree	7	<0.04	DPRALys	3.5	<95	DPRACys	6.7	<15
Bioav	5	High	AUC120	5.1	<2.26	Bioav	3.3	Low	Bioav	0.82	Low
Cfree	3.8	<0.04	DPRACys	2.1	>90	AUC120	1.7	>23	AUC120	0.44	>23
AUC120	3.6	>23	logKow	1.5	<1.9	Cfree	0.86	<0.038	Cfree	0.20	<0.04
logKow	0.43	<1.9	DPRALys	0.61	>95	logKow	0.81	<1.92	logKow	0.14	<1.9

^aThe modal value is the value x , at which probability mass function takes its maximum value.

molecular pathway triggered in KeratinoSens™ is linked to Cys-reactivity with the Keap 1 sensor protein, this assay is not a 1:1 indication of direct Cys-reactivity, which is directly measured in the DPRA-Cys. This may explain why the latent Cys-variable has a higher MI than individual reactivity DPRA-Cys and Keratinosens. Among the bioavailability parameters, C_{free} and AUC explain almost entirely the latent variable bioavailability. The relationship between epidermal disposition and sensitization potency is complex (Table 7) and more research is needed to explain it. C_{free} approximates interstitial concentration and it may be very different from the concentration at the target site where sensitization is triggered *in vivo*.

Applicability domain of the ITS-2

The ITS-2 domain includes chemicals that induce sensitization via binding to cysteine or lysine as molecular initiating events. This domain may be further defined and it may exclude certain classes of pre- and prohapten that are predicted incorrectly by the tests used in ITS-2. Separate work is ongoing to define these classes (Troutman *et al.*, 2011). The applicability domain is further discussed in (Natsch *et al.*, 2013) (Table 4 and 5).

Case Study for 1,6-Hexanediylbis[oxy(2-hydroxypropane-1,3-diyl)] Diacrylate

Finally, we present a case study in which we compare a weight of evidence expert evaluation with prediction by the ITS-2 for 1,6-Hexanediylbis[oxy(2-hydroxypropane-1,3-diyl)] diacrylate (Table 8 and 9). High reactivity of this chemical in the DPRA assay can be explained by the structural features: it is an unsaturated ester di-ester and an acrylate. As the reactivity data are in clear

conflict with TIMES, the assessor would likely conclude that the chemical possesses a sensitization hazard but would probably refrain from estimating a potency class. The assessor could only conclude that high reactivity should not be right away associated with strong sensitization potency.

In the ITS-2, TIMES, DPRA and bioavailability data are used to form the hypothesis (Table 9). The ITS-2 does not include DEREK alerts. This was a conscious choice in the ITS-2 design, as our focus is on potency prediction. In contrast to the expert evaluation, now the assessor can formulate a hypothesis that a chemical is a moderate sensitizer with low but characterized certainty (39%). If the assessor would seek to refine this hypothesis, a CD86 test has the highest MI among remaining tests. Before actually conducting the test, we can examine the impact of different CD86 results on the final hypothesis. A CD86 result combined with an MA alert would identify whether a chemical is NS, W, or M and reduce uncertainty about potency by 20% (probability for the chemical to be NS, W, or M would increase to approximately 60%).

Historically, expert evaluation has been oriented toward hazard decision, but this approach, as illustrated above, encounters difficulties with potency prediction. A structural alert is directly associated with hazard but no conclusion regarding 'certainty' and potency can be made. As shown, ITS-2 extended the utility of existing data to predict potency. This was possible because of a formal quantitative methodology, based on probability laws that allowed the assessor to form the hypothesis based on different sets of evidence, even conflicting evidence. Further, in case of limited data availability, the assessor could have difficulty identifying which missing data would be useful to generate. Often, additional data are generated only to realize later that they will not impact decision. The ITS-2 prevents the occurrence of such situations.

Discussion

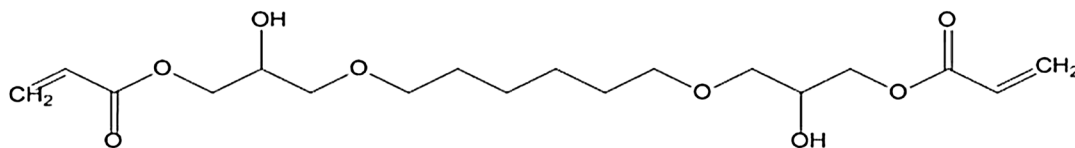
This study refined a previously published ITS framework to combine alternative test methods currently undergoing prevalidation with *in silico*-generated information to assess skin sensitization potency. The new ITS-2 was evaluated carefully based on a diverse validation set. The results indicate

Table 7. Relationship between local lymph node assay (LLNA) potency class and bioavailability variables

LLNA Potency Class	C_{free}	AUC120
NS	High	High
W	Low	Low
M	Medium	whole range
S	Medium	Low or high

Table 8. Available data for CAS# 83045-03-8.

