
Marcel Leist  
*University of Konstanz*

Nina Hasiwa  
*University of Konstanz*

Costanza Rovida  
*University of Konstanz*

Mardas Daneshian  
*University of Konstanz*

David Basketter  
*DABMEB Consultancy Ltd*

*See next page for additional authors*

Follow this and additional works at: [https://animalstudiesrepository.org/acwp_arte](https://animalstudiesrepository.org/acwp_arte)

Part of the [Bioethics and Medical Ethics Commons, Laboratory and Basic Science Research Commons](https://animalstudiesrepository.org/acwp_arte), and the [Research Methods in Life Sciences Commons](https://animalstudiesrepository.org/acwp_arte)

**Recommended Citation**

Authors
Marcel Leist, Nina Hasiwa, Costanza Rovida, Mardas Daneshian, David Basketter, Ian Kimber, Harvey Clewell, Tilman Gocht, Alan Goldberg, Francois Busquet, Anna-Maria Rossi, Michael Schwarz, Martin Stephens, Rob Taalman, Thomas B. Knudsen, James McKim, Georgina Harris, David Pamies, and Thomas Hartung

This report is available at Animal Studies Repository: https://animalstudiesrepository.org/acwp_arte/107
t4 Workshop Report*


Marcel Leist 1,2, Nina Hasiwa 1, Costanza Rovida 1, Mardas Daneshian 1, David Basketter 3, Ian Kimber 4, Harvey Clewell 5, Tilman Gocht 6, Alan Goldberg 7, Francois Busquet 1, Anna-Maria Rossi 1, Michael Schwarz 6, Martin Stephens 7, Rob Taalman 8, Thomas B. Knudsen 9, James McKim 10, Georgina Harris 7, David Pamies 7 and Thomas Hartung 1,7

1Center for Alternatives to Animal Testing, CAAT-Europe, University of Konstanz, Konstanz, Germany; 2Doerenkamp-Zbinden Chair of in vitro Toxicology and Biomedicine, University of Konstanz, Konstanz, Germany; 3DABMEB Consultancy Ltd, Sharnbrook, UK; 4Faculty of Life Sciences, University of Manchester, Manchester, UK; 5The Hamner Institutes for Health Sciences, Research Triangle Park, NC, USA; 6Eberhardt Karl University, Tübingen, Germany; 7Center for Alternatives to Animal Testing, CAAT, Johns Hopkins University, Baltimore, US; 8Cosmetics Europe, Brussels, Belgium; 9US EPA, Research Triangle Park, NC, USA; 10CeeTox, Kalamazoo, MI, USA

Summary
Since March 2013, animal use for cosmetics testing for the European market has been banned. This requires a renewed view on risk assessment in this field. However, in other fields as well, traditional animal experimentation does not always satisfy requirements in safety testing, as the need for human-relevant information is ever increasing. A general strategy for animal-free test approaches was outlined by the US National Research Council’s vision document for Toxicity Testing in the 21st Century in 2007. It is now possible to provide a more defined roadmap on how to implement this vision for the four principal areas of systemic toxicity evaluation: repeat dose organ toxicity, carcinogenicity, reproductive toxicity and allergy induction (skin sensitization), as well as for the evaluation of toxicant metabolism (toxicokinetics) (Fig. 1). CAAT-Europe assembled experts from Europe, America and Asia to design a scientific roadmap for future risk assessment approaches and the outcome was then further discussed and refined in two consensus meetings with over 200 stakeholders. The key recommendations include: focusing on improving existing methods rather than favoring de novo design; combining hazard testing with toxicokinetics predictions; developing integrated test strategies; incorporating new high content endpoints to classical assays; evolving...
1 Introduction and background

The discussion leading to this summary report started with the 7th Amendment to the Cosmetics Directive 76/768/EEC, which called for a complete ban of testing on vertebrate animals for the toxicological characterization of cosmetics ingredients in 2013. The European Commission asked experts to evaluate the availability of alternative non-animal methods. Their conclusion that alternative methods would not be available during the next 10 years (Adler et al., 2011) met with some criticism (Taylor et al., 2011), but was mostly endorsed by an independent expert group invited by CAAT also including specialists from Japan and the USA (Hartung et al., 2011). They also noted that significant advances had been made in the time between the publication of the Adler Report and the evaluation by the international group of experts. The next step, i.e., assembling experts to create a perspective for the future, was initiated by CAAT-Europe in a series of commissioned white papers on sensitization, repeated dose organ toxicity, toxicokinetics, carcinogenicity and reproductive toxicity. Important, this work addressed a broad range of chemical testing, including also the fields of drugs, pesticides and industrial chemicals in addition to cosmetics ingredients. A workshop with 35 experts discussed these white papers. This activity resulted in the extensive report A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing (Basketter et al., 2012). To involve all potential stakeholders, this

Fig. 2 Timeline of events leading to this public expert consultation report

The 7th amendment of the European Cosmetics Directive required the phasing out of animal testing to be completed by 2013. The European Commission evaluated the availability of non-animal methods and the outcome was published (Adler et al., 2011). The conclusions of that report were confirmed by independent experts (Hartung et al., 2011). In order to outline a roadmap for further development of non-animal methods for addressing systemic toxicity, an expert consortium was convened in a series of CAAT-Europe workshops to elaborate on the issue. As a result, an extensive report on the roadmap for non-animal methods for systemic toxicity testing was published (Basketter et al., 2012) and presented for a public expert consultation in 2012 in Brussels and in 2013 in Washington at an FDA-hosted event. The present report summarizes the recommendations resulting from the public expert consultation in which over 200 experts from academia, industry and regulatory authorities were involved. Grey boxes refer to actions taken by CAAT/CAAT-Europe.

