Table of Contents

t^4 Report

A Roadmap for the Development of Alternative (Non-Animal) Methods for Systemic Toxicity Testing

"The difficulty lies, not in the new ideas, but in escaping from the old ones."
John Maynard Keynes (1883-1946)

"This report, by its very length, defends itself against the risk of being read."
Winston Churchill (1874-1965)

t4 transatlantic think tank for toxicology

t4 Report*

A Roadmap for the Development of Alternative (Non-Animal) Methods for Systemic Toxicity Testing

David A. Basketter [1,§], Harvey Clewell [2,§], Ian Kimber [3,§], Annamaria Rossi [4,§], Bas Blaauboer [5], Robert Burrier [6], Mardas Daneshian [7], Chantra Eskes [8], Alan Goldberg [9], Nina Hasiwa [10], Sebastian Hoffmann [11], Joanna Jaworska [12], Thomas B. Knudsen [13], Robert Landsiedel [14], Marcel Leist [15], Paul Locke [9], Gavin Maxwell [16], James McKim [17], Emily A. McVey [18], Gladys Ouédraogo [19], Grace Patlewicz [20], Olavi Pelkonen [21], Erwin Roggen [22], Costanza Rovida [23], Irmela Ruhdel [24], Michael Schwarz [25], Andreas Schepky [26], Greet Schoeters [27], Nigel Skinner [28], Kerstin Trentz [29], Marian Turner [30], Philippe Vanparys [31], James Yager [32], Joanne Zurlo [9], and Thomas Hartung [33,§]

1. DABMEB Consultancy, Sharnbrook, UK; author whitepaper sensitization
2. The Hamner Institutes for Health Sciences, Research Triangle Park, NC, USA; author whitepaper toxicokinetics
3. Faculty of Life Sciences, University of Manchester, UK; author whitepaper sensitization
4. CAAT-Europe, University of Konstanz, Germany; author whitepaper repeated dose toxicity
5. Doerenkamp-Zbinden Chair on Alternatives to Animal Testing in Toxicological Risk Assessment, Institute for Risk Assessment Sciences, Division of Toxicology, Utrecht University, The Netherlands; respondent toxicokinetics
6. Stemina Biomarker Discovery, Madison, WI, USA; respondent reproductive toxicity
7. CAAT-Europe, University of Konstanz, Germany; scientific writer toxicokinetics
8. SeCAM, Agno, Switzerland; respondent toxicokinetics
9. CAAT, Johns Hopkins University, Bloomberg School of Public Health, Department of Environmental Health Sciences, Baltimore, MD, USA
10. CAAT-Europe, University of Konstanz, Germany; scientific writer reproductive toxicity
11. seh consulting + services, Cologne, Germany; respondent repeated dose toxicity
12. Procter & Gamble, Brussels, Belgium; respondent sensitization
13. US EPA, Research Triangle Park, NC, USA; respondent reproductive toxicity
14. BASF, Ludwigshafen, Germany; respondent carcinogenicity
15. CAAT-Europe, University of Konstanz, Germany; respondent repeated dose toxicity
16. Unilever, SEAC, Bedford, UK; respondent sensitization
17. CeeTox, Kalamazoo, MI, USA; respondent repeated dose toxicity
18. NOTOX B.V., 's-Hertogenbosch, The Netherlands; scientific writer repeated dose toxicity
19. L'Oréal, Paris, France

*a report of t4 – the transatlantic think tank for toxicology, a collaboration of the toxicologically oriented chairs in Baltimore, Konstanz, and Utrecht, sponsored by the Doerenkamp-Zbinden Foundation; participants do not represent their institutions' opinions and do not necessarily endorse all recommendations made.

§ These authors contributed equally to the underlying whitepapers.

20. DuPont Haskell Global Centers for Health and Environmental Sciences, Newark, DE, USA; respondent sensitization
21. Department of Pharmacology and Toxicology, University of Oulu, Finland; respondent toxicokinetics
22. Novozymes A/S, Denmark; respondent sensitization
23. CAAT-Europe, University of Konstanz, Germany; scientific writer sensitization
24. Animal Welfare Academy / German Animal Welfare Federation, Munich, Germany
25. Toxicology, University of Tuebingen, Germany; respondent reproductive toxicity
26. Beiersdorf, Hamburg, Germany; respondent sensitization
27. VITO, Mol, Belgium; respondent repeated dose toxicity
28. Agilent Technologies, Inc., Berkshire, UK
29. Bioservices, Planegg, Germany
30. Freelance science writer; scientific writer carcinogenicity
31. ALTOXICON BVBA, Vosselaar, Belgium; respondent carcinogenicity
32. Johns Hopkins Bloomberg School of Public Health, Environmental Health Sciences, USA; respondent carcinogenicity
33. CAAT and CAAT-EU; author introduction, conclusion and whitepapers carcinogenicity and reproductive toxicity

Summary

Systemic toxicity testing forms the cornerstone for the safety evaluation of substances. Pressures to move from traditional animal models to novel technologies arise from various concerns, including: the need to evaluate large numbers of previously untested chemicals and new products (such as nanoparticles or cell therapies), the limited predictivity of traditional tests for human health effects, duration and costs of current approaches, and animal welfare considerations. The latter holds especially true in the context of the scheduled 2013 marketing ban on cosmetic ingredients tested for systemic toxicity. Based on a major analysis of the status of alternative methods (Adler et al., 2011) and its independent review (Hartung et al., 2011), the present report proposes a roadmap for how to overcome the acknowledged scientific gaps for the full replacement of systemic toxicity testing using animals. Five whitepapers were commissioned addressing toxicokinetics, skin sensitization, repeated-dose toxicity, carcinogenicity, and reproductive toxicity testing. An expert workshop of 35 participants from Europe and the US discussed and refined these whitepapers, which were subsequently compiled to form the present report. By prioritizing the many options to move the field forward, the expert group hopes to advance regulatory science.

Keywords: skin sensitization, allergic contact dermatitis, toxicokinetics, repeated dose testing, reproductive toxicity, carcinogenicity, predictive testing, alternative approaches, risk assessment

1 Introduction

Author: Thomas Hartung[1]
Discussants: David A. Basketter, Bas Blaauboer, Robert Burrier, Harvey Clewell, Mardas Daneshian, Chantra Eskes, Alan Goldberg, Nina Hasiwa, Sebastian Hoffmann, Joanna Jaworska, Ian Kimber, Tom Knudsen, Robert Landsiedel, Marcel Leist, Paul Locke, Gavin Maxwell, James McKim, Emily A. McVey, Gladys Ouédraogo, Grace Patlewicz, Olavi Pelkonen, Erwin Roggen, Annamaria Rossi, Costanza Rovida, Irmela Ruhdel, Michael Schwarz, Andreas Schepky, Greet Schoeters, Nigel Skinner, Kerstin Trentz, Marian Turner, Philippe Vanparys, James Yager, Joanne Zurlo

1.1 Background

Two pieces of European legislation have created the pressure to develop novel approaches for systemic toxicity testing, beyond the general urge to replace animal testing as prescribed in the European Directive 2010/63/EU on the protection of animals used for scientific purposes (Hartung, 2010d; Seidle et al., 2011). This report deals with methods for the testing of all chemicals, and does not focus only on cosmetics. This activity is aimed at providing a scientific roadmap for the replacement of animal based safety testing in all domains.[2]

1.1.1 The 7th Amendment of the Cosmetics Directive

On January 15, 2003, the EU passed a law banning the testing of cosmetics and their ingredients on animals, reinforced by marketing bans with different deadlines. Known as the 7th Amendment (Directive 2003/15/EC) to the Cosmetics Directive (Directive76/768/EEC), this Directive is intended to protect and improve the welfare of animals used for experimental purposes by promoting the development and use of scientifically valid methods of alternative testing (Hartung, 2008a). The main objective of this Directive is to prohibit the testing of cosmetic products/ingredients on animals through a phased series of EU testing and marketing bans. This ban on animal testing and sales would start immediately where alternative non-animal tests are available, followed by a complete testing ban six years after the Directive became effective (i.e., in 2009). Therefore, animal experiments for cosmetic products and ingredients are completely banned, reinforced with a marketing ban in the EU since 2009, irrespective of the availability of animal-free methods, except for repeat-dose toxicological endpoints (i.e., toxicokinetics, repeated dose toxicity, skin sensitization, carcinogenicity, and reproductive toxicity) where the EU marketing ban is delayed until 2013 for tests carried out outside the EU. This ban may, however, be postponed by a new legislative act if alternative tests cannot be found.

1.1.2 Testing needs for the REACH legislation

As an enormous investment into consumer product safety, the REACH program aims to assess existing (“old”) chemicals that have previously undergone very little testing (Hartung, 2010a). Regulation (EC) 1907/2006, known as REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals), revises the Dangerous Substances Directive (67/548/EEC). The registration process has only recently begun, and the estimated testing demands are under debate (Hartung and Rovida, 2009a,b; Rovida and Hartung, 2009; Rovida et al., 2011). However, there is little doubt that systemic toxicity will account for more than 95% of the testing costs and animal use of REACH. It is clear that testing capacities are challenged and alternative approaches, especially for systemic toxicities – as called for in the legislation – might relieve such tensions.

1.2 A framework for replacing systemic toxicity testing by new approaches

The advantage and disadvantage of alternative methods lies in the reductionist character of their approach. This eases interpretation to the extent that a simpler read-out is likely to result from such an approach, but raises the issue of what aspects of the biology might be missing. Aside from abolishing useless tests (1.2.1) (which is not an alternative method but should nonetheless be considered as an option), a number of principal alternative approaches (1.2.2 to 1.2.4 and 1.2.6) were identified. These include *in vitro* and *in silico* (1.2.5), as well as combined approaches (1.2.7-1.2.8), by either mining or modeling the respective data and/or relating them back to structure and other properties of the test substance.

1.2.1 Abolition of useless tests

A cost-benefit analysis could help in making decisions to abandon tests of questionable practical utility. Such considerations may be based on reproducibility issues, lack of predictivity, lack of scientific basis, or limited contribution to regulatory decision-making. Obviously, “uselessness” is a value judgment. For animal tests, a number of limitations (Hartung, 2008b) can be evaluated in terms of whether they translate to the given test. The socioeconomic impact of wrong or missing assessments needs to be taken into consideration (Bottini and Hartung, 2009, 2010), along with other sources of information to substitute for comparison of performance characteristics with other methods. Tests that have been abolished in the past include the traditional LD_{50} test (OECD TG 401; OECD, 1987), the abnormal toxicity test for vaccines, and the ascites mouse for the production of monoclonal antibodies.

[1] The introduction text was largely part of the original whitepapers on carcinogenicity and reproductive toxicity and discussed in this context.

[2] At the workshop I. Ruhdel pointed out that from an animal protection point of view the workshop should not be seen or communicated as an activity in the context of the current discussion on possibly postponing the marketing ban on animal tested cosmetics.

1.2.2 Reduction to key events

Traditional 3Rs or alternative methods have been aimed at a one-to-one replacement of animal tests. This appears to be feasible if a key (rate determining) event can be readily identified. Examples of such attempts include key events such as mutagenicity, or possibly cell transformation, for carcinogenicity, whereas protein binding is assumed to be a prerequisite for skin sensitization. The selection of key events can be informed by the scientific understanding of the pathophysiology or through analysis of what was derived, i.e., what was actually observed in guideline tests that drove the classification (for example, which organ toxicities are actually driving regulatory decisions) or is seen in intoxicated patients (human-relevant manifestations). The obvious central question is: Can a key event for the given hazard or test concern be readily identified? The scientific challenge lies in the state of mechanistic understanding – i.e., for some toxicological endpoints a single non-animal test can be used to sufficiently characterize the adverse effects of the chemical. For other, more complex endpoints, several non-animal approaches are required to fully characterize the impact of the chemical on the relevant tissue(s).

1.2.3 Negative exclusion by lack of key property

The most prominent example of exclusion criteria (conditio sine qua non) are large molecular size or barrier models – no bioavailability/penetration, no harm. The obvious problem is the reliance on negative data (no transfer). This concept is further refined by the threshold of toxicological concern (TTC) approach (Kroes et al., 2005), where exposure (and thus resulting availability in sufficient quantity) – not absolute bioavailability – is evaluated: For non-cancer endpoints, NOAELs or, alternatively, TD_{50} (toxic dose 50%) values are collected for a large number of chemicals and their distribution is used in combination with a safety factor to set a threshold where no adverse effect is expected. TTC values have been derived for different structural classes, e.g., Cramer classes, while other TTC have been derived and subsequently refined on the basis of specific structural alerts for genotoxicity.

Similarly, many toxic endpoints rely on reactive chemistry allowing interaction with target structures. The absence of structural features allowing direct reactivity or activation via metabolism represents another example of exclusion of a hazard.

1.2.4 Optimization of existing tests

In vitro tests have no fewer limitations than their *in vivo* counterparts (Hartung, 2007a). A number of strategies may be able to improve the predictive value of existing test systems:

- extension of metabolic capacity
- organotypic 3-dimensional (co)-cultures
- more physiologic culture conditions such as homeostasis, oxygen supply, cell density
- transition from cell lines to primary cells or stem cell-derived systems
- use of human cells
- refinement and expansion of endpoints measured
- standardization and automation
- quality assurance of procedures
- appropriate statistics and prediction models
- definition of applicability domains
- extensions to address solubility issues and nanomaterials

These opportunities will differ from test to test. They can improve the predictive value of tests, making them (more) fit for purpose. Such changes, however, will typically require a (re-) assessment of the validity of the modified system.

1.2.5 In silico approaches

A number of approaches (Hartung and Hoffmann, 2009) try to link, often via structure and physicochemical descriptors, to results available for other substances to avoid testing. They are somewhat similar to what is referred to as “read-across” but in a formalized and quantitative way, using either rules, empirical correlations to parameters of interest, or other modeling exercises. For complex endpoints, such models are unlikely to be used as stand-alone replacements, but are better suited to provide valuable supporting information as part of a weight of evidence (Hartung et al., 2010) approach. They can play a key role in combination with other tools or to further optimize biological measurements. It is foreseeable that some Integrated Testing Strategies (ITS, see 1.2.7) developed in the near future actually will be *in silico* tools with biological inputs.

The basic problem is that we base our judgments on existing knowledge and its availability and quality. Surprising effects can hardly be predicted, and all quality limitations of this existing knowledge (e.g., quality of animal test data or mechanistic understanding) will translate to the estimation technique. This is not unique to modeling approaches per se, but it is important to note that the value of existing information (see 1.2.1) is again the critical starting point. While there are established measures of similarity of chemicals, these merely address structural similarity and do not consider the context of the endpoint of concern. Thus, even if we assume that we may have a fair appreciation of structural similarity, understanding whether this is key for the distribution of the chemical in the organism and its toxic mechanism is an additional consideration.

A limitation of all these techniques is that they can only be readily applied to discrete organic substances. That suggests, based on rough estimates, that some 50% of the chemicals impacted under REACH, which comprise mixtures, lack of purity, salts, metal compounds, etc., cannot be readily evaluated using modeling approaches (Hartung and Hoffmann, 2009). Furthermore, all health effects where small impurities are relevant cannot be handled with such structure-based estimation techniques: Allergic reactions (sensitization), for example, can be caused by less than 0.1% of contamination. With the same reasoning as for possible contaminants, health effects, where no thresholds can be established (carcinogenic, mutagenic, or some reproductive toxicants), should not be evaluated on the basis of the structure of the main compound only (while these contaminants are typically present in *in vivo* or *in vitro* tests). It is noteworthy that these are exactly the tests that consume the most animals and resources (>80%) under REACH.

The role of *in silico* techniques will principally be within ITS, not as stand-alone replacements. They will support other types of information, help to prioritize and – following evaluation – increasingly substitute for testing. It might be that they can serve as 2nd generation alternative methods, i.e., modeling validated *in vitro* methods, because these more simple but standardized tests allow for the generation of large datasets, which would facilitate modeling of new key events.

1.2.6 Information-rich single tests

The sensitivity of the test system, i.e., here the spectrum of interactions with xenobiotics covered by the test system, can be increased by measuring more endpoints, e.g., by omics or high-content imaging. This can be done by supervised analysis (measuring known biomarkers or hazard pathways) or in an unsupervised manner by testing for any response, which only then is interpreted as a signature of effect. Prominent examples are cell systems combined with transcriptomics, proteomics, or metabolomics. This typically will lead to signatures of toxicity (SoT), such as a reduction of information to patterns of signals associated with the hazard. Notably, identifying biomarkers from the variety of signals should shift the approach away from the more traditional (1.2.2) and (1.2.3) approaches. High-content measurements, such as image analysis, represent other technologies increasingly applied here. We should bear in mind that even the most sophisticated measurements and bioinformatics can hardly overcome the limitations of the cell systems. Therefore, the experience gained with the development and validation of alternative methods with simple endpoints is of critical importance when moving towards wholly novel technologies. Good Cell Culture Practices form only one example here (Coecke et al., 2005; Hartung, 2010b; Leist et al., 2010; Wilcox and Goldberg, 2011).

1.2.7 Integrated testing strategies (ITS)

In every case where no single property or single test system can be identified to cover a hazard, tests will need to be combined and results integrated. One key example is the combination of toxicity data on the one hand (e.g., derived *in vitro* and/or *in silico* with kinetic data (e.g., modeling) in ITS, see, e.g., DeJongh et al., 1999; Forsby and Blaauboer, 2007; Blaauboer, 2010). The purposes of combining tests can be:

- covering different mechanisms or applicability domains
- increasing the predictive value compared to a single test
- avoiding costly tests or animal tests by filtering out certain substances
- adding kinetic information to hazard evaluations
- integrating existing data

In the simplest case an ITS is a battery of tests, and any positive result is taken as an indication of toxicity, as is the case for the combined mutagenicity tests. More sophisticated combinations with interim decision points are emerging (Jaworska et al., 2011; Jaworska and Hoffmann, 2010), but accepted concepts regarding how to construct and validate them are not available. A major problem seems to be that most methods now being combined into ITS were originally developed to work as stand-alone alternatives and are now combined because they did not achieve this. The downside to this might be that they are not sufficiently complementary to make a major change in an ITS. The systematic construction of components for an ITS represents a key opportunity to advance the overall ITS approach. A very promising way of constructing a testing strategy is breaking the (patho-)physiology down to crucial elements, e.g., the different elements of the reproductive cycle (as was done for the ReProTect project) (Hareng et al., 2005) or the key processes of neurodevelopment in the series of DNT workshops. However, this still leaves open the question of how to integrate all these tests.

The concept of ITS was advanced substantially during the development of the REACH technical guidance (Schaafsma et al., 2009). Regulatory toxicology to date has been developed as a toolbox of tests, which allows the health effects of new substances (especially pre-market drugs and pesticides) to be classified before carrying out a risk assessment. Given that little was known about the inherent properties of a given chemical and minimal information about possible future uses was available, each test within the toolbox was optimized to have as few as possible false-negative results, which might represent a later safety risk. Indeed, in the absence of information it is preferable to "over-label" a possible hazard, often called the "precautionary principle." As a consequence, an unknown proportion of substances are abandoned based on false-positive test results in their development as drugs or consumer products, but this is usually accepted since similar substances with favorable profiles are available as alternatives in the test battery. Note that this situation is completely different for REACH purposes, where the same test methods need to be applied to test valuable commodity substances with a significant history of safe use.

Many tests in the toxicological toolbox are dichotomous, i.e., they can have only two outcomes (positive or negative). This suggests that when optimizing the test for few false-negatives, the number of false-positives is increased. However, even the simplest biological aspects are not dichotomous: Sex is male or female, but what about transvestites, transsexuals, hermaphrodites, castrates, Turner (only one X chromosome) or Klinefelter (XXY) syndrome? There is a grey area. When we set our thresholds, we determine the extent of grey and whether we favor false-positives or false-negatives. Due to the precautionary approach in toxicology, thresholds are set to minimize false-negatives, thus favoring false-positives. Although some non-animal test methods have prediction models with only binary outcomes (often to reflect the reference test result), this is rarely the way they are applied, and most non-animal test methods are now being designed to predict dose response information.

The "one suits all" philosophy of the animal test toolbox leads to the problem that usually only one test is available to give the final result. This means that the proportion of false-positives cannot be corrected. Even worse, if several tests allowing false-positives are combined, e.g., the mutagenicity test battery or testing in several species for repeated dose toxicity, reproductive toxicity, or carcinogenicity, a further increase in the proportion of false-positives will arise. This will be the case particularly when non-specific tests are used for relatively rare hazards (Hoffmann and Hartung, 2005). In such a case, the false-positives likely outnumber the real-positives, e.g., by tenfold in case of the cancer bioassay (see below).

The extent of false-positives also is determined by the number of replicate animals. In its most typical application (discriminating between non-responding and responding animals), the use of replicates again reduces false-negatives and increases false-positives. Similarly, multiple testing increases the number of false-positives. Setting a significance level of 95% implies that one out of 20 results is false-positive. To date, the cancer bioassay includes more than 60 endpoints, the reproductive two-generation study 80 and a 28-day repeated dose study 40 – arguably, it is difficult for any substance to test negative. The same reasoning holds true for other tests. The more tests done on a single chemical, the more likely that there is a positive result in one. A cynic might conclude that a non-toxic substance must be one that has not been tested often enough.

REACH foresees the application of the toxicological toolbox to existing chemicals of often enormous economic value. The costs of REACH have been calculated until now on the basis of the actual costs of the tests that would be required to prepare the dossiers. The consequence of false-positive classifications is largely overlooked, at least by regulatory agencies, though the potential impact is not lost on companies. The consequences include unnecessary restrictions of use and safety measures, unjustified abandoning of chemicals, or laborious follow-up studies to rule out a particular unwarranted safety concern. The only rational exit from this dilemma is through a combination of tests – a test strategy, where at least one sensitive (few false-negatives) test and one specific (few false-positive) test are combined. Integrated Testing Strategies (ITS) are needed.

There is a fundamental difference between the testing needs of new versus existing chemicals: Any new chemical represents a possible health hazard, while the longer a chemical is in use, the lower the uncertainty. After the creation of a new chemical, its utility is uncertain; while the longer it is in use, the more its economic value becomes evident. The consequence is simple: false-positive toxicological results are less and less tolerable. While we tend to accept the result of a toxicological evaluation early after generation of a chemical and uncertainty is not welcome, for advanced chemicals in broad use it is unavoidable that problematic test results be questioned.

Drug development represents a good example of the attitude toward new substances, especially since this field has pioneered and shaped our toxicological approach. Classically, i.e., when the toxicological toolbox was developed, around 10,000 substances were synthesized and evaluated to bring one product to the market. Since the bulk of the cost is generated in the clinical phase of development, and toxicological studies represent an "entry permit" for first-time testing in humans, an early and clear statement on the health hazards of a substance is most important. Typically, a broad variety of similar substances related to the lead compound under development are synthesized and often a switch to a substance with a better toxicological profile but with the possibility of a similar effect is possible. Normally there is no time to rule out false-positives. False-negatives, however, represent possible disaster (not only the worst case, when successful drugs have to be withdrawn from the market, but also when expensive clinical evaluations have to be stopped because of side-effects or the need for additional toxicological studies).

For chemicals and consumer products, the situation is, in principle, very similar. For work safety, over-labeling is not very critical, and for consumer products there is often a choice among less critical chemicals. It is telling that more than 90% of the new chemicals notified are not acutely toxic (more than 50% of the animals survive a dose of 2 g/kg); this means that such non-toxic substances mostly have been further developed to applications that reach the market and notification.

Several business impact studies have been carried out for REACH. The fundamental problem of applying tests optimized for new chemicals to *existing* chemicals, however, has so far escaped attention: How much effort will be spent to demonstrate that a result is indeed a false-positive?

Typical measures include:

- repetition of the test
- testing in a second species
- mechanistic studies
- identification of critical metabolites and possible species differences to humans
- exposure scenarios

All these measures are as costly as, or sometimes even enormously more costly than, the original test. Worse, they always leave some doubt with regard to the substance. Thus, it is critically important that the number of false-positives be limited up front. In the field of carcinogenicity, in particular, the precautionary principle produces many false-positive results. It is well known that the *in vivo* test for carcinogenicity has produced enormous numbers of false-positive results already (see below). In addition to the *in vivo* test for carcinogenicity, the current *in vitro* test battery for mutagenicity, i.e., the combination of two tests, results in a false-positive rate of 65-90% for non-carcinogenic substances. This means that the already high proportion of false-positive results from the cancer bioassay will be further increased by an enormous number of non-carcinogenic substances showing a genotoxic effect in one of the two *in vitro* tests. Furthermore, aspects of variability related to a test, e.g., inter-animal variation, or within- or between-laboratory variability, can cause false-positive results.

ITS do more than define how to test strategically; they also determine whether to test at all, as existing and non-testing information can also be integrated. There are three reasons why testing of a substance might not be necessary:

- Available information on a given substance is sufficient.
- Information on related compounds is sufficient to extrapolate.
- Exposure or uptake by the organism is so low that testing can be waived.

These three aspects have to be separated from creating new knowledge. Again, the strategic combination of individual tests is often needed. Combinations of tests are required when the performance of one test cannot suit all needs. The following aspects have to be taken into account to optimize the approach for a given purpose:

- work load and costs
- animal consumption
- certainty of result and resulting safety level
- applicability, e.g., for chemical classes

Components of ITS other than testing *in vitro* or *in vivo* are:

- Use of existing information: Possible sources of information will differ for given substances and fields. The most important questions are how to retrieve them and how to judge their quality (and, thus, their utility). Quality of science does not depend on quality measures like ISO or Good Laboratory Practice, but such quality-assurance programs safeguard proper documentation and the reliability of results. Similarly, adherence to international test guidelines is not a prerequisite for good toxicology, but it facilitates comparability and acceptance. It will be necessary to agree on criteria for each given purpose, which might benefit from the development of scoring systems for the quality of studies and possibly thresholds for acceptability.
- Extrapolation from existing information: Several ways of using information on other chemicals have to be distinguished:
  - read-across (interpolation from existing data of related chemicals), i.e., the data gap filling conducted within a category of substances
  - chemical grouping (testing of prototypic compounds out of a group of similar ones only)
  - structural alerts and rule-bases (structural characteristics that raise concerns or rule out possible hazards (SAR – structure activity relationships))
  - (quantitative) structure activity relationships, i.e., (Q)SAR (correlation of chemical characteristics – physicochemical descriptors, with activities)

The basic question is intriguing: can we use information on similar chemicals to draw conclusions for those for which we have no test results? Certainly not always. Who could possibly predict that a shift of an OH-group in a dioxin molecule changes the potency a thousand fold? The question is whether the uncertainty of such estimation techniques is larger than the uncertainty of tests and interspecies predictivity. Few formal validations have been initiated for some methods ((Q)SAR and rule-based systems). There are parallel efforts underway elsewhere to define which scientific principles and approaches are merited to confirm and justify the appropriateness of a read-across. In general, formal validations are avoided and instead concrete examples to help benchmark potential acceptance under regulatory frameworks by establishing consistent approaches dependent on context for each chemical and endpoint under consideration are needed. Some similar assessments of read-across approaches and chemical grouping will be necessary. However, concepts for validation – especially of ITS – are only emerging (Kinsner-Ovaskainen et al., 2009).

Exposure/bioavailability-based waiving represents another key decision point in many ITS. For most health effects (most likely even for cancer and reproductive toxicology), a minimum concentration must be reached in the target tissue. If this can be excluded due to exposure scenarios and/or limited uptake by the organism, it might not be necessary to conduct further testing. However, this means that the judgment is not definite but depends on chemical use (exposure scenarios and route of application). This approach is most promising for cosmetics, where clear exposure scenarios are given. It also can apply to strictly controlled intermediates when containment can be assured by appropriate risk management measures, and hence TTC type approaches can be useful to set "health benchmarks" for exposures because likely exposure scenarios can be formulated. It is worth noting that the best-established alternative approach to assess uptake is the one for skin absorption (OECD test guideline 428; OECD, 2004), again favoring applications for cosmetic ingredients. At the same time, we need ways to incorporate dermal absorption into risk assessments under REACH, rather than being forced to live with conservative 100% defaults.

When composing and validating a test strategy, it is crucial to assess the performance characteristics of all building blocks. Emerging methodologies (e.g., from Bayesian decision theory) may provide valuable tools for strategic development (Jaworska and Hoffmann, 2010).

Some principles for ITS are evident:

- Combine sensitive and specific tests; combine screening and confirmatory tests. Pertinent examples are mutagenicity tests, where the positive results of a battery of usually two *in vitro* tests (accepting a huge proportion of false-positives, i.e., 95%) are subsequently ruled out by the animal experiment.
- For rare health effects, identify the negatives; use prioritization to increase frequency of positive results.
- Assigning a test result means reducing information; combination of raw data from two tests might be more powerful than combining two final test results.
- For the mutagenicity test battery it has been shown that tests of low predictivity on their own can be combined to result in highly predictive tests (Jaworska et al., 2005).
- Allow interim decisions to obviate further testing (tiered testing strategies).
- Conduct inexpensive and/or non-animal tests first.
- Interlink tests for various health effects, e.g., using the same control groups or addressing several endpoints in one animal study (beware of multiple testing).

1.2.8 Pathways of Toxicity (PoT) and systems toxicology

Our scientific understanding of how genes, proteins, and small molecules interact to form molecular pathways that maintain cell function is evolving rapidly. Pathways that lead to adverse health effects when perturbed are referred to as Pathways of Toxicity (PoT). The exploding scientific knowledge of mode of action in target cells, tissues, and organs, driven by advances in molecular and computational tools and coupled with the concomitant development of high-throughput and high-content screening assays, enables interrogation of these PoT and provides a means to study and evaluate the effects of thousands of chemicals. A number of PoT have been identified already; however, most PoT are only partially known, and no common annotation exists. Mapping the entirety of these pathways – a project we have termed the Human Toxome – will be a large-scale effort, perhaps on the order of the Human Genome Project.

The 2007 NRC vision document, *Toxicity Testing for the 21st Century – a Vision and a Strategy* (Krewski et al., 2010), has strongly endorsed the concept of PoT. This vision embraces new high-content, high-throughput, and bioinformatics tools for identifying PoT. Europe and the US have pursued the development of new toxicological tools in very different ways (Hartung, 2010b). The NAS/NRC Tox-21c report calls for a paradigm shift in toxicology. In February 2008, several American agencies, recently joined by the FDA, announced a coalition to facilitate its implementation (Collins et al., 2008): "We propose a shift from primarily *in vivo* animal studies to *in vitro* assays, *in vivo* assays with lower organisms, and computational modeling for toxicity assessments." In *USA Today* of the same day, Francis Collins, now Director of the National Institutes of Health, stated: "(Toxicity testing) was expensive, time-consuming, used animals in large numbers, and didn't always work." In the same article, Elias Zerhouni, then Director of NIH, said: "Animal testing won't disappear overnight, but the agencies' work signals the beginning of the end." Only four years after publication of the NAS/NRC report, we have seen numerous conferences and symposia addressing the report and its implementation, the formation of an alliance of US agencies, and the development of a new EPA toxicity testing strategy in 2009. Depending on the proponent, more or less emphasis is given to technological updates, throughput of testing, costs, replacement of animal testing, or quality of toxicological assessments. There is no doubt that all aspects synergize to bring about a potentially revolutionary change (Hartung, 2008c).

Although a broad discussion has ensued on the design and feasibility of the new toxicity testing paradigm, we are only at the beginning of such a shift. Recognizing that success will require a long-term, concerted effort by many investigators working in a coordinated manner, two NIH institutes (NHGRI, NIEHS), along with EPA and FDA, entered into a formal collaboration in 2009, now known as Tox21. These partners have demonstrated high-throughput screening assays to identify toxicity pathways and are developing computational models and analysis, and informatics tools – all of which can be leveraged for this project.

