"21st Century Validation Strategies for 21st Century Tools"

On July 13-14, 2010 The Johns Hopkins Center for Alternatives to Animal Testing held a workshop at the Johns Hopkins Bloomberg School of Public Health in Baltimore. The two-day workshop, titled "21st Century Validation Strategies for 21st Century Tools," consisted of four original papers, each followed by four invited responses and a discussion period. The papers, published in this issue of ALTEX, addressed topics of interest to regulators, industry and academic scientists charged with implementing the recommendations of the new approach to toxicological testing outlined in the National Academy of Sciences 2007 report, *Toxicity Testing in the 21st Century: A Vision and A Strategy*.

Integrated Testing Strategy (ITS) – Opportunities to Better Use Existing Data and Guide Future Testing in Toxicology

Joanna Jaworska [1] *and Sebastian Hoffmann* [2]
[1] Procter & Gamble, Modelling & Simulation, Biological Systems, Brussels Innovation Center, Belgium; [2] seh consulting + services, Cologne, Germany

Summary
The topic of Integrated Testing Strategies (ITS) has attracted considerable attention, and not only because it is supposed to be a central element of REACH, the ambitious European chemical regulation effort. Although what ITSs are supposed to do seems unambiguous, i.e. speeding up hazard and risk assessment while reducing testing costs, not much has been said, except basic conceptual proposals, about the methodologies that would allow execution of these concepts. Although a pressing concern, the topic of ITS has drawn mostly general reviews, broad concepts, and the expression of a clear need for more research on ITS. Published research in the field remains scarce. Solutions for ITS design emerge slowly, most likely due to the methodological challenges of the task, and perhaps also to it its complexity and the need for multidisciplinary collaboration. Along with the challenge, ITS offer a unique opportunity to contribute to the Toxicology of the 21st century by providing frameworks and tools to actually implement 21st century toxicology data in the chemical management and decision making processes. Further, ITS have the potential to significantly contribute to a modernization of the science of risk assessment. Therefore, to advance ITS research we propose a methodical approach to their design and will discuss currently available approaches as well as challenges to overcome. To this end, we define a framework for ITS that will inform toxicological decisions in a systematic, transparent, and consistent way. We review conceptual requirements for ITS developed earlier and present a roadmap to an operational framework that should be probabilistic, hypothesis-driven, and adaptive. Furthermore, we define properties an ITS should have in order to meet the identified requirements and differentiate them from evidence synthesis. Making use of an ITS for skin sensitization, we demonstrate how the proposed ITS concepts can be implemented.

Keywords: Integrated Testing Strategy, 21st century toxicology tools, probabilistic, hypothesis-driven, adaptive framework

1 Introduction

Currently we are witnessing a tremendously increased pace of discovery in biology, especially molecular biology, that has increased our knowledge of biological systems' structure and function. New opportunities for chemical management methods are created as more mechanistic insights become available, improving our ability to address human relevance and to reduce use of full animal models. Policy changes, especially in Europe, such as the REACH and Cosmetics directives, encourage in-

creasing reliance on non-animal approaches and pose challenges to the existing frameworks for chemical safety evaluation. As a consequence, a paradigm for data evaluation was proposed: Integrated Testing Strategies (Anon, 2003, 2005, 2007).

Integrated Testing Strategies (ITS) are expected to perform better than tiered testing strategies by maximizing use of existing data and gaining a more comprehensive, mechanistic basis for decision making using *in silico*, *in vitro*, -omics and ultimately *in vivo* data, as well as exposure information (Bradbury et al., 2004; van Leeuwen et al., 2007; Ahlers et al., 2008; Schaafsma et al., 2009). Additionally, especially in the context of REACH, ITS are not only supposed to provide accurate inferences but also to be resource-effective and reduce animal testing. In essence these new data, while increasingly available, are complex and multifaceted. To fully leverage them, appropriate methodologies for their analysis and interpretation are required. Although a pressing concern, the topic of ITS has drawn mostly general reviews, broad concepts, and the expression of a clear need for more research on ITS (Hengstler et al., 2006; Worth et al., 2007; Benfenati et al., 2010). Published research in the field remains scarce (Gubbels-van Hal et al., 2005; Hoffmann et al., 2008a; Jaworska et al., 2010a). In this regard, the opinion has been expressed that no overarching scheme will be able to handle the diversity of sciences and approaches involved (van Leeuwen et al., 2007). From where we stand this is difficult to assess, but we object to such a general statement before an attempt is even made to find such as scheme.

As ITS pose unique challenges but also offer a unique opportunity, the most pressing need is to progress beyond the existing conceptual frameworks and develop transparent, structured, consistent, and causal methodological approaches (Hoffmann, 2009; Jaworska et al., 2010a), e.g. as postulated under the concept of an evidence-based toxicology (Hoffmann and Hartung, 2006; Hartung, 2009). Shortcomings of existing ITS were recently analyzed in detail by Jaworska et al. (2010a) and, in short, are as follows. The use of flow charts as the ITS' underlying structure may lead to inconsistent decisions. There is no guidance on how to conduct consistent and transparent inference about the information target, taking into account all relevant evidence and its interdependence. Moreover, there is no guidance, other than purely expert-driven, regarding the choice of the subsequent tests that would maximize information gain.

Overall, there are high expectations for ITSs and good agreement on what they are supposed to do. However, solutions for ITS design emerge slowly, most likely due to the methodological challenges of the task and perhaps also to its complexity. Along with the challenge, ITS offer a unique opportunity to contribute to the Toxicology of the 21st century by providing frameworks and tools to actually implement 21st century toxicology data in the chemical management and decision-making processes. Further, we consider ITS to have the potential to significantly contribute to a modernization of the science of risk assessment. Therefore, to advance ITS research we propose a methodical approach to their design and discuss currently available approaches as well challenges to overcome. First, a definition of ITS, which serves as a basis for this review, is presented and then contrasted with evidence synthesis. Next, we identify and discuss desired ITS elements required for a systematic – as opposed to a case-by-case – approach to ITS and present a road map leading from conceptual requirements to the operational framework. Elements of information and decision theory are introduced and their application to testing is demonstrated. We conclude with an example to illustrate how our proposals can be deployed in practice.

2 ITS – what they are and what they are not

A systematic approach to ITS starts with defining what they are. This is crucial for following our review. Furthermore, a more precise definition of ITS is also needed for the larger community in order to harmonize its use and improve scientific exchange. In narrative terms, ITS can be described as combinations of test batteries covering relevant mechanistic steps and organized in a logical, 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. We approach ITS from a system analysis perspective and understand them as decision support tools that synthesize information in a cumulative manner *and* that guide testing in such a way that information gain in a testing sequence is maximized.

