t4 Workshop Report

Integrated Testing Strategies (ITS) for Safety Assessment

Costanza Rovida1, Nathalie Alépée2, Anne M. Api3, David A. Baskette4, Frédéric Y. Bois5, Francesca Caloni6, Emanuela Corsini7, Mardas Daneshian8, Chantra Eskes8, Janine Ezendam9, Horst Fuchs10, Patrick Hayden11, Christa Hegele-Hartung12, Sebastian Hoffmann13, Bruno Hubesch14, Miriam N. Jacobs15, Joanna Jaworska16, André Kleensang17, Jon Lalko1, Robert Landsiedel18, Frédéric Leforeux19, Thomas Luechtefeld19, Monica Locatelli21, Annette Mehling18, Andreas Natsch22, Jonathan W. Pitchford23, Donald Prater24, Pilar Prieto25, Andreas Schepky26, Gerrit Schuermann27,28, Lena Smirnova20, Colleen Toole29, Erwin van Vliet30, Dirk Weisensee31 and Thomas Hartung1,20

1CAAT Europe, University of Konstanz, Germany; 2L’Oréal R&I, Aulnay, France; 3Research Institute for Fragrance Materials, Inc., Woodcliff Lake, USA; 4DABMEB Consultancy Ltd. Sharnbrook, UK; 5INERIS, DRC/VIVA/METO, Verneuil en Halatte, France; 6Università degli Studi di Milano, Department of Health, Animal Science and Biomolecular Sciences (DISFEB), Milan, Italy; 7European Society for Toxicology In Vitro, La croix Saint Ouen, France; 8National Institute for Public Health and the Environment (RIVM), Centre for Health Protection, Bilthoven, The Netherlands; 9CellSystems GmbH, Trosdorf, Germany; 10MatTek Corp., Ashland, MA, USA; 11Bayer AG, West Haven, USA; 12seh consulting + services, Paderborn, Germany; 13Cefic LRI and EPAA, Brussels, Belgium; Hubesch Consult BVBA, Sint Pieters-Leeuw, Belgium; 14Scientific Committee and Emerging Risks Unit, European Food Safety Authority, Parma, Italy; current address Centre for Radiation, Chemical and Environmental Hazards, Public Health England, UK; 15Procter & Gamble, Modelling & Simulation Biological Systems, Brussels Innovation Center, Strombeek-Bever, Belgium; 16ILS/NICEATM, Research Triangle Park, NC, USA; 17BASF SE, Ludwigshafen, Germany; 18Laboratoire de Synthèse Organique, CNRS UMR 7652, Ecole Polytechnique, Palaiseau, France; 19Center for Alternatives to Animal Testing (CAAT), Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; 20REACH Mastery, Como, Italy; 21Givaudan Schweiz AG, Dübendorf, Switzerland; 22Departments of Biology and Mathematics, University of York, UK; 23Food and Drug Administration, European Bureau, Brussels, Belgium; 24EURL ECVAM, Systems Toxicology Unit, Institute for Health and Consumer Protection, European Commission, Joint Research Centre, Ispra, Italy; 25Beiersdorf AG, Hamburg, Germany; 26UFZ Department of Ecological Chemistry, Helmholtz Centre for Environmental Research, Leipzig, Germany; 27Institute for Organic Chemistry, Technical University Bergakademie Freiberg, Germany; 28CeeTox Inc, Kalamazoo, MI, USA; 29SeCAM Services & Consultation on Alternative Methods, Agno, Switzerland

Summary

Integrated testing strategies (ITS), as opposed to single definitive tests or fixed batteries of tests, are expected to efficiently combine different information sources in a quantifiable fashion to satisfy an information need, in this case for regulatory safety assessments. With increasing awareness of the limitations of each individual tool and the development of highly targeted tests and predictions, the need for combining pieces of evidence increases. The discussions that took place during this workshop, which brought together a group of experts coming from different related areas, illustrate the current state of the art of ITS, as well as promising developments and identifiable challenges. The case of skin sensitization was taken as an example to understand how possible ITS can be constructed, optimized and validated. This will require embracing and developing new concepts such as adverse outcome pathways (AOP), advanced statistical learning algorithms and machine learning, mechanistic validation and “Good ITS Practices”.

Keywords: in vitro methods, testing strategy, Tox21c, skin sensitization, computational toxicology
1 Introduction

Regulators from different agencies worldwide as well as the regulated scientific community in general are becoming increasingly aware of the limitations of the current safety testing paradigm. Animal-based high-dose testing in typically one stand-alone guideline test is not always relevant for human exposure scenarios. New in vitro and in silico approaches, however, are limited in nature and can usually only supplement the animal test or serve as one component of a range of information sources that need to be combined in a relevant, reliable and unbiased way (Leist et al., 2014). This is exactly the purpose of the concept of Integrated Testing Strategies (ITS), also known as Integrated Approaches to Testing and Assessment (IATA) at the OECD, with the aim to combine different pieces of information / tests in a more mathematically efficient and biologically informed way. For the development of such science-based and human-relevant approaches for safety assessment of chemicals, a decision-making process needs to be adopted and accepted for regulatory purposes. In order to better understand the mechanisms and factors involved, it is now well recognized that the future of chemical safety assessment must move away from animal tests towards a combination of complementary approaches (in vitro, ex vivo, in silico, in chemico), which address functional mechanistic endpoints tied to adverse outcomes of regulatory concern. In spite of this increasing shared awareness, the way toward this goal remains unclear. There are controversies surrounding the definition of ITS, extending to how they can be implemented, validated, and reach global regulatory acceptance. Results from the different tests need to be combined in an objective and transparent way (Kinsner-Ovaskainen et al., 2009, 2012; Hartung et al., 2013a).

The principle of ITS, as they are used in this context, is to incorporate multiple data from various information streams, derived from different test methods, test method batteries, tiered test schemes, modeling techniques such as (Q)SAR (Quantitative Structure Activity Relationship), kinetics, exposure and epidemiological data, HTS (High Throughput Screening) and computational toxicology, etc. into one decision-making process (Judson et al., 2013). In this framework, the role of ITS is crucial, but the way to achieve its aim is not straightforward. There are many challenges: to accommodate the flexibility ITS require, to quantify and respond to varying levels of uncertainty, to incorporate preexisting knowledge, to assess test method availability and reproducibility, to define applicability domains of ITS components or necessary accuracy, all with the requirements of standardization that are mandatory for regulatory applications.