Although there is not yet a consensus definition for PoT (concepts range from perturbed physiological pathways to adverse outcome pathways, modes of action, or signaling cascades), the general idea is to develop a field of systems toxicology using systems biology as a "role model." Parallel developments in all fields of the life sciences will support this, but toxicology has some features that will help drive its development:

- an urgent need for change
- immediate commercial applications
- reference substances to induce toxicities
- the foundation of (pre-)validated alternative methods from $ 500+ million of research funding
- a culture of Good Laboratory Practice (GLP), Good Cell Culture Practice (GCCP), and validation (and increasingly EBT) for quality control

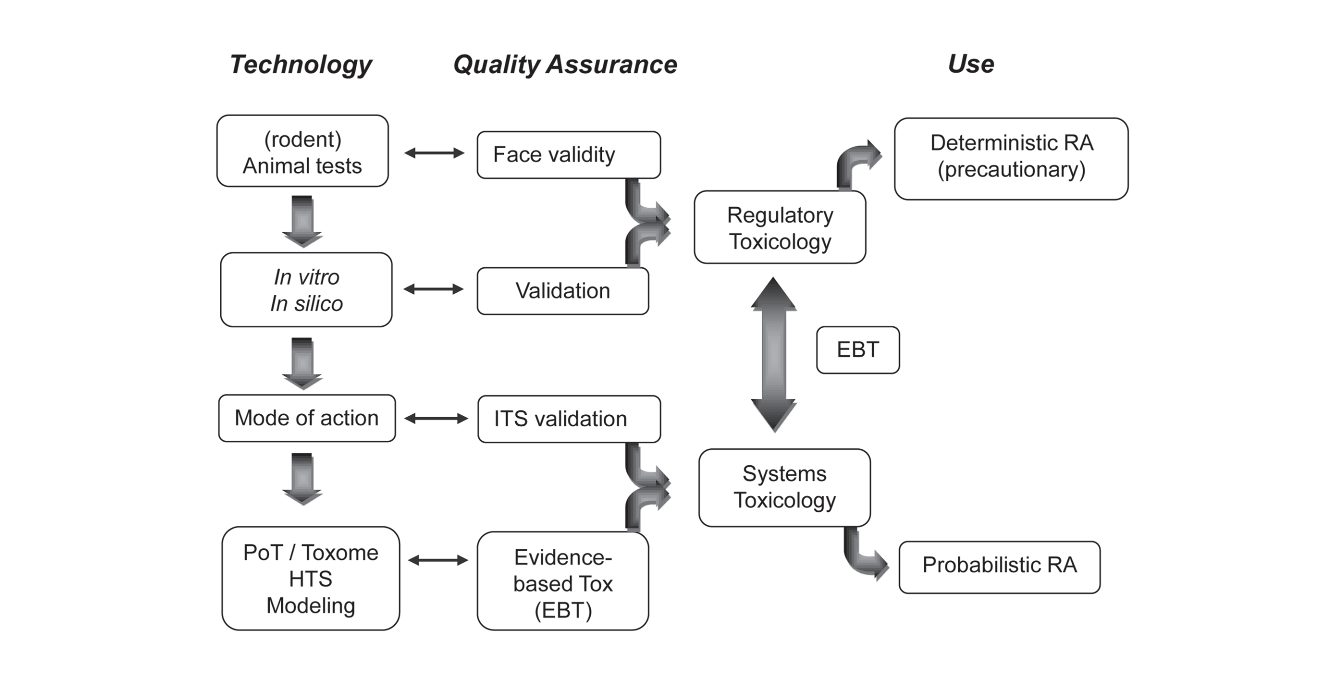

Fig. 1.1: The evolution of toxicology and its quality assurance

Toxicology is increasingly embracing the technologies of the 21st century (Bhogal et al., 2005). The discussion surrounding Tox-21c has accelerated this process, as many have started to develop and commercialize these technologies, which lend themselves to the vision's implementation (van Vliet, 2011). This parallels developments in all life sciences implementing and exploiting the new technologies. Unlike most medical questions, toxicology has the advantage of having a relatively clear start and end to the pathways, i.e., defined substances and hazards, as compared to usually multi-factorial contributors to disease and complex manifestations impacted by individual constellations of the patient.

The basic idea of Tox-21c is to change in the level of resolution. In a nutshell, biochemistry/molecular biology are used to describe phenomena versus physiology/cellular pathology, which, so far, have been used predominantly when discussing modes of action. Figure 1.1 illustrates the larger perspective on the evolution of approaches: Technologies have developed over the last century from animal to *in vitro*/*in silico* and, more recently, mode of action resolution. The concept of Tox-21c is to further refine resolution of analysis to the molecular basis of PoT. These technologies correspond to different quality assurance measures, however, where the validation of ITS (typically built from combining mode of action tests) and Evidence-based Toxicology (EBT) (Hartung, 2009b; Hoffmann and Hartung, 2006; Griesinger et al., 2007) are only emerging. The figure captures how current regulatory toxicology is formed by the earlier technologies, leading to a deterministic (point estimate), typically precautionary risk assessment. The vision is that the new tools of mode of action models, their combination in ITS, and the PoT-based emerging technologies allow the formulation of a Systems Toxicology approach. As discussed elsewhere (Hartung, 2010c), these integrated and information-rich assessments require a shift to a more probabilistic evaluation, where each and every test changes to some extent the probability of a hazard and/or its uncertainty.

The PoT approach represents the continuation of omics by reducing phenotypic characterization ("signatures") to the underlying PoT. This introduces a new quality – that of converting correlations into a hypothesis that can be tested or, in other words, validated. PoT can be manipulated (blocked, triggered) or PoT-specific assays can be designed.

We hypothesize that the number of PoT is finite. This corresponds with the idea that the number of vulnerable targets of a cell (its critical infrastructure) is finite. If this is the case, or at least if a limited number of PoT can cover a large number of agents and hazards, then a comprehensive list of PoT (the Human Toxome) (Hartung and McBride, 2011) will allow us to describe toxic effects at a new level of resolution. We will be able to annotate PoT to cell types, hazards, toxin classes, species, etc., in a manner similar to how we currently annotate (transcribed) genes. It is important to note that the Human Toxome will not be populated by a single test and a single measurement independent of its information-richness, but will require the confirmatory combination of various models and technologies. Pilot projects for endocrine disruptors, funded by NIH, and developmental neurotoxicity, funded by FDA, currently form the basis for the creation of a Pathway of Toxicity Mapping Center (PoToMaC) at Johns Hopkins University.

The identification and use of PoT is the basis for undertaking a revolutionary approach to toxicity testing. Although modern toxicology has identified many modes of action, they have largely remained isolated mechanisms that cannot be broadly applied to sufficient numbers of toxicants to warrant the establishment of dedicated toxicity tests, and they do not yet satisfy regulatory needs. This means that our proposed PoT definition and development of novel test strategies not only initiates a novel test paradigm in general, but will also benefit specific screening programs. It aims to change the general toxicity testing paradigm. The key challenges to this are:

- a harmonized definition, annotation, visualization, and sharing of PoT.
- strategies from systems biology for PoT identification and their validation.
- composition of integrated testing strategies based on these PoT with a definition of adversity and subsequent translation to a risk assessment paradigm.

Mapping the Human Toxome will be a first step towards the development of a Human Toxicology Project. In contrast to the currently used phenomenological "black box" that is animal testing, pathways of toxicity (PoT) will be identified primarily in human *in vitro* systems to provide more relevant, accurate, and mechanistic information for the assessment of human toxicological risk. The ultimate future goal is to bring together a broad scientific community to map the entirety of the Human Toxome.

The concentration at which a substance triggers a PoT will be extrapolated to a relevant human blood or tissue concentration and, finally, a corresponding dose by (retro-) PBPK (physiology-based pharmacokinetic) modeling, informing human risk assessment (Adler et al., 2011). Perhaps more importantly, if a substance does not trigger any of these PoT, it may for the first time be possible to establish the lack of toxicity (i.e., safety) of a substance at a given concentration. This project will need to combine several of the latest emerging technologies in life sciences. Transcriptomics and metabolomics currently are the most advanced technologies for pathway identification, but these are rarely combined to map pathways.

The main difference from ITS is that this approach will operate at the subcellular level and break modes of action and mechanisms down to the underlying pathways or the perturbation of physiological pathways (notably, two very different definitions). The term pathway might be misleading, as we are more likely referring to perturbations of networks. The approach only becomes meaningful if a common annotation of PoT is developed. Hence, a central repository of PoT constituting the (Human) Toxome can be created (Hartung and McBride, 2011). This might serve in the future to identify PoT associated/crucial/amplifying or pathways of defense (PoD) protecting/reversing/dampening a given hazardous effect. The link to classes of substances, cell populations, species, or resulting phenotypic changes will foster the understanding of the specific effect.

The critical question is whether there is a limited number of PoT? It is likely that the number of critical cellular infrastructures is limited, which means that the points of vulnerability, to which the PoT would converge, should also be limited.

Definition of PoT

There is no generally accepted definition of PoT. First, PoT are causal in contrast to adaptive pathways. We might define as overarching Xenobiotic Response Pathways, which include PoT, pathways of defense (PoD) and epiphenomena (EpiP), which do not affect the manifestation of the altered phenotype. Note that EpiP can still serves as biomarkers if triggered consistently with the PoT, but blocking them would not alter the manifestation of toxicity. Three proposed definitions are:

> *PoT are molecularly defined chains of not necessarily linear cellular events stretching from point of chemical interaction to perturbation of metabolic networks and phenotypic change. PoT are causal – either necessary or aggravating – and will typically have a threshold of adversity.*

Or

> *PoT are the formal description of toxic modes of action on the resolution of underlying biochemistry and molecular biology.*

Or

> *PoT are causal links between a given toxicant and its effect in a systems toxicology approach.*

These definitions distinguish PoT by molecular resolution from MoA and by causality from signatures/biomarkers. It leaves open the interactions between different PoT (synergies, leading "pacemaker" PoT, etc.) and of PoT with PoD.

Three very different approaches were taken to explore the concept: ToxCast of the US EPA (Judson et al., 2011; Kavlock and Dix, 2010) uses a broad variety of from the shelf available pathway assays to characterize biological profiles of substances in an HTS manner to associate these with their (mainly animal) toxic profile. The "Hamner approach" (Andersen et al., 2011) selected some known relevant pathways to explore the PoT concept. The approach spearheaded by CAAT (Hartung and McBride, 2011) aims for an unsupervised identification of PoT by omics technologies. The latter was just awarded an NIH Transformative Research grant, "Mapping the Human Toxome by Systems Toxicology," which aims to further define, annotate, and validate PoT as well as create a public database to share PoT from various groups and fields. The consortium includes both The Hamner Institutes for Health Sciences and ToxCast, thus raising the possibility of merging and synergizing the different approaches.

Formally developed alternative methods have one major advantage compared to the research models typically found in the literature: beside their higher degree of standardization and documentation, they need to include a prediction model, i.e., a formal algorithm for deriving predictive results. This means that the level of response indicating adversity is defined. This is rarely the case for tests, which have not been formally evaluated, and where often any significant response is taken as threshold, often rendering the systems overtly responsive. The problem of defining adversity (Boekelheide and Andersen, 2010; Boekelheide and Campion, 2010) can therefore be correlated with the thresholds of the prediction model of the alternative method they were identified in. Alternatively, methods trying to define the point of departure of biological responses are emerging (Judson et al., 2011). However, this is only a first step to finding acceptable methods to distill results from the rich datasets suitable to inform a risk assessment process. A prime example was given in 2010: The quick evaluation of dispersants used for the gulf oil spill disaster (Judson et al., 2010) shows that the new technologies can indeed deliver such information in a timely and cost-saving manner.

Need for probabilistic risk assessment

In order to make use of the novel high-content, high-throughput, and PoT information, we also need to develop ways of distilling relevant information out of the large datasets that will be produced. This requires a radical change from the past: Traditional hazard identification methods have been descriptively based or based on empirical studies, which are resource-intensive and inefficient (see above). Furthermore, empirical studies lack the capacity to detect low probability events, such as those experienced in low dose carcinogenicity. The current deterministic methods are based on point estimates, which are almost always worst-case estimates. In order to improve the transparency, consistency, and objectivity of the assessments, a need for more formal approaches to data integration has been recognized (OECD, 2009).

Three main conceptual requirements for a multi-test decision framework, based on integration of multiple pieces of evidence and a decision-theoretic setting, have recently been formulated (Jaworska et al., 2010). According to the analysis, the framework must:

- be probabilistic, in order to quantify uncertainties and dependencies;
- be consistent by allowing reasoning in both causal and predictive directions;
- support a cyclic hypothesis and data-driven approach, where the hypotheses can be updated when new data arrive.

The formal framework that potentially meets these requirements, allowing for evidence maximization and reduction of uncertainty, can be found in Probabilistic Risk Assessment Networks (PRA). These PRA methods are designed specifically for prospective analysis of the likelihood of low probability events (Greenland, 1998). PRA tools are not new to the risk assessment process (Jager et al., 2001; Verdonck et al., 2005) and they have been used mainly in the derivation of exposure assessment scenarios. The intent is to shift the emphasis of these tools to hazard identification and use PRA to analytically assess the probability that a substance could potentially cause harm. The advantage of PRA is that uncertainties are transparently taken into account, and the cautionary aspect is left to the risk management process. EPA ToxCast has started to develop a risk assessment framework based on high-throughput test systems (HTS) data (Judson et al., 2011) that has kinetic, mechanistic, and uncertainty components. Building on this approach, extending it to high-content (omics) data, and analytically combining the information within a PRA-based Bayesian network, is the logical next step.

Regulatory science is, for practical purposes, bound by the concept of classification and labeling to definitively assign a substance to hazard classes. Science, however, can only deliver probabilities (Hartung, 2010c). This is due to the nature of the underlying data: Biological objects we test are highly variable, and there are other uncertainties associated with diagnostic errors (Hoffmann and Hartung, 2005). This comforts neither the regulator nor the regulated players, as it impedes definitive hazard judgments and the resulting decisions. Tests change the pre-test to a post-test probability of hazard (Aldenberg and Jaworska, 2010; Jaworska et al., 2010; Pepe, 2004), reducing uncertainty. This new understanding analytically refines the initial hazard information. The paradigm change like this will also allow new methods to enter the regulatory arena more easily, as these refined methods are not perceived as a "game-changing," full replacement, but as changers of probabilities. With the successful PRA use in estimates and hazard judgments, its impact will grow and – we hope – eventually become central to hazard testing strategies, simultaneously reducing the costs and time associated with traditional approaches.

It will be necessary to combine the elements of high-information content methods (HIC), HTS, and ITS via PRA. The intent is to identify human hazards prospectively via efficient and effective analytical methods. The basic hypothesis of a PRA-HIC/HTS framework is that the approach provides useful information for current knowledge gaps and also better informs hazard decisions. PRA approaches, historically, have been based on traditional toxicological data (Chen et al., 2007). Here, we suggest using the data coming from HTS and HIC approaches. It is essential to develop a conceptual framework for integration of such test data coming from different sources to allow for integrated and reliable endpoint assessment, which we generally refer to as ITS. Such a decision-analytic framework will yield a more comprehensive basis upon which to guide decisions. A natural outgrowth of this approach is an increased capability to combine and reuse existing data. The integration of such probabilistic hazard information with probabilistic exposure information (van der Voet and Slob, 2007) and probabilistic dose response assessments by PBPK (Kodell et al., 2006) represent logical extensions of this approach. As a result, the goal must be to adapt HTS, HIC, and PRA to better inform hazard decisions of manufacturers and regulators.

Transition in regulatory toxicology

Developing the technologies, however, is only a first step. A possible transition to a new regulatory toxicology based on PoT represents an enormous and multi-faceted challenge (Hartung, 2009d), including:

- Testing strategies instead of individual tests:
  The new PoT approaches will be usable only in combination; however, we have no concept for composing or validating ITS.
- Statistics and multiple testing:
  Multiple testing challenges number of replicates and statistics.
- Threshold setting:
  We need to define adversity.
- What to validate against?
  Since no human data are typically available, and no animal test is replaced one-to-one, only sound science can guide us.
- How to open up regulators for change?
  The comfort zone of the regulators is a major obstacle to change.
- The global dimension:
  No method accepted in one economic area will make a change.
- Quality assurance for the new approach:
  The new technologies require QA from Good Cell Culture Practices, Good Modeling Practices, adaptation of Good Laboratory Practices to Evidence-based Toxicology.
- Validation of the new approach:
  Traditional validation is too slow, costly, and rigid to serve the new technologies.
- How to change with step-by-step developments becoming available?
  The simple incorporation of some new approaches might obscure the need for a fundamental change, but who wants to wait until a completely novel scheme is available?
- How to organize transition?
  There is a need for objective assessment, e.g., by evidence-based toxicology, to assess traditional and novel approaches.
- Making it a win/win/win situation:
  Every stakeholder will not be happy with new approaches that are more complex and more circumspect with regard to certainty of its result. We have to demonstrate the compensatory advantages of better predictivity.

2 A Roadmap for the Development of Alternative (Non-Animal) Methods for Toxicokinetics Testing

Author whitepaper: Harvey Clewell
Respondents: Bas Blaauboer, Olavi Pelkonen
Scientific writer: Mardas Daneshian
Discussants: David A. Basketter, Robert Burrier, Chantra Eskes, Alan Goldberg, Thomas Hartung, Nina Hasiwa, Sebastian Hoffmann, Joanna Jaworska, Ian Kimber, Tom Knudsen, Gavin Maxwell, James McKim, Emily A. McVey, Gladys Ouédraogo, Grace Patlewicz, Annamaria Rossi, Costanza Rovida, Irmela Ruhdel, Andreas Schepky, Greet Schoeters, Nigel Skinner, Kerstin Trentz, Marian Turner, Philippe Vanparys, Joanne Zurlo

2.1 Introduction: toxicokinetics

A recent expert panel review of the available science relevant to the 7th Amendment of the EU Cosmetics Directive's 2013 marketing ban (Adler et al., 2011) analyzed toxicokinetics, among other issues, and concluded that it would take more than five years for the development of methods for estimating *in vivo* kinetics necessary to support risk assessments based on *in vitro* assays for systemic toxicity. The proposed roadmap identifies the key research needed to support quantitative *in vitro*-to-*in vivo* extrapolation (QIVIVE) for systemic toxicity for all chemicals. The common aim of this research is to foster the development of a methodology that incorporates state-of-the-art biokinetic modeling techniques to extrapolate critical concentrations at which *in vitro* toxicity is observed to be equivalent to *in vivo* doses based on the prediction of *in vivo* target tissue dosimetry. Kinetics should not be seen as a separate endpoint; rather, it is a tool to understand *in vitro* toxicity results and properly extrapolate them to human exposure. This methodology will provide a general framework for replacement of *in vivo* animal systemic toxicity assays with alternative *in vitro* toxicity testing.

The aim of classical toxicological risk assessment is to establish safety factors for human exposure based on the evaluation of the outcome of animal tests. The principal concern is finding the dose that causes no toxicologically relevant effect in the animal studies and extrapolating to the no-effect dose in the human under the application of appropriate safety factors. Most of the efforts to replace animal testing with alternative methods have focused on the use of *in vitro* tests for topical toxicity, such as skin and eye irritation (Hartung, 2010a). In contrast to their relatively straightforward application for topical toxicity, the use of *in vitro* toxicology methods as replacements for systemic toxicity testing faces significant challenges. In particular, these studies associate an effect with a concentration in medium rather than with a dose given to the animal, making it difficult to extrapolate the findings to an intact organism. One of the most obvious differences between the situation *in vitro* and *in vivo* is the absence of processes of absorption, distribution, metabolism, and excretion (i.e., biokinetics) that govern the exposure of the target tissue in the intact organism. In addition, metabolic activation and/or saturation of specific metabolic pathways or absorption and elimination mechanisms may also become relevant for the toxicity of a compound *in vivo*. These differences may lead to misinterpretation of *in vitro* data if such information is not taken into account. Therefore, predictive studies on biological activity of a compound require the integration of data on the mode of action with data on biokinetic behavior.

QIVIVE is the process of estimating the environmental exposures to a chemical that could produce target tissue exposures in humans equivalent to those associated with effects in an *in vitro* toxicity test (e.g., an EC_{50}, a benchmark concentration, or an interaction threshold identified by a biologically based dose-response model for the toxicity pathway of concern). Using a combination of quantitative structure-property relationship (QSPR) modeling, physiologically based biokinetic (PBBK) modeling, and collection of *in vitro* data on metabolism, transport, binding, etc., QIVIVE can provide an estimate of the likelihood of harmful effects from expected environmental exposures.

Biokinetic modeling describes the dose and time-dependent absorption, distribution, metabolism, and elimination of a chemical within an organism. Biokinetic models can be divided into two general groups: data-based (classical) models and physiologically-based models (Andersen, 1991; Filser et al., 1995). Physiologically-based biokinetic (PBBK) models are especially useful for *in vitro*-to-*in vivo*, route-to-route, and animal-to-human extrapolations because they incorporate relevant anatomical structures that can be parameterized using independently derived parameters. In contrast to data-based models, PBBK modeling allows the description of the time-course of a compound's amount/concentration at the site of its action. PBBK modeling can contribute to reduction and refinement of animal studies by optimization of study design through identification of critical parameters and timeframes in kinetic behavior (Bouvier d'Yvoire et al., 2007; Clewell, 1993). In addition, PBBK models incorporating QSAR- and *in vitro*-derived parameters, coupled with *in vitro* assays of tissue/organ toxicity, have the potential to replace *in vivo* animal studies for quantitative assessment of the biological activity of xenobiotics (Blaauboer, 2001, 2002, 2003).

The overall goal of this paper is to identify the key research needs to support a viable QIVIVE capability. The research proposed in this paper is considered to be fundamental to the successful use of *in vitro* kinetic data and PBBK modeling for extrapolation of *in vitro* toxicity data to *in vivo*. This research roadmap will specifically address uncertainties in the effect of biokinetics on the estimation of systemic toxicity (both acute and subchronic) of xenobiotics from *in vitro* assays.

2.2 Overview of QIVIVE

Figure 2.1 illustrates a conceptual structure for the use of biokinetic information in the estimation of *in vivo* toxicity from *in vitro* assays. In this scheme, available *in vitro* data on the absorption, tissue distribution, metabolism, and excretion of a chemical are used to parameterize a chemical-specific biokinetic model. In many cases, current quantitative structure-property relationship (QSPR) techniques can be used to estimate chemical properties and kinetics when the specific data for that chemical is lacking. For example, simple empirical correlations have been developed for estimating the tissue partitioning of a chemical from its water solubility, vapor pressure, and octanol/water partitioning (DeJongh et al., 1997; Paterson and Mackay, 1989; Poulin and Krishnan, 1995). In addition, emerging quantitative structure-activity relationship (QSAR) techniques (e.g., knowledge-based systems) and other *in silico* models will become increasingly useful for identifying likely metabolites and predicting potential target tissues for toxicity (Barratt, 2000), so that the appropriate assays of *in vitro* effects can be selected. These target tissue assays then can provide information on the nature and concentration-response of the toxic effects of the chemical.

The complexity of the biokinetic model would depend on the physicochemical and biochemical characteristics of the chemical. A simple one-compartment description of the administered chemical may suffice for many chemicals. Even with such a simple model, it would be possible to estimate the systemic concentrations expected to result from an *in vivo* exposure to a given dose. Thus, the model could be used to relate the concentrations at which toxicity is observed in an *in vitro* toxicity assay to the equivalent dose expected to be associated with toxicity for *in vivo* exposure. Similarly, biokinetic modeling of the *in vitro* toxicity assay can provide important information on the temporal profile of cellular exposure to a free chemical, which can be used in the design of the most appropriate *in vitro* experimental protocol (Teeguarden and Barton, 2004).

The greatest challenge in parameterizing even the simplest biokinetic models is the estimation of metabolic clearance. QSAR algorithms for predicting metabolism parameters have only been developed for a limited number of chemicals, primarily volatile organic compounds that are substrates for CYP2E1 (Peyret and Krishnan, 2011). Thus, it would be necessary to perform *in vitro* assays of the dose-response (capacity and affinity) for metabolic clearance (Houston and Carlile, 1997; Kedderis, 1997; Kedderis et al., 1993; Kedderis and Held, 1996). Eventually, as data accumulates for a large number of chemicals, it may become possible to predict clearance using QSAR approaches. Qualitative prediction of whether a drug is likely to be cleared by metabolism (including the CYP isoenzyme involved) or by urinary excretion on the basis of its physicochemical properties, has recently been demonstrated (Kusama et al., 2010). Of course, there is much more extensive data on drugs than on environmental chemicals.

There are chemicals, of course, for which a one-compartment description would not be expected to be adequate: highly lipophilic compounds, for example, or compounds for which the

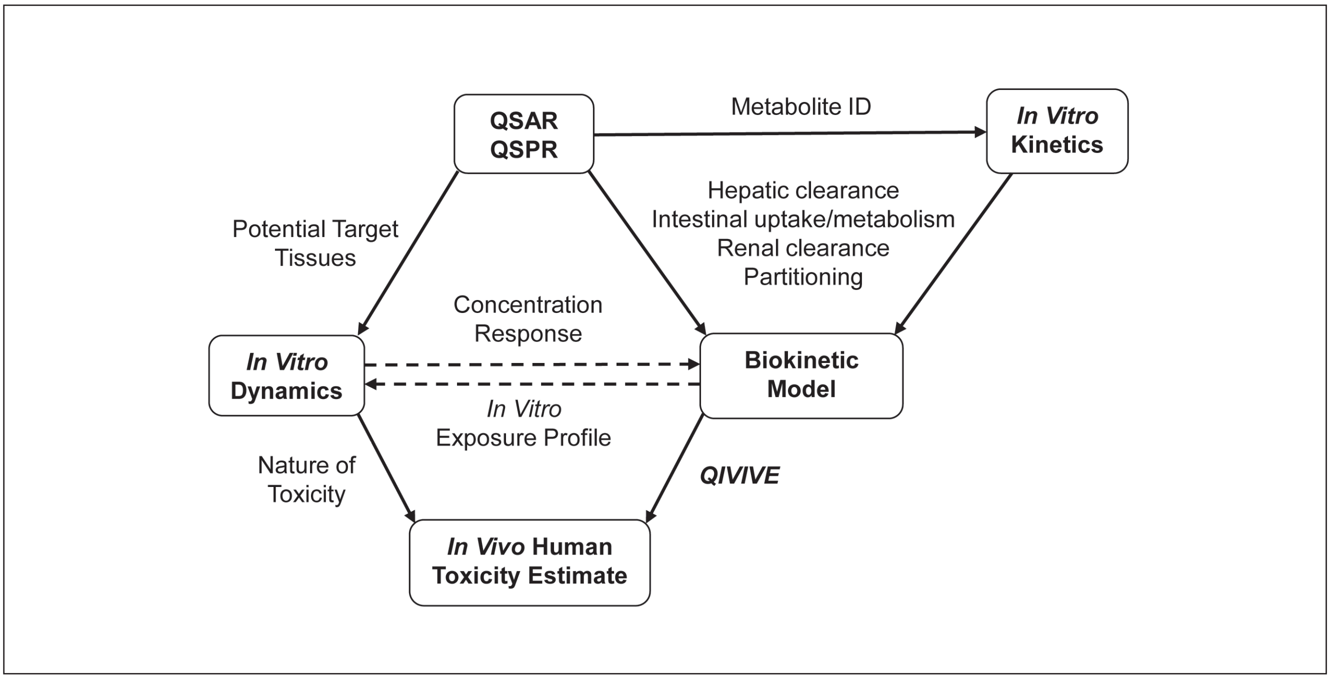

Fig. 2.1: A recommended scheme for incorporation of QSAR (QSPR) information, *in vitro* metabolism data, and biokinetic modeling in the estimation of human toxicity from *in vitro* assays
(adapted from Blaauboer et al., 2001)

toxicity results from a metabolite. The physiological mammalian structure (tissue volumes, blood flows, ventilation rate, glomerular filtration rate, etc.), however, is well characterized (EPA, 1988; Brown et al., 1997), and there is no difficulty describing tissues separately when necessary. As mentioned above, techniques exist for estimating tissue-specific partitioning for many types of compounds. Other data required would also depend on the class of chemical. For volatile chemicals, ventilatory clearance can be estimated from the blood-air partition. For water-soluble chemicals, urinary clearance can be estimated from the glomerular filtration rate or the renal blood flow (for secreted compounds). For some classes of chemicals, it would also be necessary to determine the fractional binding of the chemical to plasma proteins or the partitioning of the chemical into red blood cells.

An important underpinning of this process is that the kind of information necessary for a chemical depends on its structure and physicochemical properties. It seems reasonable to expect that chemicals could be categorized into classes based on their properties, and that this categorization would simplify the process of determining the data needed for a particular compound. This concept is illustrated in Figure 2.2.

PBBK models incorporating QSAR- and *in vitro*-derived parameters, coupled with *in vitro* assays of tissue/organ toxicity, have the potential to replace *in vivo* animal studies for quantitative assessment of the biological activity of xenobiotics (Blaauboer, 2001, 2002, 2003). Target tissues evaluated by *in vitro* assays can be included explicitly in the physiological structure of these models. The models can provide a mechanistic description of barrier functions (gut, bile, kidney, blood-brain barrier, skin, and placenta (if reproductive or developmental toxicity are under investigation)), so that the data obtained from transporter assays could be readily incorporated. Important research areas for *in vitro* methods include the development of validated, stable human hepatocyte systems, as well as *in vitro* systems for key transporters (renal, biliary, etc.). At the same time, QSAR applications need to be developed specifically to provide the kind of information needed by the PBBK models (metabolism constants, binding, etc.). Unfortunately, except in the case of drug-like compounds, the principal limitation in the development of useful QSAR applications appears to be the dearth of suitable data available for training knowledge-based systems. Nevertheless, *in silico* methods are of great interest, and some of them are under development or in the testing phase. They will gain more importance depending on the data, which will be fed in and, therefore, reliable and relevant *in silico* methods are to be expected.

The utility of an approach that integrates cell-based assays with QIVIVE has been demonstrated in the case of acute neurotoxicity for eight chemicals: benzene, toluene, lindane, acry-

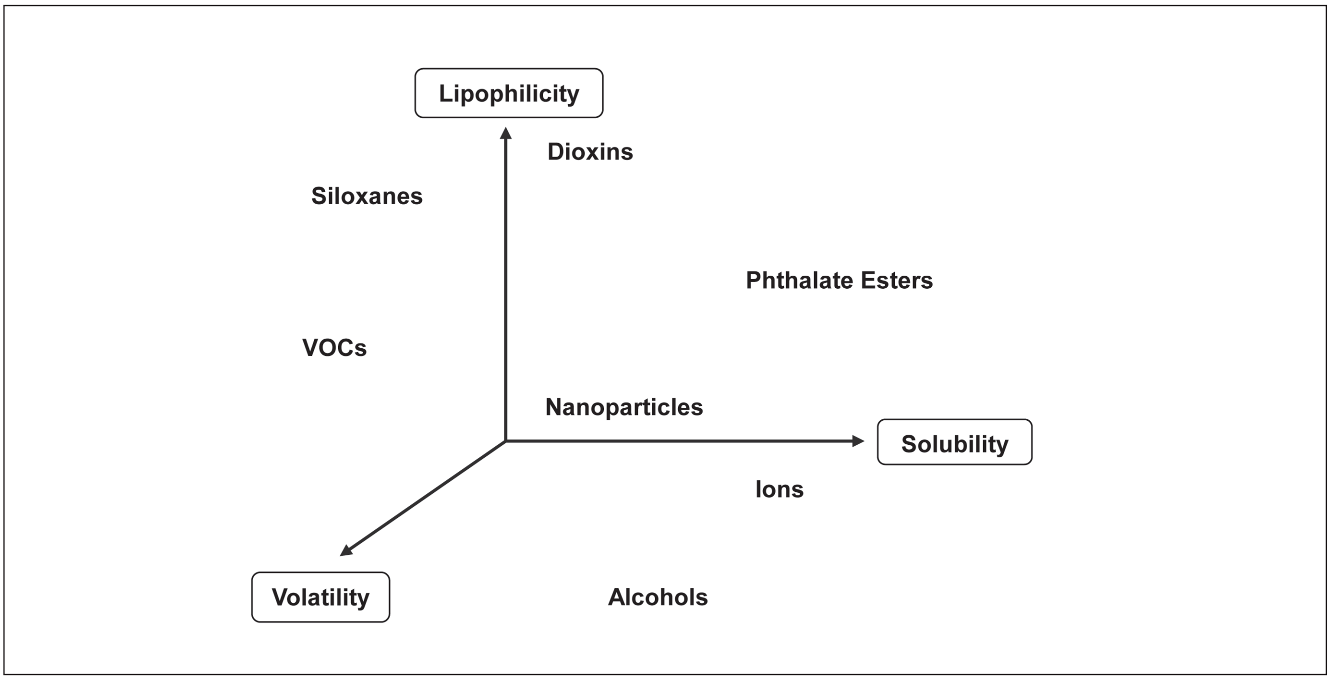

Fig. 2.2: Classification of compounds based on their physico-chemical properties
(adapted from Blaauboer et al., 2001)
In this figure, the key physicochemical properties of a compound include its volatility, water solubility, and lipophilicity. These properties can be thought of as dimensions in which compounds can be categorized. In this way, compounds with similar properties can be grouped, and data for similar compounds can be used to fill gaps in the knowledge of a particular compound. For example, a recent study evaluated the possibility of predicting the *in vivo* kinetics of volatile organic compounds (VOCs) using PBBK models derived solely on the basis of physiological data and QSPR modeling (Liao et al., 2007). The authors concluded that acceptable predictions could be made for inhalation of lipophilic VOCs, such as trichloroethylene, but that the necessary QSPR algorithms were not available for water-soluble VOCs such as acetone.

lamide, parathion/oxon, diazepam, caffeine, and phenytoin (Blaauboer, 2001). The aim of the study was the prediction of acute and subchronic neurotoxicity by integrating PBBK modeling with quantitative toxicity data obtained from non-animal studies. Specifically, the study evaluated the ability of *in vitro* neurotoxicity tests to predict the *in vivo* toxicity of the above chemicals, using PBBK models describing their biokinetic behavior to conduct QIVIVE. Model simulation of the target tissue dosimetry (i.e., the parent brain concentration) formed the basis for the prediction of the compound's systemic toxicity (Cronin et al., 2011) for different exposure scenarios (acute and subchronic). Subsequently, the neurotoxic concentrations estimated in *in vitro* tests (Kuegler et al., 2010; Crofton et al., 2011) could be compared with the brain concentrations simulated by the model. This approach allowed the authors a comparison of the toxic *in vivo* dose known from the literature with the model-predicted dose suspected to cause neurotoxicity. Overall, the results of this study showed that a reasonable prediction of the systemic toxicity could be made for six out of the eight investigated compounds. The discrepancy between the observed and estimated LOELs ranged from a factor of less than two for compounds with low toxicity, to a factor of ten for chemicals of high toxicity (Forsby and Blaauboer, 2007).