1,6-Hexanediylbis[oxy(2-hydroxypropane-1,3-diyl)] diacrylate (CAS# 83045-03-8)

DEREK
TIMES
DPRAUnsaturated ester alert
Non-sensitizer
DPRACys % remaining = 0; DPRALys % = 20

that the ITS-2 is suitable to predict potency expressed as potency classes in the LLNA. As potency can be converted easily to hazard prediction, the same ITS framework is a suitable tool to predict sensitization hazard. Using an external validation set of 21 chemicals, the ITS-2 predictions were 86% correct for potency and 95% correct for hazard. All of our success criteria established, before the study results were known, were met. The ITS-2 predicts better than (i) individual tests and (ii) equally well on the training and test sets. The ITS-2 can determine impact of additional testing.

We ascribe the success with predicting potency to the combination of: (i) ITS network structure that follows mechanistic steps of skin sensitization induction process; (ii) the generation of a large dataset used to parameterize the ITS; and (iii) a Bayesian probabilistic framework for inference and testing strategy development.

ITS-2 Network Structure

A critical consideration when integrating data into an ITS is the framework structure. It is possible to construct poorly predictive

models with good data if the structure of the model is not suitable to the problem (Thomas *et al.*, 2012). The ITS-2 structure represents, in an abstracted form, mechanistic steps during skin sensitization induction phase for which experimental tests were developed. Because of progress in the understanding of skin sensitization mechanisms, we were able to refine the ITS-2 structure compared with proof of concept in the ITS-1 Bioavailability measures are tailored to the exposure in LLNA assay. The ITS-2 considers binding to cysteine and lysine as two possible yet separate molecular initiating events. KeratinoSens™ readouts are modified by cytotoxicity, so only reactivity-related information is used in the LLNA potency predictions. Finally, LLNA potency class distribution is adjusted for chemicals with direct MA alert for which anti-inflammatory effects have been established.

In spite of originating from different motivations and independently, the BN ITS structure concept is in line with that of adverse outcome pathway (AOP). Briefly, an AOP is the sequence of events from chemical structure through the molecular initiating event to the *in vivo* outcome of interest (Ankley *et al.*, 2010). Adverse effects observed *in vivo* can be the result of many biological responses to the toxicant. Hence, AOPs are designed to avoid mixing information from multiple mechanisms (i.e. different molecular initiating events that can cause the same *in vivo* outcome through different AOPs). The BN ITS can be run by providing evidence in a sequence that mimics a single AOP but, in addition, it can also be run as a combination of parallel and sequential steps that are needed to consider multiple AOPs.

Until now, AOP development focused on documenting biology in a structured form but did not offer a quantitative framework. The Organization of Economic Cooperation and Development (OECD) identified this as a limitation that should be eliminated in the future (OECD, 2012). Determination of thresholds and response-to-response relationships used to scale *in chemico* and *in vitro* effects to *in vivo* outcomes as well as explicit accounting in variability of experimental data are deemed as methodological requirements for development of the quantitative AOP. By combining mechanistic knowledge with empirical approaches for parameter estimation, the BN ITS is capable of addressing, in part, some of the identified challenges. It provides probabilistic scaling of *in vitro* or *in chemico* to *in vivo* simultaneously considering dependencies between different tests.

In spite of the fact that the ITS-2 structure and current AOP for skin sensitization from the OECD (2012) evolved in parallel, they are fully aligned and therefore it will be easy to reapply learnings between them.

Table 9. Hypothesis development for CAS no. 83045-03-8 using the Integrated Testing Strategy (ITS)-2

Evidence	NS	W	M	S
DPRA (Cys/Lys)	2	11	27	59
Cys/Lys/B	2	15	27	56
Times - M	80	7	9	4
All w/o MA alert	29	22	39	10

Evolution of the hypothesis based on all above evidence AND different potential outcomes of the CD86 test:

CD86 < 30 ^a μM	6	3	66	26
With MA alert	15	18	62	0
CD86 < 300 μM	5	44	50	1
With MA alert	20	59	21	0
CD86 < 1025 μM	50	17	33	0
With MA alert	65	35	3	0
CD86 > 1025 μM	46	15	24	15
With MA alert	61	30	9	0

^aThe cut-offs are discretization cutoffs (see methods section) that were determined during network construction process; they are not fitted to this example.