For this reason, there is the need to develop transparent, objective and consistent ITS tools to support reliable hazard and risk assessments. These are the core conceptual ITS requirements formulated in Jaworska and Hoffmann (2010) and later reiterated by Hartung et al. (2013a).

Regarding ITS, there are still many open questions:
- What are the selection criteria for in vitro and in silico methods and the combination criteria of the methodologies for constructing ITS?
- Which statistical and/or mathematical tools are available for relevant integration of data from different sources?
- Can we adopt standards for statistical evaluation?
- How should the predictive performance of ITS be assessed and validated?
- How should the outcome of ITS be evaluated for regulatory use?

To answer these questions, a group of experts was convened, coming from many different areas of expertise and organisations, including regulatory, validation and governmental bodies (EFSA, EURLECVAM, US NICEATM, US FDA) and scientific associations (CEFIC, EPA, ESTIV) (Box 1). The present report represents the outcome of a three-day workshop sponsored and co-organized by CAAT, the Doerenkamp-Zbinden Foundation (DZF), BASF, the International Fragrance Association (IFRA), the Research Institute for Fragrance Materials (RIFM) and the European Society for Toxicology in vitro (ESTIV). This workshop was held in Ranco (Varese, Italy) on the 8th–10th of July 2013. The concepts and ideas presented here come from the individual participants and do not reflect the positions of the organizations that they are representing. The debates were on scientific

2 Background

According to the European Chemicals regulation, safety assessment of a substance is performed through the full characterisation of the risks related to its use and distribution, including physical hazards, toxicological and ecotoxicological properties. Those are combined with a detailed exposure assessment for the final definition of the risk management measures that must be implemented for a reasonable safe use, or restrictions, of the substance.

The idea of applying multiple testing strategies for hazard and safety assessment started more than twenty years ago (Basketter, 1994). The reasons why a single in vitro test may hardly replace a full in vivo test are explained by Natsch (2014), and may be summarized as follows:

1. In vivo processes usually involve a chain of events while response from one in vitro test often represent only a single or a few steps of this complex process. When the outcome of that chain of events is toxicity, they are referred to Adverse Outcome Pathways (AOP).
2. In vitro methods may represent not only a single event (a “key event” in the AOP nomenclature), but also a single mode of action. Different modes of action may cause the same toxicological effect.
3. Different classes of chemicals may require different test methods, e.g., because of the limited applicability domains of some in vitro tests.
4. ADME (Absorption, Distribution, Metabolism and Excretion) must be considered and are often not well represented by current regulatory in vitro models.

5. The outcome from an in vitro test may be limited to reflect only specific in vivo dose response range and more methods are then required to cover the full dose response range.

In spite of the common agreement that ITS hold enormous promise for effective assessment of toxicological properties of chemicals, little has been done to really apply ITS, in the sense of integration of both non-testing (QSAR, read across) and experimental assays. So far, there are very few tools to combine the methods in an objective way, and often the scientific knowledge of many toxicity mechanisms is still not available or at least not clear enough to apply a full testing strategy with the certainty that the endpoint is fully covered.

In recent years, many initiatives have taken place, and several papers have been published to foster the application of ITS.

The first legal implementation of the ITS concept was the approval of the REACH Regulation (Regulation EC 1907/2006). This ground-breaking legislation is the first time that the combined application of non-standard procedures for safety assessment was included in a legal text, as detailed in the Annex XI of that Regulation. Starting from Article 1, there is the explicit reference to the possibility of applying alternative methods to avoid new tests on animals. Annex XI of the same Regulation explains how non-standard procedures can be used, with explicit mention of the Weight of Evidence (WoE) approach, defined as the conclusion derived “from several independent sources of information leading to the assumption/conclusion that a substance has or has not a particular dangerous property, while the information from each single source alone is regarded insufficient to support this notion”. The concept was further extended in the series of guidelines that followed the REACH publication (http://echa.europa.eu/support/guidance) that explain how to use a testing strategy for each endpoint, e.g., how to use new tests for the definition of substance properties, and through the EU FP6 project OSIRIS that was devoted to developing ITS schemes for human and environmental endpoints (http://www.ufz.de/osiris). Within the scope of REACH, WoE can be defined as the organisation of existing information while the set up of ITS as the decision process that leads to performing new tests. This was due to the awareness that a vast number of chemicals have been registered within the REACH program, and that if each of them were tested in vivo then the costs in terms of animal lives and economic resources would be enormous (Hartung and Rovida, 2009; Rovida and Hartung, 2009). Mid 2014, about 12,500 individual substances have been registered within the scope of REACH and most of the data are available in a public database accessible from the ECHA website (http://echa.europa.eu/information-on-chemicals/registered-substances). This database represents a vast resource of structured information that may feed models as well as offering the possibility to check the outcome of various predictions. Methods by which registrants tried to apply non standard approaches are regularly reported by ECHA, the European Chemicals Agency (ECHA, 2014).

EURLECVAM is playing a significant role: Recently, the 3T3 Neutral Red Uptake Cytotoxicity Assay for Acute Oral Toxicity received a positive opinion by the EURLECVAM Scientific Advisory Committee (ESAC) for the identification of substances with an LD₅₀>2000 mg/kg with the caveat that, due to the limitations of the test method, results should always be used in combination with other information sources to build confidence on the negative assay results. EURLECVAM fully endorsed the ESAC opinion and further recommended the development of ITS aiming at full or at least partial identification of acute oral toxicity hazard according to the GHS (Categories 1 to 4) (http://hcp.jrc.ec.europa.eu/our_labs/eurlec-ecvam/eurlec-ecvam-recommendations/3t3-nru-recommendation). In the past, the area of skin and eye irritation yielded interesting results with proposed strategies that may fully replace in vivo testing (Scott et al., 2010). EURLECVAM organised a workshop on validation of ITS that gave the opportunity for a very interesting discussion among scientists (Kinsner-Ovaskainen et al., 2012). It was proposed that there is no need for formal validation of ITS for screening purposes, for risk assessment purposes and for hazard classification and labelling unless there is an intention to replace a test in use for regulatory purposes.

In response to both the new Regulation for cosmetics products (Regulation EC 1223/2009), which prohibits new tests on living vertebrate animals for cosmetic purposes, and to answer the specific request of consumers who are more and more responsive to animal welfare aspects, academia, cosmetics industries as well as chemical suppliers have been very active with large investments and research programs. Cosmetics Europe is now co-funding with the European Commission the cluster of project called SEURAT-1 (http://www.seurat-1.eu), which is investing significant resources to find how repeated dose toxicity studies can be replaced through alternative methods. This large project initiative combines five research projects, a central data and knowledge management project and a multidisciplinary coordination action team, trying to integrate advanced techniques in the area of stem cells, microfluidic bio-reactors, in silico modelling, etc.