This definition clearly separates ITS from tiered approaches in two ways. First, tiered approaches consider only the information generated in the last step for a decision as, for example, in current regulated sequential testing strategy for skin irritation (OECD, 2002) or the recently proposed *in vitro* testing strategy for eye irritation (Scott et al., 2010). Secondly, in tiered testing strategies the sequence of tests is prescribed, albeit loosely, based on average biological relevance and is left to expert judgment. In contrast, our definition enables an integrated and systematic approach to guide testing such that the sequence is not necessarily prescribed ahead of time but is tailored to the chemical-specific situation. Depending on the already available information on a specific chemical the sequence might be adapted and optimized for meeting specific information targets.

3 A systematic approach to ITS – defining the conceptual requirements

Building upon earlier papers delineating the conceptual requirements (Hoffmann and Hartung, 2006; OECD, 2008; Jaworska et al., 2010), ITS should be:

a) Transparent and consistent

- As a new and complex development, key to ITS, as to any methodology, is the property that they are comprehensible to the maximum extent possible. In addition to ensuring credibility and acceptance, this may ultimately attract the interest needed to gather the necessary momentum required for their development. The only way to achieve this is a fundamental transparency.
- Consistency is of similar importance. While difficult to achieve for weight of evidence approaches, a well-defined

and transparent ITS can and should, when fed with the same, potentially even conflicting and/or incomplete information, always (re-)produce the same results, irrespective of who, when, where, and how it is applied. In case of inconsistent results, reasons should be identified and used to further optimize the ITS consistency.
- In particular, transparency and consistency are of utmost importance in the handling of variability and uncertainty. While transparency could be achieved qualitatively, e.g. by appropriate documentation of how variability and uncertainty were considered, consistency in this regard may only be achievable when handled quantitatively.

b) Rational
- Rationality of ITS is essential to ensure that information is fully exploited and used in an optimized way. Furthermore, generation of new information, usually by testing, needs to be rational in the sense that it is focused on providing the most informative evidence in an efficient way.

c) Hypothesis-driven
- ITS should be driven by a hypothesis, which will usually be closely linked to the information target of the ITS, a concept detailed below. In this way the efficiency of an ITS can be ensured, as a hypothesis-driven approach offers the flexibility to adjust the hypothesis whenever new information is obtained or generated.

4 The operational framework for ITS

Definition of conceptual requirements for ITS is a prerequisite that guides the possible choices regarding methodological approaches. Since chemical risk assessment is inherently uncertain due to imperfect understanding of underlying toxicological mechanisms, variations among and between individual species used for testing, as well as those due to measurement and observational errors, a formal approach is needed to quantify this uncertainty in order to systematically reduce it. In other words the framework should allow for evidence maximization, i.e. it should guide testing in such a way that the information content can be updated stepwise and that the choice of subsequent tests is guided by the highest expected uncertainty reduction. Probabilistic methods provide a formal approach for quantifying uncertainty from heterogeneous input sources, relationships between them, and overall target uncertainty.

Further, probabilistic methods are based on fundamental principles of logic and rationality. In rational reasoning every piece of evidence is consistently valued, assessed and, coherently used in combination with other pieces of evidence. While knowledge- and rule-based systems, as manifested in current testing strategy schemes, typically model the expert's way of reasoning, probabilistic systems describe dependencies between pieces of evidence (towards an information target) within the domain of interest. This ensures the objectivity of the knowledge representation. Probabilistic methods allow for consistent reasoning when handling conflicting data, incomplete evidence, and heterogeneous pieces of evidence.

Moreover, the inference should be *hypothesis-driven*. The hypothesis-driven workflow proceeds from gathering existing data, through computational analysis, towards quantifying uncertainty in relation to the information target and, then, if required to new experimentation. A cyclic or iterative hypothesis-driven workflow cannot be incorporated into current testing strategy schemes where inference is fixed in one direction. Hypothesis-driven inference implies that causal relationships are preserved in the testing sequence. This goes beyond Hansson and Rudén (2007), who have outlined a decision-theoretic framework for the development of tiered testing systems, but who omit the idea of hypothesis-driven inference. In addition, it should be noted that hypothesis driven inference can be easily formalized in Bayesian probabilistic approach.

Currently, regulatory toxicological inference is linked to a descriptive Weight-of-Evidence (WoE) approach, which is a stepwise procedure for integration of data and for assessing the equivalence and adequacy of different types of information. This approach, associated with some fundamental problems (Weed, 2005), aims at optimal integration of information from different sources featuring various aspects of uncertainty (Ahlers et al., 2008). We acknowledge that it is a useful tool for today's chemical hazard and risk assessment. However, we doubt its use regarding ITS, primarily because it lacks a methodological basis for making transparent consistent inferences, as highlighted by Jaworska et al. (2010a) and Schreider et al. (2010). Schreider et al. (2010) mainly link credibility to transparency, but transparency is also a crucial prerequisite to improving the risk assessment process regarding the methodology used. Finally, ITS have to be assessable in order to allow overall optimization. Besides uncertainty reduction and, of course, predictive performance, relevant optimization parameters are costs, feasibility, and animal welfare, e.g. as incorporated by Hoffmann et al. (2008a).

The last conceptual requirement pertains to efficiency. Optimally efficient methods are adaptive. They optimize the solution to a specific situation and information target. In our case they need to be, e.g. chemical-specific and exposure driven. As such, this requires the departure from the check-box approach.

In summary, the outlined conceptual requirements translate into an operational framework that needs to be probabilistic, even better Bayesian and adaptive. Such a framework would be transparent and consistent by definition. Furthermore, it allows handling all kinds of uncertainty and it can be used in a rational way.

5 Elements of ITS

Having defined and described the framework of ITS, we propose to fill it with the following five elements:
1. Information target identification;
2. Systematic exploration of knowledge;
3. Choice of relevant inputs;
4. Methodology to evidence synthesis;
5. Methodology to guide testing.

The earlier presented definition separates ITS from evidence synthesis. Addressing the elements of ITS should make it even more clear that ITS go beyond evidence synthesis, especially as they present an overall, context-specific approach to guide testing. Evidence synthesis methodology will, however, be an essential element of ITS. In fact, the complexity and heterogeneity of data to be integrated will require more advanced methodologies. Evidence synthesis will be further elaborated in the section devoted to this topic, below.

5.1 Information target

A well-defined information target has to be formulated in an unambiguous and precise way as a fundamental prerequisite for any testing strategy. It is crucial to know what decision is to be informed. Only in this way can it be assured that an efficient and transparent ITS can be constructed and optimized, leading to consistent decisions. It is obvious that any change of the information target will at least require ITS-adaption. While ITS will more frequently be applied to hazard and classification and labelling (C&L), with implicit consideration of exposure, rather than to risk assessment where consideration of exposure is explicit and leads to a dose-response characterization, our framework is suitable to both.

