"Playing safe is probably the most
unsafe thing in the world.
You cannot stand still.
You must go forward"
Robert Collier (1885-1950)

Food for Thought ...
Integrated Testing Strategies for Safety Assessments

Thomas Hartung [1,2], Tom Luechtefeld [1], Alexandra Maertens [1], and Andre Kleensang [1]

[1]Johns Hopkins University, Bloomberg School of Public Health, CAAT, Baltimore, USA; [2]University of Konstanz, CAAT-Europe, Germany

Summary

Despite the fact that toxicology uses many stand-alone tests, a systematic combination of several information sources very often is required: Examples include: when not all possible outcomes of interest (e.g., modes of action), classes of test substances (applicability domains), or severity classes of effect are covered in a single test; when the positive test result is rare (low prevalence leading to excessive false-positive results); when the gold standard test is too costly or uses too many animals, creating a need for prioritization by screening. Similarly, tests are combined when the human predictivity of a single test is not satisfactory or when existing data and evidence from various tests will be integrated. Increasingly, kinetic information also will be integrated to make an in vivo extrapolation from in vitro data.

Integrated Testing Strategies (ITS) offer the solution to these problems. ITS have been discussed for more than a decade, and some attempts have been made in test guidance for regulations. Despite their obvious potential for revamping regulatory toxicology, however, we still have little guidance on the composition, validation, and adaptation of ITS for different purposes. Similarly, Weight of Evidence and Evidence-based Toxicology approaches require different pieces of evidence and test data to be weighed and combined.

ITS also represent the logical way of combining pathway-based tests, as suggested in Toxicology for the 21st Century. This paper describes the state of the art of ITS and makes suggestions as to the definition, systematic combination, and quality assurance of ITS.

Keywords: Integrated testing strategies, prioritization, predictivity, quality assurance, Tox-21c

Introduction

Replacing a test on a living organism with a cellular, chemico-analytical, or computational approach obviously is reductionistic. Sometimes this might work well, e.g., when an extreme pH is a clear indication of corrosivity. In general, however, it is naïve to expect a single system to substitute for all mechanisms, the entire applicability domain (substance classes), and degrees of severity. Still, toxicology has long neglected this when requesting a one-to-one replacement to substitute for the traditional animal test. We might even extend this to say it is similarly naïve to address an entire human health effect with a single animal experiment using inbred, young rodents… The only way to approximate human relevance is to mimic the complexity and responsiveness of the organ situation and to model the respective kinetics, i.e., the target hoped for from the human-on-a-chip approach (Hartung and Zurlo, 2012). Everything else requires making use of several information sources, if not compromising the coverage of the test. Genotoxicity is a nice example, where patches have continuously been added to cover the various mechanisms. However, here the simplest possible strategy, i.e., a battery of tests where every positive result is considered a liability, causes problems. We have seen where the inevitable accumulation of false-positives leads (Kirkland et al., 2005), ultimately undermining the credibility of *in vitro* approaches.

The solution is the "intelligent" or "integrated" use of several information sources in a testing strategy (ITS). There is a lot of confusion around this term, especially regarding how to design, validate, and use ITS.

This article aims to elaborate on these aspects with examples and to outline the prospects of ITS in toxicology. It thereby expands the thoughts elaborated for the introduction to the roadmap for animal-free systemic toxicity testing (Basketter et al., 2012). The underlying problems and the approach are not actually unique to toxicology. The most evident similarity is to diagnostic testing strategies in clinical medicine, where several sources of information are used, similarly, for differential diagnosis; we discussed these similarities earlier (Hoffmann and Hartung, 2005).

Consideration1: The two origins of ITS in safety assessments

When do we need a test and when do we need a testing strategy? We need more than one test, if:

- not all possible outcomes of interest (e.g., modes of action) are covered in a single test
- not all classes of test substances are covered (applicability domains)
- not all severity classes of effect are covered
- when the positive test result is rare (low prevalence) and the number of false-positive results becomes excessive (Hoffmann and Hartung, 2005)
- the gold standard test is too costly or uses too many animals and substances need to be prioritized
- the accuracy (human predictivity) is not satisfying and predictivity can be improved
- existing data and evidences from various tests shall be integrated
- kinetic information shall be integrated to make an *in vivo* extrapolation from *in vitro* data (Basketter et al., 2012)

All together, it is difficult to imagine a case where we should not apply a testing strategy. It is astonishing how long we have continued to pursue "*one test suits all*" solutions in toxicology. A restricted usefulness (applicability domain) was stated, but it was only within the discussion on Integrated Testing of *in vitro*, *in silico*, and toxicokinetics (adsorption, distribution, metabolism, excretion, i.e., ADME) information that such integration was attempted. Bas Blaauboer and colleagues long ago spearheaded this (DeJongh et al., 1999; Forsby and Blaauboer, 2007; Blaauboer, 2010; Blaauboer and Barratt, 1999). The first ITS were accepted as OECD test guidelines in 2002 for eye and skin irritation (OECD TG 404, 2002a; OECD TG 405, 2002b). A major driving force then was the emerging REACH legislation, which sought to make use of all available information for registration of chemicals (especially existing chemicals) in order to limit costs and animal use. This prompted the call for Intelligent TS (Anon., 2005; Van Leeuwen et al., 2007; Ahlers et al., 2008; Schaafsma et al., 2009; Vonk et al., 2009; Combes and Balls, 2011; Leist et al., 2012; Gabbert and Benighaus, 2012; Rusyn et al., 2012). The two differ to some extent as the REACH-ITS also include *in vivo* data and are somewhat restricted to the tools prescribed in legislation. This largely excludes the 21st century methodologies (van Vliet, 2011), i.e., omics, high-throughput, and high-content imaging techniques, which are not mentioned in the legislative text. The very narrow interpretation of the legislative text in administrating REACH does not encourage such additional approaches. This represents a tremendous lost opportunity, and some additional flexibility and "*learning on the job*" would benefit one of the largest investments in consumer safety ever attempted.

Astonishingly, despite these prospects and billions of Euros spent for REACH, the literature on ITS for safety assessments is still poor, and little progress toward consensus and guidance has been made. For example, two In Vitro Testing Industrial Platform workshops were summarized stating (De Wever et al., 2012): "*As yet, there is great dispute among experts on how to represent ITS for classification, labeling, or risk assessments of chemicals, and whether or not to focus on the whole chemical domain or on a specific application. The absence of accepted Weight of Evidence (WoE) tools allowing for objective judgments was identified as an important issue blocking any significant progress in the area*." Similarly, the ECVAM/EPAA workshop concluded (Kinsner-Ovaskainen et al., 2012): "*Despite the fact that some useful insights and preliminary conclusions could be extracted from the dynamic discussions at the workshop, regretfully, true consensus could not be reached on all aspects*."

We earlier commissioned a white paper on ITS (Jaworska and Hoffmann, 2010) in the context of our transatlantic think tank for toxicology (t^4) and a 2010 conference on *21st Century Validation Strategies for 21st Century Tools*. It similarly concluded: "*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)*."

It is worth noting, also, that testing strategies from the pharmaceutical industry do not help much. They try to identify an active compound (the future drug) out of thousands of substances, without regard to what they miss – but this approach is unacceptable in a safety ITS. Pharmacology screening also typically starts with a target, i.e., a mode of action, while toxicological assessments need to be open to various mechanisms, some as yet uncharacterized, until we have a comprehensive list of relevant pathways of toxicity (Hartung and McBride, 2011).

Due to them having grown from alternative methods and REACH, ITS discussions are more predominant in Europe (Hartung, 2010d). However, in principle they resonate very strongly with the US approach of toxicity testing in the 21st century (Tox-21c) (Hartung, 2009c). The latter suggests moving regulatory toxicology to mechanisms (the pathways of toxicity, PoT). This means breaking the hazard down into its modes of action and combining them with chemico-physical properties (including QSAR) and PBPK models. This implies, similarly, that different pieces of evidence and tests be strategically combined.

Consideration 2: The need for a definition of ITS

Currently, the best reference for definitions of terminology is provided by OECD guidance document 34 on validation (OECD, 2005). An extract of the most relevant definitions is given in Box 1. Notably, the term (integrated) test strategy is not defined.