The application of ITS for the assessment of skin sensitization potential of chemicals is an area of focus from EPAA (a voluntary collaboration between the European Commission and representatives from both industry and trade associations), with the aim of improving and implementing the 3Rs approach in a regulatory framework. EPAA organised two workshops to further elaborate how to apply testing strategies for skin sensitization (Basketter et al., 2013). During the ITS-focused EPAA workshop (26th September 2011, Basketter et al., 2012a), all participants agreed on the idea that ITS was necessary to improve safety assessment and not just a way to save animals. In that workshop it was also proposed to further promote the involvement of regulators, in order to expedite acceptance of the new approach. Regulatory involvement may also represent a stimulus for the industry in using new methods as soon as possible. In fact, the subsequent EPAA workshop was hosted by ECHA in Helsinki (4th February 2013), with almost 60 participants coming from industry, ECHA, EURLECVAM, OECD (Box 1) and many European Member State representatives (Basketter et al., 2013).

A quantitative WoE approach has been developed through the OSIRIS project with ITS schemes for skin sensitization, mutagenicity and carcinogenicity employing Bayesian networks (Buist et al., 2013; Rorije et al., 2013), keeping in mind that the
latter is one of several opportunities for handling situations of redundant and conflicting information. Beyond aquatic endpoints that are out of the scope of the present paper, OSIRIS project produced interesting results in the area of repeated-dose toxicity (Tluczkiwicz et al., 2013) and a paper reviews opportunities for predicting physicochemical properties in the regulatory context (Nendza et al., 2013).

The above-mentioned projects and activities provide examples of developments taking place mainly in the EU. In 2007, the National Research Council of the US National Academy of Sciences published the well-known report on Toxicity Testing in the 21st century (NRC, 2007), which explained why the classical approach to toxicology assessment was not adequate to cope with present-day needs. Traditional approaches have been found to be too time-consuming and expensive requiring also the sacrifice of many animals and, worst, most in vivo studies may not be relevant to what is really happening in humans. The revolution in this approach for toxicity testing is based on the study of the possible mechanisms of actions of the chemical substances on human cells and human gene targets. Following this concept, the EPA's ToxCast research program tests thousands of chemicals in a broad array of cellular, in vitro, biochemical, and in silico models, thereby biologically phenotyping a large number of substances via a huge number of endpoints (Tab. 1). Phase I of ToxCast tested predominantly food-use pesticides with a wealth of animal toxicity data in regulatory guideline studies, which were used to create computational models to predict endpoints such as developmental and reproductive toxicity (Martin et al., 2011; Kleinstreuer et al., 2011; Knudsen, 2012). Phase II and Phase III vastly expanded the chemical libraries to cover many untested compounds, providing the opportunity to validate these predictive signatures and prioritize environmental chemicals for potential hazard (http://epa.gov/nctc/toxcast/data.html).

The work programme of OECD is also relevant in this context, being an umbrella organisation for chemical testing harmonization representing 34 countries worldwide. OECD work on safety test guidelines and Good Laboratory Practices (GLP) is crucial and generally well accepted as a standard for regulatory purposes. The OECD Series on Testing and Assessment, No.168 (OECD No 168, 2012a,b) is particularly interesting as it states the possibility of approaching the assessment of an endpoint, in this case skin sensitization, by applying the concept of an AOP which represents the existing knowledge concerning the linkage between a molecular initiating event and an adverse outcome at the individual or population levels (Ankley et al., 2010). The OECD idea is that combinations of mechanistically based test methods within IATA are needed to be able to substitute the regulatory animal tests currently in use. Recently, OECD has published a new guidance document describing an IATA for skin corrosion and irritation (OECD No 203, 2014). This is the first well defined IATA that is being adopted by the OECD. Notably, OECD is also responsible for the so-called OECD QSAR Tool Box (http://www.qsartoolbox.org), which is a freely available software that helps risk assessors in identifying structural alerts and defining groups and similarities among chemicals. Even though not directly focused on ITS, this software represents an interesting tool that can assist users in building optimised ITS and the OECD is actively working to further develop such aspects for IATA, in the Hazard assessment and test guideline programmes.

### 3 Comparison of the European and American approaches

ITS are a very generic term, and there are different ways in which they may be constructed, e.g., with many tests/substances casting a wide net vs. a priori network construction and assay development. Their construction may differ also according to the starting point, i.e., beginning from the definition of a relevant endpoint (top down) or for screening a wide array of substances (bottom up). The EU (Baskette et al., 2012b) vs US approach (Kavlock et al., 2012) reflects such a difference: The driving force in the EU stands on new regulations that explicitly ask for application of in vitro tests before performing any new in vivo studies, with the ultimate example being the new regulation of cosmetic products and the complete ban of tests on living animals for cosmetics products and ingredients. In contrast, the US approach driven by Tox21c and the ToxCast project (Tab. 1) is one which evaluates many chemicals on a set of assays that should address the different Pathways of Toxicity (PoT) and Mode of Action (MoA) in humans. Testing is supposed to produce data that must be interpreted and then compared to a threshold determination for adversity. This approach is different than the European idea of first studying in detail the mode of action and then developing tests that mimic each step of the process.

A constructive way to compare and combine the two approaches may lie in testing a common set of chemicals, such as the OECD reference chemicals, developed for various test guidelines, and the ToxCast library. The ECHA database (http://echa.europa.eu/web/guest/information-on-chemicals/registered-substances) is particularly useful because it contains data about all registered substances within the scope of EU REACH program. Provided the data are well prepared and reliable, this will facilitate a better understanding of the applicability domain of the relevant assays.

### 4 Composition of ITS

The term ITS has been generally used when more than one test is applied in combination to characterize the toxicological effects of a substance. ITS were initially conceived with the idea of replacing in vivo tests, with the awareness that no single in vitro test can reproduce the complex interactions that occur in an intact organism. Some pioneering work was done by the ECVAM Integrated Testing Taskforce which presented ITS as a combination of toxicodynamic and toxicokinetic parameters (Blaauwboer et al., 1999). Even though that principle is still valid, nowadays the aim of ITS is definitely broader with the ambition that ITS must help elucidate the mechanism of action of chemical perturbations with respect to human health or the environment. Ideally, ITS will be the highest quality source of mechanistic information for the definition of safety assessment.
Such evolution in the concept of ITS explains why there is not a unique definition but there are many ways on how to combine tests to build a testing strategy, as explained in Figure 1. The two simplest forms of ITS can be defined as a battery of tests and tiered strategies, as explained by Jaworska and Hoffmann (2010). A battery of tests is executed in parallel and generally all tests are necessary for the definition of a specific property. The tiered strategies are intended when series of tests are applied in sequence, by following yes or no decisions, as is the example of eye irritation (Balls et al., 1999), or a chain of events, as in the case for skin sensitization (Maxwell et al., 2011; Van der Veen et al., 2014; Natsch, 2014). The most advanced Bayesian decision methods are based on such a tiered strategy but take into account the fact that test results are not perfect “yes/no” answers and within this information value there is some uncertainty; the overall uncertainty decreases as the weight of the results accumulate.