Currently, for C&L an *in vivo* animal study result usually is used as an information target. Taking the OECD sequential testing strategy for skin corrosion/irritation (OECD, 2002) as an example, the information target is dermal corrosion/irritation potential of a chemical expressed as a category (corrosive, not corrosive, irritating, not irritating) guided by the Draize test. While acknowledging that the established *in vivo* tests have served their purpose, their predictive properties and relevance for human health are largely unknown. Therefore, treating an *in vivo* animal test result as the gold standard for human risk assessment is not appropriate. It should, rather, be considered a reference test, which much better reflects its real properties and use. Confusion of the two concepts – gold standard and reference standard – created considerable problems, especially when assessing the predictive capacity of *in vitro* tests. It needs to be understood that, as the reference tests can be imperfect, a new test may have better predictive characteristics than the reference (Alonzo and Pepe, 1999; Pepe, 2003; Hoffmann et al., 2008b). In such cases, considering the information target as a reference test will allow evolution of information targets that are more human-relevant and more predictive. Therefore, in our opinion, the gold standard, although frequently used, e.g. as by Burlinson et al. (2007) or by Andersen and Krewski (2009), is a concept hardly applicable to toxicology, as it conveys wrong associations and thus should be avoided.

The problems associated with the lack of gold standards are common to other fields as well, especially the related life sciences of human and veterinary medicine. In medical diagnostic test assessment, approaches that have been developed to account for imperfect reference standards have recently been reviewed extensively (Rutjes et al., 2007; Reitsma et al., 2009). Solutions include correction for imperfectness, which may use sensitivity analysis and various approaches aiming for construction of a reference standard, including latent class analysis, composite reference standard construction, and panel diagnosis (Alonzo and Pepe, 1998 and 1999; Knottnerus and Muris, 2003; Zhou et al., 2002; Baughman et al., 2008). Interestingly, solutions in case of absence of a reference standard also are available, which might be useful for newly emerging toxicological health effects and concerns, e.g. endocrine disruption or the effects of nanomaterials. As comparable problems are encountered, similar solutions have been developed in parallel and applied to the diagnosis of animal diseases (Enøe et al., 2000; McInturff et al., 2004; Georgiadis et al., 2005). Also, in genomics the need for an appropriate reference standard for the evaluation of functional genomic data and methods has been recognized, and an approach has been proposed based on expert curation as an equivalent to panel diagnosis (Myers et al., 2006).

Acknowledging that the solutions presented above mainly address binary test outcomes, a potential benefit of those approaches when adapted to toxicology has been postulated (Hoffmann and Hartung, 2005 and 2006). The potential use of latent class analysis for the assessment of toxicological *in vitro* tests has already been highlighted (Hothorn, 2002; Hoffmann et al., 2008b). In this regard it has to be noted that latent class approaches carry the challenge that, while mathematically elegant, they require the adoption of the reference from non-observable quantities, which may be difficult to communicate.

As mentioned above, new test methods usually are assessed by comparison to a reference standard. Even today those comparisons remain one-to-one, i.e. comparing one new test with one existing test, largely disregarding the fact that more than one reference method might exist. Therefore, composite reference results obtainable by integration of existing data in a consistent and transparent manner are likely to be better suited for assessment purposes (Pepe, 2003; Hoffmann et al., 2008b). Composite reference results for reference chemicals could be generated, with or without the knowledge of expert panelists, by integrating data from agreed reference tests and, potentially, other information sources. As successful implementation of an ITS requires upfront agreement on the assessment procedure, including success criteria, definition of further reference points in addition to the information target may be required. Such an approach would be facilitated by composite reference standards, as they offer the flexibility to define reference results for each reference point.

5.2 Systematic exploration of knowledge via integrated knowledge management as a basis for ITS

With continuously increasing biological and toxicological understanding, it is evident that ITS must be frequently adapted to the current state of knowledge. But the wealth of information (and the scientific discussion about it) is overwhelming, making complete and consistent capture impossible. This explosion of biological data and the growing number of disparate data sources are exposing researchers to a new challenge – how to acquire, maintain, and share knowledge from large and distributed databases in the context of rapidly evolving research. This chal-

lenge has to be faced, as systematic exploration of knowledge is becoming critical to advance the understanding of life processes (Antezana et al., 2009; Landner, 2010).

Data can be described as incoherent bits of information without context. Only upon adding context and interpretation does data become information. Information forms the basis for the development of knowledge. Knowledge means the confident understanding of a subject with the ability to use it for a specific purpose and to recognize when that is appropriate. Regarding toxicology, with continuous advances of technologies and methodological approaches, information is no longer equivalent to knowledge (in fact the deluge of data and information may create confusion instead of knowledge). Knowledge mapping is one approach that will allow the creation of shared knowledge and the leveraging of all existing information, which can readily be updated. Knowledge mapping produces a transparent record and structure of information and knowledge available within a field. As it is mainly employed for managing complex organizations, it also should be well suited to managing the complex, diverse, and quickly evolving information in toxicology. Knowledge mapping is considered to be of particular importance for ITS where different technologies, methodological, and evidence synthesis approaches meet to aid chemical management decision making.

Knowledge maps are created by transferring tacit and explicit knowledge into graphical formats that are easily understood and interpreted by the end users. A knowledge map contains information about relevant objects and their associations and relationships. Depending on the information available, objects of a toxicological knowledge map will include but not be limited to genetic information, metabolic processing, influences of protein induction and activity, pathways, molecular targets, cellular targets, organ and whole body concepts, and reactomes (Vastrik et al., 2007; Matthews et al., 2009). The associations and relationships need to capture all known processes by which chemicals perturb the functional equilibrium of a human body in a mechanistic manner. As knowledge is continuously evolving, knowledge mapping must be a continuous effort in order to be useful. Knowledge maps can be represented by different degrees of formalization. Semantic networks, Concept maps and Bayesian networks offer interesting opportunities for knowledge mapping. Approaches that can transform knowledge maps into machine-understandable representations are particularly useful, allowing in this way the exploration of novel connections and therefore accelerating learning about the domain (Kim et al., 2002). These knowledge representation methods also provide an appropriate representation to facilitate human understanding.

The key concept for knowledge mapping is ontology. Ontologies provide a hierarchically organized vocabulary for representing and communicating knowledge about a topic in the form of terms, i.e. words or compound words in specific contexts and the relationships between them. Whether simple or complex, ontologies capture domain knowledge in a way that can be processed with a computer. The use of ontologies facilitates standard annotations, improves computational queries, and supports the construction of inference statements from the information at hand. Furthermore, ontologies are pivotal for structuring data in a way that helps users understand the relationships that exist between terms in a (specialized) area of interest, as well as to help them understand the nomenclature in areas with which they are unfamiliar. The main advantages of ontologies are:

- sharing common understanding of the information structure;
- enabling reuse of knowledge;
- making domain assumptions explicit;
- separating domain knowledge from operational knowledge;
- vanalyzing domain knowledge;
- increasing interoperability among various domains of knowledge;
- enhancing the scalability of new knowledge into the existing domain.