2.2.1 Example of a simple QIVIVE approach for parent chemical toxicity

High-throughput *in vitro* toxicity screening can provide efficient identification of the potential biological activity of chemicals. However, the concentrations at which effects are observed in the *in vitro* assays cannot be used to directly evaluate the safety of potential *in vivo* exposures without consideration of bioavailability and clearance of the chemicals (Blaauboer, 2010). Two recent studies evaluated the possibility of applying a simple QIVIVE approach to interpret the results of high-throughput assays conducted under the EPA ToxCast program (Rotroff et al., 2010; Wetmore et al., 2011). In these studies, hepatic metabolic clearance and plasma protein binding were experimentally measured for ToxCast Phase I chemicals. The experimental data were used to parameterize a simple *in vitro*-to-*in vivo* extrapolation model to estimate the human oral equivalent doses necessary to produce steady-state *in vivo* blood concentrations equivalent to *in vitro* AC_{50} (concentration at 50% of maximum activity) or LEC (lowest effective concentration) values in the *in vitro* ToxCast assays.

A simple clearance description (Wilkinson and Schenker, 1975) was used to estimate expected steady-state blood concentrations. The equation assumes zero-order uptake of a daily dose from the gut (assuming 100% oral bioavailability) with both renal and hepatic clearance. The steady-state concentration in the blood is then (see discussion in next section):

$$C_{ss} = \mathrm{ko}/[(\mathrm{GFR}\times F_{ub}) + (\mathrm{Ql}\times F_{ub}\times \mathrm{Cl}_{int}/(\mathrm{Ql} + F_{ub}\times \mathrm{Cl}_{int}))]$$

In this equation, the term GFR×F_{ub} represents the renal excretion of unbound parent compound in blood by glomerular filtration, where GFR is the glomerular filtration rate, which is about 6.7 l/h in human adults (Rule et al., 2004), F_{ub} is the fraction of the drug in the blood that is unbound (free), and ko is the input rate in mg/kg/h. The second term in the denominator is hepatic clearance, where Ql is liver blood flow (typi-

Tab. 2.1: Comparison of *in vitro*-to-*in vivo* extrapolation modeling results with *in vivo* based results
(modified from Wetmore et al., 2011)

Chemical	***In Vivo* Derived** **C_{ss}[a] (μM)**	***IVIVE*** **C_{ss}[a,b] (μM)**	***IVIVE*** **Caco-2[c]** **C_{ss}[a,b] (μM)**	***IVIVE*** **fu=0.99** **C_{ss}[a,b] (μM)**	***IVIVE*** **fu=0.99,** **Caco-2[c]** **C_{ss}[a,b] (μM)**
2,4-dichlorophenoxyacetic acid	9.05-90.05	39.25	40.34	39.25	40.34
Bisphenol-A	< 0.13	0.35	0.40	0.06	0.07
Cacodylic acid	1.80	3.06	–	3.06	–
Carbaryl	0.03	0.07	0.07	0.03	0.03
Fenitrothion	0.03	17.91	–	0.10	–
Lindane	0.46	13.21	–	0.07	–
Oxytetracycline dihydrate	0.36	2.00	0.44	2.00	0.44
Parathion	0.17	24.63	–	0.14	–
Perfluorooctane sulfonic acid	19,990	160.78	179.96	160.78	179.96
Perfluorooctanoic acid	20,120	55.34	58.19	0.40	0.40
Picloram	0.27	57.63	32.01	0.37	0.19
Thiabendazole	0.45	13.76	15.20	13.76	15.20
Triclosan	2-10	1.56	1.59	0.01	0.01

[a] C_{ss}, concentration at steady state for 1 mg/kg/day dose
[b] Predicted using the 1 μM metabolic clearance rate
[c] IVIVE performed incorporating Caco-2 data into the simulation, if available

cally on the order of 90 l/h in adults) and Cl_{int} is the intrinsic metabolic clearance for first-order conditions of metabolism in the liver at low concentrations. Hepatocellular clearance in this study was experimentally determined at 1 μM and the slope of the disappearance of the chemical over time was determined. Clearance was normalized to cell number, with the units μl/min/10^6 cells. *In vivo* intrinsic clearance was estimated by simply multiplying the *in vitro* clearance by the number of cells per gram of liver (roughly 137 x 10^6) and the weight of the liver (about 1820 g in an adult). C_{ss} calculations were performed using an arbitrary dose of 1 mg/kg/day. The Simcyp simulation platform (Rostami-Hodjegan and Tucker, 2007) was used to perform Monte Carlo analysis to simulate variability across a population of 100 healthy individuals of both sexes from 20-50 years of age. A coefficient of variation of 30% was used for intrinsic and renal clearance. Reverse dosimetry was then used to generate oral equivalent doses according to the following formula:

$$\text{Oral Equivalent Dose (mg/kg/day)} = AC_{50} \text{ or } LEC/C_{ss}$$

For a small number of these chemicals, it was possible to find *in vivo* biokinetic data to estimate a steady state concentration at an exposure of 1 mg/kg/day for comparison with the *in vitro* predictions. The results of the comparison are shown in Table 2.1.

For comparison purposes, two alternative hepatic clearance assumptions (Wilkinson and Schenker, 1975) were employed: (1) restrictive hepatic clearance (assuming only unbound chemical is available for clearance), using F_{ub} determined experimentally; and (2) non-restrictive hepatic clearance (assuming all of the chemical is available for clearance), where the F_{ub} was set to one. Riclosan is an example of a chemical that appears to have restrictive clearance, while picloram appears to have non-restrictive clearance, and the behavior of lindane appears to be intermediate between the two extremes. In general, the assumption of restrictive clearance produces a more conservative (higher) estimate of C_{ss}. In fact, these two clearance assumptions represent extremes bracketing the possible relationship between chemical disposition/transport and hepatocellular metabolism that can result in C_{ss} estimates that differ by several orders of magnitude. Hepatic clearance is complexly determined by a number of factors, including liver blood flow, the association and dissociation rates for binding of the chemical to plasma proteins such as albumin, the kinetics of hepatocellular uptake of the chemical, and the kinetics of hepatocellular metabolism. Indeed, no approaches have yet been demonstrated to predict the fraction of compound available for metabolism, even in the case of drugs.

The assumption of 100% oral bioavailability is conservative from a human health standpoint because lower absorption results in a higher oral dose required for achieving a specific C_{ss}; however, incorporation of Caco-2 assay data on bioavailability into the QIVIVE model can increase the predictivity of the C_{ss}

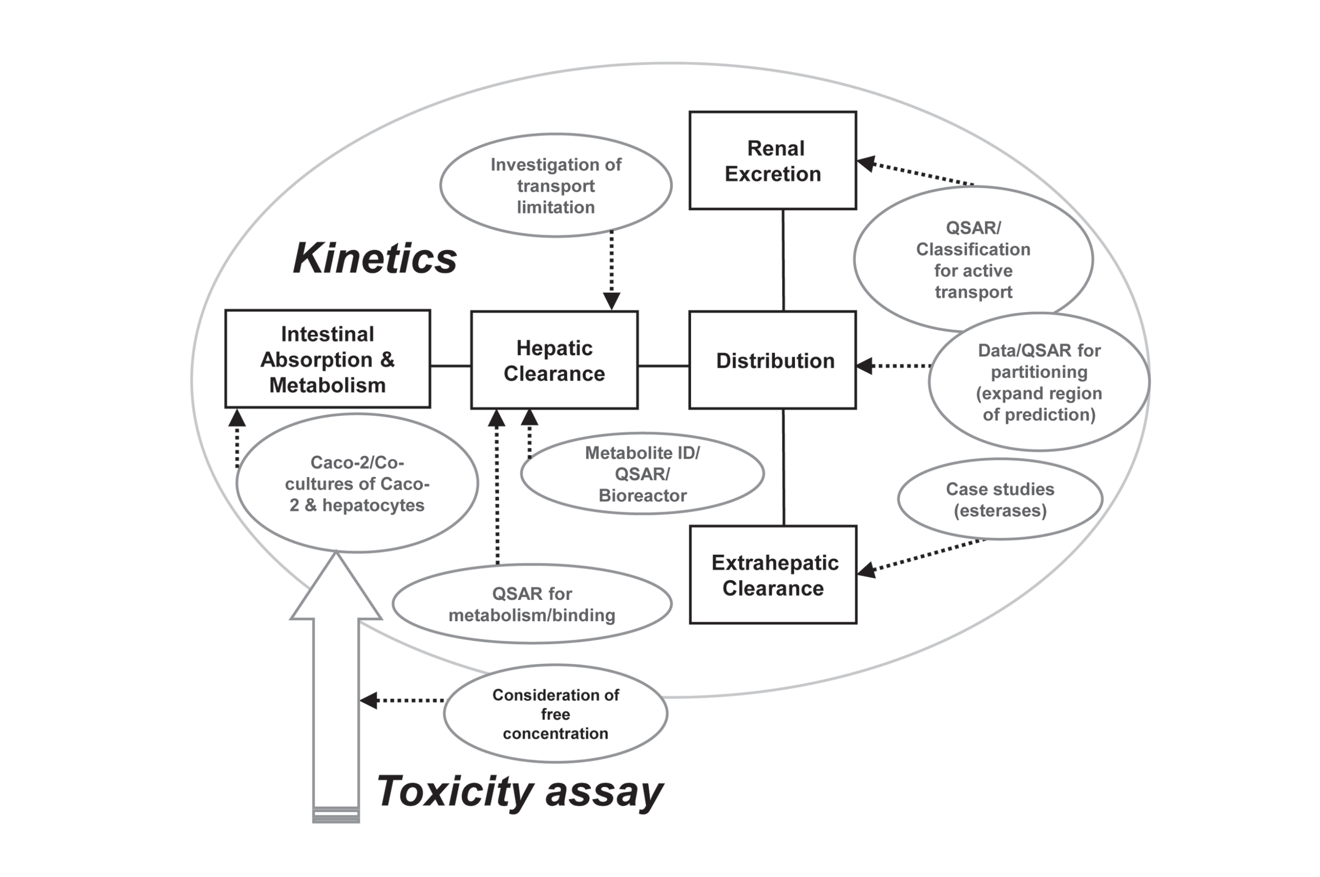
Kinetics
Investigation of transport limitation
Renal Excretion
QSAR/ Classification for active transport
Intestinal Absorption & Metabolism
Hepatic Clearance
Distribution
Data/QSAR for partitioning (expand region of prediction)
Caco-2/Co-cultures of Caco-2 & hepatocytes
Metabolite ID/ QSAR/ Bioreactor
Case studies (esterases)
Extrahepatic Clearance
QSAR for metabolism/binding
Consideration of free concentration
Toxicity assay

Fig. 2.3: Key research needed to support QIVIVE

determination, as in the case of oxytetracycline dihydrate and picloram in Table 2.1. On the other hand, the assumption that renal clearance is solely a function of F_{ub} and the GFR is not necessarily conservative, since active renal resorption would result in a higher C_{ss} at a given dose, as in the case of the two perfluorinated chemicals in Table 2.1.

Other limitations of this simple approach include:

- The analysis is predicated on the assumption that blood concentrations equivalent to the nominal *in vitro* AC_{50} or LEC values would produce equivalent responses *in vivo*. However, the concentration of free chemical in an *in vitro* assay that elicits a certain response may differ from the nominal AC_{50} value due to factors such as protein-lipid composition of the media and binding of the chemical to surfaces (Blaauboer, 2010).
- The biokinetics and bioactivity were only evaluated for the parent compound. No attempt was made to evaluate biological activities and dosimetry of metabolites.

2.2.2 Example of a QIVIVE approach for toxicity of a metabolite

A couple of publications by Punt and colleagues (Punt et al., 2008, 2009) present an example of a more sophisticated QIVIVE approach using metabolism data collected in a number of subcellular fractions. Although the intent of the study was to evaluate the relevance of carcinogenicity of estragole reported in high-dose animal studies to human exposure situations, a similar QIVIVE approach also could be applied for the interpretation of *in vitro* toxicity assays. The key metabolism parameters to be estimated were rates of multiple biotransformation reactions that determine the level of carcinogenic species (1-sulfooxyestragole) in the liver. Due to the complexity of metabolism steps involved in formation of the ultimate carcinogenic metabolite of estragole as well as detoxication of parent compound and other intermediate metabolites, the approach using a combination of subcellular fractions along with different cofactors was more valuable for the purpose of their modeling than using a more integrated system such as hepatocytes (Punt et al., 2008, 2009). By manipulating cofactors such as NADPH, UDPGA, NAD^+, and PAPS in the selected *in vitro* system of microsomes or S9, multiple steps of estragole metabolism mediated by CYPs, UGTs, dehydrogenases, and SULTs, respectively, could be characterized. The rates of those reactions were used to describe the critical metabolism pathways in estragole bioactivation and detoxication. Those reactions were described well with Michaelis-Menten kinetics and the resulting V_{max} and K_m parameters were scaled to *in vivo* based on the microsomal or S9 protein content. The interplay of these multiple reactions was integrated in the PBBK model and the simulated concentrations of two estragole metabolites in the rat and human urine were reasonably consistent with the observed *in vivo* data considering the purpose of the modeling was to evaluate the dose-dependent changes in bioactivation, not to predict the absolute dose metrics (Anthony et al., 1987; Punt et al., 2008, 2009). The estragole QIVIVE case also demonstrated that the capability to evaluate the relative importance of extrahepatic metabolism in different exposure conditions is one of the advantages of using *in vitro* metabolism data over *in vivo* experiments.

2.3 Research gaps

The subsequent sections of this paper will attempt to elucidate the key research areas needed to support QIVIVE for assessing risks on the basis of *in vitro* toxicity data, including:

- Improving the accuracy of *in vitro* toxicity assays by determining the free concentration of chemical instead of simply using the nominal concentration
- Extrapolating *in vitro* kinetic results to estimate *in vivo* clearance
- Obtaining parameters for PBBK models to perform QIVIVE

The proposed key research areas are summarized in Figure 2.3 and Table 2.2.

2.3.1 Characterization of free concentration

The free concentration of a chemical drives both its kinetics and dynamics (Mendel, 1992). The concentration of free chemical in an *in vitro* assay that elicits a certain response may differ

Tab. 2.2: Kinetics research needs to support *in vitro* based risk assessment

Research Area
Characterization of free concentration in cell-based assays • binding • metabolism • active transport
In vitro models • concurrent intestinal absorption/metabolism • dermal absorption • blood/brain barrier • hepatocyte clearance • pathway/metabolite ID/kinetics (organotypic) • renal clearance • respiratory clearance • placenta in repro and developmental models • zebrafish
Data collection development of *in silico* approaches • metabolite identification • protein binding in cell-based assays • tissue partitions (some classes of compounds) • restricted vs unrestricted hepatic clearance • metabolism rates • gut absorption/metabolism (non-druglike compounds) • transporter substrates/renal clearance
QIVIVE case studies • classes of physicochemical properties • different metabolism pathways • parent vs stable metabolite vs reactive metabolite • portal of entry vs liver vs remote toxicity
Development of generic PBPK modeling platforms • user friendly, open access • database for physiological parameters • inhalation, dermal, and oral exposure • multiple parallel metabolic pathways

from the nominal concentration (added amount of chemical divided by volume of the medium) due to factors such as protein/lipid binding in the medium (Gulden et al., 2002; Seibert et al., 2002), evaporation, precipitation, and adherence of the chemical to surfaces (Blaauboer, 2010). To determine the *in vivo* plasma concentration expected to elicit a target-tissue response similar to the cellular response in the *in vitro* assay, the free fraction must be determined in both the *in vitro* and *in vivo* exposures (Gulden et al., 2006; Gulden and Seibert, 2003; Teeguarden and Barton, 2004). To the extent that the cells in the *in vitro* assay are representative of the cells in the *in vivo* target tissue, equal free concentration in the medium and plasma will be associated with the same intracellular exposures (Gulden et al., 2001).

Protein binding can be a key determinant of disposition (Gulden and Seibert, 1997), affecting compound availability for uptake into cells *in vitro* as well as target tissues *in vivo*. For example, the use of whole serum or serum albumin in cell-based assays can greatly alter the apparent dose-response for cellular toxicity compared to serum-free media (Hestermann et al., 2000; Brunner et al., 2010). A high fraction bound also gives rise to concerns regarding potential competitive binding by other compounds that could modulate the free concentration (Teeguarden and Barton, 2004). Methodologies to estimate protein binding and approaches for the description of the kinetics of binding in biokinetic models have been areas of intense interest over the past four to five decades. Consideration of protein binding faces two parallel challenges: first, when compounds are bound in media or capillary blood, what fraction should be regarded as available for transport into cells or tissue, and, second, how does the binding influence medium/cell or blood/tissue partitioning.

In general, medium, cells, blood, and tissues all will contain free and bound forms of the compound. For equilibration, only the free compound diffuses across the medium/cell or plasma/tissue interface, and at equilibrium the free concentration on both sides of the interface is expected to be equal (except in the case of active transport). However, the equilibrium relationship of the concentration in cells or tissues compared to the medium or plasma is typically described with empirical partition coefficients based on measurements of total concentrations of the compound. Differential binding, therefore, will influence apparent partitioning. However, there are quite a number of different determinants of apparent partitioning, complicating the interpretation of such data:

- Partitioning due to lipophilicity
- Plasma binding
- Tissue binding
- Active transport
- Clearance processes
- Blood:plasma ratio

The blood:plasma ratio is needed for converting tissue:plasma partitions to tissue:blood, or fraction unbound in the plasma to fraction unbound in the blood (Yang et al., 2010).

Furthermore, the application of analytical techniques is considered a prerequisite for proper QIVIVE. The workshop participants agreed that their use (as opposed to nominal concentration) is critical and their importance not enough appreciated.

2.3.2 In vitro estimation of intestinal absorption and metabolism

To accurately predict the systemically available dose of the chemical, it is important to consider potential metabolism at the portals of entry in addition to the hepatic metabolism. Despite its importance as a modifier of oral bioavailability, intestinal metabolism has received less attention than other extrahepatic metabolism. The mucosal epithelium of the gastrointestinal (GI) tract contains substantial amounts and types of xenobiotic-metabolizing enzymes, among which CYP3A enzymes have been the focus of a great deal of research in pharmaceuticals due to their role in causing reduced oral bioavailability and as a major source of inter-individual variability resulting from variable constitutive expression of gut CYPs and potential drug-drug interactions (Paine et al., 1997). From the risk assessment point of view, other phase I and II enzymes in the GI tract should also be carefully considered in IVIVE. In addition to the liver, the GI tract is a key site for hydrolysis of a number of ester compounds of environmental concern that are used in pesticides and consumer products, including pyrethroids, phthalates, and parabens (Kluwe, 1982; Crow et al., 2007; Imai, 2006). The significance of intestinal phase II metabolism to total chemical clearance is another factor to be considered in IVIVE of metabolism. Intestinal glucuronidation of BPA demonstrates the importance of consideration of intestinal metabolism to provide key information for describing BPA biokinetics for human health risk assessment based on *in vitro* metabolism information (Mazur et al., 2010).

Compared to IVIVE of hepatic metabolism data, there are several challenges in extrapolating *in vitro* intestinal metabolism parameters to *in vivo*. First, the intestine is not a homogenous organ and therefore spatial differences are evident in distribution of metabolizing enzymes within the mucosa as well as along the length of the intestine (van de Kerkhof et al., 2007). This factor makes it difficult to interpret and extrapolate *in vitro* metabolism parameters obtained from intestinal tissue-driven *in vitro* systems such as microsomes and S9 fractions (van de Kerkhof et al., 2007). Intestinal cell lines such as Caco-2 (Karleta et al., 2010) have been used to determine absorption parameters *in vitro* (Sambuy et al., 2005), but use of these cell lines as a surrogate for metabolism in the GI tract is problematic due to differences in enzyme expression compared to human intestinal tissue (Imai et al., 2005; van de Kerkhof et al., 2007). Another complication comes from the fact that intestinal metabolism often is greatly influenced by chemical flux into the enterocytes, i.e., intestinal metabolism is closely related to the uptake/absorption process, making it difficult in terms of both measurement and interpretation of the results (Paine et al., 1997; Yang et al., 2007). More studies are warranted to develop better *in vitro* tools to predict intestinal metabolism, and then better extrapolation strategies can be developed based upon the relevant *in vitro* metabolism data for coherent extrapolation considering the interplay with chemical absorption processes in the intestine.

2.3.3 In vitro determination of dermal exposure

For environmental and cosmetic chemicals, the dermal route of exposure is highly likely. Therefore, *in vitro* assays should be further developed to predict the rate of dermal penetration and metabolism in the skin. The challenge in predicting accurate dermal uptake and metabolism is similar to that for intestinal absorption, in that absorption and metabolism are competing processes. Human skin contains both CYP enzymes (Storm et al., 1990) and esterases (Prusakiewicz et al., 2006), which can be of importance for presystemic clearance of a compound as well as for generation of toxic metabolites if the skin is a target tissue.

2.3.4 In vitro estimation of metabolism

The success of IVIVE is largely dependent on the quality and relevance of *in vitro* metabolism data (Coecke et al., 2006). There have been significant improvements in the quality of human tissue preparation in recent years, as well as parallel advances in application strategies of those *in vitro* data to predict *in vivo* kinetics (Chiba et al., 2009; Gomez-Lechon et al., 2007; Houston and Galetin, 2008). These advances have made it possible to implement QIVIVE for PBBK models during drug development (De Buck and Mackie, 2007; Pelkonen and Turpeinen, 2007; Rostami-Hodjegan and Tucker, 2007). For pharmaceutical compounds, however, the screening of new chemical entities involves evaluation of whether the candidate possesses drug-like properties, including relatively moderate metabolism and inactive metabolites. Thus, IVIVE for drug metabolism has focused largely on metabolic stability screening to inform the drug's half-life and oral bioavailability using the clearance model (Pelkonen and Turpeinen, 2007; Houston and Galetin, 2008). For this type of IVIVE, linking the total intrinsic clearance *in vitro* in conjunction with the unbound fraction in blood and the liver blood flow to predict *in vivo* clearance has been the most common practice (Fagerholm, 2007; Houston and Galetin, 2008)

Although the experience built upon drug data can be applied to the IVIVE approach for chemicals, the challenges in QIVIVE for chemicals are different from those for pharmaceuticals, primarily due to the wider range of chemical properties compared to drugs. There is also a greater need to consider the role of metabolism in determining chemical toxicity. For chemicals, IVIVE should preferably be conducted at the level of an individual enzyme/metabolic pathway primarily responsible for formation of the active species or depletion of the active parent compound instead of measuring total intrinsic clearance of the parent chemical. The apparent limitation of applying total clearance-based IVIVE to chemicals has its difficulty in describing the formation and clearance of toxic metabolite(s). Another limitation arises from dealing with the broader range of exposure concentrations and routes for chemicals compared to a narrower/targeted concentration range and oral route for drug candidates. IVIVE issues will also vary depending on the kinds of enzymes involved in chemical metabolism. Both chemical properties and knowledge of mechanism of action inform which metabolic pathways would be primarily responsible for chemical metabolism. This information can serve as criteria for categorizing chemicals into subgroups for different strategies based on primary metabolic enzymes.

Despite the fact that esterases are known to play an important role in metabolizing many environmental chemicals, including pesticides and endocrine active compounds, they have not been well studied compared to CYPs and other phase I and II enzymes. To describe the role of esterases on detoxication of the chemical, it is necessary to include extrahepatic metabolism, most representatively the metabolism in blood due to the presence of carboxylesterases and other types of esterases. Metabolism in the GI tract and skin should also be characterized to estimate esterase-mediated detoxication capacity in the body (Prusakiewicz et al., 2006).

2.3.5 Identification of the key metabolism pathways and toxic moieties

To be performed correctly, QIVIVE requires information on what the active entity would be in the target tissue based on the potential mechanisms of toxicity. Predicting primary metabolic pathways, along with the potential for producing active metabolites, could be supported by *in silico* approaches such as QSAR (Kusama et al., 2010). Knowledge built on drug data showing the role of chemical properties in metabolism, binding, and partition would help this categorization. To determine the extent and design of *in vitro* metabolism assays aided by such tools, the criteria for this classification should be based on major pathways of metabolism, since that is the key information needed in designing *in vitro* metabolism studies for IVIVE.

2.3.6 Organotypic models of in vivo hepatic function

Ensuring realistic metabolism, both qualitatively (the types of metabolites formed) and quantitatively (the relative amounts of these various metabolites) is one of the most difficult challenges in QIVIVE. One possible direction for meeting this challenge is the development of organotypic hepatic systems (bioreactors) that appropriately reflect the complexity of *in vivo* hepatic function. In principle, these systems could be used to provide data for *in silico* modeling of both kinetics and dynamics (Sung et al., 2010). However, a number of difficulties will need to be overcome: (1) developing screening analytical chemistry methods that would allow rapid evaluation of metabolites produced and excreted from the cells or cell aggregates in culture, (2) development of stable organotypic liver cultures that recapitulate *in vivo* metabolism for sequential or parallel metabolic networks, and (3) ensuring metabolic competencies, both metabolite production and parent and metabolite loss, from the tissue culture system by metabolism or routes of non-specific loss, such as renal excretion.

These organotypic hepatic cell cultures could be used to rapidly assess metabolism and confirm QSAR predictions of likely metabolites. Metabolites that were identified in some significant yield might themselves be studied in the *in vitro* test systems. In the past, analytical methods development was tedious and time-consuming. It is possible, however, that this process could be accelerated with modern methods of higher throughput analytical chemistry. The ultimate goal would be the development of tissue cultures or hepatic bioreactors (Seagle et al., 2008) that include recirculation and medium replenishment over time to mimic an *in vivo* situation.

The ability to assess metabolism by examining effluent compounds from the culture systems could be coupled with other metabolic analyses to evaluate fidelity between the *in vivo* and *in vitro* pathways. A well-designed liver bioreactor could function in a fashion similar to isolated-perfused liver preparations (Bessems et al., 2006). Analysis of metabolites produced in a bioreactor might also serve to benchmark expected metabolic pathways. Evaluation of the fidelity of the bioreactor and new organotypic systems could be verified by assessing metabolite profiles with specific test compounds, i.e., using compounds whose metabolism has already been well-studied *in vivo*.

It may be necessary to develop co-culture systems or microfluidic systems that maintain metabolism, recirculation, continuous addition of test compound, and ongoing loss from the culture system. The microfluidic, body-on-a-chip design (Maguire et al., 2009) has potential for creating custom *in vitro* toxicity evaluations for multiple cells plated onto different parts of the microfluidic plate. This system requires more development, especially to move from a laboratory research device to low to medium throughput. The system was designed based on PBPK model structures developed by Shuler and colleagues (Esch et al., 2011). Another possibility might be to have a relatively large hepatic bioreactor and to divert flow to multiple chambers with various cell types for *in vitro* testing. The cells would have continuous flow of the bioreactor fluid, and the effluent from the culture plates could be collected and re-circulated to the bioreactor. While these designs are not yet available, they are technically within reach.

2.3.7 Possible strategy to determine metabolites

Techniques to estimate the concentration of a substance at the site of action include both direct and indirect ones: biomarkers, microdialysis, imaging, mass spectrometry, and simulations by modeling (Pelkonen et al., 2008). Special emphasis should be placed on the use of advanced bioreactors (Darnell et al., 2011), including relevant cell systems, e.g., HepaRG cells, to mimic the appropriate metabolism combined with analytical methods. Mass spectrometry, in particular, has proven to be an optimal tool to determine metabolites. As Pelkonen and co-workers state (Pelkonen et al., 2009) "…in silico *or* in vitro*, in conjunction with animal data, provide useful and necessary information, on which to base the first PK studies in humans. The prerequisite is to use appropriate and up-to-date techniques and biological preparations*." The final goal would be to build a virtual human to model the whole process a compound undergoes in the human body to enhance drug development and improve risk assessment. A starting concept, taking into account available techniques, can be found in Figure 2.4.

2.3.8 In vitro estimation of renal clearance

The state of the art for *in vitro* models of renal clearance is not as advanced as in the case of liver clearance, although some progress has been made in the case of drugs (Kusuhara and Sugiyama, 2009). The relative spatial complexity of renal tubular transport systems compared to the more homogenous hepatic

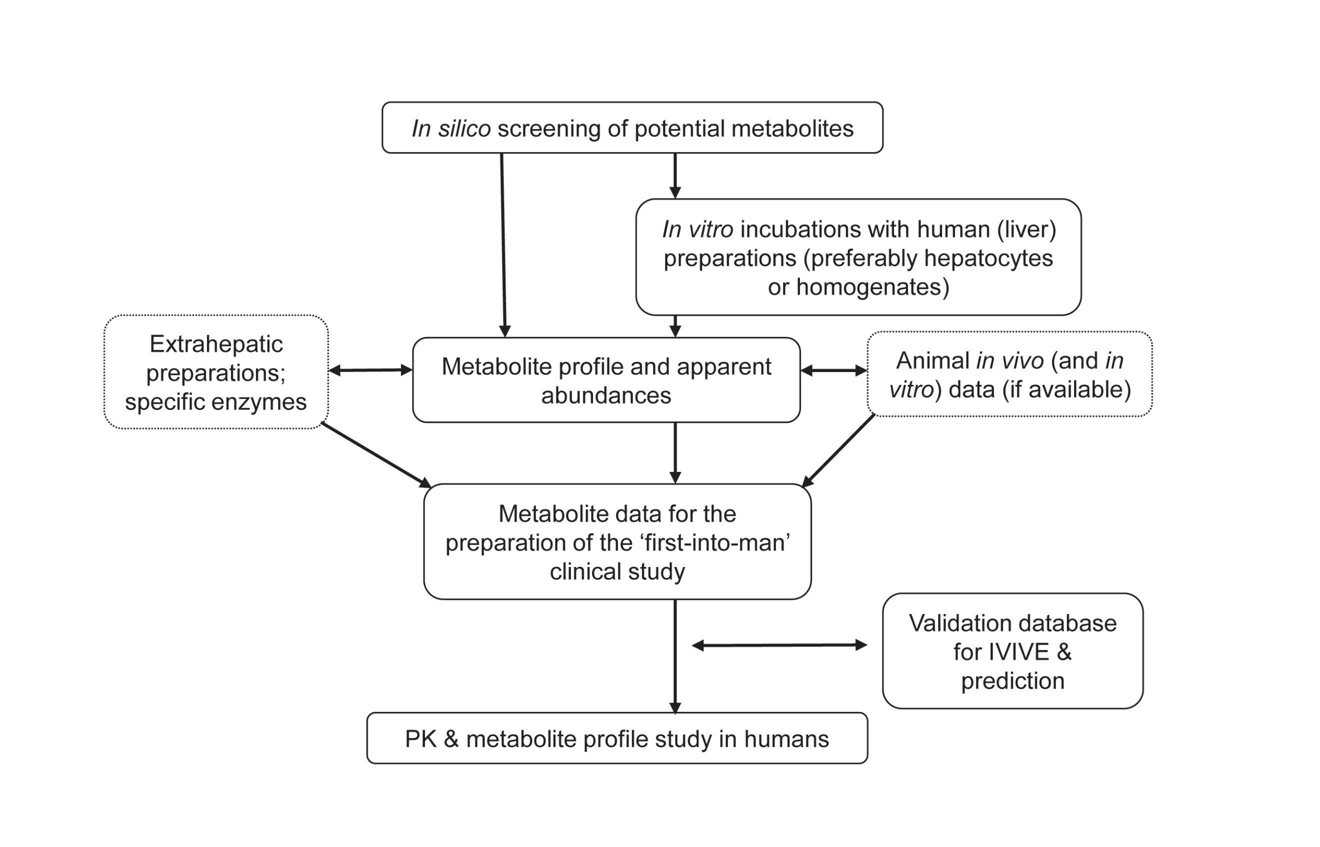
In silico screening of potential metabolites
In vitro incubations with human (liver) preparations (preferably hepatocytes or homogenates)
Extrahepatic preparations; specific enzymes
Metabolite profile and apparent abundances
Animal in vivo (and in vitro) data (if available)
Metabolite data for the preparation of the 'first-into-man' clinical study
Validation database for IVIVE & prediction
PK & metabolite profile study in humans

Fig. 2.4: Proposed strategy to assess metabolite effects in *in vitro* studies
(adapted from Pelkonen et al., 2009)

architecture greatly increases the difficulty of developing representative *in vitro* model systems. However, it should at least be possible to develop assays to identify whether a compound is a substrate for a particular transporter (Yang et al., 2009). This would provide an indication of the likelihood that a compound's renal clearance might deviate from expectations based on glomerular filtration. A similar capability could be developed for assessing biliary clearance.