Parameter Learning Based on a Large Database

A key hurdle in progressing from qualitative to quantitative inferences is the lack of high-quality databases. Therefore, large resources were allocated in this study to generate a comprehensive dataset in order to overcome limitations of ITS-1, which were due to data scarcity and data gaps. The 145 chemicals were tested in the DPRA, KeratinoSens™ and CD86 assays. This expanded the dataset by 40% compared with our proof of concept work, ITS-1. These data were complemented by *in silico* information from TIMES and the bioavailability model. The data generation created a large, diverse and high-quality dataset to establish the foundation for the ITS-2 (Natsch et al., 2013).

Probabilistic Framework

The third reason we managed to address potency is the probabilistic framework. Until now, with the exception of ITS-1, modeling techniques used in other data integration studies to predict skin sensitization were all deterministic. They included classification trees (Gerberick et al., 2004, 2007), regression models (Ashikaga et al., 2010; McKim et al., 2010; Natsch et al., 2009; Nukada et al., 2010) or rule-based approaches such as 'if two out of three positive, then positive' (Bauch et al., 2012). A deterministic technique treats every piece of information at face value and generates a single result given a set of input parameters.

Of course, the fact that deterministic models ignore uncertainty in knowledge and variability in data is not sufficient to quickly justify the use of probabilistic models in place of deterministic models. The most important aspect of the model is how it affects decision-making. Deterministic frameworks were sufficient when dealing with single tests for hazard prediction. However, in the ITS, the need for probabilistic modeling arises from the goal of placing knowledge and complex, multifaceted, and mutually dependent data in a systematic, evidence-based, reasoning framework. Therefore, probabilistic ITS frameworks are more realistic than deterministic.

Flexible and adaptive testing strategy

The practical value of using the network is far more than a prediction framework when all data are available. The network can be queried with partial evidence and is able to find testing strategy that is mechanistically interpretable and efficient. The ITS-2, as with Bayesian models in general, has the advantage over deterministic frameworks of being adaptive with respect to new information. The adaptive nature of Bayesian frameworks has important implications for decision-making. It treats the decision-making process as a dynamic process and, therefore, optimizes the sequence of decisions. Decisions made at one time may have strong effects on alternatives available at other times. Bayesian frameworks account for dynamically changing interrelationships among tests based on evidence provided. In practice, these features of BN ITS allow its use in situations that do not have all data. For example: (i) the assessor has only a chemical structure and decides to start with *in silico* information, and only then, depending on the TIMES and bioavailability-based outcome, seeks relevant *in vitro* tests; (ii) the assessor has a chemical structure and information from one or two assays, generates a hypothesis using these data, and evaluates whether further testing is needed.

In situation (i), ITS-2 adapts to generated *in silico* data via TIMES and bioavailability hypotheses regarding LLNA potency.

Depending on the hypothesis, different *in vitro* tests are recommended as confirmatory tests based on their MI values. Current data suggest that use of reactivity and CD86 simultaneously to test NS and S hypothesis is not effective, as just one piece of evidence suffices. On the other hand, combining cysteine and CD86 is an effective way to identify W and M – class sensitizers because, in general, more evidence is needed to identify these classes. As Table 4 illustrates, even with all data, the uncertainty for W and M classes remains higher than for NS and S.

In situation (ii), in order to prevent the *in vitro* hypothesis to be overshadowed by 'heavy' (or inflated as in the ITS-2) information from TIMES, the assessor can change the order of informing the network as the AOP would suggest: start with bioavailability, then use available protein reactivity data and, if needed, DC activation data. The bioavailability hypothesis and *in vitro*-based hypothesis can be combined into one, and only at the latest stage a cumulative hypothesis including TIMES is formulated. By considering dependencies between tests, the BN ITS proposes alternative tests if the most optimal test is not available. The BN ITS also identifies which tests are not effective to be used simultaneously. The fact that an *in vitro* hypothesis can be generated using any other *in vitro* test illustrates the flexibility of the BN ITS. This is a big asset in practical application, in which chemical safety data may be incomplete and the assessor must make decisions on different sets of evidence very frequently.