In biology, ontologies currently are applied in communicating knowledge as well as in database schema definition, query formulation, and annotation, i.e. investigation of current "known" facts or data. However, ontologies also can be employed to facilitate conceptual discovery, often leading to a paradigm shift. When the use of conceptual annotation grows we can expect to see a concomitant change in database retrieval strategies. It will become much more precise and complete than is currently possible. It also should allow the exploration of the relationships describing, e.g. functions, processes, and components of retrieved entries, resulting in significantly increased insight garnered from the search results (Gottgtroy et al., 2006).

The biomedical community is engaged in several activities to advance new methodologies for leveraging the semantic content of ontologies to improve knowledge discovery in complex and dynamic domains (Gadaleta et al., 2010; Weeber et al., 2005; Spasic et al., 2005). They envisage building a multidimensional ontology that will allow the sharing of knowledge from different experiments undertaken across aligned research communities in order to connect areas of science seemingly unrelated to the area of immediate interest (Caldas et al., 2009). However, more research is still needed to harmonize existing efforts so that a unique, interoperable, universal framework can arise. This will lead to future uses of computers for heterogeneous data integration, querying, reasoning, and inference, which in turn will support knowledge discovery. In toxicology, efforts so far have concentrated largely on gene ontologies (EBI, 2010; GO, 2010). For the advancement of ITS, similar efforts are urgently needed.

5.3 Strategies to identify relevant inputs to ITS

To advance ITS, knowledge mapping has to be converted to reality. Certainly we will not measure everything that is known but, ideally, only what is needed to make a decision. However, ITS should be consistent with the respective knowledge map. In order to control quality and relevance, potential pieces of ITS need to be characterized to a certain extent. For this initial step, guidance in identifying promising and/or appropriate tests is required. Initially, the driving aspect should be individual test properties, including the aspects of biological/mechanistic relevance (if known or assessable), endpoints measured,

reproducibility/reliability, applicability domains, and expected contribution to the final aim, i.e. making a decision. Most of these properties are intrinsic and, in a first step, can be addressed independently of others. However, it has to be noted that, as detailed information may not always be readily available, this preparatory work is a screening intended to support the framing and structuring of information to facilitate ITS construction. Of course, the choice ultimately will be driven by many factors, such as testing costs, animal welfare considerations, or simply test complexity/availability.

Methodological considerations

Noise in biological data can be due to various, often unavoidable causes, such as inherent variability of the biological object under study, technological limitations, measurement errors, or human mistakes. ITS require integration of these noisy data and identification of important variables among the measured variables in order, finally, to develop, improve, or adapt a strategy. If relevant, a statistically significant variable most likely will be an important variable. However, biologically important variables providing a weak signal may be crucial as well. Both need to be found and properly placed in an ITS.

Practically, simplification of the hypothesis and variable selection is usually attempted by evaluating a training set aimed at identification of the most informative variables. However, this approach has been criticized recently due to problems with overfitting. Furthermore, power is low if variables are correlated. It diminishes even further when taking the multiplicity of the testing problem into account. When dealing with biological data, in addition to noise, often only a small set of variables carries most of the information of interest. This makes classic algorithms unsuitable and leads to a high number of false positive and, to a lesser degree, false negative findings (Wu, 2009). Several new algorithms to address these problems were developed. Blanchard and Roquain (2009), e.g. proposed a false discovery rate controlling procedure for correlated variables and showed it to be much more powerful than classic procedures that control the traditional family-wise error rate. Zehetmayer and Posch (2010) developed algorithms to assess power and false negatives rate in large scale multiple testing problems. To gain further power and insight, Meinshausen (2008) developed a hierarchical testing procedure approach to address importance, not at the level of individual variables but rather at the level of clusters of highly correlated variables, and suggested that hierarchy can be derived from specific domain knowledge. Furthermore, for high dimensional noisy data probabilistic approaches to variable selection are currently being developed (Jiang and Tanner, 2008). All these new algorithms share one common feature – they are adaptive and/or hierarchical. They first evaluate data at a coarse level and then refine hypotheses via multiple iterations. In view of these recent advances, it is not entirely surprising that retrospective analyses of large biological data sets are revealing large number of false positives due to so called "fishing for significance" and need to be interpreted with caution (Boulesteix, 2010; Boulesteix and Hothorn, 2010; Boulesteix and Strobl, 2009).

While these algorithms are quite complex, finding weak signals of high importance, which may arise due to complex feedback mechanisms in biological signalling, is even trickier. For their detection, algorithms and methods enabling realistic dynamic models of information processing in cells, in particular dynamic network approaches, will be needed (Han, 2008). Rao et al. (2002) suggested applying signal processing techniques to study weak but important signals. Hong and Man (2010) give an excellent example of an important but weak signal processing in a signalling pathway.

Another way of reducing the effects of noise is to use prior knowledge about the target of interest. For example, in the learning from numerical data, Šuc et al. (2004) and Lin and Lee (2010) showed the benefits of making the learning algorithm respect the known qualitative properties of the target. As the prior knowledge, i.e. the qualitative aspects, can be extracted from knowledge mapping, it is very appealing to take advantage of knowledge mapping as it immediately provides intuitive understanding.

Data handling

It is evident that in pursuing the vision of toxicology in the 21st century and ITS as one of the relevant tools, a lot of toxicological data will have to be handled sufficiently. Data need to be stored in databases allowing easy access and efficient extraction/combination for data integration (Kavlock et al., 2008; Hardy et al., 2010).

Exploring new approaches to regulatory toxicology, such as ITS, will involve a collaborative effort. Many groups will want to and need to contribute. This requires either central data management and storage or local data storage using compatible database structures. Both entail a huge challenge, however. It will be necessary to make researchers aware of the greater scope of the effort. This concerns not only questions related to one or a few substances and/or mechanisms, but also how the data generated can help to define test methods or the substance-specific toxicology and, ultimately, how the information might be used strategically. Only if this is properly understood will researchers be able to dedicate themselves to adequate data management. Otherwise, data will be incomplete, poorly reported, biased, and ultimately of uncertain use for testing strategies.