Keywords: safety testing, animal-free testing, systemic toxicity, adverse outcome pathways

roadmap was presented in March 2012 in Brussels\(^2\) by several experts in front of about 200 stakeholders from governmental organizations, academia, industry and NGOs from all over the world (Fig. 2). A second workshop, *Scientific roadmap for the future of animal-free systemic toxicity testing*, similar in size and scope, was organized in Washington at the FDA in May 2013 to give updates on the Basketter report and scientific advances in the fields\(^3\).

Each of the lectures on the five major fields still requiring better non-animal safety testing methods (Fig. 1) was followed by one hour of intensive discussion to consolidate or improve the suggested strategies. Here we report the final outcome. This roadmap is expected to pave the way for a new toxicology that can better predict the effect of chemicals on humans, using fewer or even no vertebrate animals.

### 2 Animal models

A successful strategy to replace animal testing must take its starting point from the analysis of the current technology. The weaknesses of animal testing could then be avoided by the new approach (Fig. 3). It must be noted that rational comparisons are made difficult by the fact that almost all of the currently used animal models have never been formally validated. The rationale of their use is therefore not based on scientific data (Hartung and Leist, 2008).

Some of the problems related to animal models derive from the high doses that are tested and the multiplicity of endpoints that are measured. Experts claim, that “*Half of all chemicals, whether natural or synthetic, are positive in high-dose rodent cancer tests. These results are unlikely to be relevant at the low doses of human exposure.*” (Ames and Gold, 2000). Many of the attendees expressed the opinion that widespread knowledge on the limited value of certain animal studies frequently contributes to the decision by authorities to waive testing.

The lack of predictivity of animal models is particularly apparent from the field of drug development (Leist and Hartung, 2013). Only 8\% of drugs entering clinical phase I (first human dose) gain approval by authorities and half of them fail in phase III\(^4\). US Health and Human Services Secretary Mike Leavitt commented that “*currently, nine out of ten experimental drugs fail in clinical studies because we cannot accurately predict how they will behave in people based on laboratory and animal studies.*”\(^5\)

The conclusions are easily drawn: We need human-predictive, rapid and economical methods to evaluate whether or not a compound, no matter if chemical, drug or cosmetic ingredient, is safe for intended human use.

### 3 Future safety science and pathways of toxicity

A key step in the paradigm shift in toxicology, as far as regulatory authorities are concerned, was the 2007 US National Research Council report *Toxicity Testing in the 21\textsuperscript{st} Century – a Vision and a Strategy* (NRC, 2007; Leist et al., 2008a). It promoted the idea that the number of ways that a chemical or drug could disturb a cell is finite and can therefore be identified by appropriate screening methods. Quantitative information on the concentration-dependence of such disturbances can be used to predict the overall network of cellular regulatory reactions.

#### Table: Problems with animal experiments

<table>
<thead>
<tr>
<th>Problems with animal experiments</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>High false negatives rate</td>
<td>Compromised safety</td>
</tr>
<tr>
<td>High false positives rate</td>
<td>Economically not acceptable</td>
</tr>
<tr>
<td>Lack of validation</td>
<td>Unclear scientific basis</td>
</tr>
<tr>
<td>Low predictivity and high costs</td>
<td>Waiving praxis</td>
</tr>
<tr>
<td>Frequent waiving</td>
<td>Compromised safety</td>
</tr>
<tr>
<td>High dose – low dose extrapolation</td>
<td>Uncertainty</td>
</tr>
<tr>
<td>Metabolism different from man</td>
<td>Uncertainty</td>
</tr>
<tr>
<td>Biology different from man</td>
<td>Uncertainty</td>
</tr>
</tbody>
</table>

---

\(^2\) The forum was co-organized by Center for Alternatives to Animal testing (CAAT), Cosmetics Europe, Doerenkamp-Zbinden Foundation, European Chemical Industry Council (CEFIC), European Consensus Platform for Alternatives (ECOPA), European Society for Alternatives to Animal Testing (EUSAAT), European Society for Toxicology in vitro (ESTIV), US EPA ToxCast\(^\text{TM}\), Humane Society International (HSI), Institute for In-vitro Sciences (IIVS) and In-vitro Testing Industrial Platform (IVTIP).

\(^3\) The forum was co-organized by Agilent Technologies, American Cleaning Institute (ACI), Animal Research & Development Foundation (ARDF), American Society for Cellular and Computational Toxicology (ASCCCT), Center for Alternatives to Animal testing (CAAT), CropLife Canada, CropLife America, FDA Center for Food Safety and Advanced Nutrition (CSFAN), Grocery Manufacturer Association (GMA), Human Toxicology Project Consortium, Humane Society of the United States (HSUS), Institute for In-vitro Sciences (IIVS), National Institute for Environmental Health Sciences (NIEHS), People for the ethical Treatment of Animals (PeTA) and The Hamner Institute.