Box 1

Relevant definitions from OECD Series on Testing and Assessment No. 34 (OECD, 2005)

Adjunct test: Test that provides data that add to or help interpret the results of other tests and provide information useful for the risk assessment process
Assay: Uses interchangeably with Test.
Data interpretation procedure (DIP): An interpretation procedure used to determine how well the results from the test predict or model the biological effect of interest. See Prediction Model.
Decision Criteria: The criteria in a test method protocol that describe how the test method results are used for decisions on classification or other effects measured or predicted by the test method.
Definitive test: A test that is considered to generate sufficient data to determine the specific hazard or lack of hazard of the substance without the need for further testing, and which may therefore be used to make decisions pertaining to hazard or safety of the substance.
Hierarchical (tiered) testing approach: An approach where a series of tests to measure or elucidate a particular effect are used in an ordered sequence. In a typical hierarchical testing approach, one or a few tests are initially used; the results from these tests determine which (if any) subsequent tests are to be used. For a particular chemical, a weight-of-evidence decision regarding hazard could be made at any stage (tier) in the testing strategy, in which case there would be no need to proceed to subsequent tiers.
In silico models: Approaches for the assessment of chemicals based on the use computer-based estimations or simulations. Examples include structure-activity relationships (SAR), quantitative structure-activity relationships (QSARs), and expert systems.
(Q)SARs (Quantitative Structure-Activity Relationships): Theoretical models for making predictions of physicochemical properties, environmental fate parameters, or biological effects (including toxic effects in environmental and mammalian species). They can be divided into two major types, QSARs and SARs. QSARs are quantitative models yielding a continuous or categorical result while SARs are qualitative relationships in the form of structural alerts that incorporate molecular substructures or fragments related to the presence or absence of activity.
A screen/screening test is often a rapid, simple test method conducted for the purpose of classifying substances into a general category of hazard. The results of a screening test generally are used for preliminary decision making in the context of a testing strategy (i.e., to assess the need for additional and more definitive tests). Screening tests often have a truncated response range in that positive results may be considered adequate to determine if a substance is in the highest category of a hazard classification system without the need for further testing, but are not usually adequate without additional information/tests to make decisions pertaining to lower levels of hazard or safety of the substance
Test (or assay): An experimental system used to obtain information on the adverse effects of a substance. Used interchangeably with assay.
Test battery: A series of tests usually performed at the same time or in close sequence. Each test within the battery is designed to complement the other tests and generally to measure a different component of a multi-factorial toxic effect. Also called base set or minimum data set in ecotoxicological testing.
Test method: A process or procedure used to obtain information on the characteristics of a substance or agent. Toxicological test methods generate information regarding the ability of a substance or agent to produce a specified biological effect under specified conditions. Used interchangeably with "test" and "assay".

Following a series of ECVAM internal meetings, an ECVAM/EPAA workshop was held to address this (Kinsner-Ovaskainen et al., 2009), and it came up with a working definition: "*As previously defined within the literature, an ITS is essentially an information-gathering and generating strategy, which in itself does not have to provide means of using the information to address a specific regulatory question. However, it is generally assumed that some decision criteria will be applied to the information obtained, in order to reach a regulatory conclusion. Normally, the totality of information would be used in a weight-of-evidence (WoE) approach*." WoE had been addressed in an earlier ECVAM workshop (Balls et al., 2006): "*Weight of evidence (WoE) is a phrase used to describe the type of consideration made in a situation where there is uncertainty and which is used to ascertain whether the evidence or information supporting one side of a cause or argument is greater than that supporting the other side*." It is of critical importance to understand that WoE and ITS are two different concepts although they combine the same types of information! In WoE there is no formal integration, usually no strategy, and often no testing. WoE is much more a "*poly-pragmatic shortcut*" to come to a preliminary decision, where there is no or only limited certainty. As proponents of evidence-based toxicology (EBT) (Hoffmann and Hartung, 2006), we have to admit that the term EBT further contributes to this confusion (Hartung, 2009b). However, there is obvious cross-talk between these approaches when, for example, the quality

scoring of studies developed for EBT (Schneider et al., 2009) helps to filter their use in WoE and ITS approaches.

The following definition was put forward by the ECVAM/EPAA workshop (Kinsner-Ovaskainen et al., 2009): "*In the context of safety assessment, an Integrated Testing Strategy is a methodology which integrates information for toxicological evaluation from more than one source, thus facilitating decision-making. This should be achieved whilst taking into consideration the principles of the Three Rs (reduction, refinement and replacement).*" In line with the proposal put forward in the 2007 OECD Workshop on Integrated Approaches to Testing and Assessment, they reiterated, "*a good ITS should be structured, transparent, and hypothesis driven*" (OECD, 2008).

Jaworska and Hoffmann (2010) defined ITS somewhat differently: "*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 the current regulated sequential testing strategy for skin irritation (OECD, 2002[a]) or the recently proposed in vitro testing strategy for eye irritation (Scott et al., 2009). 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.*"

It might be useful to start from scratch with our definitions to avoid some glitches.

- The leading principle should be that a test gives one result, and it does not matter how many endpoints (measurements) the test requires. Figure 1 shows these different scenarios. A test/assay thus consists of a test system (biological *in vivo* or *in vitro* model) and a Standard Operation Protocol (SOP) including endpoint(s) to measure, reference substance(s), data interpretation procedure (a way to express the result), information on reproducibility / uncertainty, applicability domain / information on limitations and favorable performance standards. Note that tests can include multiple test systems and/or multiple endpoints as long as they lead to one result.
- An *integrated test strategy* is an algorithm to combine (different) test result(s) and, possibly, non-test information (existing data, *in silico* extrapolations from existing data or modeling) to give a combined test result. They often will have interim decision points at which further building blocks may be considered.
- A *battery of tests* is a group of tests that complement each other but are not integrated into a strategy. A classical example is the genotoxicity testing battery.

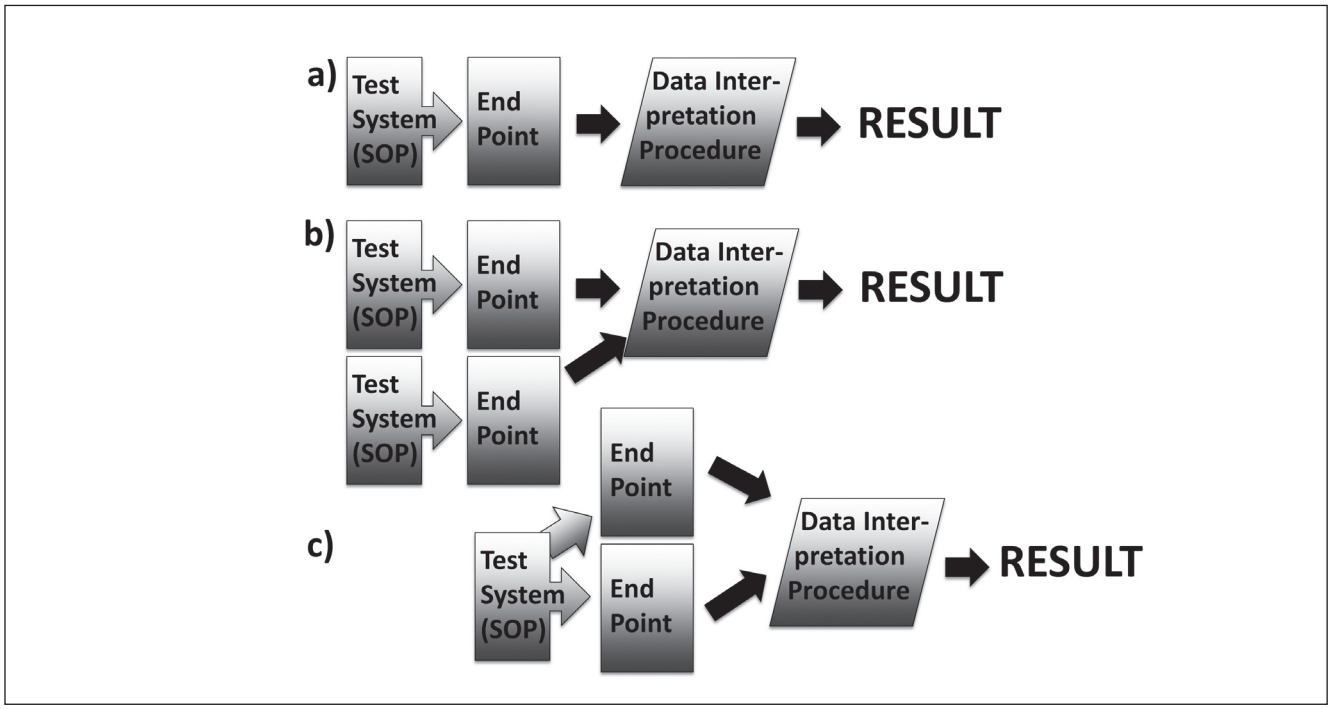

Fig. 1: Three prototypic tests
(a) a simple test with one endpoint, (b) two test systems giving a joint result, and (c) multiple endpoints (including omics and other high-content analysis)

– *Tiered testing* describes the simplest ITS, where a sequence of tests is defined without formal integration of results.
– A *probabilistic TS* describes an ITS, where the different building blocks change the probability for a test result.
– *Validation* of a test or an ITS requires a *prediction model* (a way to translate it to the *point of reference*) and the point of reference itself, which can be correlative on the basis of results, or mechanistic.

Some of these aspects are shown in Figure 2.

Consideration 3: Composition of ITS – no GOBSATT!