More generally, the ITS should be composed from building blocks with precise decision points that can halt the process when a pre-defined level of confidence is reached and should optimize the number of tests that are required accordingly. The WoE approach is usually based on existing data while ITS should prospectively address which assays need to be performed for the assessment of an endpoint or in general for the definition of the risk posed by the use of a substance. However, the ITS concept has some similarity to the WoE approach as it represents a way to combine different results to make a decision (Balls et al., 2006).

ITS are very context dependent and multiple solutions are likely to be available and desirable. Explicitly, an ITS should contain the following elements:

1. Information target identification;
2. Systematic exploration of knowledge;
3. Choice of relevant inputs;
4. Methodology to synthesise disparate evidence;
5. Methodology to guide testing.

Generally speaking, ITS should combine different building blocks which can also be based on non-test (in silico and read-across) and test methods and the final decision about the safety assessment of a substance should be based on information from more than one type of source.

All results acquired through the testing must have quantifiable confidence levels and associated uncertainties to enable the application of an objective probabilistic approach. A well-developed and widely used probabilistic modeling tool are the Bayesian Networks (BN), whose potential in ITS framework was introduced by Jaworska and Hoffmann (2010), although other techniques are also feasible (Jaworska et al., 2013). Classification algorithms, such as classification and regression trees (CART) and Random Forests, were recently used to construct testing strategies for acute oral toxicity testing (Kopp-Schneider et al., 2013; Prieto et al., 2013; Kinsner-Ovaskainen et al., 2013).

Though it is clear that not all steps need to be measured for safety assessment, setting of decision points is not easy and depends upon how much information is required to fulfill the specific needs. This is highly dependent on the purpose of the ITS, ranging from simple hazard identification and chemical prioritization to a sophisticated risk assessment with increasing numbers of tests and levels of complexity. For example, in some cases the measure of chemical reactivity, e.g., with the DPRA (Direct Peptide Reactivity Assay), can be regarded as sufficient as a yes/no answer for a simple hazard preliminary categorization of the substance (Gerberick et al., 2007). In the context of the EU FP6 project ACuteTox, the estimation of the oral LD50 dose from an effective concentration in vitro and the application of classification algorithms were used to predict official acute oral toxicity categories (Prieto et al., 2013).

Ultimately, any ITS should be designed to fit a specific purpose, by balancing the applicability domain of the tests, sufficient information, cost and experimental feasibility. Using acute oral toxicity as an example, Norlen and colleagues (2014) compared the cost-effectiveness of different approaches based on single methods (four in silico tools and one in vitro assay) and battery combinations of methods. In their paper they have nicely illustrated how to assess the cost-effectiveness of alternative methods and how to interpret the results.

Another difficulty that may arise, when a combination of tests is proposed for the prediction of an adverse effect, comes from the fact that the majority of in vitro tests are developed independently, as a “stand-alone” prediction model for hazard identification. Most of the method developers still have the hope that they might find the perfect in vitro method to fully replace the in vivo test, in spite of the fact that in reality usually a set of complementary in vitro tests are necessary to reflect complex endpoints.

Each assay that contributes to building the ITS has to be well characterized, whether it is based on the biochemical understanding of the MoA, cellular effects, or related to other aspects that for example may impact the bioavailability of the substance. Test method definitions should include a defined protocol, the precise scope of the final endpoint, knowledge on applicability domain and the variability of the measure. The proper test selection varies according to the context of the testing strategies. In some applications, e.g., for cosmetics, sensitivity, or reliable identifications of no toxicity is more important than specificity, but this can be different for other uses of the substance for example if there is relevant exposure.

Feasibility of in vitro testing may be limited; a variety of different tests may be necessary in order to cover the full chemical universe and chemical properties, for example, the applicability domain is limited in case of poorly water-soluble substances.

The definition of the predictive capacity is even more challenging: when many results are combined, each of them brings its own variability with an impact on the definition of the toxicity threshold.

Another fundamental aspect is the kinetics of the effect and the relationship between the concentrations tested in vitro, with respect to doses of the organism exposure. It is necessary to consider the absorption, distribution and the metabolism of the substance in the human organism, the metabolism and interaction among organs (Yoon et al., 2012; Jacobs et al., 2013). This
fundamental aspect is often underestimated. For example, there are accepted methods for the analysis of skin penetration, but few that consider the actual permanence of a substance in the epidermis, where the sensitization process starts (Basketter et al., 2007). In general, calculation related to the distribution in vivo of the substance and its metabolites should be considered. This aspect is under research as it is essential to underpin the so-called quantitative in vivo/in vitro extrapolation (QIVIVE) (Blaauboer et al., 2012; Bessens et al., 2014).

The way forward for ITS is outlined in Table 2. An innovative ITS should be flexible and preferably based on a probabilistic approach: assays should be selected to gain the best balance between the number of tests (i.e., effort, time and resources) and information that is obtained. All single inputs, in chemico, in silico and/or in vitro, must be combined in a way that acknowledges and statistically assesses their respective contributions. Economic considerations may have to be considered too, as usually methods that are more complex are also more expensive, requiring sophisticated equipment and technical expertise.

ITS must be adaptive, allowing straightforward omission or addition of new tests as they become available or when newly acquired knowledge yields a more effective combination. At any level, the reasoning for any selection must be transparent and objective, independent from the personal judgment of the operators.

The statistical tools used to interpret ITS outcomes must be objective and able to evaluate when the acquired knowledge has reached a sufficient level of confidence to fit the final purpose and stop the process, moving from a deterministic approach with a preconceived belief in the ability to perfectly foresee the effect, to a probabilistic prediction of the final outcome on human population. Ideally, in the future informatics tools may provide value of information analyses that identify the next tests to run to reach the final goal of optimizing costs, resources, and predictive accuracy. Even if the BN, for example, are a promising informatics tool for combining the different component of the ITS, this methodology needs formal approval, and novel network learning tools may be applied to free the process from personal judgement.