\(^5\) [http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2006/ucm108576.htm](http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2006/ucm108576.htm)
New approaches to safety testing require new strategies to stringent but flexible evaluation of the suitability and performance of methods. Methods suggested by the Evidence-based Toxicology Collaboration (European and US branches, http://www.ebtox.com) will be helpful in this process (Stephens et al., 2013). The risk classification itself is also likely to undergo fundamental changes. At present, using a very limited number of animal tests, a chemical is classified as toxic or non-toxic (deterministic risk assessment) at a given exposure level. The much richer information provided by the new approaches and the progress of safety sciences could form a basis for probabilistic risk assessment (Paparella et al., 2013). This would allow for more rational and science-based regulatory decisions by assembling information from a tailored set of tests adapted to different types of questions and scenarios of exposure and risk assessment (Fig. 5).

4 General forward strategies

While discussing the five specific toxicological domains, the experts identified issues relevant to all areas of in vitro methods. Joint knowledge management and sharing of expertise between different sectors, stakeholders and application domains were identified as important drivers for accelerated progress, in addition to accelerated test establishment and validation, and better use of computational toxicology methods (Fig. 6).

Sharing of data would hopefully lead to the creation of a human safety database, bigger and more complete than existing ones (e.g. OPENTOX, EPA, TOXBANC, IMI activities). Most importantly, it would be more accessible and should be tailored for in vitro-in vivo comparisons as well as data mining by non-specialists in bioinformatics. The focus would not be on data collection as such, but on the accessory information linked to the primary outcome data as far as mechanisms of toxicity are concerned. As there are major hurdles (e.g., intellectual property rights and industrial competitiveness issues) to be overcome, it is clear that substantial incentives must be granted to encourage industry to share their proprietary data (Fig. 7).
None of the future challenges in the field of *in vitro* toxicology can be addressed by individual test systems. The solution will rather lie in the construction of batteries of tests to be combined in integrated testing strategies (ITS). This will need to be considered right from the conception of a test, throughout its development and especially during the evaluation of its performance. The latter evaluation needs to consider the test alone, but also in the context of the added value it brings to a test battery. More research and experimentation is required on how to build ITS. One example of a flexible, yet fully quantitative approach, is the Bayesian network (McDowel and Jaworska,

---

**Fig. 5: Vision of a smooth transition from current to future toxicology in safety science**

It is envisaged that the types of test systems employed will change over the time course of the establishment of a new safety science. At present, complex test systems that are specific for organ functions and developmental stages are preferentially used. Only few programs use simple assays of elementary biochemical and cellular function (e.g., ToxcCast™ Program). Over time, more and more critical biomarkers of toxicity may be identified by the application of HCS (high-content screening) and omics technologies to the complex systems, and simple test systems may suffice to measure key processes (Rossini and Hartung, 2012). Case studies, e.g., from PBPK (physiologically based pharmacokinetic modeling) and skin sensitization fields, could be used as learning models for the transition. The principles of evidence-based toxicology and the resulting quality control will lead to an accelerated method development and validation. Over time, the goal is to shift from the present deterministic risk classification to a probabilistic risk assessment.

---

**Fig. 6: General points to consider when moving forward towards animal-free approaches for systemic toxicity testing**

**Collaboration**
- Join forces: academia, industrial sectors, regulators
- Render valuable proprietary toxicological data accessible
- Generation of curated and searchable data base

**Test systems**
- Develop better *in vitro* methods
- Optimize existing test systems
- Develop strategies to establish, validate and use integrated testing strategy
- Abolition of useless *in vivo* tests

**Computational toxicology**
- Use of systems biology approaches
- Define pathways-of-toxicity and bring to practical application
- Reduce complexity by focusing on key events
- Use information-rich endpoints as biomarkers-of-toxicity (BoT)
- Broad use of quantitative *in vitro* – *in vivo* extrapolation (qVIVE)

---

**Fig. 7: Creation of a high quality database for relating *in vivo* and *in vitro* information**

The key players (pharmaceutical, chemical and cosmetics industry, basic research and regulators) share common goals that are of high value to them. The benefit of working together should outweigh disadvantages (opening of proprietary databases). This would allow the generation of a large, high quality database of *in vitro* toxicity data. It should be publicly available, include rich data that informs on the mode of action of compounds and allows for *in vitro* – *in vivo* correlations. It should also be quality controlled and suitable for case studies. This can be achieved by joint projects and the common use of legacy data from hitherto proprietary in-house databases. The collection of human data by micro-dosing, from clinical trials and from epidemiological studies plays a major role.
2002; Jaworska and Hoffmann, 2010; Jaworska et al., 2010, 2011), which has been applied successfully in the area of skin sensitization (Fig. 8). Input of the field of machine-learning is envisaged to be very important for optimal strategic designs of ITS (Hartung et al., 2013b).