In addition to the management, the data collection also presents challenges. Prerequisite to any data synthesis, especially to a quantitative one, is to collect data in a systematic way. The importance of a systemic approach – and the consequences if not applied – was demonstrated by Rudén using risk assessments of trichloroethylene as an example (Rudén, 2001). Reasons for incomplete collection might be manifold. Among them is the fact that in toxicology reviews are traditionally narrative, opening the door for subjective, biased, not transparent, and incomplete data collection. It has been proposed that systematic reviews, as defined under evidence-based medicine (Cook et al., 1997), would be a more appropriate approach (Hoffmann and Hartung, 2006; Hartung, 2009). Potential problems in systematic reviews to be aware of include publication bias, i.e. the general tendency to publish studies demonstrat-

ing the presence of an effect, and selection bias (Horvath and Pewsner, 2004). The latter, especially, also poses a problem for toxicological reviews because selection criteria are difficult to establish when data are not assessed in a transparent and consistent way. Taking the assessment of the reliability of existing data quality/reliability in the REACH-context as an example, currently available methodology (Klimisch et al., 1997) is prone to selection bias, as it is not transparent and is open to subjectivity. Especially in this setting, however, subject reliability assessment can be expected, since such an assessment might be decisive for the use/rejection of existing data and thus is directly related to costs. As data quality is also of considerable importance for ITS, an objective assessment is needed, which allows for proper incorporation of this aspect into ITS construction and assessment. This has recently been recognized, and first attempts toward an objective assessment have been put forward (Schneider et al., 2009; Hulzebos and Gerner, 2010; Jaworska et al., 2010a). Interestingly, similar developments are taking place in ecotoxicity (Hobbs et al., 2005; Breton et al., 2009).

5.4 Evidence synthesis

In toxicological hazard and risk assessment there are many evidence synthesis information approaches. Evidence synthesis methods range from empirical approaches that are purely data-driven to phenomenological models and to explicit simulation models. They vary with regard to biological insight and degree of realism. Among evidence synthesis approaches used to fill in a hazard data gap are narrative weight of evidence, read-across, (Q)SARs, or Threshold of Toxicological Concern. Further examples of evidence synthesis approaches are the systematic reviews and meta-analysis introduced above, which have been proposed as a potentially useful methodology for test validation/assessment under the concept of an evidence-based toxicology (Hoffmann and Hartung, 2006; Hartung, 2009). Among evidence synthesis tools to assess exposure, both external and internal, there is a whole spectrum of models with different degrees of sophistication, from simple steady-state to dynamic models considering physiology and kinetics, such as PBPK models. These models have a physics law base and use a chemical's physical-chemical properties as input.

Due to poorer understanding of the pharmacodynamic effects, ITS are more applicable to hazard assessment, while pure exposure and distribution considerations can be better handled by increasingly more explicit exposure simulation models. However, we implicitly assume that inputs to ITS regarding effects do consider exposure in a concentration-effect manner, ensuring their biological relevance.

5.5 Methodology to integrate and guide testing

Once we have relevant pieces of information we need a framework for their integration. Data integration for ITS can be understood as a form of meta-analysis. However, classic meta-analytical techniques are not directly applicable, mainly due to differences in the inputs. Meta-analysis produces an estimate of the average effect seen in trials of a particular treatment (Smith, 1997) by a statistical approach that integrates the results of several independent studies considered to be combina-

Tab. 1: Comparison of BN and decision tree functionalities against identified conceptual requirements as well as selected other relevant factors related to their practical implementation.

	DTs	BNs
structure	Data centric No causal relationships between the nodes	Structure of the domain – centric Causal relationship representing knowledge map
consistent	Local optimization on individual branches, create over-complex trees that do not generalize the data well, often the same variables enter the decision tree but the order in which they enter the tree is different	Global optimization, generalize more consistently
Reason when data is incomplete	No, AND junctions	Yes, AND, OR, multiple OR junctions
Hypothesis driven	Probability of events	Probability distributions
Rational & efficient	Yes, but static	Yes, dynamic & adaptive
Communication tool	Easy for few variables but difficult for many	Compact way of a complex model representation

ble (Egger and Smith, 1997). The data inputs are homogenous in nature as the same effect is studied, only in different settings. In addition meta-analysis does not have the capacity to guide testing.

Decision trees were expected to be the method of choice, albeit some authors expressed concern about their potential size and complexity (Hartung, 2010). Despite the popularity of decision trees, it was already shown in other scientific areas (Bloemer et al., 2003) that the model structure of decision trees can sometimes be unstable due to variable correlation. This means that when carrying out multiple tests, mostly the same variables enter the decision tree, while the order of entry differs. Even if different decision trees that suffer from variable masking due to high correlation can perfectly arrive at the same final decision, the problem raises questions regarding the consistency and interpretation of the tree. Recently, Jaworska et al. (2010a) introduced an information-theoretic probabilistic framework for ITS in the form of a Bayesian network (BN). Bayesian networks can be seen as folded decision trees, which allow for a compact representation of the complex decision model and better handling of consistency by finding an optimal solution for the whole network structure. Other advantages become evident in Table 1, which presents a comparison of functionalities of BNs and decision trees regarding the identified conceptual requirements and other relevant factors related to their practical implementation.