The ITS in use to date is based on consensus processes often called "*weight of evidence*" (WoE) approaches. Such "*Good old boys sitting around the table*" (GOBSATT) is not the method of choice to compose ITS. The complexity of data and the multiplicity of performance aspects to consider (costs, animal use, time, predictivity, etc.) (Nordberg et al., 2008; Gabbert and van Ierland, 2010) call for simulation based on test data. The shortcomings of existing ITS were recently analyzed in detail by Jaworska et al. (2010): "*Though both current ITS and WoE approaches are undoubtedly useful tools for systemizing chemical hazard and risk assessment, they lack a consistent methodological basis for making inferences based on existing information, for coupling existing information with new data from different sources, and for analyzing test results within and across testing stages in order to meet target information requirements*." And in more detail in (Jaworska and Hoffmann, 2010): "*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*." Hoffmann et al. (2008a) provided a pioneering example of ITS evaluation focused on skin irritation. They compiled a database of 100 chemicals. A number of strategies, both animal-free and inclusive of animal data, were constructed and subsequently evaluated considering predictive capacities, severity of misclassifications, and testing costs. Note that the different ITS to be compared were "hand-made," i.e., based on scientific reasoning and intuition, but not on any construction principles. They correctly conclude: "*To promote ITS, further guidance on construction and multi-parameter evaluation need to be developed*." Similarly, the ECVAM/EPAA workshop only stated needs (Kinsner-Ovaskainen et al., 2009): "*So far, there is also a lack of scientific knowledge and guidance on how to develop an ITS and, in particular, on how to combine the different building blocks for an efficient and effective decision-making process. Several aspects should be taken into account in this regard, including:*
– *the extent of flexibility in combining the ITS components;*
– *the optimal combination of ITS components (including the minimal number of components and/or combinations that have a desired predictive capacity);*
– *the applicability domain of single components and the whole ITS; and*

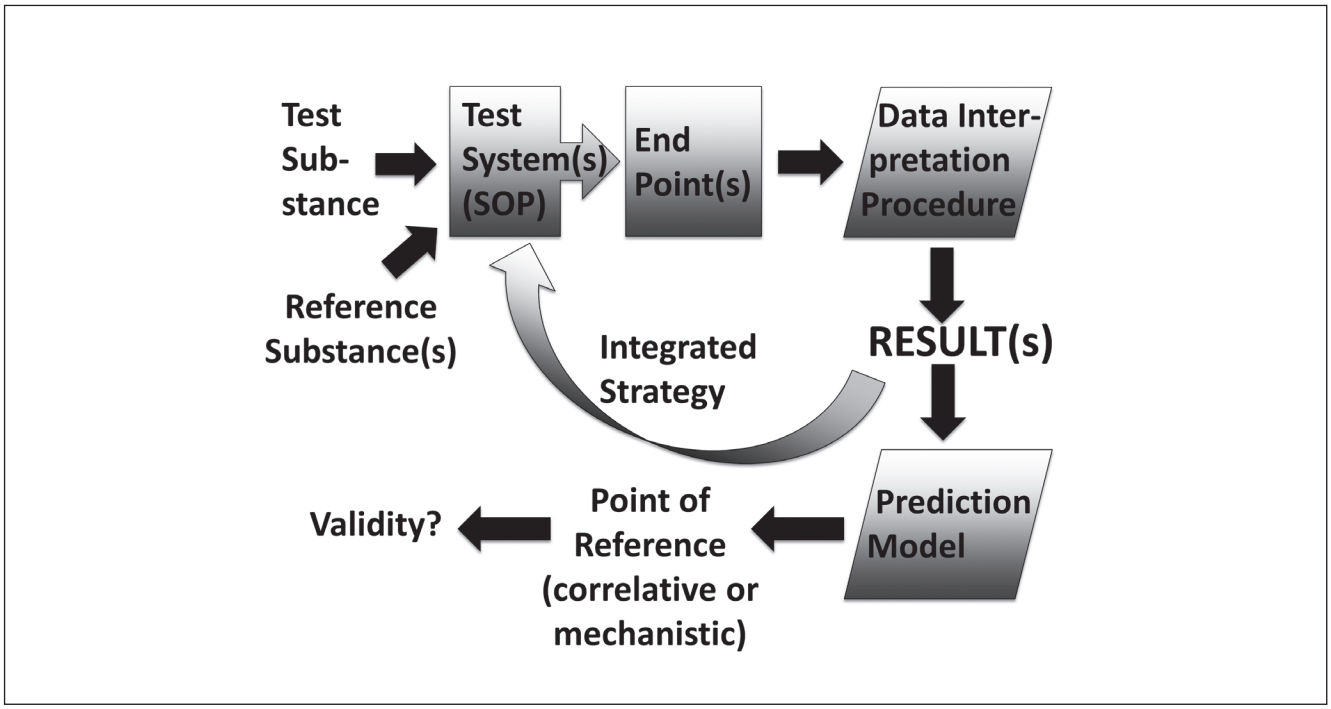

Fig. 2: Components of a test (strategy) and its traditional (correlative) or mechanistic validation

– *the efficiency of the ITS (cost, time, technical difficulties)*"
Using this "wish list" as guidance some aspects will be discussed.

Extent of flexibility in combining the ITS components:
This is a key dilemma – any validation "sets tests into stone" and "freezes them in time" (Hartung, 2007). An ITS, however, is so much larger than individual tests that there are even more reasons for change (technical advances, limitations of individual ITS components for the given study substance, availability of all tests in a given setting, etc.). What is needed here is a measure of similarity of tests and performance standards. The latter concept was introduced in the modular approach to validation (Hartung et al., 2004) and is now broadly used for the new validations. It defines what criteria a "me-too" development (a term borrowed from the pharmaceutical industry, where a competitor follows the innovative, pioneering work of another company, introducing a compound with the same active principle) must fulfill to be considered equivalent to the original one. The idea is to avoid undertaking another full-blown validation ring trial which requires enormous resources. There is some difference in interpretation as to whether there still needs to be a multi-laboratory exercise to establish inter-laboratory reproducibility and transferability as well. Note that this requires demonstrating the similarity of tests, for which we have no real guidance. It also implies, however, that any superiority of the new test compared to the originally validated one cannot be shown. For ITS components, in the same way, similarity and performance criteria need to be established to allow exchange for something different without a complete reevaluation of the ITS. This can first be based on the scientific relevance and the PoT covered, as argued earlier (Hartung, 2010b). This means that two assays that cover the same mechanism can substitute for each other. Alternatively, it can be based on correlation of results. Two assays that agree (concordance) to a sufficient degree, can be considered similar. We might call these two options "*mechanistic similarity*" and "*correlative similarity*."

The optimal combination of ITS components:
The typical combination of building blocks so far follows a Boolean logic, i.e., the logical combinations are AND, OR, and NOT. Table 1 gives the different examples for combining two tests with dichotomous (plus/minus) outcome with such logic and the consequences for the joint applicability domain and the validation need. Note that in most cases the validation of the building blocks will suffice, but the joint applicability domain will be just the overlap of the two tests' applicability domains. This is a simple application of set theory. Only if the two tests measure the same but for different substances / substance severity classes, the logical combination OR results in the combined applicability domain. If the result requires that both tests are positive, e.g., when a screening test and a confirmatory test are combined, it is necessary to validate the overall ITS outcome.

The principal opportunities in combining tests into the best ITS lie, however, in interim decision points (Figure 3 shows a simple example, where the positive or negative outcome is confirmed). Here, the consequences for the joint applicability domain are more complex and typically only the overall outcome can be validated. The other opportunity is to combine tests not with Boolean logic but with fuzzy/probabilistic logic. This means that the result is not dichotomous (toxic or not) but a probability or score is assigned. We could say that a value between 0 (non-toxic) and 1 (toxic) is assigned. Such combinations typically will only allow use in the overlapping applicability domains. It also implies that only the overall ITS can be validated. The challenge here lies mostly in the point of reference, which normally needs to be graded and not dichotomous as well.

The advantages of a probabilistic approach were recently summarized by Jaworska and Hoffmann (2010): "*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*."

The applicability domain of single components and the whole ITS:
Simple logic shows, as discussed above, that in most instances an ITS can be applied only where all building blocks applied to a substance allow so. The picture changes only if the combination serves exactly the purpose of expanding the applicability domain (by combining two tests with OR). This implies, however, that essentially the same thing is measured (i.e., similarity of tests); if tests differ in applicability domain and what they measure, a hierarchy needs to be established first. This is one of the key arguments for flexibility of ITS, as we need to exchange building blocks for others to meet the applicability domain for a given substance.