5 Strategies to improve test systems

Alternative in vitro methods have been developed for all toxicological questions, including even the most complex fields, ranging from developmental neurotoxicity (DNT) to xenobiotic metabolism (Adler et al., 2011; Basketter et al., 2012; Leist et al., 2012a, 2008b; van Thriel et al., 2012; Smirnova et al., 2014; Taylor et al., 2011). Many of these methods should be formally validated for immediate use, or they could form the basis for accelerated further development. Optimization of existing systems is an important part of the strategy to accelerate the implementation of a mostly animal-free safety science, in addition to the more time demanding development of entirely new methods (Fig. 9). One specific way to improve available tests is the incorporation of highly information-rich endpoints provided by omics technologies. The biological information flow in a cell leads from gene sequences (the code) via RNA (the messages) to enzymes and other functional proteins (the tools). Within this infrastructure small molecule metabolites may be regarded as the goods that are produced and traded. They comprise energy substrates, building blocks and signaling messengers. As there are feedback loops between all levels, the different omics technologies address these four organization levels. The disturbance of a cell by chemicals may be measured by any single technique. Combinations of more than one approach lead to a better prediction of the true human situation.
6 Specific approaches for the five toxicological endpoints still lacking validated replacement methods

A detailed strategy has been elaborated for each endpoint and described in detail (Basketter et al., 2012). The consensus meeting of the roadmap initiative highlighted specific points for immediate attention and action:

6.1 Skin Sensitization

Although the sensitization process is a disease-free state, subsequent exposures can lead to allergic contact dermatitis, the most common adverse effect of chemicals on human health. One in 5 adults suffers contact allergy to one chemical or another (Peiser et al., 2012). This area differs from others as a large amount of human data is available. Moreover, the general mechanisms of sensitization are well-defined. There are three well-developed animal models (Buehler Guinea Pig Test, Guinea Pig Maximization Test (GPMt) and Mouse Local Lymph Node Assay (LLNA)) currently used to identify chemicals with toxic potential. The LLNA, which is already a step towards refinement and reduction of the use of animals, is the preferred method for safety assessment as it provides a quantitative value (the concentration of the chemical which causes a threshold positive response (EC3)) that can determine the potency of the sensitizer.

Already over a dozen different in vitro tests to identify sensitizers have been submitted to the European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM). Currently, two of these are validated for risk assessment. These are the direct peptide reactivity assay (DPRA), based on the chemical understanding and correlation with sensitization, and the human cell line activation test (h-CLAT), based on the activation of dendritic-like cells (Bauch et al., 2011; Sakaguchi et al., 2006; Ashikaga et al., 2010). The KeratinoSense™ luciferase-reporter gene model (based on the anti-oxidant response element in the HaCat keratinocyte cell line) (Natsch, 2010; Andreas et al., 2011) has already been validated by Givaudan and is accepted by OECD. Integrated testing strategies (ITS) will be based on the use of state-of-the-art in vitro and in silico technologies. Where classical methods measure only one, or few, endpoints (e.g., metabolites or gene expression levels), the new approaches can yield thousands of data points simultaneously, and provide information on a genome-wide scale (Fig. 10) and, thus, allow insights into the reaction of a network.

![Fig. 11: Example for the concept of adverse outcome pathways (AOPs) from the area of skin allergy](image_url)

The general scheme of an AOP is illustrated in the upper panel. The AOP provides a mechanistic link between a chemical structure and the response of the organism to the chemical. At increasing levels of complexity, the xenobiotic’s action is assumed to be started by a molecular initiating event, followed by cellular and organ responses that eventually explain the effect on the organism. The middle panel gives an example by depicting the events leading to skin sensitization. Understanding the underlying pathophysiology is necessary to create a set of in vitro models for all key events. The lower panel shows an example of a specific AOP for skin sensitization. Key event 1 corresponds to the molecular initiating event. Further key events are shown and each of them may be modeled in vitro. Combination of such in vitro tests in an integrated strategy (ITS) would allow comprehensive predictions for unknown xenobiotics.
the way forward, as each assay on its own has 80% accuracy, but if combined in an ITS, 90% accuracy can be reached (Bauch et al., 2012). This level of predictivity would perform better than the validated LLNA. Thus, an ITS would fully replace the existing animal models.

Skin sensitization is a field in which several formally validated methods and ITS are expected to emerge in the near future. The reasons for this are the following: first, for skin sensitization the validation process has clear anchors: this is the only toxicological domain that is based on a formally validated animal test model (LLNA). Moreover, a large set of human data on positive control compounds is available, e.g., from diagnostic patch testing in dermatology clinics. Second, the mechanisms of skin sensitization are well understood, and the individual steps are amenable to modeling. Third, several in vitro models that seek to mimic each single step in the pathway are already available, and they now need only to be combined in an ITS. CAAT organized a workshop on ITS using the example of skin sensitization in June 2013 in Ranco, Italy; the respective report is currently being completed.

The application of the OECD-promoted concept of “adverse outcome pathways” (AOPs) to skin sensitization is relatively straightforward. Virtually all key events of the AOP already are covered by in vitro assays (Fig. 11)6. Despite this favorable situation, validation of a complete ITS for skin sensitization will require further work. It is, for instance, not yet clear how the individual tests that cover the steps of the AOP will be combined, including how much weight is given to the results of each assay and how the decision points of tiered testing would be structured. The final prediction model must be built as a whole on the assembly of tests and on the ITS rules linking them. The process of building and optimizing this overall test strategy is made difficult by the fact that the LLNA, even though it is one of the most advanced in vivo methods, can yield false-negative and false-positive results. Despite these weaknesses, and although human data are available as an alternative reference point, the LLNA is the only accepted reference for the determination of potency and for providing background data for ITS validation.