5 The example of skin sensitization

While a broad applicability of ITS is expected for many toxicological endpoints, skin sensitization was chosen here as a well-developed example and test case of how ITS can be applied in practice. The biological mechanism of skin sensitization is well known, the concepts for replacing animal testing for this hazard have been developed in a series of projects (Rovida et al., 2013; Van Loveren et al., 2008; Maxwell et al., 2011) and there are many in vitro methods that are either validated or at an advanced state of validation, as well as a validated in vivo method (LLNA) that may serve as reference (NIH, 1999).

The chains of events that trigger the sensitization response are now sufficiently understood (Fig. 2). For a chemical to induce skin sensitization, a number of events must take place and some of them are considered to be key events, essential for the adverse outcome. Physical and chemical properties of the substance are important as only low molecular weight (LMW) substances can overcome the skin barrier. As LMW substances are too small to cause an immunogenic reaction, chemical allergens must bind to extracellular and cellular skin proteins to form a complete antigen (hapten binding). Protein binding is considered as the Molecular Initiation Event (MIE) of the AOP. Following uptake of the complete antigen and in the context of danger signals primarily secreted by the activated keratinocytes, dendritic cells (DC) mature and migrate via the afferent lymph vessels to the regional lymph nodes. In the lymph node, mature DC (expressing cell surface markers such as CD80 or CD86) stimulate the activation of hapten-specific responsive T cells leading to the generation of Tc1 effector cells. Upon renewed contact with the same chemical, memory T cells are recruited to the site of contact. Interactions between T cells and antigen presenting cells can take place directly in the skin, thus initializing the inflammatory reaction (elicitation phase). This sequence of events is started by the hapten, while the substance itself may represent a pre/pro hapten, i.e., it can be either metabolized or modified by any physical-chemical agent on the skin.

This chain of events was taken by OECD to illustrate the concept of AOP (OECD No 168, 2012a,b), i.e., a description of existing knowledge concerning the linkage(s) between a molecular initiating event and an adverse outcome at the individual or population level (Fig. 3). A specific substance can be tested with different methods, each representing one or more steps in the AOP. Some of these assays are still at the early stage of development, while others are much more advanced in the validation process (Fig. 4).

For what skin sensitization regards, it may happen that existing information (in vitro, in vivo and/or human data), combined with in silico data (read-across, QSARs) can be sufficient to reach a decision for the intended aim, e.g., hazard identification. A possible step forward is a tiered approach beginning with analyzing chemical reactivity with the DPRA, followed by data from cell based assays, such as either the Keratinosens™ (Natsch et al., 2011) or the human Cell Line Activation Test (hCLAT, Sakaguchi et al., 2009). This may yield a consistent prediction and then be sufficient, or it may result in data conflict, requiring more testing to resolve the conflict and arrive at a final conclusion. Such a simple approach may answer the question of hazard identification, i.e., yes/no sensitizers, and continuing with other tests, by considering the migration of the dendritic cells or T cell reaction, may become necessary if better definition of the minimum threshold to trigger human reaction is required or to distinguish between skin and respiratory sensitizers (Rovida et al., 2013). Mathematical models have been proposed to quantitatively establish a relationship between the dose of sensitiser applied to the skin and the possible human adverse effect(Maxwell et al., 2014).

on ALternative Methods) protocol was published by ECVAM (http://ecvam-dbaln.jrc.ec.europa.eu/view_doc.cfm?id=1558&doc=prot). The performance of each method is measured in isolation, with its own reproducibility, reliability and relevance assessment, to ensure that it will be sufficiently robust for Test Guideline development. Accuracy has been calculated for the set of chemicals tested as a preliminary evaluation only. Recent OECD documents (OECD No 168, 2012a,b) explain which methods will enter the strategy, and work is being finalized for an ITS for skin sensitization. Because the OECD membership comprised of many different regulatory frameworks, precise instructions as to how to quantitatively integrate the results may need flexibility, depending upon the regulatory jurisdiction.

It has been demonstrated (Bauch et al., 2012; Natsch et al., 2013; Van der Veen et al., 2014) that the proper and appropriate combination of those methods may increase the final predictivity with respect to skin sensitization hazard identification, which may even exceed the accuracy of the LLNA. Also potency estimates for skin sensitization from ITS have been presented (Jaworska et al., 2013). This was possible because there are many chemicals whose sensitization properties are well defined. In future this can be even further expanded by considering the correspondent human response (Basketter et al., 2014). While highlighting the importance of defining a set of reference substances for a proper ITS development, this fact also highlights the difficulty in compiling a proper list of reference substances for other endpoints, such as reproductive toxicity, (developmental) neurotoxicity (Smirnova et al., 2014) and so on, when only in vivo animal data are available.

6 Validation of ITS

Validation of alternative methods usually refers to the modular approach introduced by Hartung et al. (2004). Traditionally, the validation procedure is based on the evaluation of a single method in comparison to the traditional method, which is considered as a reference standard. Validation of in vitro methods is now undertaken by EURL ECVAM in the EU and by ICCVAM (Box 1) in the US, with important contributions from JaCVAM (Box 1) in Japan and other OECD member countries, under the auspices of the OECD itself.

Establishing precise rules for ITS validation may be complex, as the ITS itself is a dynamic process that cannot be defined by strict rules and may in fact be designed to provide an evaluation of the mechanistic validity of the approach. Nevertheless, the assessment of any ITS must be rigorous, objective and transparent, with scientific validity, which must fit the purpose. This procedure, even though not considered a formal alternative method validation, must be endorsed by an official organization to facilitate regulatory acceptance and recognition by governmental institutions.

For that purpose, some specific rules need to be defined. First, each test of the strategy must be standardized, with assessment regarding reproducibility and transferability of the protocol with a clear definition of the applicability domain and the uncertainty of the measurement. Moreover, the type of information that each method delivers should be clear, whether it is to demonstrate a mechanistic action, a physical-chemical property, or to elucidate a PoT or AOP. As ITS are inherently evolutionary (i.e., undergoing constant refinement over time), the scientific validity of the approach needs to be ensured and, where applicable, with a regular peer-review process.