The BN ITS makes it possible to evaluate the impact of generating additional data on the target before testing is commenced. Therefore, it is a tool to avoid unnecessary testing. The BN-ITS can be updated regularly with new methods and data and, therefore, can become a structured depository of knowledge.

Going Forward

As the mechanistic understanding of skin sensitization progresses further, ITS will evolve accordingly. While the ITS-2 meets our current objectives for practical applications, several areas for improvement have been identified. Short-term improvements for which solutions are already emerging include: (i) discrimination between M and S sensitizers, (ii) improved detection of prehapten and prohapten over current tests, and (iii) elimination of the inflated weights for the *in silico* systems owing to overlap of the training sets. Inspection of the misclassification trends in the training set (confusion matrix in Table 3) informs us about limitations of the ITS-2 test battery. The current battery has a limited capability to discriminate M and S sensitizers, with a tendency to over predict M sensitizers. One reason is the fact that TIMES SS makes only one M-S class, thus necessitating the use of the *in vitro* tests to make the differentiation on which we fully rely. Interestingly, 14 over-predicted M sensitizers include seven directly acting MA. Anti-inflammatory effects of MA chemicals counterbalancing strong reactivity is the current explanation for their over prediction by different assays. Including the MA alert as a dampening factor into an ITS structure seems a reasonable way forward to improve correct identification of M sensitizers. A further improvement would be to increase the dynamic range, especially for reactivity assays. Incorporation of dose–response measurements as done in the PPRA (Gerberick et al., 2009) and the kinetic profiling approach (Natsch et al., 2011; Roberts and Natsch, 2009)

addresses this issue and should be considered in a further ITS evolution.

The current test battery correctly identifies the hazard for many pro- and prehapten, but there is still a trend to underestimate the potency of these chemicals. Thus, incorporating metabolic systems in either the reactivity assays or the cell-based assays will further improve potency estimations. The PPRA assay addresses this point but a complete dataset was not yet available; this test would be appropriate to include in the next generation of ITS.

Finally, expanding the database to include even more diverse chemicals will allow to identification of limitations in the current test battery and for correction of the inflated importance of *in silico*-based inputs such as TIMES in case of overlapping training sets.

Longer-term improvements are envisioned to advancing mechanistic knowledge about skin sensitization and developing tests to observe these mechanisms. Future ITS framework may consider other mechanisms that currently are not well characterized by a combination of reactivity and DC assays. A very important mechanistic step in the skin sensitization process, T-cell activation, is not addressed in ITS-2 owing to the lack of alternative experimental methods. It could be added once these methods become available to further improve accuracy of prediction and to broaden ITS applicability domain.

In addition, next to the MA alert, other dampening effects on the sensitization potential of certain chemicals, such as antioxidant effects or induction of regulatory T-cells, may be important to consider. An increased understanding of (i) such modifying factors, (ii) how they are triggered by different chemical classes, and (iii) how they can be implemented optimally in a BN ITS will help to add chemistry-based rules that will drive ITS to become even more mechanistically interpretable.

Until now our focus was on replacing the LLNA test, which while well correlated to human response, is certainly not equivalent to it (ICCVAM, 2011) and several specific false positives as well as false negatives were identified (Basketter and Kimber 2010). The ultimate goal is to develop ITS methods directly relevant for human risk assessment and tackling different exposure regimes. Such paradigms are emerging and are usually based on extrapolating adverse effects in *in vitro* tests to *in vivo* doses and specific exposures. However many practical challenges remain related to the effects and kinetic components of the extrapolation paradigm (Yoon *et al.*, 2012). Advances in understanding mechanistic steps leading to skin sensitization as well as advances in skin penetration modeling should allow application of extrapolation approach to assess skin sensitization potency in humans in the near future.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Acknowledgements

We thank G. Kasting and M. Miller for their assistance in running the skin penetration simulations. We also thank L. Foertsch, C. Ryan and R. Emter for generating experimental data used in this work.