The conclusions on the status and roadmap for skin sensitization testing are as follows:

- Many non-animal methods for skin sensitization testing have been proposed and some of them have been/ will be validated for the purpose of hazard identification. The development of non-animal methods for the evaluation of the relative skin sensitizing potency of contact allergens will require more work.
- Better measurements and tests for exposure are needed, and little is known about how to assess mixtures yet.
- Complications may arise when there is a need to test hydrophobic compounds or formulations as the proposed models may not be adequate. These problems must be tackled sooner rather than later and the applicability of each model should be assessed accordingly. This will provide opportunities for the development of other assays with other applicability domains.

- Many other areas of toxicology can follow skin sensitization as a good example where a detailed understanding of mechanisms can lead to the development of specific assays needed to identify compound toxicity.
- Computational models based on quantitative structure-activity relationships (QSARs) provide promising tools to identify sensitizers, as the toxicity of the chemical is implicit in its structure. There have been major advances in QSAR models, although studies use data from the LLNA rather than human data and have difficulties in obtaining accuracy in models for “moderate” sensitizers (Li et al., 2007).


6.2 Repeated dose toxicity
Repeated dose testing (RDT) consists in the evaluation of a chemical’s potential to cause chronic toxicity and organ-specific toxicities. Classically, tests for RDT are based on 4 (sub-acute toxicity), 13 (sub-chronic toxicity) and 26–102 (chronic toxicity) week rodent and non-rodent studies. Toxicity occurs after a chemical is absorbed into the general circulation. There is great concern about the relevance of these studies performed in animals for predicting human toxicity (Basketter et al., 2012; Chen et al., 2014; Hengstler et al., 1999; Olson et al., 2000). Different organizations (FDA, EPA, EMEA) and initiatives (REACH, TSCA and the EU Cosmetic Directive) are pushing in vitro methods in the chemical toxicity evaluation process.

RDT includes chronic adverse effects on major organs. On the one hand, the assessment of RDT requires lengthy in vivo experiments, which are difficult to model in vitro. On the other hand, inter-species differences can limit the usefulness of animal data for the prediction of human hazard in this area (Leist and Hartung, 2013). In vitro methods based on human cell lines may provide more human-relevant information (Pfaller et al., 2001). Biological models for different organs, e.g., liver, kidney, lung or brain, have been established, and new culture techniques, especially in form of 3D organoids, are expected to solve present issues about long-term culturing, absence of relevant inflammatory and immune cells (Hengstler et al., 2012) and availability of fully mature cell phenotypes. Stem cells, especially pluripotent stem cells, will be a major source of tissues and cells not available otherwise. Therefore, research on the generation of 2D cultures and 3D tissues from stem cells is of high importance. One of the approaches in this direction is the European SEURAT-1 project (following the long-term strategic target: “Safety Evaluation Ultimately Replacing Animal Testing”, http://www.seurat-1.eu). It started in 2011 with 50 million € joint funding from the European Commission and Cosmetics Europe and is focusing on the development of non-animal test systems in the field of repeat dose systemic toxicity following a case study approach based on the AOP concept. The Tox21 consortium and the US EPA’s ToxCast™ activity in the USA (Dix et al., 2007; Judson et al., 2010, 2014) as well as other activities in Europe and worldwide take similar or complementary approaches (NRC, 2007; Adler et al., 2011; Basketter et al., 2012; Judson et al., 2012; Leist et al., 2012b). Key to all these activities is the concept that most late (longer term)
In summary, repeated dose toxicity will probably be the last method to be replaced, the use of PoT and new culture systems combined with new technologies and sharing of data on pharmaceutical case studies could be the opportunity to reduce the need for such expensive and long-term studies.

6.3 Toxicokinetics and quantitative in vitro–in vivo extrapolation (qIVIVE)

To relate data from non-animal test systems to the human situation, the in vitro concentration levels need to be correlated with the real exposure in vivo. Procedures for such extrapolations (qIVIVE) have been established (Fig. 13). The starting point is a determination of the “real” toxicant concentration that a cell is exposed to. This may be different from the nominal concentration, because of evaporation, metabolism, binding to plastic or uneven distribution in cells. Next, a physiologically-based pharmacokinetic (PBPK) model would be constructed for absorption and distribution in the whole organism, followed by metabolism and excretion. In vitro test systems to predict drug effects of chemicals will be predicted from the early changes they cause in cellular signaling and regulation (Kleensang et al., 2014; Blaauboer et al., 2012). Therefore, signaling pathway identification and analysis is a crucial research necessity in toxicology, and very detailed quantitative information needs to be derived (Fig. 12) to use such data for systems biology modeling (Jennings et al., 2013; Krug et al., 2014; Hartung et al., 2012). Toxicogenomics technologies (Ramirez et al., 2013) are important tools that cover a multitude of cellular events. However, it is important to apply them to the right biological models. For instance, monocultures can hardly model the inflammatory responses frequently seen after long-term exposure to hazardous chemicals.