The problem arises when the final performance of an ITS has no defined parameters and references for comparison, as it is the case for complex endpoints such as neurotoxicity or reproductive toxicity. In those cases, a step-by-step procedure may represent the solution. The precise definition of the endpoint and the mechanistic drivers is the first step, followed by a proposed description of the AOP. The knowledge of the in vivo mechanism is essential even though it is not necessary that each and every step of the in vivo process is represented by a test. The opposite is true, i.e., the mechanistic relevance of the specific assay must be defined together with the demonstration that it fits for the purpose in the strategy, which should be clearly defined, whether it is for hazard, for potency or for the identification of a single PoT.

Validation is based on the reproducible and accurate response of the strategy when challenged with a set of chemicals with well-known characteristics and therefore the reference chemical selection is essential for a successful process. Those chemicals must have a broad array of structural characteristics, in terms of physical and chemical properties, e.g., octanol/water partition coefficient (Kow), mechanism / mode of action (if known) plus toxicological behaviors that should include the whole range of activity with both positive and negative controls. The compilation of a set of reference standards can be useful as exemplified for sensitization (Casati et al., 2009; Kolle et al., 2013) as the starting point for the development of new methods or for the comparison of two equivalent methods. Extensive validation will still require a larger number of chemicals. Active consultations with statisticians is important to decide on the proper number of chemicals and datasets needed to have statistically sufficient power, considering prevalence but also the purpose of the strategy (Kopp-Schneider et al., 2013). At this moment this is probably the main constraint, because the list of chemicals with known toxicological profiles really relevant to humans is very limited. Most of the time, only results from animal studies are available, and the relevance in translating these studies to human biology is generally questionable.

The applicability domain of the ITS must also receive special attention, being derived from the overlap of the applicability domains of each single assay composing the strategy. Foreseeing exchangeable building blocks may also help enlarging the applicability domain of the whole strategy, for example by simply including different assays for either water or non-water soluble materials.

The definition of the predictive capacity is definitely more challenging for several reasons. In particular, the ITS aim to give an answer as to the final effect rather than simply attempting to reproduce the performance of an animal test. However, the desired predictivity should be defined according to the final use (e.g., cosmetic or industrial chemicals, drugs, pesticides, etc.) and
the effect that it is intended to predict, whether it is for hazard identification, classification and labelling, potency or even the capacity of excluding a specific risk.

Another difficulty lies in the mathematical combination of tests that return many different types of results, ranging from binary outcomes to multi-dimensional continuous results, and ultimately to the explanation of a mechanism. The analysis of many chemicals with known behavior is fundamental but not always possible. While there are many data on sensitization, the data are sparse for other endpoints such as reproductive toxicity or neurotoxicity. In this sense, something like the validation of the mechanism is preferable to a validation based on the final results obtained with single chemicals (Hartung et al., 2013b).

Finally yet importantly, the evaluation of the robustness of the strategy is essential. This can be achieved through the requirement that each test must meet specific minimum intra and inter laboratory reproducibility, which need to be known and integrated in the strategy. A combination of the building blocks may influecethe overall robustness of the strategy, considering that the uncertainty of the measure may propagate when more methods are included. The final uncertainty is not only a matter of propagation of the error of each test; more careful and comprehensive sensitivity analyses are needed. Each assay may have a unique and variable impact on the final outcome and provide information that may range from dichotomous to continuous. For this reason, a straightforward approach is preferable considering that unambiguous definitions may clear the way for regulatory acceptance.

A good tool to guide the process is provided by the principles of evidence based toxicology (EBT) with calls for transparency, objectivity and consistency of approaches (Hoffmann and Hartung, 2006; Guzelian et al., 2005).

6 Discussion

In spite of the awareness of the scientific community of the need for ITS as a tool for safety assessment, its full applicability is still constrained by several factors including lack of agreement on the approach, testing methods that are not developed for the purpose and the lack of validation procedure for regulatory acceptance. To overcome these difficulties, some precise pre-requisites must be defined.

First of all, each single component of the ITS has to be defined. The purpose of each test must be clear, the protocol precise, and its robustness and variability must be carefully quantified. Even if each of those steps has to be adapted to the specific test, the general principle of scientific evidence has to be preserved.

Those pre-requisites can be immediately applied when rigid tiered strategies are used by combining validated test methods, as it is the example for eye irritation (EPA, 2009), acute oral toxicity (Prieto et al., 2013) or more recently with developmental toxicity (Sogorb et al., 2014) and reproductive toxicity (Piersma et al., 2013), even though these are still at a proposal stage. Despite the large investment of time and resources toward providing a straightforward and transparent approach, even the case of skin sensitization is complex, with the ambition of combining tests for bioavailability, mechanistic elucidation, reactivity, and potency.

Beyond assessing the validity of each single test, the whole procedure must have precise and objective characteristics for regulatory acceptability. It can be imagined that in the future, ITS will probably define the overall safety assessment of a substance, rather than the single effect, such as skin sensitization, acute systemic toxicity, developmental toxicity and so on.

From the perspective of ideal ITS as shown in Figure 5, the final results must be independent from any subjective interpretation. This can be achieved by setting minimum requirements for an ITS tool. However, the necessity for precision, objectivity and reproducibility of the ITS should not impair the necessary characteristic of ITS adaptability. Adaptability of ITS includes the possibility of introducing new tests whenever available and to adapt the ITS to the specific purpose for which it is applied. In fact, the end of the process is defined by the purpose and this can range from mechanistic elucidation to hazard identification, to a full definition of complex endpoints for specific regulatory purposes.

Learning from the example of QSAR in regulatory use, the principle for validation was defined by OECD through 5 principles (http://www.oecd.org/env/ehs/risk-assessment/37849783.pdf):

1) a defined endpoint
2) an unambiguous algorithm
3) a defined domain of applicability
4) appropriate measures of goodness-of-fit, robustness and predictivity
5) a mechanistic interpretation, if possible

The same principle could be translated to ITS validity, with some adjustments of the concepts. It should be remembered that a QSAR approach may be one of the component of the ITS and as such it must follow the principles defined by OECD.

The text below tries to develop conceptual requirements for ITS, some of them already formulated in Jaworska and Hoffmann (2010):

1) A defined endpoint
The intent of a defined endpoint in the scope of QSAR should be translated as clearly defined purpose when referring to ITS, with the awareness that more than one route is possible to get to the same result. Regarding ITS, the concept of a defined purpose is more relevant than the endpoint. This is very important as any ITS may have different levels of complexity, in accordance with the intended outcome. The defined purpose is necessary for the decision of when the ITS process can stop.
2) An unambiguous algorithm
   The algorithm that is applied to compare the different components of the ITS must be transparent and reproducible with defined tracking for any changes. The informatics and statistical tools that are applied to get to a final decision must be clearly identifiable through proper and accessible documentation. Compared to QSAR, the algorithm for ITS has another dimension of complexity as it requires proper adaptability to incorporate results from new tests, with minimal delay in adopting new approaches. The respective software programs should include transparent and documented decision rules.