The efficiency of the ITS:
Typically, efficiency refers to resources such as cost and labor. Animal use and suffering, however, lies outside its scope. How to value the replacement of an animal test is a societal decision. In the EU legislation, the term "*reasonably available*" is used to mandate the use of an alternative (Hartung, 2010a). This leaves room for interpretation, but there certainly are limits: How much more costly can an alternative method be to be reasonably available? The cost/benefit calculation also needs to include societal acceptability. However, this is missing the point: In the end, the concept of efficiency centers on predicting human health and environmental effects. What are the costs of a test versus the risk of a scandal? If we only attempt to be as good as the animal test, however, this argument has no leverage. Thus we need to advance to human relevance if we really want impact. This is difficult on the level of correlation, because we typically do not have the human data for a statistically sufficient number of substances. More and more, however, we do

Tab. 1: Test combinations and consequences for applicability domain and validation needs

Logic	Example	Joint Applicability Domain	Validation Need
Boolean			
A AND B	Screening plus confirmatory test	Overlap	Total ITS
A OR B	Different Mode of Action	Overlap	Building Blocks
	Different Applicabilty Domain or Severity Grades	Combined	Building Blocks
A NOT B	Exclusion of a property (such as cytotoxicity)	Overlap	Total ITS
IF A positive: B IF A negative: C See Figure 3	Decision points, here confirmation of result in a second test	Combined overlap A/B and overlap A/C	Total ITS
Fuzzy / Probabilistic			
p(A, B) i.e., probability as function of A and B	Combined change of probability, e.g., priority score	Overlap	Building Blocks

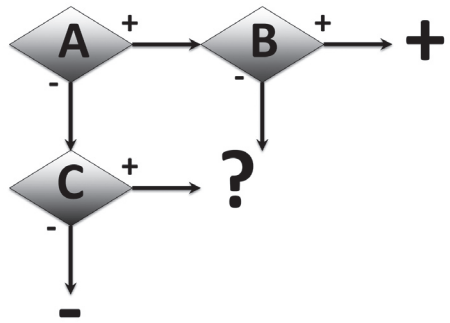

Fig. 3: Illustration of a simple decision tree, where outcomes of test A are confirmed by different second tests B or C

know the mechanisms relevant to human health effects. Thus, the efficiency with which a test system covers relevant mechanisms for human health and environmental effects is becoming increasingly important. I have called this "*mechanistic validation*" (Hartung, 2007). This requires that we establish causality for a given mechanism to create a health or environmental effect. The classical frameworks of the Koch-Dale postulates (Dale, 1929) and Bradford-Hill criteria (Hill, 1965) for assessing evidence of causation come to mind first. Dale translated the Koch postulates that need to be fulfilled to prove a pathogen to be the cause of a certain disease to ones that prove a mediator (at the time histamine and neurotransmitters) causes a physiological effect. We recently applied this to systematically evaluate the nature of the Gram-positive bacterial endotoxin (Rockel and Hartung, 2012). Similarly, we can translate this to a PoT being responsible for the manifestation of an adverse cellular outcome of substance X:

– Evidence for presence of the PoT in affected cells
– Perturbation/activation of the PoT leads to or amplifies the adverse outcome
– Hindering PoT perturbation/activation diminishes manifestation of the adverse outcome
– Blocking the PoT once perturbed/activated diminishes manifestation of the adverse outcome

Please note that the current debate as to whether a PoT represents a chemico-biological interaction impacting on the biological system or the perturbed normal physiology is reflected in using both terminologies.

Similarly, the Bradford-Hill criteria can be applied:

– Strength: The stronger an association between cause and effect the more likely a causal interpretation, but a small association does not mean that there is not a causal effect.
– Consistency: Consistent findings of different persons in different places with different samples increase the causal role of a factor and its effect.
– Specificity: The more specific an association is between factor and effect, the bigger the probability of a causal relationship.
– Temporality: The effect has to occur after the cause.
– Biological gradient: Greater exposure should lead to greater incidence of the effect, with the exception that it can also be inverse, meaning greater exposure leads to lower incidence of the effect.
– Plausibility: A possible mechanism between factor and effect increases the causal relationship, with the limitation that

knowledge of the mechanism is limited by best available current knowledge.
– Coherence: A coherence between epidemiological and laboratory findings leads to an increase in the likelihood of this effect. However, the lack of laboratory evidence cannot nullify the epidemiological effect on the associations.
– Experiment: Similar factors that lead to similar effects increase the causal relationship of factor and effect.

Most recently, a new approach to causation was proposed, originating from ecological modeling (Sugihara et al., 2012; Marshall, 2012). Whether this offers an avenue to systematically test causality in large datasets from omics and/or high-throughput testing needs to be explored. It might represent an alternative to choosing meaningful biomarkers (Blaauboer et al., 2012), being always limited to the current state of knowledge.

As a more pragmatic approach, DeWever et al. (De Wever et al., 2012) suggested key elements of an ITS:

"(1) Exposure modelling to achieve fast prioritisation of chemicals for testing, as well as the tests which are most relevant for the purpose. Physiologically based pharmacokinetic modelling (PBPK) should be employed to determine internal doses in blood and tissue concentrations of chemicals and metabolites that result from the administered doses. Normally, in such PBPK models, default values are used. However, the inclusion of values or results form in vitro *data on metabolism or exposure may contribute to a more robust out-come of such modelling systems.*

(2) Data gathering, sharing and read-across for testing a class of chemicals expected to have a similar toxicity profile as the class of chemicals providing the data. In vitro *results can be used to demonstrate differences or similarities in potency across a category or to investigate differences or similarities in bioavailability across a category (e.g. data from skin penetration or intestinal uptake).*

(3) A battery of tests to collect a broad spectrum of data focussing on different mechanisms and mode of actions. For instance changes in gene expression, signalling pathway alterations could be used to predict toxic events which are meaningful for the compound under investigation.

(4) Applicability of the individual tests and the ITS itself has to be assured. The acceptance of a new method depends on whether it can be easily transferred from the developer to other labs, whether it requires sophisticated equipment and models, or if intellectual property issues and the costs involved are important. In addition, an accurate description of the compounds that can and cannot be tested is essential in this context.

(5) Flexibility allowing for adjustment of the ITS to the target molecule, exposure regime or application.

(6) Human-specific methods should be prioritised whenever possible to avoid species differences and to eliminate 'low dose' extrapolation. Thus, the in vitro *methods of choice are based upon human tissues, human tissue slices or human primary cells and cell lines for* in vitro *testing. If* in vivo *studies be unavoidable, transgenic animals should be the preferred choice if available. If not, comparative genomics (animal versus human) and computational models of kinetics and dynamics in animals and humans may help to overcome species differences."*

This "shopping list" extends ITS from hazard identification to exposure considerations and the inclusion of existing data beyond de novo testing (including some quite questionable approaches of read-across and forming of chemical classes, for which no guidance or quality assurance is yet available). It similarly calls for flexibility, a key difference from the current guidance documents from ECHA or OECD. Compared to REACH, it calls for human predictivity and mode-of-action information in the sense of *Toxicity Testing for the 21st Century*. Similarly, an earlier report, also based on an IVTP symposium to which the author contributed, made further recommendations relating to a concept based on pathways of toxicity (PoT) (Berg et al., 2011): "*When selecting the battery of* in vitro *and* in silico *methods addressing key steps in the relevant biological pathways (the building blocks of the ITS) it is important to employ standardized and internationally accepted tests. Each block should be producing data that are reliable, robust and relevant (the alternative 3R elements) for assessing the specific aspect (e.g. biological pathway) it is supposed to address. If they comply with these elements they can be used in an ITS*."

Hoffmann et al. (2008a) added an important consideration: "*Furthermore, the study underlined the need for databases of chemicals with testing information to facilitate the construction of practical testing strategies. Such databases must comprise a good spread of chemicals and test data in order that the applicability of approaches may be effectively evaluated. Therefore, the (non-) availability of data is a caveat at the start of any ITS construction. Whilst* in silico *and* in vitro *data may be readily generated, in vivo data of sufficient quality are often difficult to obtain*." This brings us back to both the need for data-sharing (Basketter et al., 2012) and the construction of a point of reference for validation exercises (Hoffmann et al., 2008b).

The most comprehensive framework for ITS composition so far was produced by Jaworska and Hoffmann as a t^4-commissioned white paper (Jaworska and Hoffmann, 2010), see also (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.
… 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 reader is referred to the original article (Jaworska and Hoffmann, 2010) and its implementation for skin sensitization (Jaworska et al., 2011).

Consideration 4: Guidance from testing strategies in clinical diagnostics

We earlier stressed the principal similarities between a diagnostic and a toxicological test strategy (Hoffmann and Hartung, 2005). In both cases, different sources of information have to be combined to come to an overall result. Vecchio pointed out as early as 1966 the problem of single tests in unselected populations (Vecchio, 1966) leading to unacceptable false-positive rates. Systematic reviews of an evidence-based toxicology (EBT) approach (Hoffmann and Hartung, 2006; Hartung, 2009b) and meta-analysis could serve the evaluation and quality assurance of toxicological tests. The frameworks for evaluation of clinical diagnostic tests are well developed (Deeks, 2001; Devillé et al., 2002; Leeflang et al., 2008) and led to the Cochrane Handbook for Diagnostic Test Accuracy Reviews (Anon., 2011). Devillé et al. (2002) give very concise guidance on how to evaluate diagnostic methods. This is closely linked to efforts to improve reporting on diagnostic tests; a set of minimal reporting standards for diagnostic research has been proposed: Standards for Reporting of Diagnostic Accuracy statement (STARD)[1]. We argued earlier that this represents an interesting approach to complement or substitute for traditional method validation (Hartung, 2010b). Deeks (2001) summarize their experience as follows [with translation to toxicology inserted in brackets]: "*Systematic reviews of studies of diagnostic* [hazard assessment] *accuracy differ from other systematic reviews in the assessment of study quality and the statistical methods used to combine results. Important aspects of study quality include the selection of a clinically relevant cohort* [relevant test set of substances], *the consistent use of a single good reference standard* [reference data], *and the blinding of results of experimental and reference tests. The choice of statistical method for pooling results depends on the summary statistic and sources of heterogeneity, notably variation in diagnostic thresholds* [thresholds of adversity]. *Sensitivities, specificities, and likelihood ratios may be combined directly if study results are reasonably homogeneous. When a threshold effect exists, study results may be best summarised as a summary receiver operating characteristic curve, which is difficult to interpret and apply to practice*."