References

- Adler S, Basketter D, Creton S, Pelkonen O, van Benthem J, Zuang V, Andersen KE, Angers-Loustau A, Aptula A, Bal-Price A. 2011. Alternative (non-animal) methods for cosmetics testing: current status and future prospects—2010. *Arch. Toxicol.* **85**: 367–485.
- Aeby P, Sieber T, Beck H, Gerberick GF, Goebel C. 2009. Skin sensitization to p-phenylenediamine: the diverging roles of oxidation and N-acetylation for dendritic cell activation and the immune response. *J. Invest. Dermatol.* **129**: 99–109.
- Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK. 2010. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.* **29**: 730–741.
- Ashikaga T, Sakaguchi H, Sono S, Kosaka N, Ishikawa M, Nukada Y, Miyazawa M, Ito Y, Nishiyama N, Itagaki H. 2010. A comparative evaluation of *in vitro* skin sensitisation tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). *Altern. Lab. Anim.* **38**: 275–284.
- Basketter DA, Cadby P. 2004. Reproducible prediction of contact allergenic potency using the local lymph node assay. *Contact Dermatitis* **50**: 15–17.
- Basketter DA, Clewell H, Kimber I, Rossi A, Blaauboer B, Burrier R, Daneshian M, Eskes C, Goldberg A, Hasiwa N. 2012. MAIN ARTICLE-t4 report-A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing. *ALTEX* **29**: 5.
- Basketter DA, Kimber I. 2009. Updating the skin sensitization *in vitro* data assessment paradigm in 2009. *J. Appl. Toxicol.* **29**: 545–550.
- Basketter DA, Kimber I. 2010. Skin sensitization, false positives and false negatives: experience with guinea pig assays. *J. Appl. Toxicol.* **30**: 381–386.
- Bauch C, Kolle SN, Ramirez T, Eltze T, Fabian E, Mehling A, Teubner W, Ravenzwaay B, Landsiedel R. 2012. Putting the parts together: Combining *in vitro* methods to test for skin sensitizing potentials. *Regul. Toxicol. Pharmacol.* **63**: 489–504.
- Bus JS, Becker RA. 2009. Toxicity testing in the 21st century: a view from the chemical industry. *Toxicol. Sci.* **112**: 297–302.
- 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**: 221–236.
- De Wever B, Fuchs HW, Gaca M, Krul C, Mikulowski S, Poth A, Roggen EL, Vila MR. 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, Niemela J, Bailey PT, Mekenyan OG. 2005. Skin sensitization: modeling based on skin metabolism simulation and formation of protein conjugates. *Int. J. Toxicol.* **24**: 189–204.
- dos Santos GG, Spiekstra SW, Sampat-Sardjoepers SC, Reinders J, Scheper R, Gibbs S. 2011. A potential *in vitro* epidermal equivalent assay to determine sensitizer potency. *Toxicol. In Vitro* **25**: 347–357.
- Emter R, Ellis G, Natsch A. 2010. Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers *in vitro*. *Toxicol. Appl. Pharmacol.* **245**: 281–290.
- Gerberick GF, Troutman JA, Foertsch LM, Vassallo JD, Quijano M, Dobson RLM, Goebel C, Lepoittevin JP. 2009. Investigation of peptide reactivity of pro-hapten skin sensitizers using a peroxidase-peroxide oxidation system. *Toxicol. Sci.* **112**: 164–174.
- Gerberick GF, Vassallo JD, Bailey RE, Chaney JG, Morrall SW, Lepoittevin JP. 2004. Development of a peptide reactivity assay for screening contact allergens. *Toxicol. Sci.* **81**: 332–343.
- Gerberick GF, Vassallo JD, Foertsch LM, Price BB, Chaney JG, Lepoittevin JP. 2007. Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. *Toxicol. Sci.* **97**: 417–427.
- Glover F, Laguna M. 1997. *Tabu Search*. Kluwer Academic Publishers: Norwell, MA.
- Hartung T. 2010. Evidence-based toxicology—The toolbox of validation for the 21st century? *ALTEX* **27**: 253.
- Heckerman D. 1995. A Bayesian approach to learning causal networks. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 285–295.
- Ibrahim R, Nitsche JM, Kasting GB. 2012. Dermal clearance model for epidermal bioavailability calculations. *J. Pharm. Sci.* **101**: 2094–2108.
- ICCVAM. 2008. URL <http://iccvam.niehs.nih.gov/methods/immunotox/LLNA-pot/appx/LLNApotencyAppB18Jan08FD.xls> (accessed 15 April 2013).
- 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. NIH Publication No. 11-7709. National Institute of Environmental Health Sciences: Research Triangle Park, NC.
- Jain AK, Murty MN, Flynn PJ. 1999. Data clustering: a review. *ACM Computing Surveys*. **31**: 264-323.
- 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**: 157-167.