3) A defined domain of applicability
   Each test of the ITS must have a defined applicability domain. The final result should always be within the limitation posed by each component. The concept of applicability domain should be enlarged to include the purpose of the ITS in addition to the simple decision whether a specific chemical belongs to the applicability domain of the set of tests.

4) Appropriate measures of goodness-of-fit, robustness and predictivity
   This requirement is common to any scientific experimental procedure. In the case of using existing data, possibly derived from non-standard methods, some sort of adjustment to combine the results should be applied. Currently, this is typically based on expert judgment that can evaluate the validity of the data based on the available information. As more and more analyses are performed according to strict quality criteria, this will allow more independent and objective evaluations.

5) A mechanistic interpretation, if possible
   Mechanistic interpretation is the key for accepting non standard methods that usually lack points of reference. The mechanistic interpretation is commonly linked to the experimental protocol that is supposed to elucidate the specific PoT rather than the whole ITS. ITS based on a fully defined sequence of an AOP are most promising, i.e., they best reflect step by step through singular in vitro/in silico methods the key events of the AOP.

If these five principles are applied, the quality of the final ITS will be unquestionable. However, it should be recognised that we are currently working in an intermediate situation, when there is the need of applying ITS even though in some cases they are still under development.

   Regarding ITS execution, a quality system should be immediately implemented in addition to the experimental work. The concepts presented here lay the framework for defining a set of procedures that constitute Good ITS Practices (GIP), similar to what has been done for GLP, Good Manufacturing Practices (GMP), Good Cell Culture Practices (GCCP), etc. The concept of GIP was not extensively discussed during the workshop and it will not be further expanded here.

7 Conclusions and recommendations
   The route toward ITS for hazard and risk assessment is still long and difficult with undefined conditions for success. In spite of that, some steps should be immediately implemented:

1. Establishment of an international task force
   It is evident that the scientific community is pursuing different approaches. Probably, the largest difference is between US, where the approach of Tox21c has mainly resulted in large data generation and data-mining aims to identify the most informative tests (ultimately for ITS), and EU where there is the idea to fully reproduce in vitro the steps that lead to an adverse outcome in vivo. As it is always the case, both approaches have advantages and drawbacks. The creation of a supranational task force may improve the dialogue to inform and benefit from each approach, by also exploiting the results obtained on both sides. It is very important to communicate while being open minded to accept improvements, wherever they come from. The proposal is to create a transatlantic task force that regularly meets to discuss the improvements and helps to implement the lesson learnt overseas.

2. Establishment of a general database
   There are already many databases in different fields and many large companies have their own databases, plus the ECHA database, which contains information for all REACH registered substances. Unfortunately, most databases have different formats and the data is not easily extracted; this represents a hindrance, but does not prevent their use within the framework of ITS. The new platform should permit the use of data while respecting confidentiality claims and preserving intellectual property rights, including datasets which are not fully publicly visible. The usefulness of such a database should be a fundamental step to encourage scientists to populate it with their own data. Scientific funding bodies might provide economic compensation for such additional efforts to the benefit of the scientific community.

3. Definition of performance standards for ITS
   Special attention should be given to the identification of reference substances that probably represent the main bottleneck of the acceptability of ITS. The fact is that validation of alternative methods is now performed using animal tests as the reference standard, but this procedure is not applicable to ITS that strive to demonstrate the real risk for the human health. This is very important, in particular in those areas where animal models are weak, such as developmental toxicity, systemic toxicity, etc. The definition of the correct number of chemicals that must populate the set of reference substances is an aspect to be considered with
the aid of bio-statisticians. The relevant structural diversity of the chemicals selected, needs to be carefully considered at both the biological and computational model level.

4. Mechanistic validation
At the moment, mechanistic validation is the only proposed scientific rigorous possibility to overcome the problem of traditional validation vs. non-validated and non-relevant animal models. Validation of ITS should move towards a mechanistic validation to demonstrate that the crucial mechanism of damaging human health as a consequence of the action of a xenobiotic is reflected. Mechanistic validation is the answer, even though details about how to apply it are not yet defined, representing another input for discussion among experts.

5. Set up Good ITS Practices (GIP)
For quality assurance, GIP should be the very first step by setting the rules for ITS creation and implementation, creating a common language that will tremendously improve data-sharing and acceptability. A team of experts should define a list of rules for GIP, following evidence-based principles. It can thereby become a scientific assessment paining regulatory acceptance. Part of GIP should be the establishment of a standardized method to record and track versions of the IT tools that are applied, by listing all variations and improvements.

6. Training and education
Training and education are always central to the adoption of new scientific approach. (Daneshian et al., 2011). In particular, ITS are an emerging topic where consolidated experience belongs to relatively few experts, while the majority of risk assessors are still anchored to the traditional approach. Teaching the principles of ITS should be widely disseminated in universities together with the organisation of practical training courses. At the moment there are many initiatives but they all rely upon the personal initiative of certain individuals. A more harmonised and widespread program should be established to achieve a common understanding.

As a conclusion of the workshop, it is clear that much work is needed to get to the goal of completely replacing animal testing with integrated and mechanistically based testing strategies. There was, however, agreement that the fundamental elements are in place. Toxicologists agree on the general focal principles and efforts needed to develop precise sets of guidelines and practical tools that enable applicability of ITS for safety assessment.

References


http://dx.doi.org/10.14573/altex.2011.4.341


http://dx.doi.org/10.1016/j.taap.2009.02.004


http://dx.doi.org/10.1038/4601080a


http://dx.doi.org/10.1191/09603727106ht648oa


Correspondence to
Costanza Rovida, PhD
CAAT Europe
University of Konstanz
Box 600
78457 Konstanz, Germany
Phone: +39 340 4008118
e-mail: costanza.rovida@uni-konstanz.de
A set of independent assays provides the same number of results. None of them is enough for decision, but all together may lead to the final decision on the endpoint. This approach is described in Annex XI of REACH and considered as acceptable from a regulatory point of view, even if none of the tests is performed in a standard way.

The final result is defined by the sum of the results from many tests that must be all performed. This approach implies that all methods share the same applicability domain. It is considered by many scientists as the only possibility for in vitro methods to replace in vivo.