Interestingly, Schünemann et al. (2008) developed GRADE for grading quality of evidence and strength of recommendations for diagnostic tests and strategies. This framework uses "*patient-important outcomes*" as measures, in addition to test accuracy. A less invasive test can be better for a patient even if it does not give the same certainty. Similarly, we might frame our choices by aspects such as throughput, costs, or animal use.

Consideration 5: The many faces of (I)TS for safety assessments

As defined earlier, any systematic combination of different (test) results represents a testing strategy. It does not really matter if these results already exist, are estimated from structures or related substances, measured by chemico-physical methods, or stem from testing in a biological system or from human observations and studies. Jaworska et al. (2010) and Basketter et al. (2012) list many of the more recently proposed ITS. One of the authors (THA) had the privilege to coordinate from the side of the European Commission the ITS development within the guidance for REACH implementation for industry, which formed the basis for current ECHA guidance.[2] Classical examples in toxicology, some of them commonly used without the label ITS, are:

Test battery of genotoxicity assays:
Several assays (3-6) depending on the field of use (Hartung, 2008) are carried out and, typically, any positive result is taken as an alert. They often are combined with further mutagenicity testing *in vivo* (Hartung, 2010c). The latter is necessary to reduce the tremendous rate of false-positive classifications of the battery, as discussed earlier (Basketter et al., 2012). Interestingly, Aldenberg and Jaworska (2010) applied a Bayesian network to the dataset assembled by Kirkland et al. showing the potential of a probabilistic network to analyze such datasets.

ITS for eye and skin irritation:
As already mentioned, these were the first areas to introduce internationally accepted though relatively simple ITS, e.g., suggesting a pH test before progressing to corrosivity testing. The rich data available from six international validation studies,

[1] http://www.consort-statement.org/
[2] http://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment

eight retrospective assessments, and three recently completed validation studies of new tests (Adler et al., 2011; Zuang et al., 2008) make it an ideal test case for ITS development. For ocular toxicity, the OECD TG 405 in 2002 provided an ITS approach for eye irritation and corrosion (OECD 2002a,b). In spite of this TG, the Office of Pesticide Programs (OPPs) of the US EPA requested the development of an *in vitro* eye irritation strategy to register anti-microbial cleaning products. The Institute for In Vitro Sciences, in collaboration with industry partners, developed such an ITS of three *in vitro* approaches, which was then accepted by regulators (De Wever et al., 2012). ITS development has advanced greatly as a result of this test case (McNamee et al., 2009; Scott et al., 2009).

For skin irritation, we already referred to the work by Hoffmann et al. (2008a), which was based on an evaluation of the prevalence of this hazard among new chemicals (Hoffmann et al., 2005). The study showed the potential of simulations to guide ITS construction.

Embryonic Stem Cell test (EST) – an ITS?
The EST (Spielmann et al., 2006; Marx-Stoelting et al., 2009; Seiler and Spielmann, 2011) is an interesting test case for our definition of an ITS. It consists of two test systems (mouse embryonic stem cells and 3T3 fibroblasts) and two endpoints (cell differentiation into beating cardiomyocytes and cytotoxicity in both cell systems). The result (embryotoxicity), however, is only deduced from all this information. According to the suggested definition of tests and ITS, therefore, this represents a test and not an ITS. Note, however, that the EST formed a key element of the ITS developed at the end of the Integrated Project ReProTect (Hareng et al., 2005); a feasibility study showed the tremendous potential of this approach (Schenk et al., 2010).

Skin sensitization:
This area has been subject to intense work over the last decade, which resulted in about 20 test systems. As outlined in the roadmap process (Basketter et al., 2012), the area now requires the creation of an ITS. It seems that only the gridlock of the political decision process on the 2013 deadline, which includes skin sensitization as an endpoint, hinders the finalization of this important work. Since, at the same time, this represents a critical endpoint for REACH (notably all chemicals under REACH currently require a local lymph node assay for skin sensitization), such delays are hardly acceptable. It is very encouraging that BASF has pushed the area by already submitting their ITS (Mehling et al., 2012) for ECVAM evaluation. Pioneering work to develop a Bayesian ITS for this hazard was referred to earlier (Jaworska et al., 2011).

In silico ITS:
There also are attempts to combine only various in silico (QSAR) approaches. We have discussed some of the limitations of the *in silico* approaches in isolation earlier (Hartung and Hoffmann, 2009). Since they are referred to in REACH as "*non-testing methods*" they might actually be called "*Integrated Non-Testing Strategies*" (INTS). An example for bioaccumulation, already proposed to suit ITS (De Wolf et al., 2007; Ahlers et al., 2008), was reported recently (Fernández et al., 2012), showing improved prediction by combining several QSAR.

Consideration 6: Validation of ITS

Concepts for the validation of ITS are only now emerging. The ECVAM/EPAA workshop (Kinsner-Ovaskainen et al., 2009) noted only: "*There is a need to further discuss and to develop the ITS validation principles. A balance in the requirements for validation of the individual ITS components versus the requirements for the validation of a whole ITS should be considered*." Later in the text, the only statement made was: "*It was concluded that a formal validation should not be required, unless the strategy could serve as full replacement of an* in vivo *study used for regulatory purposes*." The workshop stated that for screening, hazard classification & labeling, and risk assessment neither a formal validation of the ITS components nor the entire ITS is required. We would respectfully disagree, as validation certainly is desirable for other uses, but it should be tailored to the use scenario and the available resources. The follow-up workshop (Kinsner-Ovaskainen et al., 2012) did not go much further with regard to recommendations for validation: "*Firstly, it was agreed that the validation of a partial replacement test method (for application as part of a testing strategy) should be differentiated from the validation of an* in vitro *test method for application as a stand-alone replacement. It was also agreed that any partial replacement test method should not be any less robust, reliable or mechanistically relevant than stand-alone replacement methods. However, an evaluation of predictive capacity (as defined by its accuracy when predicting the toxicological effects observed* in vivo*) of each of these test methods would not necessarily be as important when placed in a testing strategy, as long as the predictive capacity of the whole testing strategy could be demonstrated. This is especially the case for test methods for which the relevant prediction relates to the impact of the tested chemical on the biological pathway of interest (i.e. biological relevance). The extent to which (or indeed how) this biological relevance of test methods could, and should, be validated, if reference data (a 'gold standard') were not available, remained unclear.*

Consequently, a recommendation of the workshop was for ECVAM to consider how the current modular approach to validation could be pragmatically adapted for application to test methods, which are only used in the context of a testing strategy, with a view to making them acceptable for regulatory purposes.

Secondly, it was agreed that ITS allowing for flexible and ad hoc approaches cannot be validated, whereas the validation of clearly defined ITS would be feasible. However, even then, current formal validation procedures might not be applicable, due to practical limitations (including the number of chemicals needed, cost, time, etc).

Thirdly, concerning the added value of a formal validation of testing strategies, the views of the group members differed

strongly, and a variety of perspectives were discussed, clearly indicating the need for further informed debate. Consequently, the workshop recommended the use of EPAA as a forum for industry to share case studies demonstrating where, and how, in vitro *and/or integrated testing strategies have been successfully applied for safety decision-making purposes. Based on these case studies, a pragmatic way to evaluate the suitability of partial replacement test methods could be discussed, with a view to establishing conditions for regulatory acceptance and to reflect on the cost/benefit of formal validation, i.e. the confirmation of scientific validity of a strategy by a validation body and in line with generally accepted validation principles, as provided in OECD Guidance Document 34 (OECD, 2005).*

Finally, the group agreed that test method developers should be encouraged to develop and submit to ECVAM, not only tests designed as full replacements of animal methods, but also partial replacements in the context of a testing strategy."

Going somewhat further, De Wever et al. (2012) noted: "*In some cases, the assessment of predictive capacity of a single building block may not be as important, as long as the predictive capacity of the whole testing strategy is demonstrated. However, … the predictive capacity of each single element of an ITS and that of the ITS as a whole needs to be evaluated*."

Berg et al. go even further, challenging the validation need and suggesting a more hands-on approach to gain experience (Berg et al., 2011): "*Does it make sense to validate a strategy that builds upon tests for hazard identification which change over time, but is to be used for risk assessment? One needs to incorporate new thinking into risk assessment. Regulators are receptive to new technologies but concrete data are needed to support their use. Data documentation should be comprehensive, traceable and make it possible for other investigators to retrieve information as well as reliably repeat the studies in question regardless of whether the original work was performed to GLP standards*."