2.3.9 PBBK model development

The parameters in a PBPK model can be categorized into four types: exposure, physiological, partitioning, and metabolism. The exposure parameters are determined solely by the characteristics of the exposures and the physiological parameters are available from the literature (Brown et al., 1997). These types of parameters are not chemical-specific, and the values used in the evaluation of an untested compound would be the same as those used for well-characterized compounds. Partitioning and kinetic parameters, however, are chemical-specific and need to be estimated for untested compounds. A number of software platforms are available to support generic PBPK modeling for pharmaceuticals using *in vitro* metabolism data, as exemplified by the Simcyp platform (Jamei et al., 2009; Rostami-Hodjegan and Tucker, 2007). Because these generic platforms are designed to support modeling of drug compounds, their focus is on oral and intravenous exposures, and on metabolism by oxidative (CYP) and conjugative (UGT) enzymes. Effective use of these software platforms for PBPK modeling of environmental and personal care compounds would require enhancements in two areas: (1) addition of descriptions of dermal and inhalation exposure, and (2) addition of data on esterase metabolism enzymes. In the field of environmental risk assessment, PBBK models typically have been developed for individual chemicals. Although generic modeling platforms are available for some classes of compounds, e.g., MEGen (Loizou and Hogg, 2011), the development of generic models has not been as extensive as in the pharmaceutical area. A useful generic modeling platform would include the following features:

- user-friendly, open access
- database for physiological parameters
- inhalation, dermal, and oral exposure routes
- capability for multiple parallel metabolic pathways

2.3.10 Integrated testing strategies (ITS)

Toxicokinetics and the methods mentioned already should not be understood as stand-alone methods or endpoints. Kinetics is a tool to understand and modify any *in vitro* result and should be incorporated into testing strategies as a requirement for any extrapolation to *in vivo*. In general, an integrated testing strategy should consist of information about the physicochemical properties of a substance, the structure activity relationships (QSARs), *in vitro* data, and kinetic and dynamic modeling. All these factors combined should then lead to an evaluation against *in vivo* data (Dejongh et al., 1999). Experimental research, computational methods, and integrated testing strategies should be developed in an interactive way between the researcher and analysts, so they will be able to augment each other. As recently explained by Jaworska and Hoffmann (2010) via the concept of Bayesian networks, the structure of the testing strategy matters and will influence the risk assessment process. Complex networks will not lose rare but important events or small but multiple perturbations in key nodes (Jaworska and Hoffmann, 2010).

2.4 Conclusions and recommendations: toxicokinetics

The main objective of this chapter was to determine the research needs for developing a methodology to incorporate *in vitro* kinetic data into *in vivo* biokinetic models to support risk assessments based on cellular toxicity assays. The proposed methodology starts with the identification of the critical aspects of the metabolism of a compound for the intended purpose of the risk assessment. This preliminary information includes a combination of qualitative metabolism studies and selected *in vitro* toxicity assays to identify the active species and primary metabolic pathways responsible for producing and detoxifying the toxic entity. Current examples of IVIVE often rely, in part, on existing *in vivo* data. As experience with IVIVE accumulates, however, it will become increasingly possible for such information to be gained from *in silico*-based prediction tools and targeted *in vitro* kinetic studies, particularly using organotypic *in vitro* systems that better mimic *in vivo* conditions.

The current state of the art presents an excellent opportunity for development of improved *in vitro* ADME methodologies. The technologies necessary to support these initiatives are now coming to maturity, and the need for rapid toxicity testing of both drugs and commercial chemicals is becoming more acute. Recent advances in stem cell biology may allow the development of custom bioreactors with more relevant cellular components and allow the bioreactor to serve as both a metabolite generator and a test system for the toxicity and biological responses of molecules.

Recommendations: toxicokinetics

General but indispensable:

1. For the extrapolation of an *in vitro* assay to *in vivo* the measurement of the free chemical concentration is absolutely necessary.
2. *In vitro* biokinetics should be taken into consideration to improve the quality of *in vitro* toxicity data.
3. The use of kinetic parameters to correlate *in vitro* effective concentrations to a dose is absolutely essential.
4. Quality training data for the wide range of chemical property classes should be made available, particularly for "non-druglike" compounds.
5. Analytical methods and computational modeling should be taken into account and employed wherever possible.

Further research areas:

6. Improved *in vitro* models are needed, particularly in the areas of intestinal and dermal absorption and the associated presystemic metabolism. Hepatic, renal, and respiratory clearance also are of special interest. Organotypic culture kinetics and metabolite identification should be investigated. Barriers should be taken into account by appropriate *in vitro* assays.
7. The development of generic PBPK modeling platforms should be furthered. They should be user-friendly and have open access, with a database for physiological parameters. They should be able to simulate inhalation, dermal, and oral exposure, allowing description of multiple parallel metabolic pathways.
8. Standard methods for the characterization of the free concentration in cell-based assays should be developed, including the features of binding, metabolism, and active transport into the cell.
9. In the area of data collection and *in silico* approaches, metabolite identification and protein binding in cell-based assays should be addressed. Furthermore, restricted versus unrestricted hepatic clearance should be investigated. Data should be collected for hepatic and renal clearance, metabolism rates, gut absorption, and metabolism, especially for non-drug-like compounds. Also, identification of transporter substrates is a potential area of *in silico* modeling.
10. QIVIVE case studies, with special emphasis on those that did not work, should be performed for different classes of physicochemical properties, different metabolism pathways, toxicity from parent versus stable metabolite versus reactive metabolite, and portal of entry versus liver versus remote toxicity.

Special research areas:

11. The very promising areas of *in vitro* bioreactors and the microfluidic human-on-a-chip should be further developed and standardized.
12. High-Throughput Toxicity Screens combined with kinetics data should be further investigated.
13. An equivalent to the Lipinsky rules for drugs should be developed for chemicals.

3 A Roadmap for the Development of Alternative (Non-Animal) Methods for Skin Sensitization Testing

Authors whitepaper: Ian Kimber, David A. Basketter
Respondents: Joanna Jaworska, Gavin Maxwell, Grace Patlewicz, Erwin Roggen, Andreas Schepky
Scientific writer: Costanza Rovida
Discussants: Bas Blaauboer, Robert Burrier, Harvey Clewell, Mardas Daneshian, Chantra Eskes, Alan Goldberg, Thomas Hartung, Nina Hasiwa, Sebastian Hoffmann, Tom Knudsen, Paul Locke, James McKim, Emily A. McVey, Gladys Ouédraogo, Olavi Pelkonen, Annamaria Rossi, Irmela Ruhdel, Greet Schoeters, Michael Schwarz, Nigel Skinner, Kerstin Trentz, Marian Turner, Philippe Vanparys, James Yager, Joanne Zurlo

3.1 Introduction: skin sensitization

Allergic contact dermatitis resulting from skin sensitization is an important occupational and environmental health problem. Many hundreds of chemicals are known to cause skin sensitization, and allergic contact dermatitis is the most common manifestation of immunotoxicity in humans. It is important, therefore, that skin sensitization hazards/risks of new chemicals and products be evaluated accurately. In fact, toxicologists have methods and models that provide a reliable basis for the identification of skin sensitizing chemicals, assessment of relative sensitizing potency, and the development of effective risk assessments. However, current practices rely heavily on animal models of skin sensitization, particularly the local lymph node assay (LLNA), the preferred method for safety assessment.

There are now compelling reasons to develop novel approaches to skin sensitization testing that do not require the use of experimental animals. There has been – and continues to be – a very substantial investment in achieving this goal. Several promising methods are now undergoing validation. However, no non-animal test has yet been formally validated. Against that background, the purpose of this chapter is to provide a partisan (some might say biased!) view of the current development of alternative methods for assessment of skin sensitizing activity, to reflect on some of the challenges that we face in delivering novel testing strategies, and to provide a view of what is needed to improve and accelerate progress towards this objective.

A quick glance at the title of this chapter will trigger a feeling of *déjà vu* in many readers. Indeed, our own reaction would have been "Does the scientific literature really need another review of skin sensitization and the development of alternative methods for hazard and risk assessment?" The answer must be no – surely there is nothing new to say. There are already sufficient – or more than sufficient – reviews and overviews available. Some of those are cited here, and collectively they provide a comprehensive record of past and present activity in this area (Adler et al., 2011; Aeby et al., 2010; Basketter and Maxwell, 2007; Basketter, 2008; Basketter et al., 2007; Basketter and Kimber, 2011; Bauch et al., 2011; De Silva et al., 1995; dos Santos et al., 2009; Hartung et al., 2011; Kimber et al., 1999b, 2001, 2010, 2011; Martin et al., 2010; Maxwell and Mackay, 2008; Patlewicz et al., 2007; Reuter et al., 2011; Ryan et al., 2005; Vandebriel and van Loveren, 2010). However, our intention here is not to provide yet another scholarly review article. It is rather to provide a personal perspective with regard to the development of alternative methods for skin sensitization testing, how the current landscape appears to us, and what we believe are the requirements for (and distractions from) achieving real success in addressing this challenge. This is not, therefore, a consensus document. One important aim is to be provocative, to excite argument and discussion. If what follows is at odds with the views of others, we apologize in advance for any perceived criticism and ask only that the article be viewed as a stimulus for informed debate.

We have chosen to address this issue by tackling a number of relevant questions for an assessment of the current state of the art of alternative methods for skin sensitization testing and what future needs might be. These are as follows:

1. *What is it that we are really trying to achieve – what will success look like?*
2. *Is the international scientific community marshaled in the right way to make real progress in this area?*
3. *What should be the future research imperatives?*
4. *Skin sensitization testing in vitro – can we do it already?*
5. *Is hazard identification alone good enough?*
6. *What needs to change?*

3.2 What is it that we are really trying to achieve – what will success look like?

There is a genuine need to identify chemicals that have the potential to cause skin sensitization and allergic contact dermatitis (ACD) and to accurately assess the likely risks to human health. Many hundreds of chemicals cause skin sensitization, and ACD is a common occupational and environmental disease (Febriana et al., 2011). Historically, the identification of contact allergens relied on the use of guinea pig tests (Buehler, 1965; Magnusson and Kligman, 1969). More recently, the murine local lymph node assay (LLNA) has found favor as a method that, compared with guinea pig tests, offers important animal welfare benefits (Kimber et al., 2002, 2011). The LLNA provides a generally robust and reliable means of identifying chemicals that have the potential to cause skin sensitization but also permits a characterization of relative sensitizing potency, which is required for the development of accurate risk assessments (Api et al., 2008; Basketter et al., 2007; Rovida, 2011). However, it has to be acknowledged that, as with all test methods, the LLNA is not without limitations.

There are strong scientific, ethical, and legislative reasons why it is important to ensure that opportunities to Reduce, Refine, and Replace the use of experimental animals in research and investigative studies are exploited quickly and effectively. Probably the most appropriate code of practice when considering the use of animals in research is to pose a number of questions: (a) is the issue being addressed legitimate and important? (b) is it possible to address the question effectively without the use of experimental animals? If the answer to the latter question is no, then the third question is; (c) what is the best and most appropriate experimental design to ensure that the principles of the 3Rs are adhered to and a robust answer achieved?

Translating this to the theme of this article, the assumption is that it is not currently possible to conduct an evaluation of the skin sensitizing activity of chemicals without the use of animals (either guinea pig tests or the LLNA) that is fully accepted by regulators. Although this assumption will be tested further under question 4, it is certainly the case that no validated, non-animal methods for the identification of skin sensitizing chemicals are currently available. As a consequence, and in line with the scientific, ethical, and legislative imperatives (such as the 7th Amendment to the Cosmetics Directive in the EU) that are driving interest in developing non-animal methods, there has been a very substantial investment in alternative strategies for skin sensitization testing.

A detailed survey of the many approaches is unnecessary here, and more information is available from the review articles cited above. The important point, however, is that alternative methods, in most cases, are based upon an attempt to identify biological properties or structural motifs of chemicals that are believed to be required for the acquisition of skin sensitization. The palette of strategies that has been considered is based on the understanding that for a chemical to cause skin sensitization a number of things have to be achieved, or biological/biochemical hurdles must be cleared. These include, but are not limited to: (a) the chemical gaining access to the viable epidermis, (b) the stable association of chemical with protein to form a complete antigen, (c) the activation, mobilization, and migration of cutaneous dendritic cells (DC) for transport of antigen to regional lymph nodes, and (d) the activation within lymph nodes of responsive T lymphocytes. Another approach has been to apply a systems approach to model *in silico* the key chemical and biological pathways that drive the induction of sensitization (Maxwell and Mackay, 2008).

For the most part, the alternative methods being explored (*in vitro* and *in silico*) are based upon evaluation of the ability of a chemical to provoke one (or more) of these required events or processes. That strategy appears appropriate, but it is crucial to bear in mind that the ability of a chemical to clear any one of those biological or (bio)chemical hurdles does not necessarily mean that it should be classified as a skin sensitizer. Does the ability of a chemical to cause covalent bonds with protein necessarily signify that it will cause sensitization? Similarly, is there any reason to believe that if a chemical is capable *in vitro* to cause the activation of DC that this property alone will be sufficient to translate into skin sensitizing activity *in vivo*? Of course, failure to negotiate any one of the hurdles will lead to the absence of sensitizing activity.

It will be clear that the move from an animal model to a non-animal method is fraught with difficulties and challenges. Put simply, mice used within the LLNA represent integrated biological models that incorporate, in a fully coordinated and physiologically relevant way, all the events and processes that are needed for the acquisition of sensitization. So, if a chemical is positive in the assay, then the assumption is that it has successfully achieved all that a chemical must accomplish to drive sensitization.

In a similar vein, it is estimated currently that approximately 25% of contact allergens must be activated by air oxidation, or within the skin, to acquire the chemical reactivity necessary to associate with proteins. It has proven difficult to achieve effective incorporation within *in vitro* models of appropriate and adequate oxidative/metabolic capacity.

The move from a fully integrated biological (animal) model is, therefore, necessarily going to be challenging. Nevertheless, that is the aim: to develop a method that will allow characterization of skin sensitizing activity without the use of animals and possibly with a higher level of confidence.

It is worth reflecting briefly on what is meant by the above phrase: "*characterization of skin sensitizing activity*." The first step in any toxicological investigation is the identification of hazard, and in the context of this article, to discriminate between chemicals that do, and chemicals that do not, have the potential to cause skin sensitization. Although that is an important step (and the aspect for which formal validation is required), it is not sufficient for addressing the likelihood that exposure to the chemical under any given set of circumstances will result in an adverse effect. For an effective risk assessment, an appreciation of relative potency is necessary. In the context of skin sensitization, this equates to the amount of chemical encountered on a skin surface that is required for the induction of sensitization. This, of course, is true for all forms of toxicological evaluation, but it is particularly important with regard to skin, because it is known that contact allergens vary by up to five orders of magnitude in relative skin sensitizing potency.

A measure of relative potency is provided by the LLNA because it is known that the readout used (the proliferation of draining lymph node cells) not only has a causal relationship with the acquisition of sensitization but also correlates quantitatively with sensitizing activity (Kimber and Basketter, 1997). Achieving an understanding of relative potency with a full *in vitro* approach is going to be very challenging. This is a theme that we will return to in Chapter 2.6. Drawing together the elements discussed above, the answer to the first general question we posed is as follows: *The aim is to develop a non-animal method(s) that will provide a means of identifying contact allergens with an accuracy at least equivalent to, or approaching that of, the preferred animal models*. While that is the minimum requirement, ideally, any novel method should improve upon the performance and reliability of the LLNA (considered the best animal model) and simultaneously provide an accurate assessment of relative potency.

Success, therefore, will be the development of a non-animal approach that provides a basis for reliable hazard and risk assessments of at least comparable accuracy to those afforded by the

LLNA. If that is what success looks like, then a frequently asked question is whether it is reasonable to expect that a single *in vitro* approach or test method will provide the required level of certainty. Some speculate that it will be necessary to develop a suite of methods that collectively provide a basis for making judgments about sensitizing potential. This sounds sensible, but relying on a battery of assays may prove to be technically demanding and experimentally unwieldy.

3.3 Is the international scientific community marshaled in the right way to make real progress in this area?

The most important opportunities in applied toxicology, including the development of alternative test methods applicable in an industrial context and for regulatory purpose, derive from an investment in "pure" research and an improved understanding of relevant cellular and molecular mechanisms. We mentioned above that there has been a huge investment, particularly in Europe, directed at promoting the development of alternative methods for skin sensitization testing. There is no suggestion that the motives driving this investment are anything other than laudable, but it is relevant to question whether the focus of that investment is the most likely to deliver the necessary breakthroughs. There has been too much emphasis placed on supporting applied research, often focused narrowly on the design and development of new methods. In recent years, too many new approaches have been proposed. In most cases, however, few of them have proven useful for a workable, full-replacement strategy.

Why have we seen a change of emphasis in skin sensitization research from characterization of the mechanisms through which cutaneous immune and inflammatory responses are induced and orchestrated, to new test development? A number of factors have influenced this:

– Everyone wants to develop a test. Even a superficial survey of platform and poster presentations at toxicology conferences reveals an ever-increasing number of papers describing attempts to develop "alternative" test methods for the identification of skin sensitizing chemicals. Naturally, many of these communications have merit, but not all. It is evident that some investigators do not have a clear understanding of what "alternative" is really required nor what is required by those charged with making decisions about the safety of new chemicals or products.
– Research follows the money, and it is clear that many investigators have found it necessary to develop more research themes focusing on applying current knowledge for test development to attract funding.
– In certain commercial sectors, and particularly the cosmetics industry (because of the deadlines imposed by European regulation), there is a very clear and pressing need to develop non-animal methods so that new innovation can be supported when it is no longer possible to use animal tests.

A case is not being made that all currently supported research in skin sensitization is without value. Indeed, there have been important achievements. However, there is now an imbalance between fundamental mechanistic research and research driving the application of the test by industry and regulatory authorities on the one hand, and research applying established knowledge focused on the design of new test methods. If so many investigators are applying their skills to the development of alternative predictive tests, where will the transformational research that will provide the basis for really innovative developments in the future come from?

As an illustration, one area of investigation that has attracted considerable interest in the context of new approaches to skin sensitization testing is dendritic cell (DC) biology. There are a variety of proposed test methods based on the use of cultured DC, or DC-like cell lines, the theory being that exposure of such cells to skin sensitizing chemicals, but not to non-sensitizers, will provoke functional or phenotypic changes that will serve as biomarkers of sensitizing activity and as readouts for cell-based assay systems (An et al., 2009; Arkusz et al., 2010; Ashikaga et al., 2010; Ouwehand et al., 2010; Python et al., 2007; Johansson et al., 2011). There is no doubt that some of these assays perform rather well and show real promise. Nevertheless, the approach is predicated on a (largely unproven) assumption that the impact of chemical allergens on DC or DC-like cells in culture (usually at cytotoxic concentrations) is reflective of the changes induced in epidermal Langerhans cells (LC) and dermal DC during the initiation of sensitization in intact skin. However, despite much enthusiasm for this approach, and despite considerable investment that has supported a wide range of proposed assay methods, very little is known about how chemicals cause changes in cultured DC, or what relevance, if any, this has to the acquisition of sensitization. In the rush to develop a *test* or a *new method* intriguing and important questions about fundamental mechanistic biology are ignored – or at least put aside.

3.4 What should be the future research imperatives?

There needs to be a realignment of "pure" and "applied" research in skin sensitization, with increased emphasis on exploring some of the important uncertainties and intriguing unknowns. Among the many issues that remain to be clarified are the following:

– The balance achieved in the skin and regional lymph nodes between the immunostimulatory, promotional, and regulatory signals delivered by discrete populations of DC, and how that balance impacts the acquisition of sensitization.
– The role of regulatory T cells (T_{reg} cells) in controlling and constraining the induction of skin sensitization, the elicitation of ACD, and the relationship of T_{reg} cells with effector T lymphocytes in determining the net vigor and quality of immune responses to contact allergens.
– The influence of the ways in which haptens interact with target proteins on the development of skin sensitization.

This list is merely indicative, certainly not exhaustive. Other investigators doubtless will be drawn to other research questions. However, the common thread in the examples highlighted above is that they each have the potential to inform our understanding of the factors that govern sensitizing potency. The balance achieved between stimulatory and regulatory signals from DC, the balance achieved between effector T lymphocytes and T_{reg} cells, and the impact of the kinetics and selectivity of the interaction of chemical allergens with skin proteins are all strong candidates with regard to influencing sensitizing potency.

Other exciting research themes could also be identified. There is no shortage of relevant challenging and exciting areas of research in skin sensitization. Addressing issues such as those listed above will not necessarily lead directly and immediately to the identification of alternative test methods, but there can be no doubt that the increased understanding of relevant immunobiological mechanisms that would result from such research would drive innovation and open up the development of new strategies.

Sensitizing potency, however, is currently the key challenge. It is now understood that contact allergens vary by up to five orders of magnitude with respect to their relative skin sensitizing potency. In practical terms, this means that with potent chemical allergens only very low levels of exposure are required for the development of sensitization, whereas with weak allergens repeated high-level exposure may be required for sensitization to develop. The phenomenon is clear, but we really do not understand why this is – and what specific factors govern potency. The question is of great academic interest but is also of considerable importance in the development of new test methods. If an *in vitro* assay is going to provide useful information about relative potency, then there will need to be readouts that correlate quantitatively with sensitization and are reflective of dose-response relationships.

At a relatively simplistic level it is clear that the extent to which sensitization is acquired is associated with the vigor of T lymphocyte responses in regional lymph nodes draining the site of exposure to the inducing chemical allergen (Kimber and Dearman, 1991; Kimber et al., 1999a). This is not unexpected, because skin sensitization is mediated by T lymphocytes, and the greater the level of proliferation in draining lymph nodes the larger will be the pool of antigen-responsive T cells. However, as alluded to above, it has to be acknowledged that, in addition to the extent of clonal expansion, the effectiveness of skin sensitization will likely be impacted by the quality of the T lymphocyte response. One qualitative aspect of that response is the balance between effector T lymphocytes that will drive the elicitation of ACD, and T_{reg} cells that will down-regulate and constrain sensitization. In addition, the overall effectiveness of sensitization may be influenced by the "breadth," or clonal diversity, of the T lymphocyte response.

One can conclude, therefore, that the main influence on the effectiveness of skin sensitization will be the quantity and quality of the T lymphocyte response generated. However, this does not provide any indication of the events induced following encounter with a contact allergen that shapes the response. There are several factors that individually, or in concert, may impact the vigor and quality of the T lymphocyte response. These include: (a) the speed with which the chemical reaches the viable epidermis, (b) the nature of "danger signals" elaborated, (c) the kinetics of association with target proteins, (d) the promiscuity of chemical interaction with proteins, either in terms of number of proteins with which adducts are made, and/or at the level of amino acid selectivity, (e) the kinetics of LC and DC activation and mobilization, (f) the balance achieved between activated epidermal LC and activated dermal DC, and (g) the amount of antigen delivered to draining lymph nodes.

Again, the options listed above are not exhaustive, and there may be a number of other events that influence overall potency. Nevertheless, it is the case that we really have little understanding of how events induced in the skin and regional lymph nodes in the minutes and hours following topical encounter with a contact allergen shape the T lymphocyte response that will be induced. Tackling this question will not be easy, but a research investment in this area might elucidate the pivotal events and processes that determine the effectiveness of immune responses to contact allergens – and may also provide an appreciation of how the vigor of immune responses in general is controlled. Certainly, an understanding of the key events that impact on skin sensitizing potency would be of enormous value in considering novel approaches to testing that would deliver not only hazard identification but also an assessment of relative potency. In conclusion, therefore, the proposal is that there should be a greater investment in tackling some of the important questions remaining about the way in which the acquisition of skin sensitization is induced and orchestrated. In addition, a high priority should be given to investigation of the chemical, biochemical, and immunological events that determine the relative potency of skin sensitizing chemicals.

3.5 Skin sensitization testing *in vitro* – can we do it already?

There are currently no formally validated methods for the (hazard) identification of contact allergens using non-animal methods. But an important and interesting question is this: Leaving aside considerations of validation and regulatory acceptance, are we already in a position to make sound judgments regarding the skin sensitizing potential of chemicals? If we were to effectively marshal our collective know-how about skin sensitization, together with access to data generated by selected *in vitro* approaches, would it be possible to achieve something approaching a 90% overall accuracy of prediction of skin sensitizing activity – a performance in line with the LLNA or better? Such a hypothetical scenario would have two parts.

The first of these would be to bring together a small group of seasoned investigators with experience of skin sensitization and making judgments about the sensitizing potential of chemicals. This expert panel would include those with expertise in QSAR and aligning sensitizing potential with structural motifs and physicochemical properties. That expertise could be supplemented by access to one or more expert systems that seek to predict skin-sensitizing activity as a function of chemical structure.

The second element would be the availability of data generated by selected *in vitro* tests, albeit test methods that have not yet been validated (although those mentioned are in the latter stages of formal evaluation). There are several assay systems from which to choose, including the following: peptide binding assays (Gerberick et al., 2004, 2009; Troutman et al., 2011) and based upon the Nrf2/Keap 1 electrophile-sensing pathway the KeratinoSens assay (Emter et al., 2010); and the CeeTox Assay (McKim et al., 2010), as well as a variety of cellular assays based on the use of cultured DC or DC-like cells (Aeby et al., 2010; Ashikaga et al., 2010; Reuter et al., 2011; Sakaguchi et al., 2006; Johansson et al., 2011). It would be interesting to evaluate, prospectively and with an unbiased set of chemicals, how such an expert panel (with access to QSAR models and data derived from selected *in vitro* tests) would fare compared with the LLNA or with human ACD.

The foregoing is not a *cri de coeur* for the use of unvalidated tests in the safety assessment process (which would be inappropriate). Rather, it should be viewed as a reflection of how near we perhaps are to being able to identify skin sensitizing chemicals without recourse to animal experiments – if there is a willingness to align the experience and expertise which is already available with the outputs of selected test methods.

3.6 Is hazard identification alone good enough?

If resources are marshaled carefully, the identification of skin sensitizing hazards, without the need for animal tests, should be a realistic goal. Hazard identification might be sufficient to satisfy regulatory requirements (and, of course, where there is no skin sensitization hazard, further work will be unnecessary). However, just as safety evaluation cannot be completed solely on the basis of exposure data, the absence of information about relative potency for identified skin sensitization hazards will not support the development of accurate risk assessments or enable meaningful risk management. This is important because over several decades of the implementation of regulatory identification of skin sensitization hazard, there is no evidence of any impact on the clinical burden of allergic contact dermatitis. One could go so far as to argue that efforts to develop effective risk assessments based largely or solely on exposure data are doomed to failure, as it is rarely possible to link specific exposures with the development of allergic contact dermatitis.

One answer, therefore, is to make a greater research investment in the expectation that a more complete understanding of the immunology and biochemistry of skin sensitization will disclose the pivotal events in determining potency. Another approach is to consider how information deriving from currently available non-animal models for sensitization testing might be used to rank chemical allergens according to potency. Previous exercises have explored how, in theory at least, it might be possible to derive an estimate of overall potency by integrating information from two or more of several *in vitro* approaches. One strategy explored was to assign chemicals scores on the basis of whether there was a structural alert, and on relative activity in *in vitro* tests configured to measure the ability of chemicals to: (a) gain access to the viable epidermis, (b) form stable associations with peptides or proteins, (c) stimulate the activation/maturation of DC or DC-like cells, or (d) provoke proliferative responses by cultured T lymphocytes. In some cases the scores were binary (that is 1 or 2; for epidermal bioavailability and structural alerts). For other readouts a scale of 0 to 5 was used. Based on this paradigm, the relative sensitizing potential of a chemical would be calculated as the product of individual scores (Basketter and Kimber, 2009; Jowsey et al., 2006).

The above is a pragmatic approach and has not been tested adequately in practice; there has been only a single partial attempt – by Natsch et al. (2009). Nor does it claim to distinguish between threshold events and those that bear a direct quantitative relationship with sensitizing potency. Nevertheless, it does at least provide a framework for how some assessment of potency might be informed by the use of readouts from *in vitro* tests combined with appropriate SAR analyses.

It is relevant here also to highlight that a number of other strategies have been proposed for informing skin sensitizing potency without recourse to animals. These include the development and deployment of appropriate mathematical models (Maxwell et al., 2011; Maxwell and Mackay, 2008), the development of an Integrated Testing Strategy (ITS) for skin sensitization in the form of a Bayesian Network (Jaworska et al., 2011; Maxwell et al., 2011; Maxwell and Mackay, 2008) and a tiered approach combining a keratinocyte-based test for identifying skin sensitizers and an epidermal equivalent-based potency test (dos Santos et al., 2011; Galbiati et al., 2011).

Notwithstanding the crafting of theoretical frameworks for considerations of potency that may or may not work in practice, the answer to the question posed in this section is that a non-animal solution to hazard identification alone (although being a significant achievement) is insufficient for a full safety evaluation and risk assessment. This view is clearly reflected by the recent expert review of a European Expert Group (Adler et al., 2011).

3.7 What needs to change?

Against the background of the issues described above, we have listed what we believe to be the most important changes that are required to promote the development of effective non-animal methods for the assessment of skin sensitizing activity. In no particular order the key issues are:

- The need for a realignment of skin sensitization research so that there is greater emphasis on exploring basic mechanistic aspects in the expectation that this will yield information and understanding that, in time, will provide a platform for real innovation and, hence, new ground-breaking solutions.
- Such a research investment would provide a much clearer understanding of the factor(s) that serve to determine the potency of skin sensitizing chemicals.
- The need to bring a greater realism to some within the scientific community who are seeking to develop novel test methods. It needs to be understood and appreciated that for a new method to be valuable it has to be technically robust, perform reliably, and offer the required level of predictive performance.
- In tandem with the above, there needs to be a greater willingness among some test developers to take a more dispassionate approach to the evaluation of putative tests. There is a need for a critical evaluation of the strengths and limitations of novel methods compared with existing *in vivo* models, even with their limitations.
- For the evaluation and validation of new methods (of whatever type) there is a need to evaluate specificity, sensitivity,

and overall accuracy with a gold standard dataset. In this case, that translates into a dataset that is populated by chemicals where there is sound evidence for the presence or absence of significant skin sensitizing potential in humans.
- Finally, there is a critical need for a general acknowledgement that the complete replacement of animal methods (such that safety assessments remain at least as effective as they are currently) requires that alternative approaches inform both hazard identification and assessment of potency. At present, formal validation activity addresses only the first of these.

3.8 Conclusions and recommendations: skin sensitization

The purpose of this chapter was to provide a critical and partisan appraisal of the current landscape with regard to skin sensitization testing. There is no doubt that there have been considerable achievements. Peptide binding assays continue to evolve and appear very promising. Cellular assays based on induced responses by DC, DC-like cells, and other cell types have considerable momentum currently, and there are three such assays currently undergoing formal validation in Europe. Efforts continue with the development and evaluation of SAR paradigms. New opportunities based upon an appreciation of the activation of the Keap 1/Nrf2 pathway are being explored (McKim et al., 2010). The KeratinoSens assay recently has completed an inter-laboratory evaluation and has been submitted for formal validation (Andreas Natsch, personal communication). In addition, an inter-laboratory evaluation of a tiered-approach combining the IL-18 assay (Galbiatti et al., 2011) and the epidermal equivalent potency test (dos Santos et al., 2011) is currently ongoing. So progress continues, and in all likelihood it soon will be possible to configure testing strategies based on accumulated expertise and experience combined with data from those *in vitro* and *in silico* approaches that are found to perform well, to identify skin sensitizing hazards without the use of animals. The aim will be to ensure that such predictive approaches are at least as accurate or probably better than the LLNA – and that also should be achievable.

However, as highlighted elsewhere, there is more to achieve and more that needs to be achieved. An increased investment will be needed in research focused on providing a more detailed understanding of the cellular and molecular mechanisms through which skin sensitization is induced and orchestrated. The dividends of that research investment will provide the momentum for truly innovative solutions to the unaddressed challenges and that will inform our understanding of the biological/biochemical bases that determine relative potency.

Two examples serve to illustrate the point. Work by Stefan Martin (University of Freiburg) and others has provided new insights into the role of the innate immune system in skin sensitization, and interactions between inflammatory reactions and adaptive and innate immune responses. This research will help define the danger signals and cofactors that are required for the effective acquisition of sensitization (Lass et al., 2010; Martin et al., 2008, 2011; Martin and Jakob, 2008; Schmidt et al., 2010; Weber et al., 2010).

A second area where there has been considerable progress has been in defining the phenotype, function, and impact on the induction of skin sensitization of discrete functional subpopulations of cutaneous DC. Of particular interest is the interplay between epidermal LC and dermal DC (Bobr et al., 2010; Clausen and Kel, 2010; Kaplan, 2010; Kaplan et al., 2008; Kimber et al., 2009; Noordegraaf et al., 2010). Our increased understanding of the roles played by skin DC not only in initiating but also in orchestrating cutaneous immune responses to contact allergens may pave the way to more sophisticated and more informative DC-based assay systems.

An investment in high quality research addressing important questions will always pay important dividends – and that holds true for skin sensitization and our need to drive new innovation in safety assessment.