- Jaworska J, Harol A, Kern PS, Gerberick GF. 2011. Integrating non-animal test information into an adaptive testing strategy - skin sensitization proof of concept case. *ALTEX* **28**: 211-225.
- Jaworska J, Hoffmann S. 2010. Integrated Testing Strategy (ITS) - Opportunities to better use existing data and guide future testing in toxicology. *ALTEX* **27**: 231-242.
- Jowsey IR, Basketter DA, Westmoreland C, Kimber I. 2006. A future approach to measuring relative skin sensitising potency: a proposal. *J. Appl. Toxicol.* **26**: 341-350.
- JRC. 2005. REACH and the Need for Intelligent Testing Strategies. Institute for Health and Consumer Protection, European Commission. Ispra, Italy.
- Kimber I, Basketter DA, Gerberick GF, Ryan CA, Dearman RJ. 2011. Chemical allergy: translating biology into hazard characterization. *Toxicol. Sci.* **120**(Suppl 1): S238-268.
- Kjærulff UB, Madsen AL. 2012. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Springer: New York, NY.
- Kumuthini J, Jouffe L, Bessant C. 2007. A Novel Graph Optimisation Algorithm for the Extraction of Gene Regulatory Networks from Temporal Microarray Data In: *Book*. Edited by Verlag S; 2007. pp. 28-37.
- Lalko JF, Kimber I, Gerberick GF, Foertsch LM, Api AM, Dearman RJ. 2012. The direct peptide reactivity assay: selectivity of chemical respiratory allergens. *Toxicol. Sci.* **129**: 421-431.
- Lambrechts N, Vanheel H, Nelissen I, Witters H, Van Den Heuvel R, Van Tendeloo V, Schoeters G, Hooyberghs J. 2010. Assessment of chemical skin-sensitizing potency by an in vitro assay based on human dendritic cells. *Toxicol. Sci.* **116**: 122-129.
- Maxwell G, Mackay C. 2008. Application of a systems biology approach to skin allergy risk assessment. *Altern. Lab. Anim.* **36**: 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**: 171-192.
- Mehling A, Eriksson T, Eltze T, Kolle S, Ramirez T, Teubner W, van Ravenzwaay B, Landsiedel R. 2012. Non-animal test methods for predicting skin sensitization potentials. *Arch. Toxicol.* **86**: 1273-1295.
- 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**: 284-292.
- 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**: 106-121.
- Natsch A, Haupt T, Laue H. 2011. Relating skin sensitizing potency to chemical reactivity: reactive Michael acceptors inhibit NF-kappaB signaling and are less sensitizing than S(N)Ar- and S(N)2- reactive chemicals. *Chem. Res. Toxicol.* **24**: 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.* DOI: 10.1002/jat.2868
- Nukada Y, Foertsch L, Ashikaga T, Sakaguchi H, Nishiyama N, Gerberick F. 2010. Predicting allergy potential by battery evaluation system using two in vitro skin sensitization tests. *Contact Dermatitis* **63**: 79-80.
- OECD. 2002. *Test Guideline 429: The Local Lymph Node Assay*. Organization for Economic Cooperation and Development: Paris.
- OECD. 2012. The Adverse Outcome pathway for Skin Sensitization Initiated by Covalent Binding to Proteins. URL [http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono\(2012\)10/part1&doclanguage=en](http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2012)10/part1&doclanguage=en) (accessed 15 April 2013).
- Pearl J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann: Burlington, MA.
- Pepe MS. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: New York, NY.
- Python F, Goebel C, Aeby P. 2007. Assessment of the U937 cell line for the detection of contact allergens. *Toxicol. Appl. Pharmacol.* **220**: 113-124.
- Roberts DW, Natsch A. 2009. High throughput kinetic profiling approach for covalent binding to peptides: Application to skin sensitization potency of Michael acceptor electrophiles. *Chem. Res. Toxicol.* **22**: 592-603.
- Thomas RS, Black MB, Li L, Healy E, Chu TM, Bao W, Andersen ME, Wolfinger RD. 2012. A Comprehensive Statistical Analysis of Predicting In Vivo Hazard Using High-Throughput In Vitro Screening. *Toxicol. Sci.* **128**: 398-417.
- Troutman JA, Foertsch LM, Kern PS, Dai HJ, Quijano M, Dobson RL, Lalko JF, Lepoittevin JP, Gerberick GF. 2011. The incorporation of lysine into the peroxidase peptide reactivity assay for skin sensitization assessments. *Toxicol. Sci.* **122**: 422-436.
- Yoon M, Campbell JL, Andersen ME, Clewell HJ. 2012. Quantitative in vitro to in vivo extrapolation of cell-based toxicity assay results. *Crit. Rev. Toxicol.* **42**: 633-652.