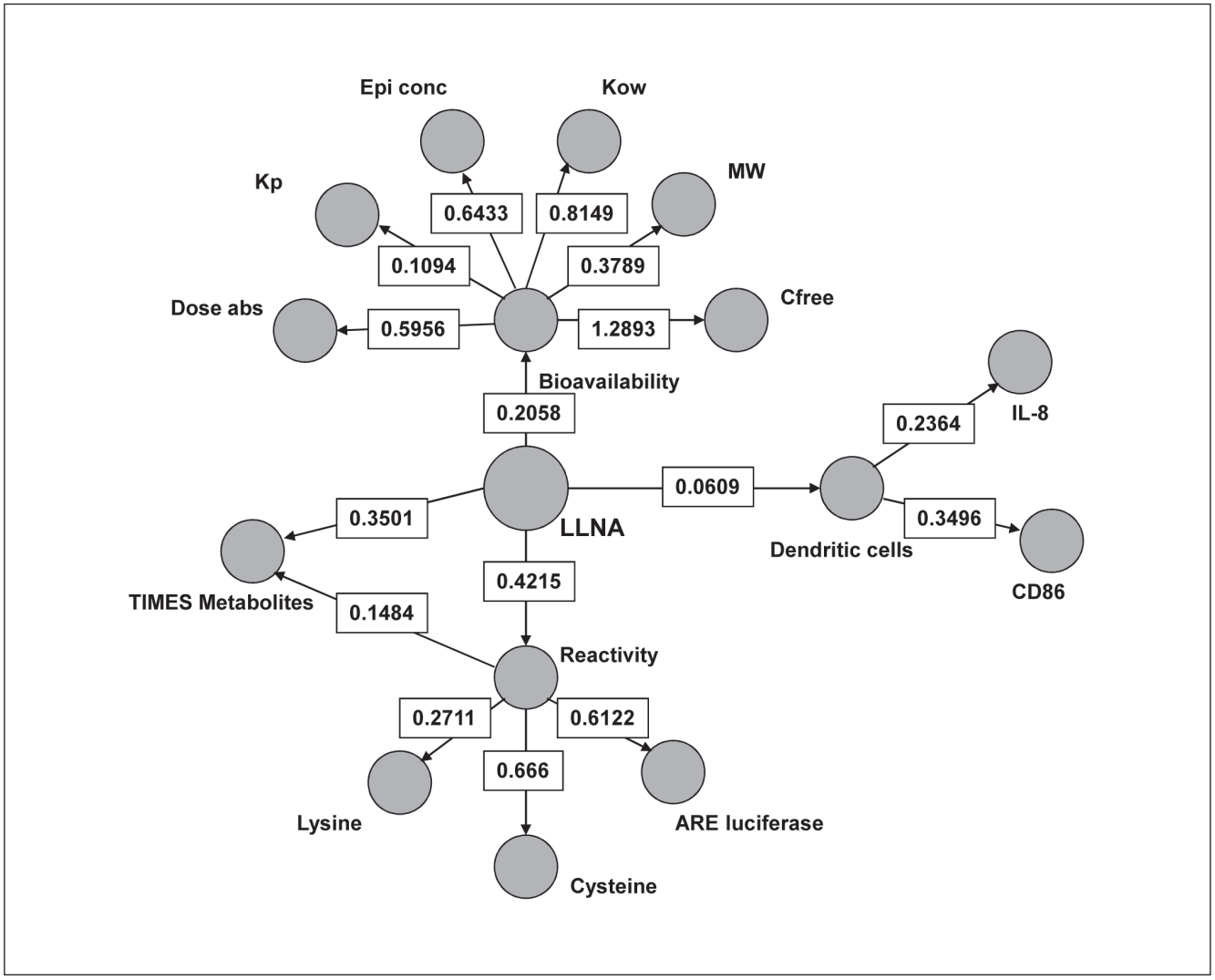

Fig. 1: Bayesian network ITS structure to reason about skin sensitization.
The LLNA potency is the information target of this ITS. The bioavailability, reactivity and dendritic cells nodes are latent (unobservable) variables and represent combined evidence from the observable tests (nodes) connected to them. The latent variables that can be interpreted as major lines of evidence connect to the information target. The nodes are characterized by a probability distribution. The arcs are characterized by conditional probability tables. Arcs are tagged with mutual information values that change upon providing evidence to the network.

The Bayesian network methodology is a formal approach for evidential reasoning that has been proven useful in many different domains, including medical diagnosis and testing and bioinformatics. A BN is a quantitative approach allowing for consistent, transparent, and reproducible inferences and decisions suited to combining information from multiple, heterogeneous sources. It is able to handle noise data with varying degrees of uncertainty. The probabilistic approach allows differentiation of relations that are known with a high level of certainty and those that are more speculative. It resolves conflicting evidence, reasons consistently given different and incomplete data sets. The Bayesian network formulation offers flexibility that can be used to express knowledge on major lines of evidence and on specific evidence on the level of individual tests. This functionality is very important because we expect ITS to have hierarchical structures. BN can also be used for guiding adaptive testing strategies based on dynamic calculations of Mutual Information and Value of Information (Pearl, 1988). A "one step look-ahead hypothesis" approach is used to identify information that has the highest potential to refine hypothesis variables. Reduction in the certainty of the evidence synthesis outcome related to conditional dependence between tests can be demonstrated and taken into account while assessing information gains from multiple assays. Although a BN currently seems an appealing solution to ITS design, it certainly is not the only solution, and others will emerge as use of ITS become more established.

6 Skin sensitization example

Using above considerations, Jaworska et al. (2010b) developed an ITS for skin sensitization in the form of a Bayesian network. The structure of the developed network reflects the current knowledge mapping about skin sensitization and includes the three key processes of dermal penetration, reaction with proteins, and dendritic cell activation. The framework combines prior biological knowledge with heterogeneous experimental evidence from twelve *in silico*, *in chemico* and *in vitro* tests and generates a probabilistic hypothesis about the skin sensitization potency of a chemical in the local lymph node assay (LLNA), which has served as the reference standard (Fig. 1). Inputs to bioavailability have been generated *in silico* and include molecular weight, the octanol/water partition coefficient (Log Kow), and calculated variables related to penetration based on a dynamic skin model (dose absorbed systemically, free chemical concentration in the skin, and maximum concentration in the epidermis). Inputs characterizing reactivity include data from *in chemico* tests such as reactivity with lysine, cysteine peptides and ARE-luciferase reactivity. Finally, the evidence related to dendritic cells is based on CD86 expression and IL-8 production of the human lymphoma cell line U937. Jaworska et al. (2010b) demonstrated how to use the BN both as a purely evidence synthesis tool as well as a tool to guide testing strategies. Moreover, BN-based ITS present an approach toward reduction and refinement postulated by Bessems (2009) and others. In particular, it can separate chemicals with well known potency, especially on the extremes, from chemicals for which more evidence needs to be generated by providing information target uncertainty distributions.

7 Conclusions

Future chemical management faces the challenge of dealing with a wealth of multifaceted information, all of which might support associated decision making. To be able to handle the resulting complexity, more structure and reduced heuristics are needed in decision making. ITS hold the promise of addressing this need and have the potential to significantly contribute to a modernization of risk assessment science. Along with the challenges, ITS have a unique opportunity to contribute to the Toxicology of the 21st century by providing frameworks and tools to actually implement 21st century toxicology data in the chemical management and decision making processes.

ITS development requires a conceptually consistent and transparent framework for data integration and efficient guidance of the testing. In order to make a real impact, more research on ITS operational frameworks is required. Operational frameworks will provide methodologies to identify decision-relevant information in the potentially unmanageable heap of information that we may be able to generate. It is essential that their structure is adaptive, recognizing that while preserving consistency, ITSs are supposed to flexibly accommodate substance-specific, exposure-related information as well as new mechanistic knowledge.

Because high-throughput datasets may suffer from technical and biological noise or from various technical biases and biological shortcomings, improved statistics are needed for the separation of signal from noise, as well as for better data integration annotating biologically relevant relationships. The logical interpretation of the complex signal propagation leading to an observed effect is not easily comprehensible. Therefore, computational modelling can be expected to play a crucial role in predicting the output from the signal input or system perturbation to obtain a more comprehensive, less technically biased and more accurate view of the true effect.

Due to its complexity, progress in ITS research requires a combined expertise from several life science fields, leveraging tools, methodologies, and technologies that were not traditionally used by toxicologists. Building such multidisciplinary teams presents another, organizational, challenge.