Recommendations: skin sensitization

1. Recent investments in the development of alternative *in vitro* methods for skin sensitization hazard identification have resulted in the design of a substantial number of potential assays. Those that show promise should be evaluated as soon as possible.
2. Progress in the development and refinement of *in silico* approaches to skin sensitization testing (mathematical modeling and computational chemistry) should be accelerated.
3. The main priority now is to develop non-animal methods for assessment of skin sensitizing potency of contact allergens. In this context it is important to identify biomarkers or chemical signatures that are quantitatively associated with the acquisition of skin sensitization.
4. The ability of existing *in vitro* tests, QSAR methods, and other testing strategies to inform skin sensitizing potency, in addition to identifying skin sensitizing hazards, should be investigated.
5. New strategies for potency assessment based on approaches such as: (a) an appreciation of the balance between promotional and regulatory signals by skin DC, (b) an understanding of the impact of the vigor, quality, and breadth of T cell responses on the development of sensitization, (c) the design of appropriate mathematical models, and (d) integrated testing systems should be explored.
6. An investment in developing a more detailed understanding of the cellular and molecular events that initiate, orchestrate, and control immune responses to skin sensitizing chemicals should be encouraged.
7. An investment in activities facilitating the application of the emerging tests by industry and regulatory authorities and assessing the limitations and strengths of the tests before full validation should be considered.

4 A Roadmap for the Development of Alternative (Non-Animal) Methods for Repeated Dose Testing

Author whitepaper: Annamaria Rossi
Respondents: Sebastian Hoffmann, Chantra Eskes, Marcel Leist, James McKim, Greet Schoeters
Scientific writer: Emily A. McVey
Discussants: David A. Basketter, Bas Blaauboer, Robert Burrier, Harvey Clewell, Mardas Daneshian, Alan Goldberg, Thomas Hartung, Nina Hasiwa, Joanna Jaworska, Ian Kimber, Tom Knudsen, Paul Locke, Gavin Maxwell, Gladys Ouédraogo, Grace Patlewicz, Olavi Pelkonen, Costanza Rovida, Irmela Ruhdel, Andreas Schepky, Michael Schwarz, Nigel Skinner, Kerstin Trentz, Marian Turner, Philippe Vanparys, James Yager, Joanne Zurlo

4.1 Introduction: repeated dose toxicity

This chapter provides an overview of possibilities for replacing animals in repeated dose toxicity (RDT) testing and recommendations to improve and speed up the process of that replacement. The importance of RDT testing in the safety evaluation of chemicals, agrochemicals, pharmaceuticals, and cosmetics cannot be overestimated. RDT evaluation examines the potential for chronic toxicity and for organ-specific toxicities not seen in acute testing. The present testing schemes are based on rodent or non-rodent studies performed for 4 weeks (subacute toxicity), 13 weeks (subchronic toxicity), or 26-102 weeks (chronic toxicity). The tests, as they currently exist, are often followed up with further testing to more clearly define the nature of initial findings. Nevertheless, the false positive and false negative rate with respect to human adverse effects may be as high as 50% (Olson et al., 2000). The focus of toxicity testing in general should be to capture all potential toxicants and to assess their hazard. Species differences imply that the use of animals to assess toxicity is probably not the most efficient or detailed way to the end of safe and effective chemicals, pharmaceuticals, and cosmetics. Moreover, statistical issues, extrapolation from high to low doses, and other difficulties reviewed many times contribute to the weaknesses of animal-based safety testing. This chapter provides information on current and potential future *in vitro* and *in silico* approaches for assessing the major endpoints used in repeated dose toxicity. It also contains suggestions for improving and implementing those tests, and it outlines a roadmap forward in the field of alternatives to RDT testing.

The objective of RDT testing is to assess the potential hazard of a chemical after long-term exposure. The goal of such studies is to define a No Observed Adverse Effect Level (NOAEL) for the compound in question. Currently, the testing is usually performed in a rodent (usually rat) and potentially a non-rodent mammalian species (non-human primate or dog). The mechanistic and conceptual basis for RDT may be broad, and it is not well understood for many compounds. In some cases, it may be due to a build-up of toxic substance(s) in one or more sensitive areas of the body. In other cases, the changed compound concentration is not a driving factor. In such cases, defense mechanisms may be exhausted, the tissue may be altered by regulations and counter-regulations, or immunological reactions involving the specific or non-specific immune system may be triggered. Besides assessing obvious signs of toxicity and organ specific toxicity, a number of other endpoints are evaluated, including body weight, hematological parameters, urinary constituents, and histopathology of each organ system. RDT testing is thought to be extremely important in toxicity testing, as it is considered to model repeated exposures to lower doses of a compound, which is more likely to occur in a real-world situation than short term exposure to high doses. Moreover, this approach also offers the opportunity to assess recovery in between dosing. Toxicities not seen in acute testing or in reproductive toxicity testing may be revealed by RDT tests.

Regulatory risk assessments for chemicals, pharmaceuticals, and cosmetics, including REACH, TSCA and the FDA and EMA guidances, respectively, require RDT testing as an integral part of assessing the potential risks of a substance. The EU Cosmetics Directive (Cosmetic Directive 76/768/EEC), by adopting its 7th Amendment (2003/15/EC), has already instituted an animal testing ban, and as of January 1, 2013, a marketing ban will go into effect for any new substance tested on animals. Thus, the need for alternative methods is clear. Besides these regulatory reasons, animal testing is considered ethically questionable by many, and it is expensive. Most importantly, the present animal-based regulatory tests do not provide specific information on human hazard, and they fail to provide a mechanistic rationale that would explain toxicity and allow science-based predictions. This problem is currently circumvented by the introduction of safety factors, and the need to move to more specific and useful results for toxicity testing is clear.

The classical 1:1 replacement approaches have found their limit when faced with the problems of assessing RDT. The adverse effects can be based on a complex web of disturbances in multiple target tissues, and the interplay between various pathways and systems requires new modeling approaches, integrating multiple models, pharmacokinetic parameters, and a large battery of mechanistic tests designed to elucidate PoT.

In 2010, with the 2013 deadline looming, experts in the various areas of toxicology were invited by the European Commission and stakeholders (such as industry, non-governmental organizations, EU States, and the Commission's Scientific Committee on Consumer Safety (SCCS)) to analyze the status of alternative methods and to estimate the time necessary to achieve a full replacement of animals in the cosmetic industry. An extensive and detailed document (Adler et al., 2011) summarized their conclusions. In that document, virtually all of the available *in vitro* and *in silico* methods were considered, and the unanimous conclusion was that seven to nine years will be required before there will be replacement of animal testing for skin sensitization and five to seven years to finalize methods to predict toxicokinetics. A timeline for full animal replacement could not be set for repeated dose toxicity, carcinogenicity, and reproductive toxicity. An additional group of experts in the field of alternative methods has reviewed the outcome of this report, and they came to the same conclusion (Hartung et al., 2011). For this reason, this chapter will not list the existing tests and methods again but rather will discuss current gaps of knowledge and ways forward to arrive at an alternative testing strategy within a reasonable time period.

Although the above documents focused specifically on testing for cosmetics, many of the underlying themes apply to all areas of toxicity testing, including pharmaceuticals, agrochemicals, and industrial chemicals. The tests requested by the various regulatory bodies for RDT evaluation are in most aspects equivalent. Also, independent of the commercial product area, the mechanisms that are the basis of toxic effects are expected to be the same.

This chapter aims to place the available methods (extensively covered in previous papers) in a chronology, providing a roadmap for replacement of RDT testing based on what can and should be done now and what will require more time and effort. Thus, the methodologies reviewed below are in order of current feasibility, with particular note of steps that must be taken to move each methodology closer to fruition. The final section summarizes the findings of this group of authors and provides recommendations for moving forward.

4.2 Create new alliances

It is in the interest of every party, including chemical, pharmaceutical, food, and cosmetics companies, to decrease the number of animals used, not only for ethical reasons but also from a budgetary point of view. This is particularly the case when regulators request a more expensive second species study in repeated dose toxicity testing. To combine efforts in the challenging task of reducing the use of animals to predict for repeated dose toxicity (and, in general, any type of toxicity) would dramatically increase the possibility of success. Each company (regardless of the area of business), and each of the different regulatory agencies, has a database of information on toxicity, as well as investigative data that could advance the development and validation of *in vitro*/*in silico* systems immensely; inform pathways of toxicity based on *in vivo* data, and speed the process of reducing and replacing animals in repeated dose toxicity testing. This collaboration and data gathering initiative is something that could begin today and quickly show measurable results. The suggestion to create a consortium with a vision of generating such a database may sound naïve in the present competitive environment. However, this type of exercise could lead to a clear and definitive advance in the prediction of toxicities without further use of animals. Setting up such a consortium would require strong regulatory pressures and incentives, particularly with regard to breaking down of barriers to data sharing. It would be well worth the effort, however, to come closer to the larger goal of animal-free toxicity testing.

There are several examples of this type of initiative already: The U.S. Environmental Protection Agency (U.S. EPA) has recognized the relevance of collecting high quality regulatory *in vivo* data and making it accessible for cross-chemical computational toxicology analysis to create an *in silico* assessment for chemical compounds. They have created a U.S. EPA ToxRef Database that profiles chemicals based on chronic toxicity. This database has already proven to be very valuable (Liebsch et al., 2011; Martin and Jakob, 2008). Another example of relevant collaboration is the Innovative Medicine Initiative (IMI) project eTox[1]. This consortium aims to create a drug safety database from industry legacy toxicology reports and public data that will allow *in silico* prediction of toxicities.

Linking these types of efforts across different business and product areas could confirm common pathways of toxicity or help identify new ones. The OpenTox[2] initiative could be a starting point; however, OpenTox currently only works on open collaborations, communications, and advisory boards. To function optimally, and have more impact in creating predictive assays and tools, it would need to add more rigorous experimental data sharing.

4.3 Integrated testing strategies

It is clear that, at present, we cannot screen for repeated dose toxicity using only alternative methods. However, the implementation of decision trees and tiered approaches, i.e., integrated testing strategies (ITS), will contribute greatly to the reducing the use of animals (Grindon et al., 2008; McKim, 2010; Vanhaecke et al., 2011). This is something that can begin now and be modified as more and more alternative tests become available. Approaches such as the decision tree proposed by Vanhaecke et al. (2011) for predicting liver toxicity represent a very good start. The authors propose integrating computational and cell-based toxicity information, along with pre-existing data, to arrive at a theoretical NOAEL and assess an acceptable margin of safety. Similar strategies could be developed for all relevant organ systems, and be implemented in a variety of sectors, to achieve meaningful validation. However, such ITS will need to be broadly discussed among different experts in the toxicities assessed and in the technologies used. Before implementing the ITS, a very careful evaluation of the assays is needed. For example, Vanhaecke et al. (2011) suggests us-

[1] http://www.imi.europa.eu

[2] http://www.opentox.org

ing stem-cell-derived hepatocytes, which currently are not the best model to screen for liver toxicity (Guguen-Guillouzo et al., 2010). The use of reference compounds appropriate to each organ system to test the ITS would be particularly useful in such a validation.

In some cases, such as in reducing animal use for REACH in the chemical industries, decision trees and tiered approaches may be well characterized and ready to be implemented. However, in other areas, such as in the pharmaceutical industry, these approaches are still in a validation phase and have yet to prove their ability to predict and manage risk at an early stage. Technologies and assays used in any ITS would be regularly reviewed and revised, of course, to ensure their continuous development and improvement.

Of note, and rarely mentioned, is the fact that ITS can and should incorporate human data, including epidemiological, genetic, and medical/clinical data, whenever applicable. As for data from *in vitro* and *in silico* systems, standards must be set to ensure the use of quality and comparable data in each system, bearing in mind that the overarching goal is to predict human toxicity. The use of this data cannot and should not be ignored in developing testing approaches.

4.4 Signaling pathway identification and analysis

In 2007, the United States National Academy of Sciences (NAS) published a report *Toxicity Testing in the 21st Century – a Vision and a Strategy*, which envisioned a new approach to toxicology (NRC, 2007). The report called for the application of new and advanced technologies and biological knowledge to move toxicology forward. One focus of the report was the identification of common pathways of toxicity. New frontiers of science, such as systems biology, bioinformatics, high-throughput screening, high-content screening, transcriptomics, proteomics, and metabolomics, applied in a synergistic manner, will improve our understanding of the cellular pathways involved in toxic processes and thereby improve our ability to predict toxicity. Understanding alterations in signaling pathways, and the role of these alterations in the activation of toxic effects, is crucial not only to elucidate how and why a toxic effect is occurring but also to extrapolate the effect to and from an *in vitro* system. The task of identifying relevant pathways, understanding them, and demonstrating correlation with toxicity is a challenging one, particularly if we consider that often it is the interaction between pathways and their localization that results in toxicity or protection of a cell (Latta et al., 2000; Volbracht et al., 1999).

Cells receive, process, and respond to information and signals through many distinct molecular pathways, which permit them to function properly. Often, components of several different pathways interact, resulting in signaling networks to create these cellular responses. A toxicological response often is induced by a disruption of these signals. For example, protracting a signal for a long period of time could lead to a toxicological response rather than a physiological one. Similarly, the strength of the signal could be enhanced by a toxicant and, therefore, unbalance the normal cellular responses or disrupt feedback loops. In considering the analysis of signaling pathways, it should be taken into account that often it is not a new pathway that is activated but rather a normal signal that is disrupted by the toxicant (see Fig. 4.1). Therefore, it is necessary to have an idea of the threshold that must be crossed to exert a toxic effect. Together with the identification of the pathways themselves, a quantification of the major players in the pathways should be performed. The identification of common pathways of toxicity, and the quantification of the signals therein, should continue to be the major focus of toxicological research, as not only will it move the field forward in understanding, it will assist in the development and validation of the *in vitro* and *in silico* systems we need to replace animals in repeated dose toxicity testing (Stokes and Wind, 2010).

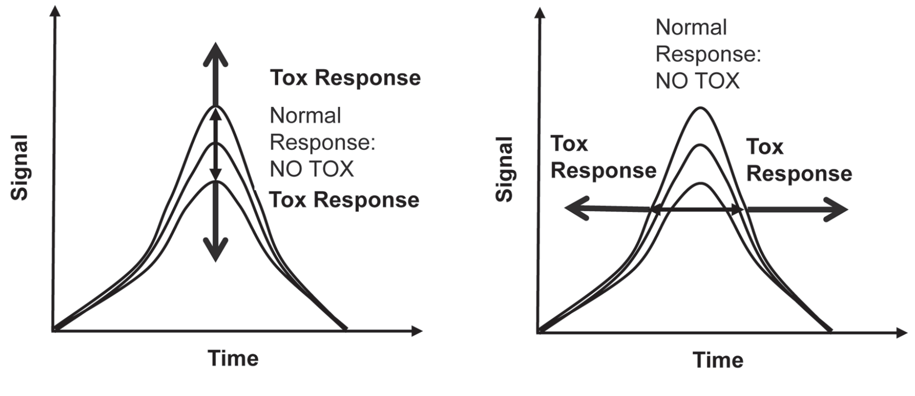

Fig. 4.1: Pathway of Toxicity (PoT) alterations – alterations in the magnitude or timing of signaling

There are a number of mechanisms by which these pathways can be investigated – many of them pioneered in medical research in order to identify pathway alterations that result in human pathologies. These include technologies from knockout yeast and mice, to antibodies and perhaps RNA interference (RNAi) (Moffat and Sabatini, 2006). The concept of each of these technologies is that a specific member of the pathway in question can be perturbed, which can lead to identification of important pathway members, how they interact, and to what degree.

RNA Interference (RNAi) and other interventions to define pathways of toxicity (PoT)

RNAi technology is based on the concept that by introducing a sequence-specific iRNA that will lead to a post-transcriptional gene-silencing process, one can identify important pathway interactions and mechanisms. This tool is important because knocking down (KD) the gene(s) inhibits the entire pathway, allowing identification of the proteins that play a role in signaling through post-translational modifications and monitoring of their role in signaling (Virshup and Shenolikar, 2009). Different specific RNAi sequences can result in different percentages of KD in the cells, giving hints as to the threshold of inhibition or activation needed to create a toxic effect in the signaling pathway. In other words, this technology is based on a "quantitative" indication of the signaling pathway.

RNAi in combination with bioinformatics tools can provide even more knowledge on the signaling pathways involved in toxicity. There are several examples of software and algorithms created to correlate the KD pathway from an RNAi experiment with other pathways (Kaderali et al., 2009). Expanding the genetic analysis of the KD genes can achieve a similar result: A transcriptomic evaluation of gene changes within a pathway can provide an overview not only of the specificity of the inhibition, but also of possible correlations among pathways that could highlight new toxicological interactions, particularly if included in a bioinformatics analysis.

RNAi has proven useful in areas such as in identifying pathways involved in cancer and apoptosis induction (MacKeigan et al., 2005). However, applying this technology directly to the *in vitro* systems mentioned above might not be immediately feasible, since thus far RNAi methods are well established in cell lines and in dividing cells but are difficult to use in non-dividing primary cultures and ESC-derived cells. Luckily, the signaling pathways involved can be investigated in non-primary, non-ESC cell lines, and the discoveries made there can be applied to the *in vitro* technologies to improve predictability.

RNA interference has been described here in more detail as an example for a pathway-mapping tool. Other approaches are available and are being developed. One involves the selective inhibition of pathways by chemicals. These act especially fast, and new pathways have been identified in this way (Falsig et al., 2004b; Lotharius et al., 2005; Lund et al., 2005). A big advantage is that they can also be combined for more sophisticated pathway mapping (Falsig et al., 2004b). The major disadvantage is that often they do not selectively inhibit only a specific pathway, and a more detailed analysis of pathways inhibition should be performed. Knocking out genes is another approach. Different strategies have been followed to do this systematically in murine and human cells. One modern approach is the generation of haploid stem cells, and the knockout of every gene of the genome in such cells, which then can be differentiated to any given cell type. More traditional approaches use cells from humans with different mutations, and a sophisticated variant uses such cells together with a derived cell type in which the mutation has been repaired. A completely different and complementary strategy useful for PoT mapping is the visualization of pathway activities by cellular reporter constructs.

Case studies

To bring together new technologies and existing toxicological knowledge, extensive case studies will take a central role, and their importance cannot be overestimated. One important basis will be an assembly of compounds causing RDT. Among these, the ones not identified in acute studies need to be identified. Many such examples are known from the pesticides field (Spielmann and Gerbracht, 2001; van Ravenzwaay, 2010). A first selection would involve those that result in human toxicity, or for which the animal toxicity is seen in the range of human exposure. The case study, then, would examine whether toxicity would have been predicted correctly by alternative methods. The cases where this approach failed are especially interesting. This should give an incentive for the establishment of methods that fill the gap. It would be worthwhile to promote such studies in EU-funded research consortia. The road to the future in this area leads through learning from the past. Such efforts need support as goal-oriented applied research. They are not funded by classical scientific funding bodies, which appear to consider them too applied and not sufficiently innovative. Unfortunately, it is not broadly accepted that such work forms the basis for large innovations in the field of risk assessment.

4.5 *In vitro* methods

4.5.1 A brief overview of in vitro models

The need for *in vitro* systems, which can address all areas covered in RDT testing, is obvious. This section provides a brief overview of the available technologies and highlights some barriers and considerations. The cellular tools currently available are primary cultures or established cell lines from animals or humans (Skelin et al., 2010).

Primary cultures obtained from animals have three major limitations, the most obvious one from the point of view of reducing animal use being that animals are used. The number of animals used to perform experiments with primary cultures is fewer than for *in vivo* testing, but often a significant number of animals must be sacrificed to obtain a primary culture, particularly for difficult-to-culture cell-types. A second limitation of primary cultures is the short life span of the cultures. With the exception of some neuronal systems (Viviani, 2006) that have a relatively long life *in vitro*, many cultures have a lifespan of 2 to maximum 14 days (Volz et al., 1991). Additionally, even if the survival of the culture can be increased using improved culturing methods, some of the relevant function and signaling pathways of the cells can be lost (Hartung, 2007a). For example, hepatocyte cultures are relevant, not only to predict one of the most common types of toxicity observed when testing chemicals and drugs, but also to properly predict metabolism and pharmacokinetics – key parameters necessary to properly forecast repeated dose toxicity. Several efforts have been made in improving this relevant cellular system and, at present, the cultures can survive for several weeks. However, during this time the cultures lose their metabolizing capacity (Miranda et al., 2009) and therefore lose value in predicting for repeated dose toxicity. A third limitation, which may also be seen as an opportunity, is that these systems will be predictive (potentially) of animal toxicity, rather than human toxicity, and assessment factors for interspecies variation would still be necessary in the final risk assessment (Falsig et al., 2004a; Lund et al., 2006). Since the goal is to predict human toxicity, continuing to use animal systems may not be ideal. However, these animal systems may be seen as an interim step in the process of full animal replacement, in that it will be easier to validate/assess the value of these tests by comparing the *in vitro* results to already available *in vivo* data from the current animal tests, and these tests may be easier/quicker to develop than the respective human systems.

Human primary cultures are perhaps the most relevant system for *in vitro* screening from the standpoint of species specificity and maintenance of the optimal genetic profiles and signaling pathways. However, a major disadvantage of human primary cultures is the poor availability of human samples (often derived from cadavers or cancer patients), resulting in little control over the phenotypes selected for screening.

An alternative to primary cultures is immortalized *human cell lines*. While these cells have the advantage of being easy to culture and the ability to increase screening throughput, they may have altered signaling pathways, and in some cases their metabolism is changed (more glycolytic energy generation). In most cases they lose xenobiotic-metabolizing capacity (Hartung, 2007a). Conditionally-immortalized cells, or cells in which the immortalizing transgene can be deactivated, may offer a compromise solution (Lotharius et al., 2005; Scholz et al., 2011).

Stem cells can be classified into three major categories, according to derivation: embryonic stem cells (ESC), adult stem cells (ASC) and induced pluripotent stem cells (iPSC). Adult stem cells (ASCs) comprise, e.g., Mesenchymal Stem Cells (MSCs), are present in somatic tissues and have characteristics of multipotent adult progenitor cells. They are not able to differentiate into all cell types of the organism. However, it should be taken into consideration that both bone marrow mesenchymal stem cells (BMSC) and adipose-derived mesenchymal stem cells (ADMSC), when properly differentiated, have potential for hepatic and neuronal differentiation (Banas et al., 2007; Gimble and Guilak, 2003). This could be an interesting area to explore further, particularly considering the accessibility of adipose tissue from surgical operations.

Embryonic stem cells (ESC) are isolated from the inner cell mass of 5-6-day-old blastocysts (Davila et al., 2008) and are fully pluripotent, meaning they are capable of giving rise to most tissues of the organism, including germ line cells. Under proper differentiation they should be capable of generating all the cell types present in an organism.

In 2006, Takahashi and Yamanaka published a breakthrough in stem cell biology: mouse somatic cells that could be reprogrammed back to pluripotent stem cells. The era of induced pluripotent stem cells (iPSC) had begun (Takahashi and Yamanaka, 2006). One year later, the same group showed that human somatic cells can also be reprogrammed into pluripotent stem cells by transduction of four defined transcription factors: Oct3/4, Sox2, Klf4, and c-Myc. The derived cells had the same morphologic, genetic, and epigenetic characteristics as stem cells (Takahashi et al., 2007). However, before considering the use iPSC for clinical or toxicological purposes, the issue of the mechanism of reprograming must be solved, since this process implies the use of viral transduction (which leads to a safety concern for clinical applications) and the activation of transcription factors and oncogenes present also in cancer stem cells (a concern for both clinical and toxicological use) (Jaenisch, 2009). That said, various papers already have been published reporting that reprogramming of somatic cells can be achieved without using viral delivery of reprogramming factors and evaluating the relevance of c-Myc and Klf4 in this process. A reliable methodology for doing so across many labs would increase the potential for these cells in both clinical and toxicological applications (Cox and Rizzino, 2010; Jaenisch, 2009). Thus, iPSCs may have great potential for predicting toxicity. They can be a source of potentially all tissues derived from various human populations with different pharmacogenomics profiles and a variety of genetic variabilities.

Stem cells represent a cellular system that has several advantages compared to stabilized cell lines and primary cultures, including normal genetic profile, normal growth, uniform cellular physiology, and pharmacology (McNeish, 2007). They have a number of unique features that make them attractive and potentially valuable for toxicological screening (Ameen et al., 2008; Davila et al., 2008; Jensen et al., 2009; McNeish, 2004):

a) Stem cells divide and renew themselves for a long period of time, and therefore they can provide an almost unlimited supply of cells. Since, like all *in vitro* dividing cells, stem cells can accumulate mutations, karyotyping the cells is necessary after long periods in culture to confirm genetic normality prior to use in testing.
b) Stem cells are pluripotent and therefore potentially able to differentiate into any human tissue. This opens the possibility of creating different cell types from the same organ in one culture. For example, an ideal *in vitro* liver toxicity system would have a "liver-like" organ that includes not only hepatocytes but also all other relevant cells, such as Kupffer cells,

stellate cells, and cholangiocytes. Multiple cell types in each organ system are undoubtedly important in various types of toxicity, so a wider variety of cells in the *in vitro* system could provide a better picture of potential toxicity.

c) Stem cells can represent genetic diversity. This is particularly true if induced pluripotent stem cells (iPSC) are used (see note on types of stem cells above).

d) Under the appropriate culture and assay conditions, the throughput and predictivity of *in vitro* assays using cell culture would be increased significantly through the use of stem cells.

However, the limitations of the system should not be neglected. Stem cell biology is a young science, and so far the culture of stem cells is not trivial. Additionally, when stem cells differentiate into different cellular systems, the differentiation rarely occurs in 100% of the population, and not all of the cells are in the same stage of full differentiation (Ameen et al., 2008). For example, hepatocytes, when properly differentiated to produce hepatic endoderm cells or hepatocyte-like cells, present characteristics of fetal hepatocytes and do not express fully active cytochrome P450 signals (Greenhough et al., 2010). Similarly, stem cell-derived cardiomyocytes resemble human heart tissue but variably and with gene expression that is not the same as in adult heart tissues, indicating that additional differentiation protocols are needed (Asp et al., 2010).

So far, only a limited number of cell types have been differentiated, compared to the variety of potential cell types within an organism. Limited phenotypes and functional data are available for the embryonic stem-derived cells, with few exceptions. More research and investigation is needed to determine the state of maturation and functionality of the different cellular types. When these characteristics can be verified, the possibility of applying *in vitro* stem cell-derived models in predictive strategies in toxicology will increase dramatically. This is not possible in the next couple of years, but, based on the data available so far, it may be possible in three to seven years.

4.5.2 Specific considerations for in vitro methods

The following constitute special issues that must be borne in mind during the development, validation, and implementation of cellular test methods:

Culture methods

The application of techniques such as 3D culture systems and co-culture has great potential for toxicity testing. Several examples confirm the relevance of *3D culture models* to improve the structure and the prediction rate, not only for toxicity but also in screening for pharmacological assays (Dash et al., 2009; Lan and Starly, 2011; Meng, 2010; Nakamura et al., 2011; Toh et al., 2009). Similarly, *co-culture methods* will give a better idea of the relevance of interaction and crosstalk between the different cell types. Co-culture systems have been proven particularly relevant in prediction of inflammatory effects and the physiological interaction between signaling pathways (Boraso and Viviani, 2011; Scharf et al., 1996; Tukov et al., 2006). The further development of these methodologies in concert with the cell-systems that constitute them will be important in predicting complex toxicities.

Endpoints

The main toxicological endpoints for *in vitro* technologies have been the classical markers for cell death, such as membrane permeability, intercellular energy levels, glutathione levels, and other general endpoints that represent a high level of toxicity. However, it is important to remember that toxicity is first induced by the malfunctioning of cells, from which significant cell dysfunction and death follow, i.e., if we consider compounds that act on the cytoskeleton or on exocytosis, we must consider cytoskeletal component alteration or enzyme release as a significant endpoint, rather than just cell death. In this way, the substitution of classical "toxicological" endpoints with functional ones is a way in which classical toxicity prediction could be improved, especially when predicting organ-specific toxicities. For example, compounds that are cardiotoxicants often are found to be cytotoxic in hepatocytes or other types of cell cultures. This information may be useful for acute toxicity but less relevant for the assessment of organ-specific toxicity in the heart. In this case, it is more relevant to consider the contractive capacity of cells, rather than the induction of apoptosis.

Another good example of the concept of more specific toxicological endpoints improving the overall quality of *in vitro* testing in general is that better *in vitro* ADME (absorption, distribution, metabolism, and excretion) prediction significantly reduced the attrition percentage for compounds in development in the pharmaceutical industry (Kola and Landis, 2004). Specific studies of ADME-related mechanisms led to the development of good predictive systems with the most indicative endpoints. Recently, more investment has been made in developing new approaches to investigate more sophisticated and meaningful endpoints. High-content screenings, platforms for biomarker detection, TaqMan Low-Density Arrays, and new technologies for the assessment of phosphorylated proteins are examples of technologies that allow the investigation of a wider variety of toxicological pathway endpoints.

In vitro exposure

Kinetics and biodistribution are two key factors that must be included in the evaluation of repeated dose toxicity. However, *in vitro* screenings often do not consider the actual (as opposed to nominal) *in vitro* concentration, bioavailability, and degradation of compounds. Frequently, synthesized compounds are not stable at 37°C and/or bind to plastics or media proteins, factors that often are not considered or accounted for in *in vitro* tests. The importance of this cannot be overestimated, as it is crucial for data interpretation. Although it is labor intensive, the detection of the real free concentration and measurement of the stability and availability of the compounds in an *in vitro* system cannot be neglected.

Assay validation

Not all parties were in agreement with the timelines envisioned in the EC's expert panel report (Adler et al., 2011). Some claim that the date of 2013 for animal replacement is still possible (Balls and Clothier, 2010; Spielmann, 2010; Taylor et al., 2011). Taylor and Casalegno, in particular, claim that several alternative methods are available where the percentage of prediction is above 80% (Carfi et al., 2007; Duff et al., 2002; Huang et al., 2009; Inoue et al., 2007; Langezaal et al., 2002; Pessina et al., 2001). Although these are all very promising examples, seldom more than ten compounds were tested in these assays. They must be more appropriately validated with a larger number of compounds, while still achieving a high percentage of prediction to be more universally accepted. To that end, it is worth mentioning that a good validation, particularly for the complex endpoints of repeated dose toxicity, should include a sufficient number of compounds, ideally representing a variety of classes. To increase the number of classes of compounds that can be used for validation of common tests, again, collaboration between different industries and entities is the ideal.

Each validation must be tailored to the system being tested, and certain agreements must be set for all tests for a certain type of toxicity (Hartung, 2007b). For example, for organ-specific toxicities using *in vitro* cellular assay tests, it must be decided "What constitutes a heart?" and "What constitutes a liver?" More broadly, what cell types, gene expression, and physiological markers must be set in order for a system to appropriately represent the organ in question? Thus, comparison directly to current endpoints and markers may be necessary at first, but a true assay validation must be tailored to the test or testing scheme in question, particularly for repeated dose toxicity.

4.6 *In silico* prediction

The value of bioinformatics, *in silico* technologies, and systems biology in analyzing the data, identifying new pathways, and predicting toxicity is inarguable. Many of the aforementioned reports and reviews on the replacement of animal tests summarize the state of the science for *in silico* methodologies for repeated dose toxicity testing, so we will not provide a summary here. However, as we work toward the goal of *in silico* models and methodologies as a key part of toxicity testing, it is of extreme importance to recognize that the quality of the data used to create predictive *in silico* models significantly affects the quality of the system itself. If low quality data are used, the system is designed to fail. When designing *in silico* methods using *in vivo* data, it is vital to have data from well-designed experiments that indicate the time course of the toxicity and that will correlate pathology with molecular and mechanistic endpoints. If *in silico* methods are developed on the basis of *in vitro* data, the quality and predictivity of the experiments become even more important. For example, basing an *in silico* model for pathway analysis on data from tumor cell lines would be suspect, since these cell lines often have altered signaling pathways. Another example is the use of RNAi data: it is essential that the appropriate cell line was used to derive the data, and only the pathway in question was affected by the interference.

4.7 Conclusions and recommendations: repeated dose toxicity

It is most likely that a decade or more will be required before the gaps can be appropriately filled. That said, the authors recommend a step that can be taken immediately: the implementation of more stringent and appropriate ITS for testing. Another step that could move forward immediately, and that could improve ITS testing schemes, as well as *in silico* and *in vitro* technologies, is a frank and complete gathering and assessment of the repeated dose toxicity data that already exists for a wide variety of compounds, and the use of these data for case studies investigating the needs and pitfalls for new assays. This would require the collaboration of a variety of entities, including commercial, governmental, and non-governmental. The benefits that could be reaped by such a concerted effort in data gathering and sharing clearly outweigh the difficulties. In order to improve the predictivity of current *in vitro* and *in silico* tests, and even the current tests, the identification of pathways of toxicity must continue to move forward. This requires the use of new technologies in the field of omics and systems biology, combined with new cell models and evaluation strategies based on chemical inhibitors or gene inactivation. One particular issue that must be addressed is the setting of guidelines as to what constitutes an appropriate model for each organ system, i.e., what makes a heart a heart, a liver a liver, etc. In this context, it will be important to consider how immunological and inflammatory reactions can be incorporated in such organ systems.

Recommendations: repeated dose toxicity

The following steps are suggested to replace animal testing for repeated dose toxicity in an appropriate and timely fashion:

1. **Joint task force:** A joint effort toward a toxicity database to gather all current data on a wide variety of compounds would greatly improve the quality and speed of new test development and validation. Organization of this effort should begin immediately. The data should be used to support case studies designed to identify test requirements and pitfalls, as well as for test evaluations.
2. **Tiered testing systems and decision trees (ITS):** Although it clearly is not yet possible to replace *in vivo* testing completely, we can refine and reduce the number of animals used today. Implementing decision trees, tiered approaches and/or applying screening strategies is possible immediately, and these can be modified as more and more non-animal tests become available. In addition to existing animal data, data from *in vitro* tests and data from *in silico* systems, as well as human data (epidemiological and clinical/medical), can be integrated into these types of approaches. These data should not be ignored! The time to act on this is now, for all types of compounds.