What is the problem? If we follow the traditional approach of correlating results, we need good coverage of each branch of the ITS with suitable reference substances to establish correct classification. Even for these very simple stand-alone tests, however, we are often limited by the low number of available well-characterized reference compounds and how much testing we can afford. However, such an approach would be valid only for static ITS anyway, and it would lose all the flexibility of exchanging building blocks. The opportunity lies in the earlier suggested "*mechanistic validation*." If we can agree that a certain building block covers a certain relevant mechanism, we might relax our validation requirements and also accept as equivalent another test covering the same mechanism. This does not blunt the need for reproducibility assessments, but a few pertinent toxicants relevant to humans should suffice to show that we at least identify the liabilities of the past. The second way forward is to stop making any test a "game-changer": If we accept that each and every test only changes probabilities of hazard, we can relax and fine-tune the weight added with each piece of evidence "on the job." It appears that such probabilistic hazard assessment also should, ideally, be compatible with probabilistic PBPK modeling and probabilistic exposure modeling (van der Voet and Slob, 2007; Hartung et al., 2012). This is the tremendous opportunity of probabilistic hazard and risk assessment (Thompson and Graham, 1996; Hartung et al., 2012).

Consideration 7: Challenges ahead

Regulatory acceptance:
A key recommendation from the ECVAM/EPAA workshop (Kinsner-Ovaskainen et al., 2009) was: "*It is necessary to initiate, as early as possible, a dialogue with regulators and to include them in the development of the principles for the construction and validation of ITS*." An earlier OECD workshop in 2008 (OECD, 2008) made some first steps and posed some of the most challenging questions addressing:

- how these tools and methods can be used in an integrated approach to fulfill the regulatory endpoint, independent of current legislative requirements;
- how the results gathered using these tools and methods can be transparently documented; and
- how the degree of confidence in using them can be communicated throughout the decision-making process.

With impressive crowd-sourcing of about 60 nominated experts and three case studies, a number of conclusions were reached:

- *"There is limited acceptability for use of structural alerts to identify effects. Acceptability can be improved by confirming the mode of action (e.g.,* in vitro *testing,* in vivo *information from an analogue or category).*
- *There is a higher acceptability for positive (Q)SAR results compared to negative (Q)SAR results (except for aquatic toxicity).*
- *The communication on how the decision to accept or reject a (Q)SAR result can be based on the applicability domain of a (Q)SAR model and/or the lack of transparency of the (Q)SAR model.*
- *The acceptability of a (Q)SAR result can be improved by confirming the mechanism/mode of action of a chemical and using a (Q)SAR model applicable for that specific mechanism/mode of action.*
- *Read-across from analogues can be used for priority setting, classification & labeling, and risk assessment.*
- *The combination of analogue information and (Q)SAR results for both target chemical and analogue can be used for classification & labeling and risk assessment for acute aquatic toxicity if the target chemical and the analogue share the same mode of action and if the target chemical and analogue are in the applicability domain of the QSAR.*
- *Confidence in read-across from a single analogue improves if it can be demonstrated that the analogue is likely to be more toxic than the target chemical or if it can be demonstrated that the target chemical and the analogue have similar metabolization pathways.*
- *Confidence in read-across improves if experimental data is available on structural analogues "bracketing" the target substance. The confidence is increased with an increased*

number of "good" analogues that provide concordant data.
- *Lower quality data on a target chemical can be used for classification & labeling and risk assessment if it confirms an overall trend over analogues and target.*
- *Confidence is reduced in cases where robust study summaries for analogues are incomplete or inadequate.*
- *It is difficult to judge analogues with missing functional groups compared to the target; good analogues have no functional group compared to the target and when choosing analogues, other information on similarity than functional groups is requested.*"

Taken together, these conclusions address more a WoE approach and the use of non-testing information than actual ITS. They still present important information on the comfort zone of regulators and how to handle such information for inclusion into ITS. Note that the questions of documentation and expressing confidence were not tackled.

Flexibility by determining the Most Valuable (next) Test:
A key problem is to break out of the rigid test guideline principles of the past. ITS must not be forced into a scheme with a yearlong debate of expert consensus and committees. Too often, technological changes to components, difficulties with availability and applicability of building blocks, and case-by-case adaptations for the given test sample will be necessary. For example, the integration of existing data, obviously at the beginning of an ITS, already creates a very different starting point. Chemico-physical, structural properties (including read-across or chemical category assignments) and prevalence also will change the probability of risk even before the first tests are applied. In order to maintain the desired flexibility in applying an ITS the MVT (most valuable test) to follow needs to be determined at each moment. Such an approach should have the following features:

1. Assess, finally, the probability of toxicity from the different test results.
2. Determine most valuable next test given from previous test results and other information.
3. Have a measure of model stability (e.g., confidence intervals) and robustness.

Assessing the probability of toxicity for given tests can be done by machine learning tools. Generative models work best for providing the values needed to find a most valuable test given prior tests. One simple generative model would predict probability of toxicity using a discriminative model (e.g., Random Forests (Breimann, 2001)), and test probability via a generative model (e.g., Naive Bayes). A classifier for determining risk of chemical toxicity must have the following traits:
- Outputs: unbiased and consistent probability estimates for toxicity (e.g., by cross-validation).
- Outputs: probability estimates even when missing certain results (both Random Forests and Naive Bayes can handle missing values).
- Reliable and stable results based on cross-validation measures.

The MVT identification based on previous tests is not a direct consequence of building a toxicity probability estimator. To find MVTs we need a generative model capable of determining test probabilities. One simple and effective way to determine the MVT is via the same method that decision trees use, i.e., an iterative process of determining which tests gives the most "information" on the endpoint. Information gain can be calculated given a generative model. To determine the test that gives the most information, we can find the test that yields the greatest reduction in Shannon entropy. This is basically a measure that quantifies information as a function of the probability of different values for a test and the impact those values have on the endpoint category (toxic vs. non-toxic). The mathematical formula is:

$$H(T) = -\sum_{i=0}^{n} p(T_i) * p(toxicity|T_i) * log(p(toxicity|T_i))$$

Where T is the test in question and *p*(Ti) signifies the probability of a test taking on one of its values (enumerated by *i*). To determine the most valuable test we need not only the toxicity classifier but also the probability estimates for every test as a function of all other tests. To determine these transition probabilities we need to discretize every test into the *n* buckets shown in the above equation.

We can expect that users applying this model would want to determine probabilities of toxicity for their test item within some risk threshold in the fewest number of test steps or minimizing the costs. When we start testing for toxicity we may want to check the current level of risk before deciding on more testing. For example, we might decide to stop testing if a test item has less than 10% chance of being toxic or a greater than 90% chance. Finding MVTs from a generative model has an advantage over directly using decision trees. Unfortunately, decision trees cannot handle sparse data effectively. The amount of data needed to determine n tests increases exponentially with the number of tests. By calculating MVTs on top of a generative model we can leverage a simple calculation from a complex model that is not as heavily constrained by data size.

Combining the ITS concept with Tox-21c:
As discussed above, Tox-21c relies on breaking risk assessment down into many components. These need to be put together again in a way that allows decision making, ultimately envisioned as systems toxicology by simulation (Hartung et al., 2012). Before this, an ITS-like integration and possibly a probabilistic condensation of evidence into a probability of risk are the logical approaches. However, there are special challenges: Most importantly, the technologies promoted by Tox-21c, at this stage mainly omics and high-throughput (Hattis, 2009), are very different from the information sources promoted in the European ITS discussion. We see how the ITS discussion is crossing the Atlantic, however – in the context of endocrine disruptor testing, for example (Willett et al., 2011). They are so data-rich that, from the beginning, a data-mining approach is necessary, which means that the weighing of evidence is left to the computer. Not all regulators are comfortable with this.

Our own research is approaching this for metabolomics (Bouhifd et al., in press); using endocrine disruption as a

test case might illustrate some of the challenges of the high-throughput, systems biology methods and omics technologies. Metabolomics – defined as measuring the concentration of "all" low molecular weight (<1500 Da) molecules in a system of interest – is the closest "omics" technology to the phenotype, and it represents the upstream consequences of whatever changes are observed in proteomic or transcriptomic studies. Small changes in the concentration of a protein, which might be undetectable at the level of transcriptomics or proteomics, can result in large changes in the concentrations of metabolites – changes that often are invisible at one level of analysis (i.e., co-factor regulation of an enzyme), are more likely to be apparent in a metabolic profile. By taking a global view, metabolomics provides clues to the systemic response to a challenge from a toxin and does so in a way that provides both mechanistic information and candidates for biomarkers (Griffin, 2006; Robertson et al., 2011). In other words, metabolomics offers both the possibility of seeing the high-level pattern of altered biological pathways while drilling down for relevant mechanistic details.

Metabolomics produces many of the same challenges as other high-content methods – namely, how to integrate the surfeit of data into a meaningful framework, but at the same time, it has some unique challenges. In particular, metabolomics lacks the large-scale, integrated databases that have been crucial to the analysis of transcriptomic and proteomic data. As was the case in the early years of microarrays, we are still without established methods to interpret data. Exploring data sets via several methods (Sugimoto et al., 2012) (ORA, QEA, correlation analysis, and genome-scale network reconstruction), hopefully, will provide some guidance for future toxicological applications for metabolomics and help us to better understand the puzzle as well as to develop and provide new perspectives on how to integrate several "-omics" technologies. At some level, metabolomics remains, at this stage, a process of hypothesis generation and, potentially, biomarker discovery, and as such will be dependent on validation by other means.