For repeated dose toxicity, two different approaches are taken in the development of alternative methods: (a) substitution of animals by a battery of relatively complex surrogate models that reflect important features of target tissues and organs. They often use “apical” phenotypic endpoints (e.g., cell death markers) as readouts; (b) an integrated and tiered systems biology approach based on mechanistic endpoints and using the vast knowledge on biological regulation and homeostasis. Pathways-of-toxicity (PoT), emerging from such approaches, will guide hazard evaluation and risk assessment when combined with toxicokinetics modelling (Hartung and McBride, 2011; Boekelheide and Andersen, 2010). The two types of approaches may also be combined.

![Fig. 12: Illustration of the different deviations of signals (physiological cellular responses) that need to be measured by modern in vitro methods](image)

The normal cellular response is shown in blue. This is meant to symbolize any cellular function, such as a muscle contraction, an electrical signal in neurons or the regulation of glucose. Red and white curves exemplify different toxic responses. The examples show that key parameters need to be measured at high temporal and spatial resolution and over many concentrations to be sure the whole range of toxicological reactions is covered. “Toxicity” is in many cases not a simple decrease or absence of a response, but too much or wrong timing can be equally problematic.

![Fig. 13: Schematic explanation of quantitative in vitro–in vivo extrapolation (qIVIVE)](image)

The qIVIVE procedure is considered a pivotal step in the use of in vitro data for the risk assessment process. In vitro toxicity assays provide a benchmark concentration (BMC), i.e., a concentration above which a chemical is considered to be toxic in this system. The BMC is used as the point of departure (POD) for further qIVIVE steps. It allows the calculation of the corresponding human plasma concentration (PC). By taking into account in vitro data on metabolic conversion, human physiology and metabolic parameters, the human equivalent dose can be estimated. This is the starting point of the risk assessment process.
Inductive toxicity testing has shown high background variability (even among untreated control animals) and is characterized by low species concordance (ToxCast™, for example, showed 60% concordance between rat and rabbit studies; and 56% concordance between zebrafish and rat). In some cases, to overcome low sensitivity, studies in a second species may be requested by regulators. However, the two-species approach increases both the cost of the studies and the false-positive rates dramatically (Hartung, 2009). For this reason, in 2009, a revision of the REACH legislation reduced the use of a second species.

Further progress in this area would be accelerated by regulatory steps that preclude the use of in vivo data unless they come from a formally validated model and therefore have a known predictivity (Carney et al., 2011). Uses of the zebrafish assay (Selderslaghs et al., 2011; Padilla et al., 2012; Truong et al., 2014), the embryonic stem cell test (eSt) (van Dartel and Piersma, 2011; Seiler and Spielmann, 2011) and further developments on the basis of ReProtect test systems (Piersma, 2010) could immediately fill the gap until assays based on human cells become available. The field is developing very dynamically, and, especially in the area of developmental neurotoxicity, many new test systems are emerging (Zimmer et al., 2014; Smirnova et al., 2014; Leist et al., 2013; Bal-Price et al., 2012).

New assays will need to be assembled into an advanced test battery using concepts of ITS design (Fig. 15).

The ReProTect project assembled 35 European partners from academia, SMEs (Small-Medium Enterprises) and governmental institutes in order to develop in vitro reproductive toxicity approaches (http://www.reprotect.eu/). The scientific problem of identifying non-animal test methods in this field was addressed (Hareng et al., 2005). The project was based on a battery of in vitro methods that covered different steps of the reproductive cycle (Fig. 16). In a so-called “feasibility study” conducted at the end of the project, 10 blinded chemicals were tested by the consortium. Effects on 3 endpoints, namely male fertility, female fertility and embryotoxicity were predicted. The results of the feasibility study demonstrated that the vast majority of the predictions made were correct (Schenk et al., 2010).

Reproductive toxicity, including developmental toxicology, is a particularly difficult field as far as animal-to-human predictions are concerned (Knudsen et al., 2011; Makris et al., 2011). Reproductive toxicity aims to assess possible hazard to the reproductive cycle, with a high interest in the early stages of embryonic development (embryotoxicity). Tests like the two-generation study are among the most costly and require up to 3,200 animals per substance (Hartung, 2008; Rovida and Hartung, 2009). This makes it impossible to test the enormous amount of chemicals present in the market, leading to a lack of information on reproduction and development toxicity of tens of thousands chemicals. Moreover, animal-based tests offer little mechanistic insight into a chemical’s toxic mode-of-action (MoA) (Knudsen et al., 2013; Knudsen, 2013). Animal reproductive toxicity testing has shown high background variability (even among untreated control animals) and is characterized by low species concordance (ToxCast™, for example, showed 60% concordance between rat and rabbit studies; and 56% concordance between zebrafish and rat). In some cases, to overcome low sensitivity, studies in a second species may be requested by regulators. However, the two-species approach increases both the cost of the studies and the false-positive rates dramatically (Hartung, 2009). For this reason, in 2009, a revision of the REACH legislation reduced the use of a second species.