3. **Understand signaling pathways:** Understanding the molecules and pathways involved in toxicological events is crucial for progress in toxicology. This is probably the most important activity for future success in replacing animals for RDT testing. We should consider:
   a. The signaling pathways involved in toxicity may be normal signals that are altered in the duration or magnitude of response. Therefore, a quantification of the signal is of great relevance. For this reason, two different concepts are followed initially. The identification of PoT is a more long-term goal. An ITS based on high-throughput mapping of PoT and their disturbance in simple systems may eventually yield a good toxicity prediction. In the meantime, while not all quantitative relationships of the network of PoT are known, and while it is still unclear why chemicals affect one cell type more than another, more complex systems will be employed to arrive at more apical endpoints (Zimmer et al., 2011). The two approaches will be complementary and require a parallel development for some time.
   b. Tools such as RNAi or chemical interference, which are often implemented to aid in understanding signaling pathways in various diseases, could help toxicologist understand the signaling pathways involved in toxicity.
4. **Considerations for development and validation of *in vitro* systems:** A large number of potentially useful *in vitro* cellular assays are available, and each of them has advantages and disadvantages. It needs to be considered that:
   a. All *in vitro* systems have limitations, and the choice of which to use will depend on the question asked. This is particularly important in the nearer future, with the use of complex test systems. Only these experiments, and comparison with high-throughput approaches, can show whether the complex systems eventually can be replaced with simple assays, as in the ToxCast program.
   b. More sophisticated methods will probably decrease the throughput, but, at present, they will most likely provide more long-term and stable systems. They may, for the foreseeable future, be better at predicting more complex organ toxicity (e.g., 3D-systems and co-cultures), particularly inflammatory and fibrotic processes.
   c. Appropriate endpoints must be chosen for each test and test system: what do we want to know and what toxicity are we trying to predict? Omics approaches will get rid of this problem, as many endpoints can be evaluated simultaneously (Henn et al., 2009).
   d. Real free concentration and stability of the compounds during the exposure *in vitro* is of major relevance for evaluating the actual toxic dose. Overall, the modeling and prediction of compound concentrations will play a key role for QIVIVE.
   e. For the complex models of biological processes, a significant number of known positive and negative compounds are required to evaluate the performance of the system. The selection of compounds should consider the applicability domain and different chemical classes, as well as modes of actions. The creation of a reference list of compounds for which information on mechanisms of toxicity and potency is readily available would speed the validation process immensely for all new testing systems.
5. **Considerations for the development and validation of *in silico* models:** It is extremely important to be sure of the quality of data used to build *in silico* models. Specific criteria to evaluate the robustness and quality of the experimental data used in the development of *in silico* models should be developed and agreed upon in order to address this issue and to assist in design and validation of high quality *in silico* models.

5 A Roadmap for the Development of Alternative (Non-Animal) Methods for Carcinogenicity Testing

Author whitepaper: Thomas Hartung
Respondents: Robert Landsiedel, Philippe Vanparys, James Yager
Scientific writer: Marian Turner
Discussants: David A. Basketter, Bas Blaauboer, Robert Burrier, Harvey Clewell, Mardas Daneshian, Chantra Eskes, Alan Goldberg, Nina Hasiwa, Sebastian Hoffmann, Joanna Jaworska, Ian Kimber, Tom Knudsen, Paul Locke, Gavin Maxwell, James McKim, Emily A. McVey, Gladys Ouédraogo, Grace Patlewicz, Olavi Pelkonen, Annamaria Rossi, Costanza Rovida, Irmela Ruhdel, Andreas Schepky, Greet Schoeters, Michael Schwarz, Nigel Skinner, Kerstin Trentz, Joanne Zurlo

5.1 Introduction: carcinogenicity

In April 2010, the US President's Cancer Panel published the report "*Reducing Environmental Cancer Risk*" (Reuben, 2010). Although the report acknowledges that "*overall cancer incidence and mortality have continued to decline in recent years*" (see also Fig. 5.1), it states "*the true burden of environmentally induced cancer has been grossly underestimated. With nearly 80,000 chemicals on the market … un- or understudied and largely unregulated, exposure to potential environmental carcinogens is widespread.*" This situation must be considered in the context that life expectancy has tripled (Kirkwood, 2008) during the period in which these chemicals were introduced.

At the same time, the possible health risks posed by chemicals are of considerable concern to the general public (Entine, 2011), which fuels the demand for safety testing of chemicals. Surveys conducted by Eurobarometers in 2005 and 2010 asked Europeans the question of how likely they consider the possibility that environmental chemicals damage their health. In both years, around 18% of respondents considered this "very likely" and 43% "fairly likely" (Eurobarometer 73.5 from 06/2010 and 64.1 from 09-10/2005). In strong contrast, the degree of contribution of chemical exposure to the overall cancer rate has been estimated at only 4% for occupational exposure, 2% for pollution, less than 1% for industrial products, and 1% for medicines and procedures (Doll and Peto, 1981). These estimates, however, are outdated and, for example, did not take into account the interactions of multiple factors.

It is not the purpose of this paper to take a position in any of these debates but rather to address the issue of how to best test chemicals for carcinogenic potential, given the potential of these chemicals to exert health effects. At the same time, we have to ask ourselves whether traditional precautionary methods used to inform the general public about the risks that chemicals may pose have served us well. We believe it is likely that revamping our testing paradigms by basing them on updated and rigorously tested science and leaving the precautionary aspect explicitly to the risk management process would better serve both those involved in carcinogenicity testing and the public.

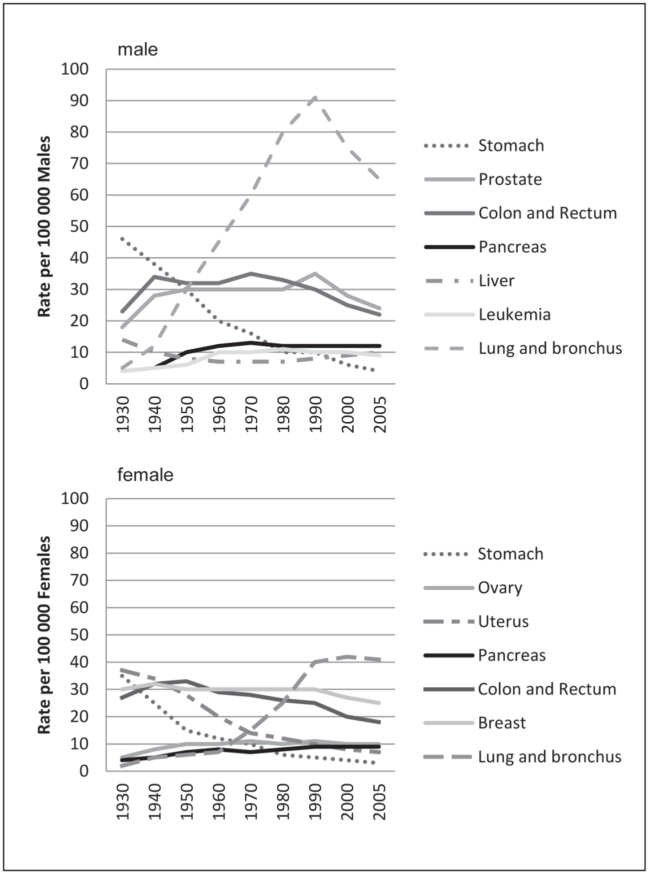

Fig. 5.1: Cancer mortality in the US over time
Annual age-adjusted cancer death rates among males and females for selected cancers, US 1930-2006. Adopted from (Jemal et al., 2010). Rates are adjusted to the 2000 US standard population. Due to changes in International Classification of Diseases (ICD) coding, numerator information has changed over time. Rates for cancers of the lung and bronchus, colon and rectum, and liver are affected by these changes.

Our current basic understanding of chemical carcinogenesis (Loeb and Harris, 2008; Luch, 2005; Oliveira et al., 2007; Williams, 2001; Wogan et al., 2004) is shown in Figures 5.2 and 5.3. These figures demonstrate a multi-step process in which several mechanisms – genotoxic and non-genotoxic –

contribute to cancer initiation and promotion. The potential of chemicals to interfere with repair and defense mechanisms, as well as detoxification and excretion, further contribute to this complexity.

An ideal carcinogenicity testing system would take all of these factors into account. Unfortunately, such a system does not exist. In this paper, we assess the available tools for carcinogenicity testing, introduce emerging tools that could transform this testing paradigm, and discuss the potential we see for these novel methodologies.

Definition of carcinogenicity[1]: "*Chemicals are defined as carcinogenic if they induce tumors, increase tumor incidence and/or malignancy, or shorten the time to tumor occurrence. Benign tumors that are considered to have the potential to progress to malignant tumors are generally considered along with malignant tumors. Chemicals can induce cancer by any route of exposure (e.g., when inhaled, ingested, applied to the skin, or injected), but carcinogenic potential and potency may depend on the conditions of exposure (e.g., route, level, pattern, and duration of exposure).*"

5.2 Application of the framework to carcinogenicity testing

We have applied the assessment framework presented in Chapter 1 to analyze various options as potential alternatives to the cancer bioassay (OECD TG 451; OECD, 2009), which is conducted as a 2-year bioassay in rats and mice and is currently the only accepted test for carcinogenicity.

Testing with the cancer bioassay in two species can involve 600-800 animals, the histopathological examination of more than 40 tissues per animal, and costs approximately € 1 million per chemical and species (Vanparys et al., 2011). This bioassay is obviously time-consuming and expensive, and uses large numbers of animals. In addition, the assay's predictivity for humans has been challenged (Knight et al., 2006a,b,c; Shanks et al., 2009). Thus, while protection against potential carcinogenic effects of environmental chemicals is a key desire of the public, this assay is not suitable for broad use, nor is it broadly used.

5.2.1 Abolition of useless tests

The concept that genotoxicity is the first and foremost mechanism of chemical carcinogenicity is rarely challenged. However, there are little or no epidemiological data that support the hypothetical existence of widespread chemical carcinogenesis. Not only has average age increased continuously over the last 150 years (Kirkwood, 2008), during which period about 100,000 chemicals were introduced into our environment, but age-adjusted cancer rates did not increase over this time period (Jemal et al., 2009). Furthermore, exposure to mutagens did not correlate with oncomutations in people (Thilly, 2003).

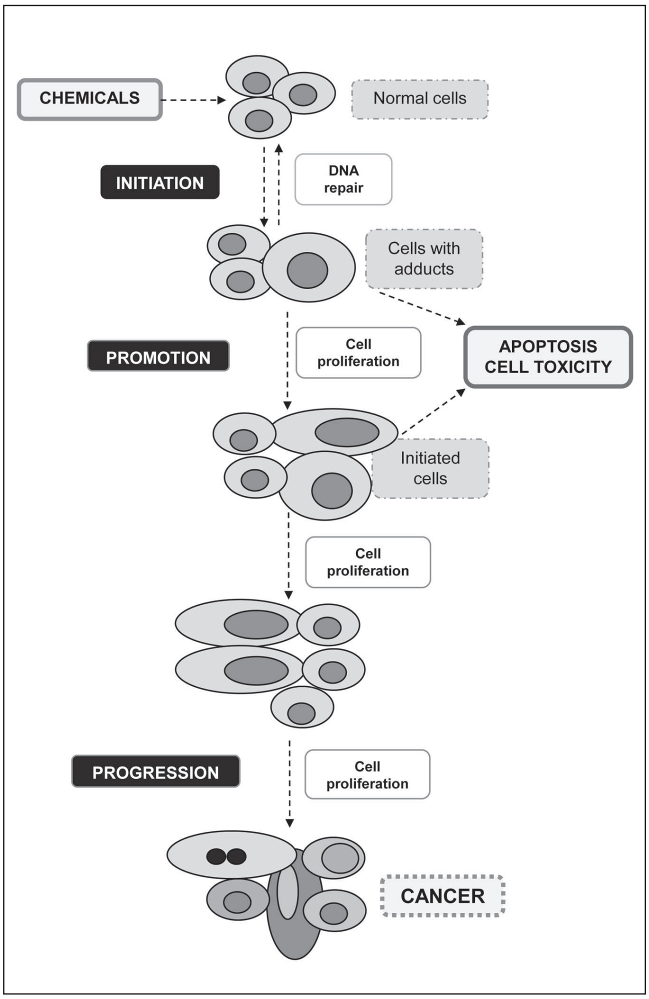

Fig. 5.2: Chemical carcinogenesis stages and the occurrences involved in each one
(modified from Oliveira et al., 2007)

It is important to note that carcinogenicity testing was developed as a result of historical cases of adverse effects, and the test models currently in place were developed with the existing knowledge at that time. However, the fact that there has been much scientific progress relevant to this field since then, combined with the degree of public concern about potential chemical carcinogenicity, has led us to focus this paper on carcinogenicity testing.

Standardization of protocols

The cancer bioassay is astonishingly young, given the importance of the health effect in question: the standardized protocol was suggested by the US National Cancer Institute in 1976 and adopted by OECD in 1981. The ICH (International Council on Harmonisation of Technical Requirements for the Registration of Pharmaceuticals for Human Use) only adopted the test for use in pharmaceuticals in 1997.

[1] http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r7a_en.pdf (last accessed 08.09.2011)

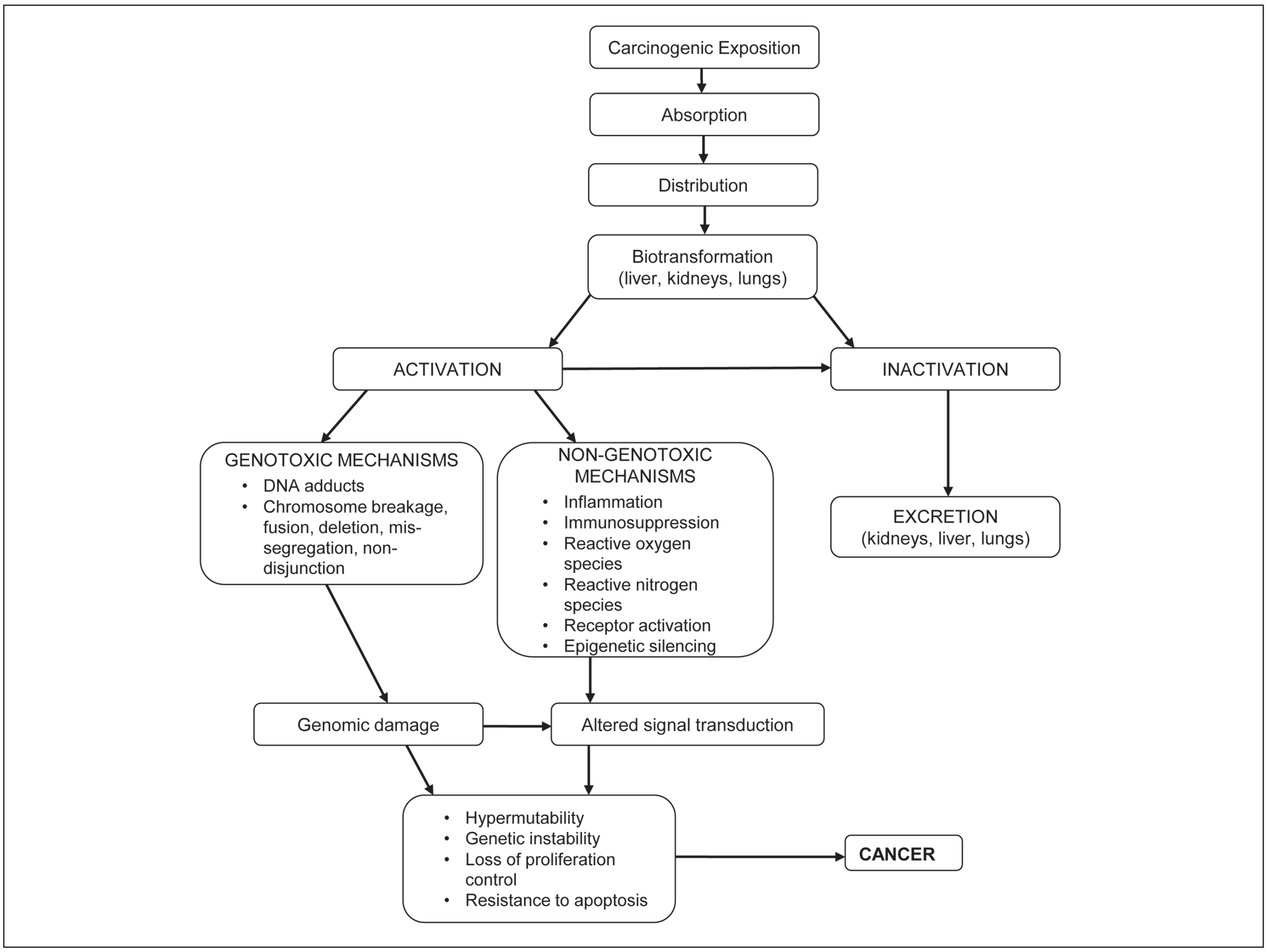
Carcinogenic Exposition
Absorption
Distribution
Biotransformation (liver, kidneys, lungs)
ACTIVATION
INACTIVATION
GENOTOXIC MECHANISMS
• DNA adducts
• Chromosome breakage, fusion, deletion, mis-segregation, non-disjunction
NON-GENOTOXIC MECHANISMS
• Inflammation
• Immunosuppression
• Reactive oxygen species
• Reactive nitrogen species
• Receptor activation
• Epigenetic silencing
EXCRETION (kidneys, liver, lungs)
Genomic damage
Altered signal transduction
• Hypermutability
• Genetic instability
• Loss of proliferation control
• Resistance to apoptosis
CANCER

Fig. 5.3: Metabolic activation of chemical compounds and genotoxic and non-genotoxic effects of carcinogens
(modified from Oliveira et al., 2007)

Although it is in many respects a well-standardized protocol, it has been criticized as having poorly defined endpoints and a high level of uncontrolled variation. Suggestions for aspects of the test that could be optimized include proper randomization, blinding, better necroscopy work, and adequate statistics (Freedman and Zeisel, 1988). However, 20 years after its adoption by OECD, the most recent test guidelines (OECD, 2009) still do not make randomization and blinding mandatory, and the guideline statistics do not control for multiple testing, despite the fact that about 60 endpoints are assessed in the assay. Furthermore, the data analysis is ill-defined: "*When applicable, numerical results should be evaluated by an appropriate and generally acceptable statistical method*."

Reducing the duration of the assay to 18 months has also been suggested (Davies et al., 2000), although others contradicted the applicability of this option (Haseman et al., 2001).

In addition, the assay has not been standardized for animal strains, with the only definition being that "*young healthy adult animals of commonly used laboratory strains should be employed*." This is contrary to evidence that strain standardization appears to be a most critical factor. Even when the same strain is used, there appear to be problems with standardization that hamper the use of historical control groups (Haseman et al., 1997). In this study, the most commonly used strains showed strong weight gain and changes in some tumor incidences that resulted in reduced survival over just one decade, which was attributed to the intentional or inadvertent selection of breeding stocks with faster growth and easier reproduction. Other factors that have been suggested that could possibly influence the bioassay protocol over time include caging protocols, diet, environmental factors, genetic drift, study duration, and survival differences.

An analysis of 1,872 individual species/gender group tests in the US National Toxicology Program (NTP) showed that 243 of these tests resulted in "equivocal evidence" or were judged as "inadequate studies" (Seidle, 2006), suggesting the protocol as it stands is not robust. The two-species paradigm also has been challenged (Alden et al., 1996) by studies showing that rats are more sensitive, and regulatory action is rarely taken on the basis of bioassay results in mice (Van Oosterhout et al., 1997; van

Ravenzwaay, 2010). It is estimated that $ 1-2 million and up to 1,000 mice over a 3-year period would be saved by eliminating the mouse section of each chemical test (Alden et al., 1996).

Reproducibility
Gottmann et al. (2001) compared 121 replicate rodent carcinogenicity assays from the two sections (National Cancer Institute/National Toxicology Program and literature) of the Carcinogenic Potency Database (CPDB) to estimate the reliability of these experiments. They found a concordance of 57% between the overall rodent carcinogenicity classifications from both sources; this result did not substantially improve when species, sex, strain, and target organ information was considered. They concluded: "*These results indicate that rodent carcinogenicity assays are much less reproducible than previously expected, an effect that should be considered in the development of structure-activity relationship models and the risk assessment process*." Ironically, cell transformation assays (CTA, discussed in more detail below) appear to reproduce the cancer bioassay better than it reproduces itself. Thus, it appears likely that the existing bioassay would fail any validation investigation that a replacement test would be subjected to.

Potency correlation between species
This is not a classical validation criterion, but it is part of the Bradford-Hill criteria to support associations. The apparent correlation between potency of carcinogens in mice and rats has been shown to be largely an artifact (Bernstein et al., 1985).

Interspecies and organ site correlation
Concordances of 57% were reported between mouse and rat bioassays. Better correlations that were previously reported (71% rat to mouse, 76% mouse to rat) were driven by the abundance of strong mutagens studied, which are typically positive in all sexes, many species, and several organs (Gray et al., 1995). An analysis of bioassays in rats, mice, and hamsters, as well as comparisons with humans for known carcinogens, has shown that the likelihood of a chemical that induces tumors in one species in a certain organ also inducing tumors in another species in the same organ is less than 50% (Gold et al., 1991, 1998).

Sex specificity
A critical appraisal of the role of sex hormones (endocrine status) on species susceptibilities in chemical carcinogenesis (Toth, 2002) concluded: "*There are compelling indications, particularly in the fields of physiology and metabolism, to conclude the limited usefulness of the various animal species in sex hormone research. The findings allow only restricted inferences for the human species*."

Scientific relevance
The first critical issue is that of high-dose to low-dose extrapolation. The use of maximum tolerated doses appears to be the source of many artifacts. Jay Goodman, Michigan State University, is cited (Schmidt, 2002) as saying: "*If we're dealing with a situation in which the likely human exposure is in the same ballpark, then these (dosing regimens) may be applicable. But doses that are hundreds to thousands of times higher than normal exposures (such as those often given during animal testing) might be carcinogenic simply because they overwhelm detoxification pathways. In these cases, we see tumors along with gross histopahologic evidence of tissue damage*." However, dose regimens are defended by others (Bucher, 2000), and many substances test positive for carcinogenicity only at maximum tolerated doses, including some accepted human carcinogens. These results also might be interpreted as species differences that are hidden by high-dose artifacts at the expense of many false-positives.

Predictivity of point of reference (human cancer)
An analysis by Pritchard et al. (2003) suggested 69% predictivity of human carcinogenicity for the two-species cancer bioassay, which ironically dropped to 65% when it was combined with *in vitro* genotoxicity test findings (Pritchard et al., 2003). This contrasts with an analysis by Knight et al. (2006a,b), who showed that in 58% of cases considered by the EPA, they deemed results from a positive cancer bioassay as insufficient for assigning human carcinogenicity, even though the EPA was far more likely to assign this classification than the IARC. A previous comparison of known human carcinogens, as classified by the IARC mainly based on epidemiology, with corresponding animal data found an unconvincing correlation (Freedman and Zeisel, 1988): "*The research reports of the cancer community (even taken at face value) do not sustain the conventional argument for the validity of the qualitative extrapolation ... We remain sympathetic to the idea that animal data have some predictive value for carcinogenicity in humans ... But the evidence for such propositions is surprisingly weak*." It is also worth noting that the most typical sites of tumor formation in humans do not correspond to those in rodents (Anisimov et al., 2005), as shown in Table 5.1:

Tab. 5.1: Most common spontaneous cancers in humans and rodents
(adapted from Anisimov et al., 2005)

Cancer	Mice	Rats	Humans
Breast carcinoma	+	+	+
Lung carcinoma	–	–	+
Prostate	–	+	+
Colon	–	–	+
Skin	–	–	+
Stomach	–	–	+
Liver	–	–	+
Endometrial carcinomas	–	+	+
Leukaemia / lymphoma	+	–	+
Thyroid	–	+	+
Bladder	–	–	+

In the absence of human data, it might be considered reasonable to use data from tests in nonhuman primates for comparative purposes. Cancer bioassays in nonhuman primates were carried out on 37 compounds within 34 years (Takayama et al., 2008); the results were "... *Inconclusive in many cases*," but carcinogenicity was shown unequivocally for four of them.

Taken together, the cancer bioassay is "... *Often not relevant to human carcinogenesis risk assessment*." (Ward, 2007).

Specificity

About 50% of all chemicals tested in the cancer bioassay test positive (see Tab. 5.2), and 53% of 301 chemicals tested by the NTP were positive, with 40% of these positives classified as non-genotoxic (Ashby and Tennant, 1991). It is sometimes claimed that this high positive rate is due to the testing of suspicious substances, especially in early years of identification of mutagens. Of substances tested in the NTP simply because of exposure considerations, 80% were found not to be carcinogenic (Fung et al., 1995). In contrast, Johnson identified 60% of 128 high production volume chemicals to be rodent carcinogens (Johnson, 2003). A similarly high proportion, around 50% positives, can be found in various databases for pharmaceuticals (MacDonald, 2004).

Pharmaceuticals are rapidly discontinued when they are found to be possibly genotoxic, but also many non-genotoxic ones test positive in the cancer bioassay (Silva Lima and Van der Laan, 2000). "*The database compiled from the 'Physician's Desk Reference' (PDR), including registered pharmaceuticals only, also provides a good illustration of rodent tumor findings being irrelevant to humans*" (Davies and Monro, 1995; Silva Lima and Van der Laan, 2000). Over two decades, 101 out of 241 substances entered the market despite positive cancer bioassays, presumably primarily as a result of the positive bioassay testing being considered irrelevant compared to the medical benefit of the compounds (Davies and Monro, 1995). It is not known how many chemicals were rejected over the same period (Davies and Monro, 1995). An early analysis of 20 putative human non-carcinogens found 19 false-positives, suggesting only 5% specificity (Ennever et al., 1987). The inappropriateness of rodent carcinogenicity assays as currently performed has been examined by Roe (1987), who notes that: "*There can be no sense in testing chemicals for carcinogenicity in rats maintained under conditions such that 50-100% of them (the control animals) develop pituitary and mammary tumors, etc. There is no identifiable population of humans for which such rats could constitute a model*." The implications of these observations for risk assessment have been noted by Bridges (Bridges, 1988). However, others see even this as an underestimate (Sobels, 1987): "... *Carcinogenicity is expressed to a different extent in different species of rodents, so that bioassay results in only two rodent species are likely to underestimate the proportion of chemicals with carcinogenic potential*."

Tab. 5.2: Proportion of chemicals evaluated as carcinogenic
(modified from Ames and Gold, 2000; Gold et al., 2005)

	Proportion	%
Chem. tested in rats and mice	379 / 648	58%
- natural	86 / 165	55%
- synthetic	293 /493	59%
Chem. tested in rats or mice	751 / 1456	52%
- Natural pesticides	41 / 75	52%
- Commercial pesticides	79 /198	55%
- Chemicals in roasted coffee	23 / 32	72%
- Mold toxins	15 / 25	60%
Drugs (PDR)	117 / 241	49%
Drugs (FDA)	125 / 282	44%

Sensitivity

Assessing the sensitivity of the cancer bioassay is made difficult by the fact that most human carcinogens were designated as such, to a large extent, by animal tests (with the discussed problematic interspecies correlation), and those typically missed are not identified by other means. There are strong claims that all known human carcinogens are detected with the cancer bioassay (Huff, 1999; Rall, 2000), but this could be considered a self-fulfilling prophecy, as most of these classifications are based on animal experiments. However, not all known human carcinogens can be modeled in animals (Silbergeld, 2004). For example, there is

- no animal model of cigarette smoke-induced lung cancer,
- no rodent leukemia induced by benzene, and
- no genetic point mutations in animals induced by arsenic.

This situation does not necessarily represent a contradiction, as these agents are positive for carcinogenicity in other organs or by other modes of action. However, achieving the right classification but for the wrong reason is a questionable outcome. Furthermore, the current testing situation leads to an enormous number of false-positives; Rall suggests that only one in ten compounds is truly carcinogenic (Rall, 2000).

Despite all of these false-positives, cases of human carcinogens that are not detectable in animals remain, e.g., the anticonvulsant diphenyl-hydantoin (phenytoin) is classified as carcinogenic to humans but showed no carcinogenic effect in experimental mice and rats (Anisimov et al., 2005). Ennever and Lave (2003) also have discussed the chemical combination of aspirin/phenacetin/caffeine, which is classified as a human carcinogen but tested negative in both rodent species (Ennever and Lave, 2003). Johnson (2001) presents a list of the known human carcinogens that have been tested in the NTP rat bioassay prior to 2000 (Johnson, 2001): "*The list contains 10 different chemicals, counting the various forms of asbestos as one, the three nickel compounds as one, and the 10 benzidine-like compounds as one ... (Of) the 13 individual chemicals tested in four sex-species groups, two chemicals were positive in four groups, one was positive in three groups, six were* *positive in two groups, one was positive in one group, and three were positive in no (0) groups. Only two human carcinogens (thiotepa and benzene) are bona fide trans-species carcinogens. Thus, for NTPRB-tested chemicals, it is not evident that human carcinogens necessarily demonstrate clear trans-species carcinogenic effects.*" These examples clearly contradict claims of 100% identification of known human carcinogens. It is also worth noting that an early assessment of the bioassay suggested only 46% sensitivity based on 19 human carcinogens (Salsburg, 1983).

The fact that rats and mice predict each other only about 57% does not fit with an assumption that 100% of human carcinogens are detected, as it is fair to assume that humans are not better predicted by either species than they predict each other. These figures of 57% concordance between species, 10% real human carcinogens, and 53% positives in the rat, combine to give a sensitivity of 100% with a specificity of 47%. Lave et al. previously arrived at an estimate of 70% sensitivity as well as specificity, assuming 10% real human carcinogens (Lave et al., 1988).

If the same calculation is performed with the assumption that 20% of all chemicals are carcinogenic in humans, this results in 75% sensitivity with 53% specificity. Interestingly, when we use the suggested 28% positives in rat, if non-suspicious chemicals are tested the result is 0% sensitivity and 65% specificity. Thus, whatever assumption is used, the assay does not perform well by any standard.

In a telling modeling exercise, Gaylor (2005) showed that increasing the number of animals per group from 50 to 200 would result in statistically significant ($p<0.01$) dose-responses for 92% of substances tested (Gaylor, 2005). This shows how the inherent variability of the test produces false-positives and reduces specificity using the current data analysis process.

Applicability domain

An applicability domain, i.e., the part of the chemical universe where the cancer bioassay is applicable to make sufficiently correct predictions, has not been established for the rodent cancer bioassay. Occasional reference is made to a better prediction of (strong) genotoxic substances, but these substances are exactly the ones filtered out by the *in vitro* genotoxicity testing battery and are unlikely to be tested in the bioassay.

Performance standards

Performance standards have been introduced for test methods as a guide to demonstrating that a given variant of a test is equivalent to the originally validated test. No such performance standards exist for the bioassay, although they would be very helpful for evaluation of alternative test methods. Bucher reported a discordance rate of 13/38 for the transgenic approach (Bucher, 1998), or a level of agreement of 68%, which barely differs from the 65% shown by the Salmonella mutation test. In response, Johnson showed the fingerprint pattern of organ sites affected (Johnson, 1999), concluding "*... It seems unlikely that transgenic models could ever replicate or faithfully emulate the carcinogenic response observed in natural whole animals.*" More extensive evaluations were conducted by an ILSI/HESI committee (Cohen et al., 2001). An article on these efforts (Schmidt, 2002) resulted in a discussion of whether this is "bashing the cancer bioassay" (Johnson and Huff, 2002). While possibilities for improving the animal test are outside the scope of this paper, the discussion shows the difficulties of using the cancer bioassay as point of comparison.

These comments, taken together, indicate that the cancer bioassay – although it has never been formally assessed – appears to have severe limitations. Furthermore, the assay would not stand up to the assessment criteria that any potential replacement test would have to fulfill. However, these limitations are not fully understood by many who use the assay for validation of alternative methods or regulatory purposes.

It appears timely to address these limitations before embarking on the expensive process of developing and validating replacement strategies that would only then be measured against this wrongly-considered "gold standard" test. It might be debated whether this represents a case for formal invalidation (Balls et al., 2006; Balls and Combes, 2005), but an approach based on the principles of evidence-based toxicology (Hartung, 2010c) seems to be more appropriate in this scenario than formal validation. A formal assessment of the assay would allow widespread dissemination – and encourage acceptance – of the evidence for the assay's limitations.