One critical problem for metabolomics is that while a more-or-less complete "parts list" and wiring diagrams exist for genomic and proteomic networks, knowledge of metabolic networks is still relatively incomplete. Currently, there are three non-tissue specific genome-scale human metabolic networks: Recon 1 (Rolfsson et al., 2011), the Edinburgh Human Metabolic Network (EHMN) (Ma et al., 2007), and HumanCyc (Romero et al., 2005). These reconstructions are "first drafts" – in addition to genes and proteins of unknown function, as well as "dead end" or "orphaned" metabolites that are not associated with specific anabolic or catabolic pathways. Furthermore, the networks are not tissue-specific. Many toxicants, including endocrine disruptors, exhibit tissue-specific toxicity, and a cell or tissue-specific metabolic network (Hao et al., 2012) should provide a more accurate model of pathology than a generic, global human metabolic network. In the long term, a well-characterized, biochemically complete network will help make the leap from pathway identification to a parameterized model than can be used for more complex simulations such as metabolic control analysis, flux analysis, and systems control theory to clarify the wiring diagram that allows the cell to maintain homeostasis and to determine where, within that wiring diagram, there are vulnerabilities.

Steering the new developments:
At this stage, no strategic planning and coordination for the challenge of ITS implementation exists. This was noticed in most of the meetings so far, e.g., (Berg et al., 2011): *"… there was a clear call from the audience for a credible leadership with the capacity to assure alignment of ongoing activities and initiation of concerted actions, e.g. a global human toxicology project*." The Human Toxicology Project Consortium[3] is one of the advocates for such steering (Seidle and Stephens, 2009). There is still quite a way to go (Hartung, 2009a). While we aim to establish some type of coordinating center in the US at Johns Hopkins (working title PoToMaC – Pathway of Toxicity Mapping Center), no such effort is yet in place in Europe. We suggested the creation of a *European Safety Sciences Institute* (ESSI) in our policy program, but this discussion is only starting. It is evident, however, that we need such structures for developing the new toxicological toolbox, along with a global collaboration of regulators of the different sectors, to finally revamp regulatory safety assessment.

References

Adler, S., Basketter, D., Creton, S., et al. (2011). Alternative (non-animal) methods for cosmetics testing: current status and future prospects – 2010. *Arch Toxicol 85*, 367-485.

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

Aldenberg, T. and Jaworska, J. S. (2010). Multiple test in silico weight-of-evidence for toxicological endpoints. *Issues Toxicol 7*, 558-583.

Anon. (2005). REACH and the need for Intelligent Testing Strategies [EUR 21554 EN]. Institute for Health and Consumer Protection, EC Joint Research Centre, Ispra, Italy. http://reach-support.com/download/Intelligent%20Testing.pdf

Anon. (2011). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. http://srdta.cochrane.org/handbook-dta-reviews

Balls, M., Amcoff, P., Bremer, S., et al. (2006). The principles of weight of evidence validation of test methods and testing strategies. The report and recommendations of ECVAM workshop 58. *Altern Lab Animal 34*, 603-620.

Basketter, D. A., Clewell, H., Kimber, I., et al. (2012). A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing – t[4] report. *ALTEX 29*, 3-91.

Benfenati, E., Gini, G., Hoffmann, S., and Luttik, R. (2010). Comparing in vivo, in vitro and in silico methods and inte-

[3] http://htpconsortium.wordpress.com

grated strategies for chemical assessment: problems and prospects. *Altern Lab Animal 38*, 153-166.

Berg, N., De Wever, B., Fuchs, H. W., et al. (2011). Toxicology in the 21st century – working our way towards a visionary reality. *Toxicol In Vitro 25*, 874-881.

Blaauboer, B. J. (2010). Biokinetic modeling and in vitro-in vivo extrapolations. *J Toxicol Environ Health B Crit Rev 13*, 242-252.

Blaauboer, B. and Barratt, M. (1999). The integrated use of alternative methods in toxicological risk evaluation. *Altern Lab Animal 27*, 229-237.

Blaauboer, B. J., Boekelheide, K., Clewell, H. J., et al. (2012). The use of biomarkers of toxicity for integrating in vitro hazard estimates into risk assessment for humans. *ALTEX 29*, 411-425.

Bouhifd, M., Hartung, T., Hogberg, H. T., et al. (in press). Review: Toxicometabolomics. *J Appl Toxicol*.

Breiman, L. (2001). Random Forests. Statistics Department, University of California. Machine Learning. http://oz.berkeley.edu/users/breiman/randomforest2001.pdf

Combes, R. D. and Balls, M. (2011). Integrated testing strategies for toxicity employing new and existing technologies. *Altern Lab Animal 39*, 213-225.

Dale, H. H. (1929). Croonian lectures on some chemical factors in the control of the circulation. *Lancet 213*, 1285-1290.

De Wever, B., Fuchs, H. W., Gaca, M., et al. (2012). Implementation challenges for designing integrated in vitro testing strategies (ITS) aiming at reducing and replacing animal experimentation. *Toxicol 26*, 526-534.

De Wolf, W., Comber, M., and Douben, P. (2007). Animal use replacement, reduction, and refinement: Development of an integrated testing strategy for bioconcentration of chemicals in fish. *Integr Environ Assess Manag 3*, 3-17.

Deeks, J. J. (2001). Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *Brit Med J 323*, 157-162.

Devillé, W. L., Buntinx, F., Bouter, L. M., et al. (2002). Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol 2*, 9.

Dejongh, J., Forsby, A., Houston, J. B., et al. (1999). An Integrated Approach to the Prediction of Systemic Toxicity using Computer-based Biokinetic Models and Biological In vitro Test Methods: Overview of a Prevalidation Study Based on the ECITTS Project. *Toxicol In Vitro 13*, 549-554.

Fernández, A., Lombardo, A., Rallo, R., et al. (2012). Quantitative consensus of bioaccumulation models for integrated testing strategies. *Environ Intern 45*, 51-58.

Forsby, A. and Blaauboer, B. (2007). Integration of in vitro neurotoxicity data with biokinetic modelling for the estimation of in vivo neurotoxicity. *Hum Exp Toxicol 26*, 333-338.

Gabbert, S. and van Ierland, E. C. (2010). Cost-effectiveness analysis of chemical testing for decision-support: How to include animal welfare? *Hum Ecol Risk Assess 16*, 603-620.

Gabbert, S. and Benighaus, C. (2012). Quo vadis integrated testing strategies? Experiences and observations from the work floor. *J Risk Res 15*, 583-599.

Griffin, J. L. J. (2006). The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philosophical Transactions of the Royal Society B: Biological Sciences 361*, 147-161.

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.

Hao, T., Ma, H.-W., Zhao, X.-M., and Goryanin, I. (2012). The reconstruction and analysis of tissue specific human metabolic networks. *Molec BioSyst 8*, 663-670.

Hareng, L., Pellizzer, C., Bremer, S., et al. (2005). The Integrated Project ReProTect: A novel approach in reproductive toxicity hazard assessment. *Reprod Toxicol 20*, 441-452.

Hartung, T., Bremer, S., Casati, S., et al. (2004). A modular approach to the ECVAM principles on test validity. *Altern Lab Anim 32*, 467-472.

Hartung, T. (2007). Food for thought ... on validation. *ALTEX 24*, 67-73.

Hartung, T. (2008). Food for thought ... on alternative methods for cosmetics safety testing. *ALTEX 25*, 147-162.

Hartung, T. (2009a). A toxicology for the 21st century – mapping the road ahead. *Toxicol Sci 109*, 18-23.

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

Hartung, T. (2009c). Toxicology for the twenty-first century. *Nature 460*, 208-212.

Hartung, T. and Hoffmann, S. (2009). Food for thought ... on in silico methods in toxicology. *ALTEX 26*, 155-166.

Hartung, T. (2010a). Comparative analysis of the revised Directive 2010/63/EU for the protection of laboratory animals with its predecessor 86/609/EEC – a t^4 report. *ALTEX 27*, 285-303.

Hartung, T. (2010b). Evidence-based toxicology – the toolbox of validation for the 21st century? *ALTEX 27*, 253-263.

Hartung, T. (2010c). Food for thought ... on alternative methods for chemical safety testing. *ALTEX 27*, 3-14.

Hartung, T. (2010d). Lessons learned from alternative methods and their validation for a new toxicology in the 21st century. *J Toxicol Environ Health B Crit Rev 13*, 277-290.

Hartung, T. and McBride, M. (2011). Food for thought ... on mapping the human toxome. *ALTEX 28*, 83-93.

Hartung, T. and Zurlo, J. (2012). Alternative approaches for medical countermeasures to biological and chemical terrorism and warfare. *ALTEX 29*, 251-260.

Hartung, T., van Vliet, E., Jaworska, J., et al. (2012). Food for thought ... systems toxicology. *ALTEX 29*, 119-128.