Further progress in this area would be accelerated by regulatory steps that preclude the use of in vivo data unless they come from a formally validated model and therefore have a known predictivity (Carney et al., 2011). Uses of the zebrafish assay (Selderslaghs et al., 2011; Padilla et al., 2012; Truong et al., 2014), the embryonic stem cell test (eSt) (van Dartel and Piersma, 2011; Seiler and Spielmann, 2011) and further developments on the basis of ReProtect test systems (Piersma, 2010) could immediately fill the gap until assays based on human cells become available. The field is developing very dynamically, and, especially in the area of developmental neurotoxicity, many new test systems are emerging (Zimmer et al., 2014; Smirnova et al., 2014; Leist et al., 2013; Bal-Price et al., 2012). The new assays will need to be assembled into an advanced test battery using concepts of ITS design (Fig. 15).

The ReProTect project assembled 35 European partners from academia, SMEs (Small-Medium Enterprises) and governmental institutes in order to develop in vitro reproductive toxicity approaches (http://www.reprotect.eu/). The scientific problem of identifying non-animal test methods in this field was addressed (Hareng et al., 2005). The project was based on a battery of in vitro methods that covered different steps of the reproductive cycle (Fig. 16). In a so-called “feasibility study” conducted at the end of the project, 10 blinded chemicals were tested by the consortium. Effects on 3 endpoints, namely male fertility, female fertility and embryotoxicity were predicted. The results of the feasibility study demonstrated that the vast majority of the predictions made were correct (Schenk et al., 2010).
In the follow-up European project ChemScreen (http://www.chemscreen.eu) 12 chemicals were tested for embryotoxicity in a final performance test. The battery correctly detected 11 out of 12 compounds tested. The consortium concluded that “this study illustrates added value of combining assays that contain complementary biological processes and mechanisms, increasing predictive value of the battery over individual assays” (Piersma et al., 2013).

In silico models (e.g., the US EPA’s Virtual Embryo project) have also shown potential application in the reproductive toxicology field. It is expected that ultimately a computer model that simulates cellular function in the growing embryo can be used to determine the effects of teratogens. Some promising first results come from an in silico modeling platform: A novel multi-cellular agent-based model (ABMs) of vasculogenesis using the CompuCell3D (http://www.compucell3d.org/) modeling environment supplemented with semi-automatic knowledgebase creation has been developed by EPA. Dynamic cell ABMs have been shown to simulate complex developing systems and, consequently, display a potential to simulate adverse effects (Klein-streuer et al., 2013; Hester et al., 2011; Shirinifard et al., 2013) and aberrant tissue fusion (Ray and Niswander, 2012).

6.5 Carcinogenicity
At present, the carcinogenicity hazard of chemicals is determined by a costly and lengthy animal test, the “cancer bioassay”, although its relevance for human health is seriously doubted (Al- den et al., 1996; Knight, 2007; Gottmann et al., 2001). Results of more than 3,500 cancer bioassays, which cost about € 800,000 per substance and species, are publically available: 53% of all substances tested were positive, suggesting an enormous false-positive rate, but still some accepted human carcinogens were not positive (Basketter et al., 2012). The correlation of findings in rats and mice is less than 60%, even less if the site of cancer in the organism is considered. The experts suggested a thorough evaluation of the test, taking into account the principles of evidence-based toxicology (Hoffmann and Hartung, 2006b). This might lead to an abolition of the in vivo assay (Fig. 17). Chemical carcinogenicity may be based either on genotoxic or non-genotoxic (epigenetic) mechanisms (Oliveira et al., 2007). Alternative methods for the determination of genotoxicity have been in use for over 40 years. An ITs has been suggested to combine such available methods (Pfuhler et al., 2010; Aldenberg and Ja-worska, 2010). Testing for non-genotoxic carcinogens has proven more difficult, but good results have been obtained recently by various cell transformation assays (Vanparys et al., 2011). A combination of mutagenesis assays, tests for DNA damage, cell transformation assays and targeted tests for frequent epigenetic

**Fig. 16: Examples of a test battery addressing a highly complex toxicological endpoint**

The reproductive cycle with its four main phases is the target of reproductive toxicants. The FP6 EU project ReProTect established an in vitro test battery for reproductive toxicity testing covering the reproductive cycle with a series of individual tests. Each test system covers a small part of the reproductive cycle. The names of the different tests are depicted outside and inside the circle, indicating which part of the developmental process is modeled. For full explanation see http://axlr8.eu/axlr8-2010-progress-report.pdf

**Fig. 17: Roadmap for animal-free carcinogenicity predictions**

In the area of carcinogenicity the experts suggested, in addition to the points summarized in Figures 4 and 6, to include several specific measures and research lines to be followed.

**Roadmap to carcinogenicity predictions**

- Generate appropriate biological material (e.g., human cells)
- Controlled culture conditions (defined genetic backgrounds)
- New technologies (operator-independent, automation)
- Quality assurance
- Validation of state-of-the-art → abolition of current test
- Optimize battery for genotoxic carcinogens → new assays of repair, recombination ...
- ITs including "non-genotoxic" modes of action, epigenetics, PoT basis
- Further evaluation of "CTAs" (cell transformation assays)
mechanisms (e.g., nuclear receptor activation) will most likely form the basis for a future ITS. Most elements are available in some form, but they will require further development and optimization for satisfactory predictivity (Fig. 17).