In line with these suggestions, the REACH guidance by ECHA is already quite cautious in its recommendations for use of the cancer bioassay (ECHA, 2008): "*A carcinogenicity study may, on occasion, be justified. If there are clear suspicions that the substance may be carcinogenic, and available information (from both testing and non-testing data) are not conclusive in this, both in terms of hazard and potency, then the need for a carcinogenicity study should be explored. In particular, such a study may be required for substances with a widespread, dispersive use or for substances producing frequent or long-term human exposures. However, it should be considered only as a last resort.*"

5.2.2 Reduction to key events

This approach aims to replace *in vivo* testing with stand-alone alternative methods. Carcinogenicity traditionally is seen as the combination of genotoxicity leading to mutation and subsequent promotion of the mutation. The development of an *in vitro* genotoxicity test battery would be aimed at reducing these hazards to key events. This rather simplistic view, and the testing approach that would result, is unconvincing, as many mutagens are not carcinogenic and substances do not exist in isolation in real life; rather, people are exposed to complex mixtures of substances, including incomplete carcinogens, such that there are situations in which compounds that are either only genotoxic or only promoting complement one other. Furthermore, there is increasing evidence that many modifying factors influence organotropy, growth rate, metastasis, resistance to immune reactions and treatment, etc.

The Ames test is the best standalone predictor of the rodent bioassay of the traditional genotoxicity test battery, with about 60% sensitivity (Kirkland et al., 2005). Earlier, Gold et al. (1998) reported that out of 465 chemicals, 45% were found to be mutagens by the Ames test, 63% were carcinogens, and 72% were either, i.e., 79% of mutagens were carcinogens, but 43% of carcinogens were not mutagens, and 25% of the non-carcinogens were mutagens. When genotoxicity assays are combined, sensitivity inevitably increases but specificity falls: When all three tests were performed, 75-95% of non-carcinogens gave positive (i.e., false-positive) results in at least one test in the battery (Kirkland et al., 2005). For marketed drugs (which typically exclude substances found to be genotoxic during development), no particularly strong concordances were seen between the 29% positive for genotoxicity and the 38% with positive or equivocal findings in the cancer bioassay (Snyder and Green, 2001). These results raise strong questions as to whether such tests, alone or in combination, can really help to determine the carcinogenic potential of substances.

Cell transformation (Berwald and Sachs, 1963, 1965), i.e., lack of growth inhibition in the case of confluency, has been suggested as a key event reflecting mutagenicity and some initial effects on cell replication, reduced apoptosis, DNA repair, oncogene activation, suppressor gene inactivation, and epigenetic effects. The value of these assays has been long discussed (Combes et al., 1999), leading to parallel test guideline development at OECD and prevalidation at ECVAM (Vanparys et al., 2011). A key piece of the validation exercise was the detailed review document (DRP) prepared by OECD summarizing existing experience with the assay and some additions to this dataset (Mascolo et al., 2010; Mazzotti et al., 2002). The data presented in the DRP were considered at an OECD Expert Consultation Meeting in 2006. Overall, it was concluded that the SHE and BALB/c 3T3 assays had a strong ability to detect rodent carcinogens, with a good positive and negative predictive capacity and sensitivities and specificities in the 80% range. Unfortunately, detailed information on the validation, its peer-review by ECVAM's Scientific Advisory Committee (ESAC) (with only the statement now available for public discussion), and the subsequent conclusions of OECD are still not available (according to personal communications it was decided end of 2011 to proceed with the OECD guideline for the SHE assay but not for the Balb/c 3T3 assay), although a recent review sheds some light (Creton et al., 2011). A parallel Japanese validation study on an improved assay has been published in the meantime (Sakai et al., 2011).

Perhaps the greatest concern, however, relates to the lack of understanding of the mechanisms by which CTAs operate (Ashby, 1997; Farmer, 2002). It is puzzling, for example, how the 3T3 variant of the assay, which has limited metabolizing capacity (Colacci et al., 2011), can so well reflect *in vivo* carcinogenicity in rats, while activation of xenobiotics to form reactive substances is considered a key event for genotoxicity. The question, therefore, might be turned around: given the high false-positive rate of the cancer bioassay (as discussed below), does the CTA actually reflect the false-positives of an organism overwhelmed with maximum tolerated doses, where metabolism contributes little more?

The CTA validation shows the limitations of traditional validation studies. With costs of about € 15,000 per substance tested in one laboratory, the number of chemicals that can be included in a ring trial is extremely small, and the complexity and duration of the protocol results in transferability and reproducibility issues. A feasibility study conducted to assess some aspects of reproducibility showed that the CTA could be used for decision making when combined with retrospective assessments of its predictive value, as made possible by a modular approach to validation (Hartung et al., 2004). This points out the need to transition to novel types of objective assessments (Hartung, 2010c). However, such retrospective analysis of existing data requires a level of transparency of the process that typically is not provided.

The CTA represents a prime opportunity to replicate results of the traditional animal-based approach by reducing the tests to a key event. A recent evaluation based on 141 studies showed that the SHE-7 variant of CTA predictions of rodent carcinogenicity gave a sensitivity of 88%, specificity 77%, accuracy 85%, positive predictivity 89%, and negative predictivity 75% (Benigni and Bossa, 2011). More importantly, the detailed review paper of OECD 2006 indicated for the CTA a sensitivity of 90% of class I (known human carcinogens) and 95% of class II (possible/probable human carcinogens) (Long, 2007; OECD, 2007).

The CTA can and must undergo further optimization with regard to:
- transition to human cells
- addition of metabolic competence
- automation, especially of foci reading
- possible transition to biomarkers of cell transformation measurements
- statistics (Ponti et al., 2007)

The CTA also represents an interesting opportunity for pathway of toxicity (PoT) mapping, as discussed below.

A focus on key events, as just described, could also be applied to non-genotoxic mechanisms of carcinogenicity, such as immunosuppression, inflammation, and hormonal activity (Tab. 5.3). This corresponds to some extent with the type of information currently considered in weight of evidence approaches, especially in REACH, but it is more likely to form the basis

Tab. 5.3: Mechanisms of non-genotoxic carcinogenicity

1. Chronic cell injury
2. Immunosuppression
3. Increased secretion of trophic hormones
4. Estrogenic activity
5. Receptor activation
6. Block of gap junctional intercellular communication
7. Inflammation
8. Oxidative stress, reactive oxygen and nitrogen species (ROS and RNS)
9. DNA methylation
10. CYP450 induction

of integrated testing strategies (ITS) than stand-alone replacements. Notably, REACH guidance by ECHA lists a number of *in vitro* tests that add weight of evidence (Tab. 5.4).

5.2.3 Negative exclusion by lack of key property

The current use of the *in vitro* genotoxicity battery follows a negative exclusion approach, i.e., substances showing no genotoxic potential are considered of low carcinogenic potential. The limitations of this approach, namely a high false-positive rate of the combined use of these assays, are well known (Blakey et al., 2008; Kirkland et al., 2007) and addressed elsewhere (Benfenati et al., 2009; Kirkland et al., 2007; Pfuhler et al., 2009, 2010a). It appears that the cancer bioassay produces far too many false-positive results when compared to human hazards (Ames and Gold, 1990), the mutagenicity testing *in vivo* further adds genotoxic substances that are not carcinogens, and this is further aggravated by the over-predictive *in vitro* battery.

This approach is meant to be precautionary, but is it sufficiently accurate? Certainly, we have to call it prohibitive, as it excludes large parts of the chemical universe from many uses. A debate intended to improve genotoxicity testing has started (Goodman et al., 2007; Lorge, 2009), with the aim of also incorporating Tox-21c approaches and new technologies (Elespuru et al., 2009).

Other properties and tests might exclude carcinogenic potential. The concept of “no penetration, no harm” offers some opportunities, for example. Large molecular weight typically is accepted as an indication of no harm, although fiber toxicity, as seen with asbestos and now, increasingly, with nanoparticle toxicology, might challenge this (Hartung, 2010e; Hartung and Sabbioni, 2011). The major problem in this approach is its reliance on negative data (no uptake). This concept is further refined by the threshold of toxicological concern (TTC) approach, where exposure in sufficient quantity, rather than bioavailability

Tab. 5.4: *In vitro* tests adding weight of evidence for carcinogenicity assessments according to REACH guidance by ECHA
http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r7a_en.pdf (last accessed 08.09.2011)

“**in vitro** ***cell transformation assay results:*** *such assays assess the ability of chemicals to induce changes in the morphological and growth properties of cultured mammalian cells that are presumed to be similar to phenotypic changes that accompany the development of neoplastic or pre-neoplastic lesions* in vivo *(OECD, 2006). The altered cells detected by such assays may possess other targeted mechanisms of action, or can subsequently acquire, the ability to grow as tumours when injected into appropriate host animals. As* in vitro *assays, cell transformation assays are restricted to the detection of effects of chemicals at the cellular level and will not be sensitive to carcinogenic activity mediated by effects exerted at the level of intact tissues or organisms.*

mechanistic studies*, e.g. on:*

- *cell proliferation: sustained cell proliferation can facilitate the growth of neoplastic/pre- neoplastic cells and/or create conditions conducive to spontaneous changes that promote neoplastic development.*
- *altered intercellular gap junction communication: exchange of growth suppressive or other small regulatory molecules between normal and neoplastic/pre-neoplastic cells through gap junctions is suspected to suppress phenotypic expression of neoplastic potential. Disruption of gap junction function, as assessed by a diverse array of assays for fluorescent dye transfer or the exchange of small molecules between cells, may attenuate the suppression of neoplastic potential by normal cells.*
- ***hormone- or other receptor binding:*** *a number of agents may act through binding to hormone receptors or sites for regulatory substances that modulate the growth of cells and/or control the expression of genes that facilitate the growth of neoplastic cells. Interactions of this nature are diverse and generally very compound specific.*
- *other targeted mechanisms of action*
- ***immunosuppressive activity:*** *neoplastic cells frequently have antigenic properties that permit their detection and elimination by normal immune system function. Suppression of normal immune function can reduce the effectiveness of this immune surveillance function and permit the growth of neoplastic cells induced by exogenous factors or spontaneous changes.*
- ***ability to inhibit or induce apoptosis:*** *apoptosis, or programmed cell death, constitutes a sequence of molecular events that results in the death of cells, most often by the release of specific enzymes that result in the degradation of DNA in the cell nucleus. Apoptosis is integral to the control of cell growth and differentiation in many tissues. Induction of apoptosis can eliminate cells that might otherwise suppress the growth of neoplastic cells; inhibition of apoptosis can permit pre-neoplastic/neoplastic cells to escape regulatory controls that might otherwise result in their elimination.*
- ***ability to stimulate angiogenesis or the secretion of angiogenesis factors:*** *the growth of pre- neoplastic/neoplastic cells in solid tumours will be constrained in the absence of vascularisation to support the nutritional requirements of tumour growth. Secretion of angiogenesis factors stimulates the vascularisation of solid tumour tissue and enables continued tumour growth.”*

or no bioavailability, is employed. Exposure issues, as formalized in TTC approaches, are limited by the no-threshold philosophy for cancer hazards, which is under continual discussion (Calabrese, 2009; Crebelli, 2000; Kirsch-Volders et al., 2000; Lutz, 2000; Morelli, 2000; Neumann, 2009; Rhomberg, 2011). However, for some genotoxic mechanisms, such as aneugenic activity leading to tumors, thresholds are already accepted. It is also important to note that the bioavailability of a compound to cells in *in vitro* culture is often much higher than its bioavailability in tissue.

Ironically, the non-threshold concept and its broad acceptance might have been flawed from the beginning (Calabrese, 2011). It appears that this is actually an example of unsuitable statistics, where deterministic calculations are used to handle rare events (carcinogenic effects at very low concentrations); a switch to probabilistic methods (see Chapter 1) might resolve this. For a lay audience, Taleb has explained the concept in his bestseller *The Black Swan* (Nassim, 2010).

Notably, this testing is applied very differently in different sectors, allowing for example, TTC approaches for food contaminants (Barlow and Schlatter, 2010; Kroes et al., 2004, 2005; Munro et al., 2008; O'Brien et al., 2006; Pratt et al., 2009), or margin of exposure (MOE) approaches (Benford et al., 2010), in which differences between actual human exposure and the point of departure of toxicity in animal experiments are used. Some authors have an alternative way of stating this discrepancy: "*Analysis also indicates that many ordinary foods would not pass the regulatory criteria used for synthetic chemicals*" (Ames and Gold, 2000; Silva Lima and Van der Laan, 2000). As an extreme example, two of the authors have shown that using the same regulatory approach for alcohol as for TCDD (dioxin) based on carcinogenic potential in rodents would allow a person to drink one beer in 345 years (Ames et al., 1990).

A similarly pragmatic approach allowing TTC would be likely to help in a large number of cases for cosmetics and other consumer products, without even the need to introduce new test approaches (Kroes et al., 2007). It should be noted that recent refinements include genotoxicity data, thus bridging to actual test data (Felter et al., 2009). The question that arises is how this approach can be validated. Indeed, this might actually be a case that is better suited to an evidence-based toxicology (EBT) evaluation (Hartung, 2010c) than a prospective ring trial.

Many toxic endpoints, especially in genotoxicity, rely on reactive chemistry allowing interaction with target structures – the absence of structural features allowing direct reactivity or activation via metabolism represents another example of exclusion of a hazard. Rather simple approaches give valuable information (Pelkonen et al., 2009), but it seems unlikely that this is sufficient for exclusion of a hazard. Still, assays like the peptide reactivity assay might be explored as to their predictive value for carcinogenicity.

5.2.4 Optimization of tests

Both genotoxicity tests (Speit, 2009) and the CTA (Combes et al., 1999) leave room for optimization, as discussed in part earlier. This might improve their value as stand-alone tests as well as test blocks in an ITS. It is worth noting that the goals of such optimization might be very different: for an ITS, the goal is not necessarily the best prediction or highest sensitivity for each test component, but value added in complementing the other blocks of the strategy. Many earlier developments might need to be revised when tests are now considered for ITS instead of stand-alone applications.

A number of strategies might be able to improve the predictive value of existing test systems:

- extension of metabolic capacity
- organotypic 3-dimensional (co)-cultures
- more physiologic culture conditions such as homeostasis, oxygen supply, cell density
- transition from cell lines to primary cells or stem cell-derived systems
- use of human cells, preferably primary cells and possibly a battery of different human cell types, ideally derived from stem cells
- human cells that have wildtype p53 and are DNA-repair competent
- use of genetically stable cells
- refinement and expansion of endpoints measured
- restriction of maximally used concentrations
- standardization and automation
- quality assurance of procedures
- appropriate statistics and prediction models
- definition of applicability domains
- better understanding of the mechanism of action
- knowing the weaknesses and strengths of systems for the development of new models

5.2.5 In silico approaches

Due to the enormous costs involved, public interest, and the availability of *in vivo* data (especially from the NTP), carcinogenicity testing has been subjected to intense *in silico* modeling. For carcinogenicity prediction, however, the use of these models is rather limited. Benigni and Bossa summarized (Benigni and Bossa, 2006): "*Overall, it can be concluded that predictions for the individual chemical cannot be taken at face value and cannot replace the experiments, when necessary. Their main role is to complement the information of different nature and from different sources*." Other reviews have come to similar conclusions (van Leeuwen and Zonneveld, 2001): "*Let us now return to the original question: do current models contribute to the aim to relate exposure to carcinogenic response? ... We think the answer is at best 'to some extent'. The models used are not overly realistic for the purpose of data description, because they ignore essential processes*." And (Benigni, 2004): "*Study of the structure of the chemicals generates predictions with limited reliability for the individual chemicals*" but Benigni sees enormous value for priority setting for testing. A key prerequisite for improving the available models will be to generate larger homogenous datasets for modeling (Patlewicz et al., 2003).

An impressive discrepancy currently exists between studies employing external evaluations, such as the Predictive Toxicology Challenge (PTC), and internal validation results. For the PTC a training set of 509 compounds from the NTP was used, with results for carcinogenic effects (Helma and Kramer, 2003).

The US FDA used a test set with data from 185 substances. Fourteen groups submitted 111 models, but only five were better than random at a significance level of p=0.05, with accuracies of predictions between 25 and 79% (Toivonen et al., 2003). Two previous comparative exercises by the NTP had challenged models with 44 and 30 chemicals prospectively, i.e., with chemicals which were to be tested only (Benigni and Giuliani, 2003). The accuracy of *in silico* predictions in the first attempt was in the range of 50-65%, while the biological approaches attained 75%. The results in the second attempt (Benigni and Zito, 2004) ranged from 25 to 64%. In remarkable contrast, mere internal validations can show results of 75-89% predictivity for carcinogenicity (Matthews et al., 2006; Julien et al., 2004).

It is worth documenting that, although REACH guidance is overtly positive about (Q)SAR in respect to other tests, it reserves a reluctant tone for discussing the use of (Q)SARs in carcinogenicity testing: "*The capacity for performing the standard rodent cancer bioassay is limited by economic, technical, and animal welfare considerations, such that an increased emphasis is being placed on the development of alternative, non-animal testing methods. However, carcinogenicity predictions through use of non-testing data currently represent an extreme challenge due to the multitude of possible mechanisms. Prediction of carcinogenicity in humans is especially problematic.*"

However, it is important not to dismiss *in silico* options. While the REACH assessment is rather skeptical with regard to standalone *in silico* solutions, they have broad applicability in ITS or when combining *in vitro* and *in silico* techniques for a standalone test. As a recent consensus report concluded (Benfenati et al., 2009): "In silico *methods can be used for priority setting, mechanistic studies, and to estimate potency. Ultimately, such efforts should lead to improvements in application of* in silico *methods for predicting carcinogenicity to assist industry and regulators and to enhance protection of public health.*"

A worthy summary of the situation was given as early as 1994 by John Ashby: "*The accurate prediction of chemical carcinogenicity can only be achieved by a balanced consideration of the following factors: the chemistry and metabolism of the test agent, the interaction between toxicity and genetic toxicity, the possibility of non-genotoxic events that trigger subsequent non-targeted mutagenesis, the difference between activities observed* in vitro *and* in vivo*, and the possible inadequacy and/or partiality of all datasets and observations. Extrapolation of activities within a series of congeners is usually possible, but predictions across different chemical classes/ mechanisms of carcinogenicity are difficult. Artificial intelligence systems can be used to predict one or more of the above parameters given adequate learning sets, but the hope for a single, coherent and self-contained method of predicting all instances of carcinogenicity is unreal. The future of carcinogen/ mutagen prediction lies with data-rich artificial intelligence systems based on known mechanistic principles used selectively within the context of chemical and biological human insight. The major current obstacle to progress is the assumption that mutagenicity and carcinogenicity are unitary phenomena that can be learned and predicted by artificial intelligence systems operating in isolation.*"

5.2.6 Information-rich single tests

The advent of "omics" (genomics, proteomics, etc.), image analysis systems, and other high-content measurement systems has introduced new opportunities for pattern recognition: instead of choosing a more or less meaningful endpoint to represent the response of a biological system, a multitude of signals can be recorded, hopefully including meaningful ones among many meaningless ones. The art lies in filtering the former, but the availability of bioinformatics tools for this purpose is increasing. The advantage of this approach is that the most informative biomarkers can be chosen in an unbiased way, independent of our initial understanding of a system. These can be individual endpoints as well as patterns, which we call "signatures of toxicity" (SoT). When combined with existing biological knowledge, such as our understanding of pathways from biochemistry, cell physiology, molecular biology or toxicology, these signatures can ultimately be translated for assessing perturbations of the living system, i.e., using a systems biology approach. At the level of signatures, this is a simple correlative approach with many limitations, including that epiphenomena cannot be distinguished from causal factors. For example, repair responses will correlate with damage, but obviously do not cause it. As a result, early response genes are typically seen when transcriptomics is used in toxicology. These approaches bear the risk of being non-specific and uninformative about the mode of action. However, some of these limitations might be overcome as our understanding of pathways of toxicity increases (see below). As Adler et al. (2011) note: "*The mechanisms by which non-genotoxic carcinogens cause tumors are in most cases related to tissue- and species-specific disturbances in normal physiological control, gene expression patterns implicated in cellular proliferation, survival, and differentiation (Baylin and Ohm, 2006; Esteller, 2007; Widschwendter and Jones, 2002).*" This is almost a definition of systems toxicology.

For carcinogenicity, *in vitro* transcriptomics approaches are emerging (Guyton et al., 2009; Jacobs, 2009; van Kesteren et al., 2011; Vinken et al., 2008). These have been applied initially to genotoxic carcinogens (Tweats et al., 2007), but the approach makes just as much sense for the non-genotoxic mechanisms, and there are early indications that genotoxic and non-genotoxic effects can be discriminated (Magkoufopoulou et al., 2011; Plant, 2008). However, others have found that not even genotoxic carcinogens that do not function via DNA adducts can be identified (Benigni et al., 2010), although this identification seems to be possible in short-term animal tests (Fielden et al., 2008, 2011; Waters et al., 2010).

5.2.7 Integrated testing strategies (ITS)

Our understanding of chemical carcinogenesis is continuously improving (Cohen and Arnold, 2011). This means a comprehensive testing strategy, designed to complement an optimized genotoxicity testing toolbox, must integrate more and more mechanisms and modes of action. There are already suggestions, however, for how to generate ITS for genotoxicity (Pfuhler et al., 2010b; Aldenberg and Jaworska, 2010).

A relatively simple ITS combining the Ames test with the CTA for Ames-negative substances resulted in impressive predictions of the cancer bioassay and reduced *in vivo* testing needs by 90% (Benigni and Bossa, 2011) and the number of CTA by almost 50%. Even more tests were avoided (95%) when structural alerts were combined with the CTA.

Several ITS have been proposed, but their composition has been based primarily on the expertise and opinion of their respective proponents, as no principles for ITS composition are available (Cohen, 2004; Combes et al., 2007), see Figures 5.4 and 5.5. There is already an ITS suggested in the ECHA guidance to industry (Fig. 5.6).

Lave and Omenn started to model the combination of the cancer bioassay with a screening test as early as 1988 (Lave et al., 1988). In addition to sensitivity and specificity, the prevalence of the hazard among the substances studied is key for such calculations. Taking into account the societal costs of misclassification, they suggest that the screening test employed must be either the most accurate or the least expensive.

Specific considerations for non-genotoxic carcinogens
Although different mechanisms may be involved in the carcinogenic action of non-genotoxic compounds, several common characteristics may be defined (Silva Lima and Van der Laan, 2000):

1. *Specificity*
   In contrast to genotoxic compounds, which usually result in tumor development in several organs of the same animal species and even in several animal species, non-genotoxic carcinogens might be more specific with respect to their tumorigenic potential, as they frequently induce tumors in only mice or rats, one sex, and, in most cases, in one or few organs.
2. *Existence of a threshold*
   Often tumorigenic effects only occur when high doses of a compound are used in order to produce prolonged interference with normal physiological control and modifications of cellular proliferation patterns, i.e., threshold doses exist. In addition to the identification of the existence of a threshold, it is crucial to assess the mechanism by which high doses exert a carcinogenic effect.
3. *Reversibility*
   Tumorigenic/carcinogenic effects of non-genotoxic substances are observed only when a compound is continuously applied over extended periods and may be at least partially reversed after administration of the compound is discontinued.

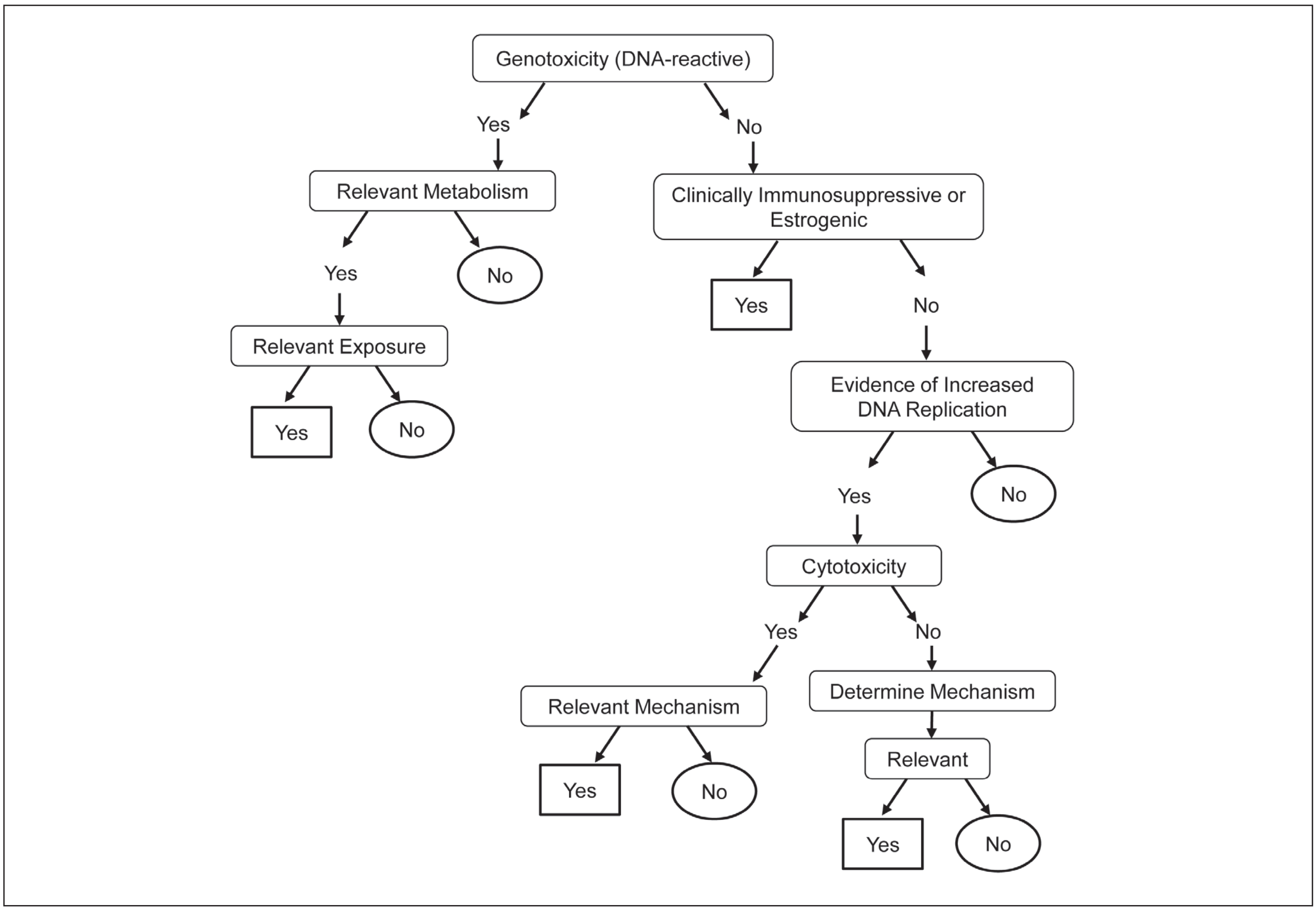

Fig. 5.4: Proposed guide for evaluating the potential carcinogenicity of chemicals
(modified from Cohen, 2004)
Each box poses an evaluation to be performed. If the sequence results ultimately in a NO that is in a circle, there is no (or negligible) carcinogenic risk in humans. If the sequence results ultimately in a YES that is in a square, it poses a presumptive carcinogenic risk.

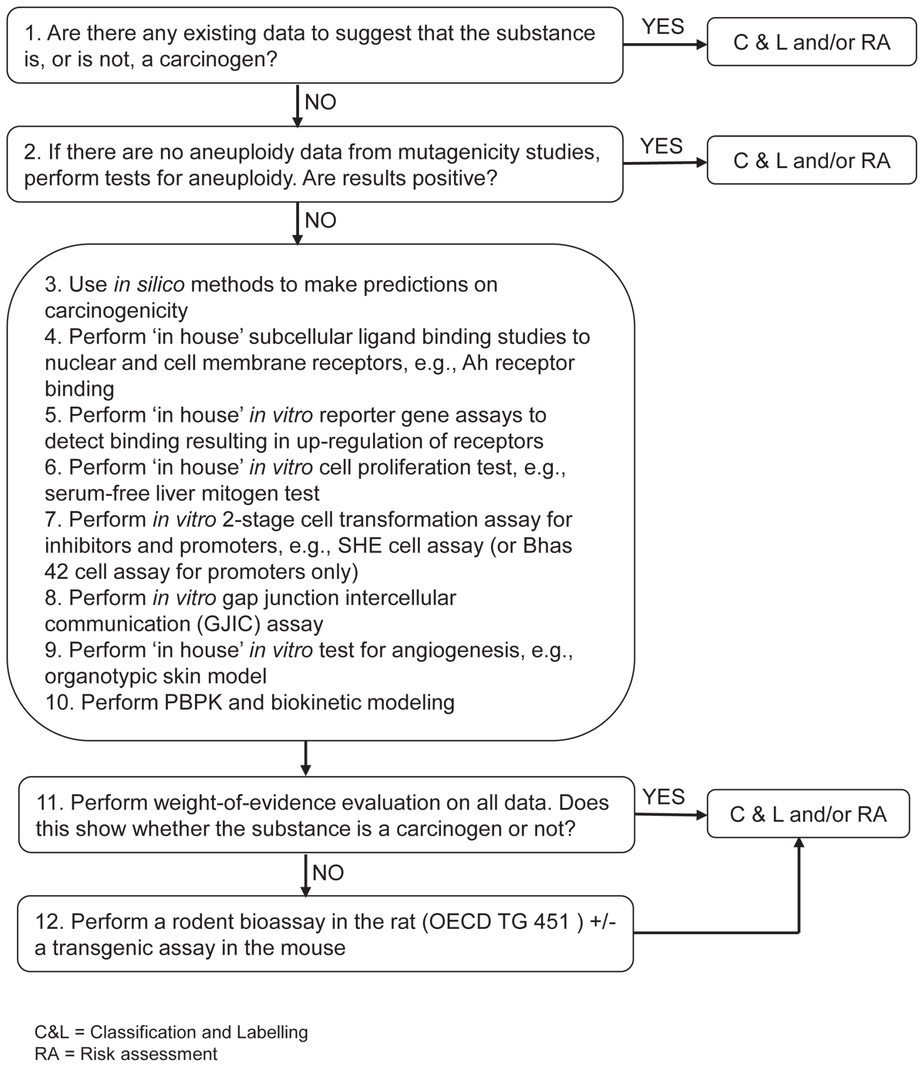
1. Are there any existing data to suggest that the substance is, or is not, a carcinogen?
YES
C & L and/or RA
NO
2. If there are no aneuploidy data from mutagenicity studies, perform tests for aneuploidy. Are results positive?
YES
C & L and/or RA
NO
3. Use *in silico* methods to make predictions on carcinogenicity
4. Perform 'in house' subcellular ligand binding studies to nuclear and cell membrane receptors, e.g., Ah receptor binding
5. Perform 'in house' *in vitro* reporter gene assays to detect binding resulting in up-regulation of receptors
6. Perform 'in house' *in vitro* cell proliferation test, e.g., serum-free liver mitogen test
7. Perform *in vitro* 2-stage cell transformation assay for inhibitors and promoters, e.g., SHE cell assay (or Bhas 42 cell assay for promoters only)
8. Perform *in vitro* gap junction intercellular communication (GJIC) assay
9. Perform 'in house' *in vitro* test for angiogenesis, e.g., organotypic skin model
10. Perform PBPK and biokinetic modeling
11. Perform weight-of-evidence evaluation on all data. Does this show whether the substance is a carcinogen or not?
YES
C & L and/or RA
NO
12. Perform a rodent bioassay in the rat (OECD TG 451) +/- a transgenic assay in the mouse
C&L = Classification and Labelling
RA = Risk assessment

Fig. 5.5: Decision tree testing strategy for carcinogenicity
(redrawn with permission from Combes et al., 2007), C&L = Classification and Labelling, RA = Risk Assessment

These characteristics suggest that different mechanisms are involved and call for complementing the genotoxic test battery with assays that address pertinent non-genotoxic mechanisms (Tab. 5.4).

For example, literature surveys showed that 38 out of 48 endocrine-disrupting chemicals (79%) studied were positive in at least one cancer bioassay (Choi et al., 2004; Dietrich, 2010). A number of endocrine disruptor assays have been developed for the respective screening programs, and some of them have even been validated and accepted at OECD level. Mode of action based tests also are available for many other mechanisms, or they can be easily adapted from tests designed for other purposes; for example, inflammation represents another non-genotoxic mechanism (Emmendoerffer et al., 2000; Ohshima et al., 2003) for which ample *in vitro* testing systems are available. The same holds true especially for immunosuppression (Carfi et al., 2007; Galbiati et al., 2010; Gennari et al., 2005; Langezaal et al., 2001; Lankveld et al., 2010) but also for chronic cell injury, increased secretion of trophic hormones, oxidative stress, and reactive oxygen and nitrogen species (ROS and RNS). An assay for CYP450 induction in cryopreserved human hepatocytes also is currently under prevalidation (Abadie-Viollon et al., 2010; Richert et al., 2010). Despite the positive fact that existing *in vitro* tests are already available, and appropriate, for testing for a number of non-genotoxic mechanisms, there are other non-genotoxic mechanisms that may contribute to carcinogenicity and are much more difficult to assess *in vitro*, such as immunesurveillance of cancer cells.