Hattis, D. (2009). High-throughput testing – the NRC vision, the challenge of modeling dynamic changes in biological systems, and the reality of low-throughput environmental health decision making. *Risk Anal 29*, 483-484.

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

Hill, A. B. (1965). The environment and disease: association or causation? *Proc R Soc Med 58*, 295-300.

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., Cole, T., and Hartung, T. (2005). Skin irritation: prevalence, variability, and regulatory classification of existing in vivo data from industrial chemicals. *Regulat Toxicol Pharmacol 41*, 159-166.

Hoffmann, S. and Hartung, T. (2006). Toward an evidence-based toxicology. *Hum Exp Toxicol 25*, 497-513.

Hoffmann, S., Saliner, A. G., Patlewicz, G., et al. (2008a). 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. (2008b). Points of reference in the validation process: the report and recommendations of ECVAM Workshop 66. *Altern Lab Anim 36*, 343-352.

Jaworska, J. and Hoffmann, S. (2010). Integrated Testing Strategy (ITS) – Opportunities to better use existing data and guide future testing in toxicology. *ALTEX 27*, 231-242.

Jaworska, J., Gabbert, S., and 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, P. S., et al. (2011). Integrating non-animal test information into an adaptive testing strategy – skin sensitization proof of concept case. *ALTEX 28*, 211-225.

Kinsner-Ovaskainen, A., Akkan, Z., Casati, S., et al. (2009). Overcoming barriers to validation of non-animal partial replacement methods/Integrated Testing Strategies: the report of an EPAA-ECVAM workshop. *Altern Lab Anim 37*, 437-444.

Kinsner-Ovaskainen, A., Maxwell, G., Kreysa, J., et al. (2012). Report of the EPAA-ECVAM workshop on the validation of Integrated Testing Strategies (ITS). *Altern Lab Anim 40*, 175-181.

Kirkland, D., Aardema, M., Henderson, L., et al. (2005). Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. *Mutat Res 584*, 1-256.

Leeflang, M., Deeks, J. J., and Gatsonis, C. (2008). Systematic reviews of diagnostic test accuracy. *Annals Internal Med 49*, 889-897.

Leist, M., Lidbury, B. A., Yang, C., et al. (2012). Novel technologies and an overall strategy to allow hazard assessment and risk prediction of chemicals, cosmetics, and drugs with animal-free methods. *ALTEX 29*, 373-388.

Ma, H., Sorokin, A., Mazein, A., et al. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol 3*, 135.

Marshall, M. (2012). How to prove cause leads to effect. *New Scientist of 29 Sep 2012*, 8-9.

Marx-Stoelting, P., Adriaens, E., Ahr, H.-J., et al. (2009). A review of the implementation of the embryonic stem cell test (EST). The report and recommendations of an ECVAM/ReProTect Workshop. *Altern Lab Anim 37*, 313-328.

McNamee, P., Hibatallah, J., Costabel-Farkas, M., et al. (2009). A tiered approach to the use of alternatives to animal testing for the safety assessment of cosmetics: Eye irritation. *Regulat Toxicol Pharmacol 54*, 197-209.

Mehling, A., Eriksson, T., Eltze, T., et al. (2012). Non-animal test methods for predicting skin sensitization potentials. *Arch Toxicol. 86*, 1273-1295.

Nordberg, A., Ruden, C., and Hansson, E. (2008). Towards more efficient testing strategies – Analyzing the efficiency of toxicity data requirements in relation to the criteria for classification and labelling. *Regulat Toxicol Pharmacol 50*, 412-419.

OECD (2002a). TG 404 ''Acute dermal irritation/corrosion'', adopted April 24th 2002, including a Supplement to TG 404 entitled. A sequential testing strategy for dermal irritation and, corrosion. (pp. 11-14). http://www.oecd-ilibrary.org/environment/test-no-404-acute-dermal-irritation-corrosion_9789264070622-en

OECD (2002b). TG 405 ''Acute eye irritation/corrosion'', adopted April 24th 2002, including a Supplement to TG 405 entitled. A sequential testing strategy for eye irritation and, corrosion. (pp. 9-13). http://www.oecd-ilibrary.org/environment/test-no-405-acute-eye-irritation-corrosion_9789264185333-en

OECD (2005). Series on Testing and Assessment No. 34. Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment. http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2005)14&doclanguage=en

OECD (2008). Series on Testing and Assessment No. 88. Workshop on integrated approaches to testing and assessment. http://www.oecd.org/chemicalsafety/testingofchemicals/40705314.pdf

Romero, P., Wagg, J., Green, M. L., et al. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol 6*, R2. doi: 10.1186/gb-2004-6-1-r2

Robertson, D. G., Watkins, P. B., and Reily, M. D. (2011). Metabolomics in toxicology: preclinical and clinical applications. *Toxicol Sci 120, Suppl 1*, S146-170.

Rockel, C. and Hartung, T. (2012). Systematic review of membrane components of Gram-positive bacteria responsible as pyrogens for inducing human monocyte/macrophage cytokine release. *Front Pharmacol 3*, 56.

Romero, P., Wagg, J., Green, M. L., et al. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol 6*, R2.

Rolfsson, O., Palsson, B. Ø., and Thiele, I. (2011). The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC Syst Biol 5*, 155.

Rusyn, I., Sedykh, A., Low, Y., et al. (2012). Predictive modeling of chemical hazard by integrating numerical descriptors of chemical structures and short-term toxicity assay data. *Toxicol Sci 127*, 1-9.

Schenk, B., Weimer, M., Bremer, S., et al. (2010). The ReProTect feasibility study, a novel comprehensive in vitro approach to detect reproductive toxicants. *Reprod Toxicol 30*, 200-218.

Scott, L., Eskes, C., Hoffmann, S., et al. (2009). 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.

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). "ToxR-tool", a new tool to assess the reliability of toxicological data. *Toxicol Lett 189*, 138-144.

Schünemann, H. J., Oxman, A. D., Brozek, J., et al. (2008). Rating Quality of Evidence and Strength of Recommendations: GRADE: Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Brit Med J 336*, 1106-1110.

Seidle, T. and Stephens, M. L. (2009). Bringing toxicology into the 21st century: A global call to action. *Toxicol In Vitro 23*, 1576-1579.

Seiler, A. E. M., and Spielmann, H. (2011). The validated embryonic stem cell test to predict embryotoxicity in vitro. *Nature Protoc 6*, 961-978.

Spielmann, H., Seiler, A., Bremer, S., et al. (2006). The practical application of three validated in vitro embryotoxicity tests. The report and recommendations of an ECVAM/ZEBET workshop (ECVAM workshop 57). *Altern Lab Anim 34*, 527-538.

Sugihara, G., May, R., Ye, H. et al. (2012). Detecting causality in complex ecosystems. *Science 338*, 496-500.

Sugimoto, M., Kawakami, M., Robert, M., et al. (2012). Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Curr Bioinform 7*, 96-108.

Thompson, K. M. and Graham, J. D. (1996). Going beyond the single number: Using probabilistic risk assessment to improve risk management. *Hum Ecolog Risk Assessm 2*, 1008-1034.

van der Voet, H. and Slob, W. (2007). Integration of probabilistic exposure assessment and probabilistic hazard characterization. *Risk Anal 27*, 351-371.

Van Leeuwen, C. J., Patlewicz, G. Y., and Worth, A. P. (2007). Intelligent Testing Strategies. In C. J. Van Leeuwen and T. G. Vermeire (eds.), *Risk Assessment of Chemicals. An Introduction*. 2nd edition (467-509). Dordrecht, The Netherlands: Springer.

van Vliet, E. (2011). Current standing and future prospects for the technologies proposed to transform toxicity testing in the 21st century. *ALTEX 28*, 17-44.

Vecchio, T. J. (1966). Predictive value of a single diagnostic test in unselected populations. *New Engl J Med 274*, 1171-1173.

Vonk, J. A., Benigni, R., Hewitt, M., et al. (2009). The use of mechanisms and modes of toxic action in integrated testing strategies: the report and recommendations of a workshop held as part of the European Union OSIRIS Integrated Project. *Altern Lab Anim 37*, 557-571.

Willett, C. E., Bishop, P. L., and Sullivan, K. M. (2011). Application of an integrated testing strategy to the U.S. EPA endocrine disruptor screening program. *Toxicol Sci 123*, 15-25.

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.

Zuang, V., Eskes, C., and Griesinger, C. (2008). ECVAM key area topical toxicity: Update on activities. *AATEX 14, Spec Issue*, 523-528. http://altweb.jhsph.edu/wc6/paper523.pdf

Acknowledgements

The support by NIH transformative research grant "Mapping the Human Toxome by Systems Toxicology" (RO1ES020750) and FDA grant "DNTox-21c Identification of pathways of developmental neurotoxicity for high throughput testing by metabolomics" (U01FD004230) is gratefully appreciated.

Correspondence to

Thomas Hartung, MD PhD
Center for Alternatives to Animal Testing
Johns Hopkins Bloomberg School of Public Health
615 North Wolfe Street
W7032, Baltimore, MD 21205, USA
e-mail: thartung@jhsph.edu