7 Evaluation of test system performance

Evaluation of test system performance has classically considered three aspects (Fig. 18): (1) the technical reliability of the test; (2) the scientific background, rationale and scope; (3) the correlation of test data with a gold standard (e.g., animal data). The latter point has also been called predictivity (Hartung et al., 2004; Hoffmann and Hartung, 2006a; Moore et al., 2009). The validation procedure has until now followed very strict and rigid rules in the field of chemical testing. This has led to high costs and long delays before new assays were introduced. Moreover, the definition of assay predictivity on the basis of animal data has proven to be problematic because of the shortcomings of the in vivo experiments. Therefore, new validation concepts have to be considered. For instance, high throughput screening assays need to be treated differently from other tests as, e.g., ring trials cannot be performed when certain robotics equipment is available only in one place (Judson et al., 2013). In cases where predictivity cannot be determined from correlation studies, en-

---

**Fig. 18: New validation approach for novel toxicity tests**

At present, test validation relies on three pillars: reliability, scientific basis and predictivity. Predictivity has in practical terms been determined by the correlation of in vitro test results with animal data. This approach is not possible for many of the toxicity domains discussed here and many of the assays that are developed are part of a test battery. Future validation must therefore rely on two pillars: Even more focus is required on test quality (reliability). Moreover, the scientific basis of a test needs to be broadened to provide a rationale for the predictive capacity of the test, not based on statistical correlation but based on scientific (mechanistic) explanations.

---

**Fig. 19: Vision for the future of toxicity testing**

The current approach is first to test unknown chemicals in animal tests. This limits overall throughput and leads to a high rate of false positives and false negatives. Mixtures are hardly ever tested because of the limitation of resources. High costs in combination with a low predictivity lead to many cases of “no testing”. Mechanistic studies are only carried out in few cases of particular interest to identify the factors causing the toxicity. The new approach, suggested here, is based on 21st century in silico and in vitro methods identifying PoT. This will in most cases lead to an amount of data that is sufficient to decide whether a substance is toxic (positive) or non-toxic (negative) for the intended scenario. Only in few cases, when not enough information can be obtained, will animal tests be performed as an additional source of information. Good information can be provided on all chemicals and, due to the high throughput of the approach, also on mixtures.
tirely different approaches need to be considered. Mechanistic validation, i.e., focus on the scientific background and consistency of an assay (Hartung et al., 2013a; Leist et al., 2012a), is the most promising general option. For specialized assays, e.g., within a test battery or as the basis of a screening approach, the test performance will also need to be judged according to its fit for the specific purpose (Fig. 18). The concept of flexible evaluations of test system performance has been pioneered in the field of drug discovery. Fit-for-purpose evaluation and rating of test predictivity based on the scientific rationale of a test are commonplace in this field, and this experience could help to establish faster and more efficient assay evaluation procedures for chemicals, cosmetics and pesticides. In the discussion on test system predictivity it is important to keep in mind that test reliability will always be a necessary, absolutely mandatory condition, whatever evaluation process is chosen. The progress and acceptance of non-animal based testing will depend on this criterion and its strict implementation in the field (Leist et al., 2010).

8 The future of toxicity testing

The participants of the roadmap consensus symposium envisaged that two key features will distinguish the future from the present toxicity testing (Fig. 19). First, the present animal-based testing (sometimes followed by in vitro tests to supply mechanistic information) will be substituted by ITS using in vitro and in silico approaches (sometimes followed by animal tests, where further data are needed). Second, according to the vision for a new toxicology, data will be generated for every chemical and possibly also for important mixtures. This contrasts with the present situation, in which hardly any data is available on many chemicals (Crofton et al., 2012) and many tests are being waived. Moreover, a large part of the available data lacks mechanistic background and consistency controls. It therefore cannot be used to supply information concerning adverse outcome pathways. Despite the many shortcomings and, in particular, a lack of formal validation, animal data are still being used as gold standard. This contributes to an underestimation of the success of non-animal methods. The expectations regarding scientific validity and predictivity are usually higher for alternative methods than for the respective in vivo models. This is a key issue for the roadmap initiative, as the primary goal is to provide methods that are as good as animal models (not better). This goal might already have been reached in some areas. The future goals would then be to further improve the quality of safety testing beyond that of animal experiments.

As detailed already in the NRC report on the vision for toxicity testing in the 21st century (NRC, 2007; Leist et al., 2008a), the radically new approach to risk assessment has large economic and scientific advantages over the present animal-based system. The roadmap outlined in this overview highlights important steps towards this goal. All experts agreed that considerable work is still needed, but there was also strong consensus that already an impressive advance has been achieved and that the goal is well worth the required efforts.

References


Acknowledgements

This work was supported by the Doerenkamp-Zbinden Foundation, the State of Baden-Württemberg, Germany, the German BMBF and the Centers for Alternatives to Animal Testing in the US and Europe.

Correspondence to

Marcel Leist, PhD
University of Konstanz
Box M657
78457 Konstanz
Germany

e-mail: marcel.leist@uni.konstanz.de