References

Ahlers, J., Stock, F. and Werschkun, B. (2008). Integrated testing and intelligent assessment – new challenges under REACH. *Environ. Sci. Pollut. Res. 15*, 565–572.

Alonzo, T. A. and Pepe, M. S. (1999). Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat. Med. 18*, 2987-3003.

Alonzo, T. A. and Pepe, M. S. (1998). Assessing the accuracy of

a new diagnostic test when a gold standard does not exist. *UW Biostatistics Working Paper Series*, working paper 156.

Andersen, M. E. and Krewski, D. (2009). Toxicity testing in the 21st century: bringing the vision to life. *Toxicol. Sci. 107*, 324-330.

Anon (2007). Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union L396*, 1-856, 2007.

Anon (2005). REACH and the need for intelligent testing strategies. [EUR 21554 EN], 32pp. Ispra, Italy: Institute for Health and Consumer Protection, EC Joint Research Centre.

Anon (2003). Directive 2003/15/EC of the European Parliament and the Council of 27 February amending Directive 76/786/EEC on the approximation of the laws of the member States relating to cosmetic products. *Official Journal of the European Communities L66*, 26-35.

Antezana, E., Kuiper, M. and Mironov, V. (2009). Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief Bioinform. 10*, 392-407.

Baughman, A. L., Bisgard, K. M., Cortese, M. M. et al. (2008). Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. *Clin. Vaccine Immunol. 15*, 106-114.

Benfenati, E., Gini, G., Hoffmann, S. and Luttik, R. (2010). Comparing in vivo, in vitro and in silico methods and integrated strategies for chemical assessment: problems and prospects. *ATLA 38*,153-166.

Bessems, J. G. M. (2009). Opinion on the usefulness of in vitro data for human risk assessment. Suggestions for better use of non-testing approaches. *RIVM brief rapport 320016002*, 25 p.

Blanchard, G. and Roquain, E. (2009). Adaptive FDR control under independence and dependence. *J. Mach. Learn Res. 10*, 2837-2871.

Bloemer, J. M. M., Brijs, T., Swinnen, G. and Vanhoof, K. (2003). Comparing complete and partial classification for identifying customers at risk, *International Journal of Research in Marketing 20 (2),* 117-131.

Boulesteix, A.-L. (2010). Over-optimism in bioinformatics research. *Bioinformatics 26*, 437-439.

Boulesteix, A.-L. and Hothorn, T. (2010). Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinformatics 11*, 78.

Boulesteix, A.-L. and Strobl, C. (2009). Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction. *BMC Med. Res. Methodol. 9*, 85.

Bradbury, S., Feijtel, T. and Van Leeuwen, K. (2004). Meeting the scientific needs of ecological risk assessment in a regulatory context. *Environ. Sci. Technol. 38*, 463a-470a.

Breton, R. L., Gilron, G., Thompson, R. et al. (2009). A new quality assurance system for the evaluation of ecotoxicity studies submitted under the New Substances Notification Regulations in Canada. *Integr. Environ. Assess. Manag. 5*, 127-137.

Burlinson, B., Tice, R. R., Speit, G. et al. (2007). In Vivo Comet Assay Workgroup, part of the Fourth International Workgroup on Genotoxicity Testing. *Mutat. Res. 627*, 31-35.

Caldas, J., Gehlenborg, N., Faisal, A. et al. (2009). Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics 25*, i145-153.

Cook, D. J., Mulrow, C. D. and Haynes, R. B. (1997). Systematic reviews: synthesis of best evidence for clinical decisions. *Ann. Intern. Med. 126*, 376-380.

Egger, M. and Smith, G. D. (1997). Meta-Analysis. Potentials and promise. *BMJ 315*, 1371-1374.

Enøe, C., Georgiadis, M. P. and Johnson, W. O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med. 45*, 61-81.

EBI – European Bioinformatics Institute. http://www.ebi.ac.uk (accessed Oct 05, 2010).

Gadaleta, E., Lemoine, N. R. and Chelala, C. (2010). Online resources of cancer data: barriers, benefits and lessons. Br*ief Bioinform.*, Epub ahead of print on March 25.

Georgiadis, M. P., Johnson, W. O. and Gardner, I. A. (2005). Sample size determination for estimation of the accuracy of two conditionally independent tests in the absence of a gold standard. *Prev. Vet. Med. 71*, 1-10.

GO (2010). "Gene Ontology", retrievable from the Internet: http://www.geneontology.org (accessed Oct 05, 2010).

Gottgtroy, P., Kasabov, N., MacDonell, S. (2006). Evolving ontologies for intelligent decision support. In Sanchez E. (ed.), Fuzzy Logic and the Semantic Web. *Capturing Intelligence 1*, 415-439.

Gubbels-van Hal, W. M., Blaauboer, B. J., Barentsen, H. M. et al. (2005). An alternative approach for the safety evaluation of new and existing chemicals, an exercise in integrated testing. *Regul. Toxicol. Pharmacol. 42*, 284-295.

Han, J. (2008). Understanding biological functions through molecular networks. *Cell Res. 18*, 224-237.

Hanson, S. O. and Rudén, C. (2007). Towards a theory of tiered testing. *Regul. Toxicol. Pharmacol. 48*, 35-44.

Hardy, B., Douglas, N., Helma, C. et al (2010). Collaborative development of predictive toxicology applications. *J. Cheminform.*, accepted 3 June 2010.

Hartung, T. (2010). Evidence based-toxicology, the toolbox of validation for the 21st century? *ALTEX*, current issue.

Hartung, T. (2009). Food for thought… on evidence-based toxicology. *ALTEX 26*, 75-82.

Hengstler, J. G., Foth, H., Kahl, R. et al. (2006). The REACH concept and its impact on toxicological sciences. *Toxicology 220*, 232-239.

Hobbs, D. A., Warne, M. S. and Markich, S. J. (2005). Evaluation of criteria used to assess the quality of aquatic toxicity