According to the results of the first step, it is decided which is the following test to be performed. It is an open system as there is no precise combination of tests, which may be an in vitro method, a QSAR evaluation. Any of the tests can return either a number, or a mechanistic elucidation or a yes/no answer.

The three schemes are not always well separated. In some cases, not all tests from a battery (Fig. 1b) are necessary for the definition of the endpoint, while in some cases the path of a tiered strategy (Fig. 1c) is considered as mandatory, resembling more as a battery of tests. Stopping and decision making is triggered when a predefined knowledge level is reached.
Fig. 2: Scheme of the sequence of events that may induce skin sensitization potential
From Rovida et al. 2013
Fig. 3: Flow diagram of the pathways associated with skin sensitization
From OECD No 168a-b, 2012
Fig. 4: Adverse outcome pathway with potential non-animal methods for contact hypersensitivity
References provide explanations on methods name, principle and protocol. The methods listed here are nonexhaustive and are meant as examples only; other methods are available and may also be applicable.
Fig. 5: Scheme for an optimal ITS
Each step should be ruled by precise and objective decisions.

The end of the process is set by:
- Regulatory request
- Required accuracy
- Level of concern
- Purpose of use
- Cost

Some elements are not independent, as the purpose of use may have an impact on the level of concern, the required accuracy should not be independent from regulatory request (and cost) and so on.
Box 1: List of acronyms of organizations with corresponding websites

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
<th>website</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAAT-US</td>
<td>Center for Alternatives to Animal Testing</td>
<td>caat.jhsph.edu</td>
</tr>
<tr>
<td>CAAT-Europe</td>
<td>Center for Alternatives to Animal Testing – Europe</td>
<td>cms.uni-konstanz.de/leist/caat-europe</td>
</tr>
<tr>
<td>CEFIC</td>
<td>European Chemical Industry Council</td>
<td><a href="http://www.cefic.org">www.cefic.org</a></td>
</tr>
<tr>
<td>DZF</td>
<td>Doerenkamp-Zbinden Foundation</td>
<td><a href="http://www.doerenkamp.ch/en/?id=10">www.doerenkamp.ch/en/?id=10</a></td>
</tr>
<tr>
<td>ECHA</td>
<td>European Chemical Agency</td>
<td>echa.europa.eu</td>
</tr>
<tr>
<td>EFSA</td>
<td>European Food Safety Authority</td>
<td><a href="http://www.efsa.europa.eu">www.efsa.europa.eu</a></td>
</tr>
<tr>
<td>EPA</td>
<td>Environmental Protection Agency</td>
<td><a href="http://www.epa.gov">www.epa.gov</a></td>
</tr>
<tr>
<td>EPAA</td>
<td>European Partnership for Alternative Approaches to Animal Testing</td>
<td><a href="http://www.epaa.eu.com">www.epaa.eu.com</a></td>
</tr>
<tr>
<td>ESAC</td>
<td>EURL ECVAM Scientific Advisory Committee</td>
<td>see EURL ECVAM</td>
</tr>
<tr>
<td>ESTIV</td>
<td>European Society of Toxicology in vitro</td>
<td><a href="http://www.estiv.org">www.estiv.org</a></td>
</tr>
<tr>
<td>EURL ECVAM</td>
<td>European Union Reference Laboratory for alternatives to animal testing</td>
<td>ihcp.jrc.ec.europa.eu/our_labs/eurl-ecvam</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
<td><a href="http://www.fda.gov">www.fda.gov</a></td>
</tr>
<tr>
<td>ICCVAM</td>
<td>Interagency Coordinating Committee on the Validation of Alternative Methods</td>
<td>ntp.niehs.nih.gov</td>
</tr>
<tr>
<td>IFRA</td>
<td>International Fragrance Association</td>
<td><a href="http://www.ifraorg.org">www.ifraorg.org</a></td>
</tr>
<tr>
<td>NICEATM</td>
<td>National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods</td>
<td><a href="http://www.niehs.nih.gov/research/atniehs/dntp/assoc/niceatm/">www.niehs.nih.gov/research/atniehs/dntp/assoc/niceatm/</a></td>
</tr>
<tr>
<td>NIH</td>
<td>National Institute of Health</td>
<td><a href="http://www.nih.gov">www.nih.gov</a></td>
</tr>
<tr>
<td>OECD</td>
<td>Organisation for Economic Co-operation and Development</td>
<td><a href="http://www.oecd.org">www.oecd.org</a></td>
</tr>
<tr>
<td>RIFM</td>
<td>Research Institute for Fragrance Materials</td>
<td><a href="http://www.rifm.org">www.rifm.org</a></td>
</tr>
<tr>
<td>RIVM</td>
<td>National Institute for Public Health and the Environment</td>
<td><a href="http://www.rivm.nl">www.rivm.nl</a></td>
</tr>
</tbody>
</table>
### Table 1: List of work packages in the ToxCast and Tox21 HTS Projects

Further details at [http://www.epa.gov/ncct/toxcast/chemicals.html](http://www.epa.gov/ncct/toxcast/chemicals.html)

<table>
<thead>
<tr>
<th>Set</th>
<th>Chemicals</th>
<th>Assays</th>
<th>Endpoints</th>
<th>Completion</th>
<th>Available to Tox21 partners</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToxCast Phase I</td>
<td>293</td>
<td>~600</td>
<td>~1100</td>
<td>2011</td>
<td>03/2013</td>
</tr>
<tr>
<td>ToxCast Phase II</td>
<td>767</td>
<td>~600</td>
<td>~1100</td>
<td>03/2013</td>
<td>10/2013</td>
</tr>
<tr>
<td>ToxCast Phase IIIa</td>
<td>1001</td>
<td>~100</td>
<td>~100</td>
<td>Ongoing</td>
<td>2015?</td>
</tr>
<tr>
<td>E1K (endocrine)</td>
<td>880</td>
<td>~50</td>
<td>~120</td>
<td>03/2013</td>
<td>10/2013</td>
</tr>
<tr>
<td>Tox21</td>
<td>8,193</td>
<td>~25</td>
<td>~50</td>
<td>Ongoing</td>
<td>Ongoing</td>
</tr>
</tbody>
</table>
Tab. 2: Comparison between the possible composition of ITS strategy as it can be immediately applied and the innovative ITS that should be pursued for a more efficient safety assessment

<table>
<thead>
<tr>
<th>ITS for hazard characterization (REACH)</th>
<th>ITS for full safety assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prescribed</td>
<td>Flexible</td>
</tr>
<tr>
<td>Deterministic</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>Classification</td>
<td>Fit for purpose</td>
</tr>
</tbody>
</table>