SUPPLEMENTARY INFORMATION B

Bayesian Network Construction

The structure of the ITS-2 is knowledge driven, while the probabilistic relationships between the variables were learned from data. The network development consisted of the following steps: 1) transforming the training set into discrete variables, 2) latent variables structure learning, and 3) final model learning. The network was constructed using the BayesiaLab software (www.bayesia.com).

A BN is a graphical model that efficiently encodes the joint probability distribution for a set of variables (Heckerman 1995; Pearl 1988). The BN is composed of nodes representing domain variables and arcs connecting these nodes. An arc represents dependence between the nodes connected by the arc, while the absence of an arc demonstrates independence between the nodes. The parameters of a BN are conditional probabilities that characterize arcs. Each node has to be provided with a table of conditional probabilities, where the variable in question is conditioned by its immediate predecessors in the network. They are estimated using the relative frequencies of all combinations of variable states as in the training data and organized as conditional probability tables (CPTs).

Methodology to Guide Testing

Value of Information (VoI) measures and a one-step look-ahead hypothesis were used as the methodologies to guide testing. The one-step look-ahead hypothesis calculates the VoI from all possible individual information sources and chooses the one that will maximize the information gain about the target variable. 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 generation of new data. In this study, we use a relative mutual information (MI) index between variables X and Y to measure VoI. It measures amount the amount of uncertainty in Y that is removed by knowing X. This corroborates the intuitive meaning of MI as the amount of information (that is, reduction in uncertainty) that knowing X variable provides about the Y. The relativity refers to the percent of entropy of the parent node Y reduced by knowledge of X.

Data Preparation

The upper limits for the KEC1.5 and KEC3 were set to the next highest values increased by 10%, that is, to 2026 μM and 1310 μM , respectively. The upper limit for CD86 was set to the next highest value increased by 10%, i.e., to 7404 μM . A decision tree algorithm was used to discretize DPRACys, DPRALys, log Kow, and AUC120, KEC1.5, KEC3, and CD86 variables. For the variables Ctot and C_{free} the k-means algorithm (with four intervals) was used (Jain *et al.*, 1999). Each variable's density curve, representing the density of probability versus the variable's values, was inspected. Additional cutoffs were added manually at the density curves' inflection points. The final cutoff values for each continuous variable are given in the Supporting Information (Table A1). The final number of intervals is four for the variables KEC1.5, KEC3, CD86, DPRACys, log Kow, AUC120, Ctot, and C_{free}, and three for the variables IC50 and DPRALys.

Construction of Latent Variables

Two latent variables, cysteine and bioavailability, were created in order to cluster the manifest variables related to cysteine reactivity (DPRACys, KEC1.5, and KEC3) and to

bioavailability (AUC120, C_{free} , log Kow). BayesiaLab's proprietary Multiple Clustering algorithm calculated three for the latent variable cysteine and four for bioavailability as the optimal number of classes, and then clusters for the manifest variables in Naïve Bayes subnetworks (within each cluster, the latent variable is the parent of each manifest variable).

Network Structure Learning

Since the network topology represents biological knowledge, the following arcs were fixed prior to final network structure learning: the arcs linking LLNA to bioavailability, cysteine, CD86, DPRALys, and TIMES, the arcs linking IC50 to KEC1.5 and KEC3 that reflect the cytotoxicity correction of reactivity measured in the KeratinoSens™ assay, and the arcs linking cysteine to CD86 and TIMES. Forbidden arcs were defined to prevent any direct arc between LLNA and the reactivity manifest variables KEC1.5, KEC3, and DPRACys and between IC50 and variables other than KEC1.5 and KEC3. To make the prediction more robust, the probability density estimation was smoothed. The number of occurrences within each cell of the CPTs with zero was reset to “virtual occurrences” of one, and then the joint probability density was estimated. The taboo learning algorithm (Glover and Laguna 1997; Kumuthini *et al.*, 2007) was chosen for its suitability to learn a network with a structure defined by expert knowledge.