data. *Integr. Environ. Assess. Manag. 1*, 174-180.
Hoffmann, S. (2009). Aspects of test assessment. *Hum. Exp. Toxicol. 28*, 95-96.
Hoffmann, S. and Hartung, T. (2006). Towards an evidence-based toxicology. *Hum. Exp. Toxicol. 25*, 497-513.
Hoffmann, S. and Hartung, T. (2005). Diagnosis: toxic! – Trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations. *Toxicol. Sci. 85*, 422-428.
Hoffmann, S., Saliner, A. G., Patlewicz, G. et al. (2008). A feasibility study developing an integrated testing strategy assessing skin irritation potential of chemicals. *Toxicol. Lett. 180*, 9-20.
Hoffmann, S., Edler, L., Gardner, I. et al. (2008). Points of reference in the validation process: the report and recommendations of ECVAM Workshop 66. *ATLA 36*, 343-352
Hong, D. and Man, S. (2010). Signal propagation in small-world biological networks with weak noise. *J. Theor. Biol. 262*, 370-380.
Horvath, A. R. and Pewsner, D. (2004). Systematic reviews in laboratory medicine: principles, processes and practical considerations. *Clin. Chim. Acta 342*, 23-39.
Hothorn, L. A. (2002). Selected biostatistical aspects of the validation of in vitro toxicological assays. *ATLA 30, Suppl. 2*, 93-98.
Hulzebos, E. and Gerner, I. (2010). Weight factors in an Integrated Testing Strategy using adjusted OECD principles for (Q)SARs and extended Klimisch codes to decide on skin irritation classification. *Regul. Toxicol. Pharmacol. 58(1),* 131-144
Jaworska, J., Gabbert, S. and Aldenberg, T. (2010a). 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, P. and Gerberick, F. (2010b). Bayesian adaptive testing strategy based on integrated non-animal information – skin sensitization test case. *Tox. Sci.,* submitted.
Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist. 36*, 2207-2231.
Kavlock, R. J., Ankley, G., Blancato, J. et al. (2008). Computational toxicology – a state of the science mini review. *Tox. Sci. 103*, 14-27.
Kim, H.-G., Fillies, C., Smith, B. and Wikarski, D. (2002). Visualizing a dynamic knowledge map using semantic web technology. In *Engineering and Deployment of Cooperative Information Systems* (130-140). Berlin: Springer.
Klimisch, H.-J., Andreae, M. and Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul. Toxicol. Pharmacol. 25*, 1-5.
Knottnerus, J. A. and Muris, J. W. (2003). Assessment of the accuracy of diagnostic tests: the cross-sectional study. *J. Clin. Epidemiol. 56*, 1118-1128.
Lander, A. (2010). The edges of understanding. *BMC Biol. 8,* 40-44.
Lin, W.-Y. and Lee, W.-C. (2010). Incorporating prior knowledge to facilitate discoveries in a genome-wide association study on age-related macular degeneration. *BMC Res. Notes 3*, 26.
Matthews, L., Gopinath, G., Gillespie M. et al. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res. 37*, D619-622.
McInturff, P., Johnson, W. O., Cowling, D. and Gardner, I. A. (2004). Modelling risk when binary outcomes are subject to error. *Stat. Med. 23*, 1095-1109.
Meinshausen, M. (2008). Hierarchical testing of variable importance. Biometrika 95, 265-278.
Myers, C. L., Barrett, D. R., Hibbs, M. A. et al. (2006). Finding function: evaluation methods for functional genomic data. *BMC Genomics 7*, 187.
OECD (2008). Workshop on integrated approaches to testing and assessment. OECD Environment Health and Safety Publications. *Series on Testing and Assessment No. 88*. Paris: OECD.
OECD (2002). OECD guideline for testing of chemicals No. 404: Acute dermal irritation/corrosion. *Organisation for Economic Cooperation and Development, Paris*, 1-13.
Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco: Morgan Kaufmann Publishers.
Pepe, M. (2003). The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press.
Rao, C. V., Wolf, A. and Arkin, P. (2002). Control, exploitation and tolerance of intracellular noise. *Nature 420*, 231-237.
Reitsma, J. B., Rutjes, A. W., Khan, K. S. et al. (2009). A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J. Clin. Epidemiol. 62*, 797-806.
Rudén, C. (2001). The use and evaluation of primary data in 29 trichloroethylene carcinogen risk assessments. *Regul. Toxicol. Pharmacol. 34*, 3-16.
Rutjes, A. W., Reitsma, J. B., Coomarasamy, A. et al. (2007). Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol. Assess. 11*, iii, ix-51.
Schaafsma, G., Kroese, E. D., Tielemans, E. L. et al. (2009). REACH, non-testing approaches and the urgent need for a change in mind set. *Regul. Toxicol. Pharmacol. 53*, 70-80.
Schneider, K., Schwarz, M., Burkholder, I. et al. (2009). "ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicol. Lett. 189*, 138-144.
Schreider, J., Barrow, C., Birchfield, N. et al. (2010). Enhancing the credibility of decisions based on scientific conclusions: transparency is imperative. *Toxicol. Sci. 116*, 5-7.
Scott, L., Eskes, C., Hoffmann, S. et al. (2010). A proposed eye irritation testing strategy to reduce and replace in vivo studies using Bottom-Up and Top-Down approaches. *Toxicol. In Vitro 24*, 1-9.
Smith, G. D., Egger, M. and Phillips, A. N. (1997), Meta-analysis. Beyond the grand mean? *BMJ 315*, 1610-1614.
Spasic, I., Ananiadou, S., McNaught, J. and Kumar, A. (2005).

Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform. 6*, 239-251.

Šuc, D., Vladusic, D. and Bratko, I. (2004). Qualitatively faithful quantitative prediction. *Artif. Intell. 158*, 189-214.

van Leeuwen, C. J., Patlewicz, G. Y. and Worth, A. P. (2007). Intelligent testing strategies. In C. J. Van Leeuwen, T. G. Vermeire (eds.), *Risk assessment of chemicals. An introduction* (467-509, second ed.). Dordrecht, The Netherlands: Springer Publishers.

Vastrik, I., D'Eustachio, P., Schmidt, E. et al. (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biol. 8*, R39.

Weeber, M., Kors, J. A. and Mons, B. (2005). Online tools to support literature-based discovery in the life sciences. *Brief Bioinform. 6*, 277-286.

Weed, D. L. (2005). Weight of evidence: a review of concept and methods. *Risk Anal. 25*, 1545-1557.

Worth, A. P., Bassan, A., de Bruijn, J. et al. (2007). The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods. SAR QSAR Environ. *Res. 18*, 111-125.

Wu, W. (2009). On false discovery control under dependence. *Ann. Stat. 36*, 364-380.

Yoav, B. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat. 29*, 1165-1188.

Zehetmayer, S. and Posch, M. (2010). Post hoc power estimation in large-scale multiple testing problems. *Bioinformatics 26*, 1050-1056.

Zhou, X., Obuchowski, N. and McClish, D. (eds.) (2002). Methods for correcting imperfect standard bias. In *Statistical methods in diagnostic medicine* (359-395). New York: John Wiley & Sons.

Acknowledgement

The funding of the Doerenkamp-Zbinden Foundation and European Union 6th Framework OSIRIS Integrated Project (GOCE-037017-OSIRIS) is gratefully acknowledged.

Correspondence to

Joanna Jaworska, PhD, Principal Scientist
Procter & Gamble, Modelling & Simulation
Biological Systems, Brussels Innovation Center
Temselaan 100
1853 Strombeek-Bever, Belgium
e-mail: jaworska.j@pg.com