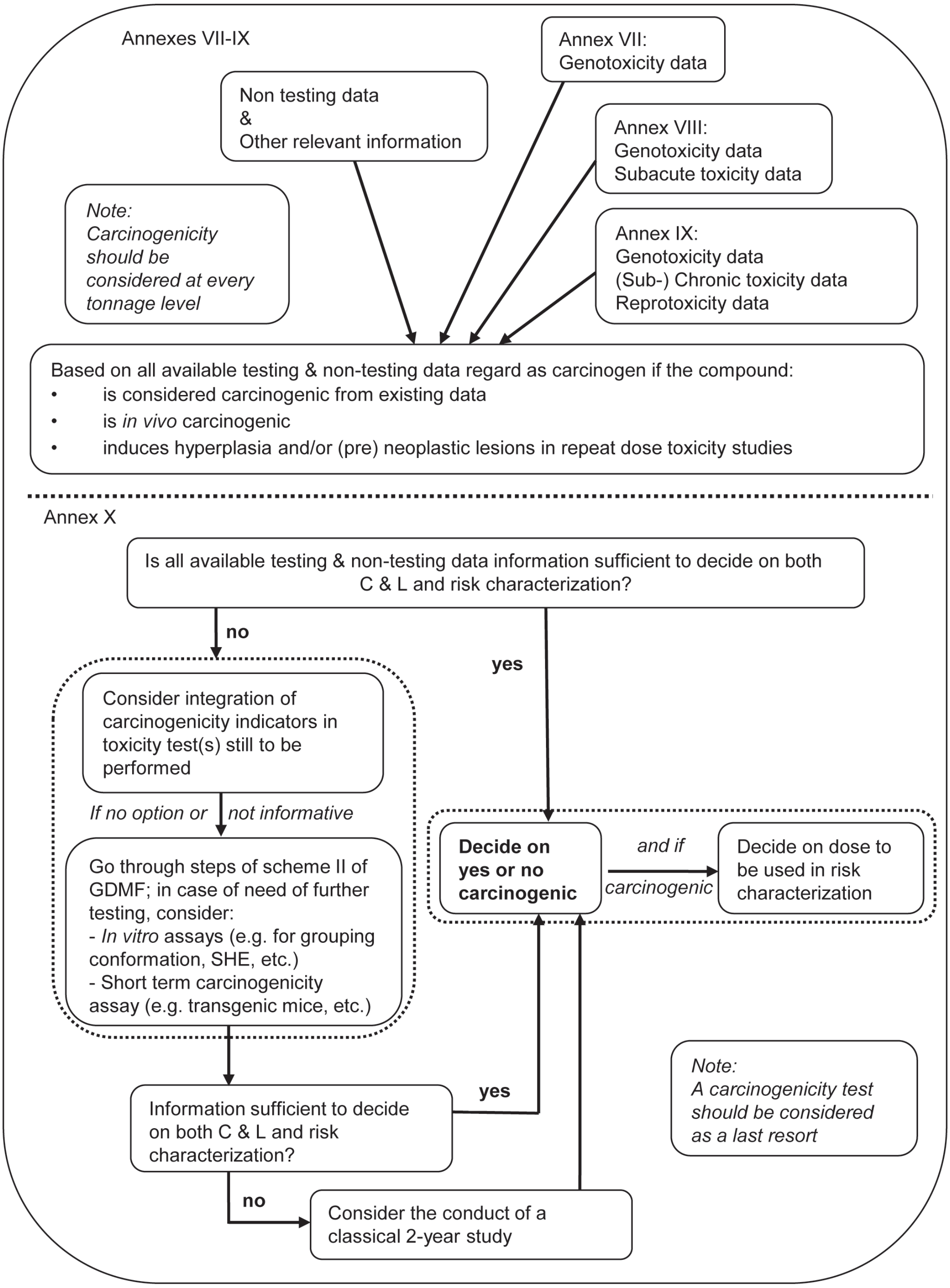
Annexes VII-IX
Annex VII: Genotoxicity data
Non testing data & Other relevant information
Annex VIII: Genotoxicity data Subacute toxicity data
Note: Carcinogenicity should be considered at every tonnage level
Annex IX: Genotoxicity data (Sub-) Chronic toxicity data Reprotoxicity data
Based on all available testing & non-testing data regard as carcinogen if the compound:
• is considered carcinogenic from existing data
• is in vivo carcinogenic
• induces hyperplasia and/or (pre) neoplastic lesions in repeat dose toxicity studies
Annex X
Is all available testing & non-testing data information sufficient to decide on both C & L and risk characterization?
no
yes
Consider integration of carcinogenicity indicators in toxicity test(s) still to be performed
If no option or not informative
Go through steps of scheme II of GDMF; in case of need of further testing, consider:
- In vitro assays (e.g. for grouping conformation, SHE, etc.)
- Short term carcinogenicity assay (e.g. transgenic mice, etc.)
Decide on yes or no carcinogenic
and if carcinogenic
Decide on dose to be used in risk characterization
yes
Information sufficient to decide on both C & L and risk characterization?
Note: A carcinogenicity test should be considered as a last resort
no
Consider the conduct of a classical 2-year study

Fig. 5.6: Integrated testing strategy for carcinogenicity
(modified from ECHA, 2008)

Jaworska and Hoffmann have defined a framework for ITS that will inform toxicological decisions in a systematic, transparent, and consistent way (Jaworska and Hoffmann, 2010). They reviewed conceptual requirements for ITS and presented a roadmap to an operational framework that should be probabilistic, hypothesis-driven, and adaptive, as well as outlining properties an ITS should have in order to meet the identified requirements and differentiate them from evidence synthesis. We strongly recommend that an ITS framework along these lines should be applied to a battery of mode-of-action tests. An example of this process in the context of sensitization testing was recently published (Jaworska et al., 2011).

5.2.8 Pathways of Toxicity (PoT) and systems toxicology

The concept of PoT, as detailed in Chapter 1, is being pioneered in the EPA ToxCast project. Phase 1, focusing mainly on pesticides and off-the-shelf available pathway assays in HTS platforms, has delivered impressive first results supporting the concept of PoT for carcinogenicity and genotoxicity as well (Knight et al., 2009; Martin et al., 2009a). The current expansion to more substances and substance classes, as well as PoT assays, represents a prime opportunity to explore this approach. This approach, however, is limited by the use of known PoT and available tests. Unsupervised identification of PoT by mapping the human toxome is the logical complement to this approach. We strongly recommend that a genotoxicity and carcinogenicity branch of this activity be developed. In both cases there are (pre-)validated tests, as discussed above, and human-relevant reference substances available. This process will lead to new approaches in carcinogenicity testing, especially when it is combined with the HTS approaches of ToxCast using similar substance sets.

Studies of cancer biology already have identified 12 signaling networks that are important in oncogenesis. Almost all cancers show perturbations in molecules in one or more of these pathways. Although these networks do not represent specific PoT, they may be useful starting points from which to look for biomarkers and identify potentially carcinogenic PoT. However, it is important to note that these pathways, as they are understood at the moment, are not necessarily predictive of carcinogenicity, as perturbations in the pathways often arise only as a later outcome of a mutagenic effect. Also, some non-genotoxic carcinogenic effects, such as immune surveillance escape, may not implicate one of these 12 pathways.

5.3 Conclusions and recommendations: carcinogenicity

The cancer bioassay is a two-year test conducted in rats or mice and is currently the only accepted test for carcinogenicity. Testing a single chemical compound using the cancer bioassay requires the use of at least 600 animals and costs approximately € 1 million – yet the assay is estimated to have a concordance of only 57% between rats and mice and to predict 9 “innocent” chemicals as being carcinogenic for each one it correctly identifies.

The cancer bioassay is a “one size fits all” assay and is, by definition, problematic, as a testing assay can be either specific or sensitive – but not both. We believe that an alternative to the bioassay, in the form of an integrated testing strategy (ITS) using animal-free tests – which is then subjected to probabilistic risk assessment – would provide better information on the carcinogenic risks of new and existing chemicals to regulatory agencies and the public. In this chapter, we present a roadmap for how this might be achieved.

We start with a summary analysis of the cancer bioassay, but we stress that this is by no means a complete and objective assessment of the assay. Indeed, our primary conclusion from this paper is a strong recommendation that such an assessment be carried out. Only when this is objectively conducted will it be possible to move forward effectively in the development of alternative testing strategies.

We continue with an analysis of the assessment framework presented in Chapter 1 with respect to carcinogenicity. We do not believe that it will be possible to find a standalone, single *in vitro* assay for carcinogenicity testing, and that reduction of carcinogenicity to a key event or negative exclusion by lack of a key property is too simplistic for carcinogenicity testing. While the cell transformation assay (CTA) provides a surprisingly high reproducibility of results compared to the bioassay, these findings should be considered with caution, and we feel the CTA needs further evaluation.

Our vision for an alternative to the cancer bioassay is an ITS that uses a combination of *in vitro* and *in silico* techniques to assess both genotoxic and non-genotoxic carcinogenicity mechanisms. Furthermore, we suggest that testing should be separated from risk analysis, and that the latter should be done in a probabilistic, rather than deterministic, manner. Finally, we feel strongly that genotoxicity and carcinogenicity pathways of toxicity (PoT) should be investigated as part of the newly established Human Toxome project, and this will feed new information into the carcinogenicity ITS.

The complexity of potential targets and interactions for systemic hazards prompted the use of whole animal test models to mirror as many of these targets and interactions as possible. We increasingly understand, however, that these tests inevitably bring with them a high number of differences in these targets and their interactions. As far as available data can determine, the correspondence between different animal species for the cancer bioassay is not better than 57% in rats versus mice, and there is no reason to assume that any of them predicts humans better than they predict each other. Reproducibility issues, small group sizes, and poor statistics further limit the reproducibility of these assays. With the express purpose of erring on the side of safety, animal models have been rendered more sensitive (precautious) by high-dose testing, with an overemphasis on any positive (i.e., toxicity) findings and the two-species paradigm.

This situation results in two major problems:

- There is no way to model the complexity of the hazard with simple systems.
- The results (where available) from animal tests as such do not qualify to validate novel approaches against them.

It is not realistic that any *in vitro* or *in silico* tool at this stage can be fully predictive of a human systemic toxicity. The questions that must be addressed, however, are how close can we come to this and how can we get closer to achieving that goal, especially by combining multiple approaches. *In vivo* alternatives to the bioassay do exist (especially the shorter assays in transgenic animals or when replacing the bioassay with a genotoxicity testing battery combined with a 28- or 90-day animal test), but they are beyond the scope of this paper.

Recommendations: carcinogenicity

The major recommendations from this report are:

1. It appears that the cancer bioassay has severe limitations. Assorted data as to its validity are available, but many of the analyses are relatively old. An objective evaluation of the test using EBT approaches is warranted. This will document the limitations of the assay and allow a more critical assessment of when – and indeed whether – it should be used. A general feeling in the expert panel was that the assay qualifies for invalidation. A better understanding of the assay's limitations also will be informative for interpretation of assay results in cases where it is still used. We also feel that an objective evaluation will provide a helpful impetus for the search for alternative approaches.
2. It is clear that the cancer bioassay must not be the point of reference for validation exercises in future approaches. While the assay may continue to be used in some cases until alternatives are available, these alternatives must not be compared to the bioassay.
3. The CTA and the genotoxicity test battery merit further optimization and evaluation in order to positively or negatively filter substances for carcinogenic potential. Larger datasets will also benefit modeling attempts. Although some evaluations of the CTA have shown it to be a useful alternative to the bioassay, these should be treated with caution, as it is difficult to understand how a CTA assay can reproduce animal tests better than animal tests reproduce themselves. To have the CTAs better accepted, it would be good to have the applicability domain (chemical classes, etc.) retrospectively determined on the basis of the information in the detailed review document (DRP) prepared by the OECD.
   Furthermore, its predictivity of human carcinogens should also be addressed. OECD is currently planning a further review of the CTA, and the findings of this process should be carefully considered during the development of an ITS as an alternative to the bioassay. The suggested evaluation of the bioassay will have important implications for this review of the CTA assays.
4. Such optimization should include the combination with high-content measures, *in silico* analysis, and automation for HTS.
5. Carcinogenicity qualifies for ITS development with a number of assays representing non-genotoxic mechanisms lending themselves to evaluation. A "CarcinoTect" evaluation, in a similar manner to the ReProTect process that has been conducted for reproductive toxicology, may be a good starting point for the development of a carcinogenicity ITS.
6. With (pre-)validated cell systems and ample reference compounds, especially from the IARC process, PoT identification represents a key priority to accelerate Tox-21c. PoT identification requires the complement of probabilistic condensation of the information generated.

6 A Roadmap for the Development of Alternative (Non-Animal) Methods for Reproductive Toxicity Testing

Author whitepaper: Thomas Hartung
Respondents: Robert Burrier, Thomas B. Knudsen, Michael Schwarz
Scientific writer: Nina Hasiwa
Discussants: David A. Basketter, Bas Blaauboer, Harvey Clewell, Mardas Daneshian, Sebastian Hoffmann, Joanna Jaworska, Ian Kimber, Paul Locke, Gavin Maxwell, James McKim, Emily A. McVey, Gladys Ouédraogo, Grace Patlewicz, Olavi Pelkonen, Annamaria Rossi, Costanza Rovida, Irmela Ruhdel, Andreas Schepky, Kerstin Trentz, Marian Turner, Philippe Vanparys, Joanne Zurlo, James Yager

6.1 Introduction: reproductive toxicity

Developmental and reproductive toxicity was not in the foreground of safety assessments for many years after the shock of the thalidomide disaster (Kim and Scialli, 2011) had died down. More recently, the European REACH legislation, which is extremely demanding in this field (Breithaupt, 2006; Hartung and Rovida, 2009a; Rovida and Hartung, 2009; van der Jagt et al., 2004; Rovida, 2010), has stirred discussion again, notably because tests like the two-generation study are among the most costly and require up to 3,200 animals (two-generation study) per substance. Another driving force is the European ban on testing for cosmetics ingredients (Hartung, 2008a). A series of activities by ECVAM, including several workshops, have tackled this challenge and will be condensed here. The Integrated Project ReProTect (Hareng et al., 2005) was one of its offspring, pioneering several alternative approaches.

Reproductive toxicity aims to assess possible hazard to the reproductive cycle, with certain emphasis on embryotoxicity. Only 2-5% of birth defects can be associated with chemical and physical stress (Mattison, 2010). This includes mainly the abuse of alcohol and other drugs. For the assessment of the prevalence of effects on mammalian fertility, the available database is even more limited.

This roadmap paper also has benefitted from the recent discussions, including the recent detailed analysis of the 2013 marketing ban for cosmetic ingredient testing in Europe (Adler et al., 2011; Hartung et al., 2011; Mattison, 2010). In addition, some activities under the auspices of ILSI/HESI and the US ToxCast project have helped to clarify opportunities and challenges.

This paper will not always distinguish clearly between developmental and reproductive toxicity, simply considering developmental effects (teratogenicity) as the key concern within reproductive toxicity (which obviously also includes aspects of fertility and other impairments of the reproductive cycle). Developmental processes are especially difficult to assess (Knud-

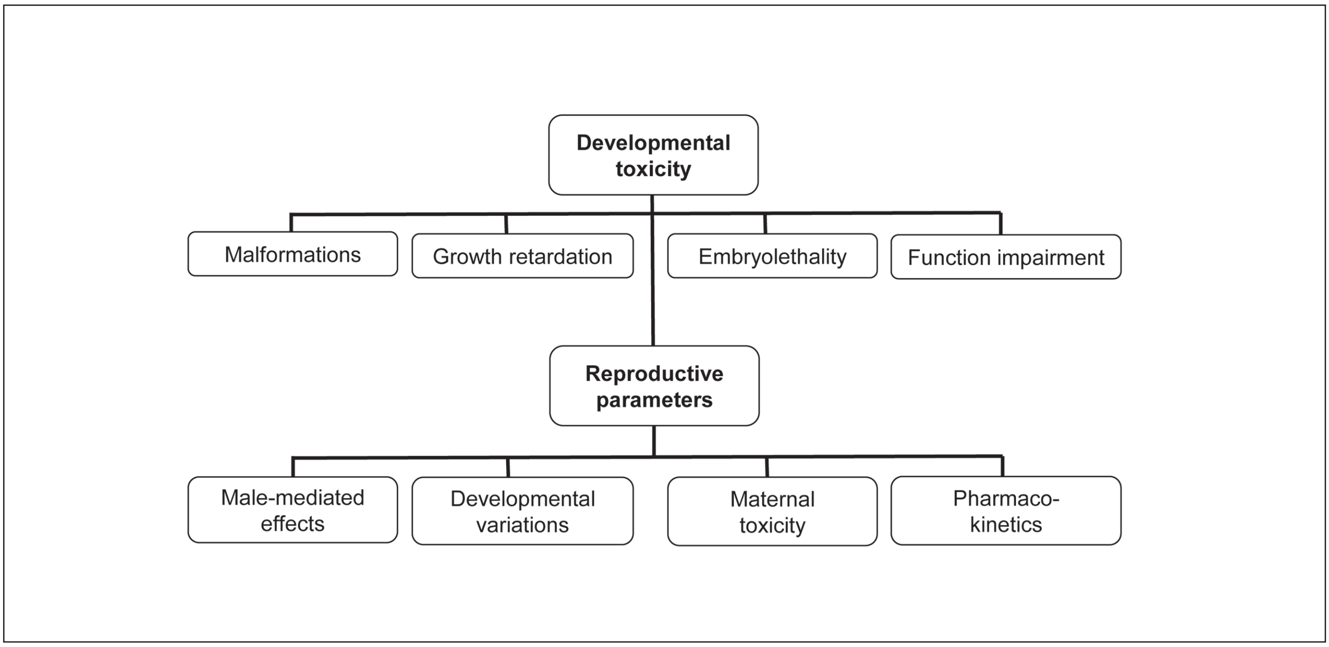

Fig. 6.1: Principal manifestations in developmental toxicity
(modified from Pellizzer et al., 2005)

sen et al., 2011), as the timing of processes creates windows of vulnerability, the process is especially sensitive to genetic errors and environmental disruptions, simple lesions can lead to complex phenotypes (and vice versa), and maternal effects can have an impact at all stages.

6.1.1 Current testing

The treatment of one or more generations of rats or rabbits with a test chemical is the most common approach for identifying chemically induced adverse effects on reproduction (Fig. 6.1). For evaluating developmental toxicity, test guidelines were designed to detect malformations in the developing offspring, together with parameters such as growth alterations and prenatal mortality (Collins, 2006). Developmental toxicity tests are considered mainly as screening tests (especially for REACH (Rovida et al., 2011)). The shorter and less complex "screening" tests, which combine reproductive, developmental, and (optionally) repeated dose toxicity endpoints into a single study design, are variants.

As a result of these studies (Tab. 6.1), a No Observed Effect Level (NOEL) is determined. These data then are extrapolated from animal studies to humans. In this process, safety factors are applied. This safety factor is normally 100, i.e., 1% of the dose that did not cause any adverse effects is considered safe in humans (acceptable daily intake values). The value of 100 is a common default as a safety factor (based on the assumption that 10 is an estimate of interspecies and another 10 of intra-species differences), but justifiable deviations are possible in both directions.

Reproductive toxicity testing has not been developed for, nor been largely applied to, chemicals in general – which is often overlooked – but has been used predominantly for pharmaceuticals and pesticides. Pharmaceuticals are designed for oral, high-dosage, effect-driven use, while chemicals, if at all, will typically affect the human body in a low-dose, long-term manner. Therefore, adapting the risk assessment of pharmaceuticals to chemical effects might not be appropriate. Despite that, the latter approach was introduced for chemicals several decades ago, but it held true only for new chemicals at a certain production volume. Very few new chemicals, however, are produced in high enough volumes to trigger such testing. Thus, experience with the predictive value and performance in general for ordinary chemicals is more than limited. So are the laboratory capacities available to carry out testing. Bremer et al. (2007) showed that in both the New Chemicals Database and the US EPA HPV database, any given reproductive toxicity test has been used for less than 3% of the notified substances (Bremer et al., 2007a) (Fig. 6.2).

Fleischer has demonstrated the limited testing facilities and a lack of sufficient scientific/technical know-how (Fleischer, 2007): A survey including 28 major independent and corporate laboratories in Europe indicated that only 11 offer two-generation studies with a capacity of 28 substances per year. This total suggests a capacity to carry out about 50 parallel, two-generation studies in Europe, each lasting about two years. Thus, every year 25 new substances can be included. The majority of this testing capacity is employed for drugs and pesticides. Only about three general chemicals per year have been tested in two-generation studies since the introduction of the Dangerous Substances Directive in 1981 (Fleischer, 2007). Thus, testing of hundreds or even thousands of chemicals in the context of REACH will overwhelm available test capacities. This calls for adequate prioritizing to make best use of these limited resources as well as for the use of any other means to satisfy the information requirement by way of an alternative and integrated testing strategy.

Tab. 6.1: Harmonized guidelines used in screening and testing for developmental and reproductive toxicities for EU, US EPA, and OECD

	OECD (Organisation of Economic Co-operation and Development)	OPPTS (Office of Prevention, Pesticides, and Toxic Substances), US EPA	EU Method
Prenatal developmental toxicity	TG 414 (2001)	OPPTS 870.3700 (U.S. EPA, 1998)	B 31
Reproduction and fertility effects	TG 416 (2001)	OPPTS 870.3800 (U.S. EPA, 1998)	B 35
One-generation reproductive toxicity study	TG 415 (1983)		B 34
Reproduction/developmental toxicity screening test	TG 421 (1995)	OPPTS 870.3550 (U.S. EPA, 2000)	
Combined repeated dose toxicity study with the reproduction/ developmental screening test	TG 422 (1996)	OPPTS 870.3650 (U.S. EPA, 2000)	
Developmental neurotoxicity	TG 426 (2007)	OPPTS 870.6300 (U.S. EPA, 1998)	
Extended one-generation reproductive toxicity study	TG 443 (2011)		

The expense and animal use associated with reproductive toxicity testing is questionable when considering that reproductive toxicity is most probably an event with a low frequency in the universe of industrial chemicals. An independent expert panel of industrial reproductive toxicologists has concluded that, in all likelihood, less than 5% of industrial chemicals possess properties that could be harmful to the developing child. We have found, by reviewing the New Chemical Database of the ECB, that 15 two-generation studies have led to only one R60 classification, whereas 58 one-generation studies have led to three classifications (Bremer et al., 2007a).

Publically available data on reproductive toxicity are very rare. Less than 5% of dossiers in the US EPA HPV database or the EU New Chemical Database (not public) contain any data in this field (Bremer et al., 2007a). Knudsen et al., have analyzed available data (Knudsen et al., 2011) in various databases: "*NIEHS' National Toxicology Program (NTP) online database, for example, provides developmental effects data on only about 3% of the listed chemicals (70 of 2,330). Other databases are similarly sparse for developmental (or reproductive) effects including FDA's Center for Drug Evaluation and Research publicly accessible database (16.3%; 58 of 355 listed compounds), and FDA's Center for Food Safety and Nutrition database (27.2%; 312 of 1,146 listed compounds; provided in Leadscope Databases (Leadscope) (Chihae Yang and Ann Richard, personal communication; see also Singh et al., 2010). The EPA Integrated Risk Information System (IRIS) contains comprehensive reviews for 553 environmental chemicals (as of April 2010), and identifies the most sensitive or 'critical effect' as the basis for setting safe exposure levels to protect the public health. The critical effect is the first observed effect deemed adverse that is likely to occur in the most sensitive species as the dose rate of an agent increases (IRIS, 2010). Less than 2% of 533 IRIS assessments report the critical effect for the derivation of a noncancer reference value (i.e., a safe exposure level) as being a developmental (5 of 553) or reproductive (4 of 553) effect (http://www.epa.gov/IRIS/). This may be due to other effects being more sensitive, but more likely due to a lack of de-*

	EPA HPV Program Database	ECB Database
Total number of substances	2200	4407
OECD 414	41	58
OECD 415	8	55
OECD 416	10	14
OECD 421/422	43	Not done
Other	144	Not done

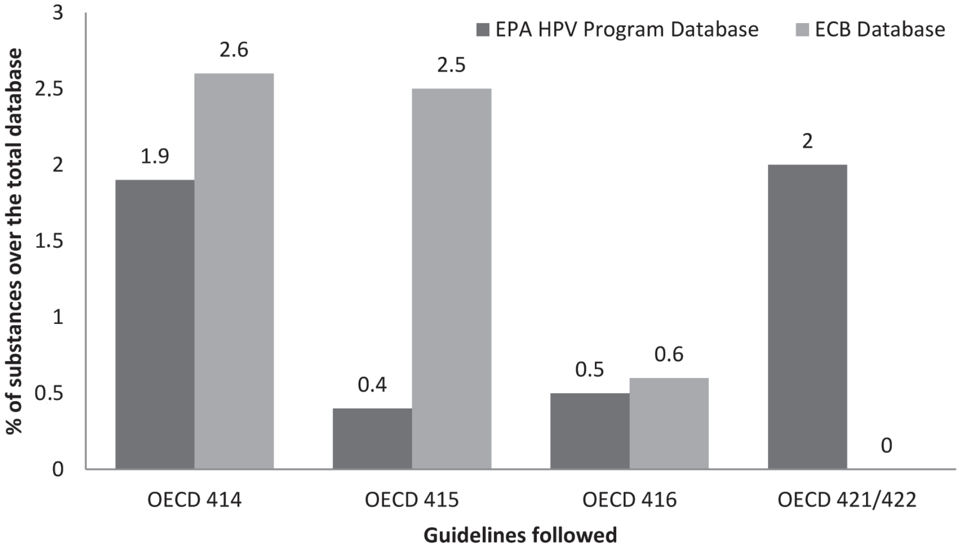

Fig. 6.2: Summary of existing data from the US-EPA HPV database and the ECB database fulfilling the standard information requirements of REACH
(modified from Bremer et al., 2007a)

velopmental and/or reproductive effects data, which contributed to an increased uncertainty in the database for the choice of the critical effect, and resulted in a lower reference value in 85% of the cases where an uncertainty factor for an inadequate database was used. Finally, in one of the largest data compilations from multiple resources to-date, EPA's Aggregated Toxicology Resource (ACToR) identified available developmental toxicity data for less than 30% of the 9,912 chemicals in commerce or of environmental interest, out of a chemical domain of 418,513 generic chemicals (Judson et al., 2009)."

It needs to be stressed that the described effects do not automatically point to impaired mammalian reproduction, but only to observed histopathological effects. The prevalence of reproductive toxicity is, most probably, lower than this query demonstrates.

To overcome low sensitivity, regulatory bodies often request testing in a second species. It should be stressed that the sensitivity of the test design requesting two species is still unknown. But the consequence of requesting two species is dramatic: By assuming a maximum prevalence of 5% for developmental toxicity in the universe of industrial chemicals, and by requesting additional testing in another species in case of a negative first study, the number of animals needed for developmental toxicity testing is nearly doubled. Fortunately, in a 2009 amendment to REACH, the original consideration of a second species was removed, though the respective guidance for developmental screening by ECHA has not yet been adapted (Rovida et al., 2011). In addition, a side-effect of requesting a second species that is often overlooked in the current testing practice but that will have a high impact on large testing programs, is the increase in the rate of false-positives, and therefore the unwanted restrictions of valuable substances (Hartung and Rovida, 2009a; Hartung, 2009a).

Many regulatory agencies have recognized the need for a transformative shift and have initiated research programs to achieve the vision and goals laid out by the NRC (Leist et al., 2008b; NRC, 2007). These include the NIEHS NTP *Roadmap for the 21st Century from 2004* (National Toxicology Program, 2004) and the *FDA Critical Path Initiative* (Woodcock and Woosley, 2008; Woosley and Cossman, 2007) of the same year. EPA created the National Center for Computational Toxicology (NCCT) in 2005 and launched the ToxCast research program in 2006; in 2009, the NRC vision was largely adopted as EPA's toxicity testing paradigm (EPA, 2009). The OECD initiated a Molecular Screening for Characterizing Individual Chemicals and Chemical Categories Project in 2007, published a monograph on a 2007 Workshop on Integrated Approaches for Testing and Assessment, and actively utilizes Test Guideline Committees and a QSAR Expert Group to ensure global harmonization and validation of any new approaches. What is most astonishing is the fact that we see more US and international activities than European contributions, though at this moment the highest demand for change is created by European legislations; efforts in the EU are mainly carried out by research consortia between academia and industry, with typically only long-term perspectives for transition into regulatory use.

6.1.2 Framework for replacing systemic toxicity by novel approaches

This framework is presented in more detail in Chapter 1. The following approaches to overcome animal testing for a given area were identified:

1. Abolition of useless tests
2. Reduction to key events
3. Negative exclusion by lack of key property
4. Optimization of existing tests
5. *In silico* approaches
6. Information-rich single tests
7. Integrated testing strategies (ITS)
8. Pathways of Toxicity (PoT) and Systems Toxicology

The distinction between (2) and (3) was made to stress that identifying positive or negative substances for a given hazard represents different approaches with different requirements as to prediction models, statistics, etc. Note that this framework remains largely on the level of hazard identification. Dose-response considerations and quantitative extrapolation to humans are not considered.

6.2 Application of the framework to reproductive toxicity testing

6.2.1 Abolition of useless tests

Every model has limitations – this holds true for *in vivo* (Hartung, 2008b), *in vitro* (Hartung, 2007a), and *in silico* (Hartung and Hoffmann, 2009) approaches. It is rare that a model contributes so little that it should be abandoned. In particular, it is impossible to predict whether a model cannot be improved to make a useful contribution in the future. The proposal by Balls and Combes (2005) to formally invalidate useless tests was the topic of a joint FRAME/ECVAM workshop (Balls et al., 2006). The participants finally agreed that invalidation makes sense only for prescribed regulatory tests, since the potential remains for further development and possible inclusion into the regulatory toolbox. Even though reproductive toxicity testing is not likely to be a candidate for abolition, it is worthwhile to apply criteria that typically are used for novel tests to illustrate the performance of the traditional tests.

The weaknesses of current developmental toxicity safety assessment were recently summarized as follows (Carney et al., 2011):

- Large numbers of animals required
- High cost per compound (>$ 100,000 per study)
- Long time requirements to evaluate each compound
- Capacity gap: cannot keep pace with increasing demands to evaluate existing and new chemicals, as well as mixtures
- Maternal toxicity: can confound data interpretation
- Fundamental knowledge of developmental biology for current animal models (e.g., rat, rabbit, monkey) is sparse relative to mouse or lower organisms
- Uncertainty regarding interpretation of low incidence findings
- Large amount of effort placed on the evaluation of minor skeletal variations with little impact on risk assessment

- Use of high doses that sometimes far exceed human exposure levels

At the same time there is increasing doubt as to the usefulness of the 2nd generation for testing of substances. Janer et al. (2007) have shown in a retrospective analysis that this made no relevant contribution to the regulatory decision-making. US EPA obtained similar data (Martin et al., 2009a) supporting the development of an extended one-generation study (TG 443, OECD; OECD, 2011), originally proposed by the ACSA initiative. Though of lesser relevance here, this shows that (elements of) study protocols can indeed be useless and warrant critical assessment.

Another way of asking the question of relevance is whether the test is more sensitive (responsive at lower concentrations) for reproductive toxicity than the maternal toxicity, i.e., repeated-dose toxicity. For this comparison, Martin et al. (2009b) analyzed data in ToxRefDB for 254 chemicals tested in both multigeneration and 2-year chronic studies, and 207 chemicals tested in both multigeneration and 90-day subchronic studies: "*For the majority of chemicals, potency values between the multigeneration, chronic, and subchronic studies were comparable, with a general linear relationship falling within ten-fold of each other. However, for four chemicals ... that caused parental or reproductive effects in the multigeneration study, there was no systemic toxicity observed in either the chronic or subchronic studies. For another five chemicals ... potencies for the most sensitive multigeneration endpoints were more than 10-fold greater than for the most sensitive effects in chronic studies. Of these five chemicals only thiamethoxam was more potent based solely on reproductive endpoints, that is, testicular atrophy*." This means with an assessment factor of 10, the hazard of reproductive toxicity might be covered for 99.8% of substances.

The assessment here will be based on the most common criteria for validation (Hartung et al., 2004).

Standardization of protocols

The protocol has recently been critically reviewed by Holson et al. (2006) and more recently by Carney et al. (2011), who conclude: "*Developmental toxicity safety assessment is mainly a descriptive science designed to detect adverse developmental outcomes, namely teratogenicity, intrauterine death, intrauterine growth retardation, and functional deficits. Evaluation of teratogenicity requires detailed examinations of fetal morphology, including external features, internal organs and tissues, and assessment of more than 200 bones of the fetal skeleton. These assessments have evolved over time, such that very subtle changes (often called variations) can be detected, in addition to (real malformations).*

The descriptive nature of these fetal examinations brings with it some critical challenges … One is that the evaluation criteria and nomenclature for fetal morphology has been difficult to standardize across different laboratories. Although this problem would seem to be easily remedied, it has been difficult because individual laboratories have built up large volumes of historical data based on their own criteria, and they also may use different animal strains and evaluate fetuses on different days of gestation. Fetal examinations also are very time consuming and labor intensive, and require a significant investment in examiner training in fetal morphology, coupled with extensive proficiency testing.

One issue with skeletal evaluation is the interpretation of minor skeletal variations and their impact on risk assessment. This issue was the subject of a previous ILSI-HESI expert panel project … (Daston and Seed, 2007). Depending on the laboratory's evaluation scheme, a large number of individual skeletal variations often are recorded and some occur at a very high incidence (sometimes >80%), even in control animals. Many laboratories distinguish between several subtly different degrees of ossification of individual bones, leading to a large volume of statistical analyses and evaluation of corresponding historical control data (reviewed in Carney and Kimmel, 2007). Although the practice of recording minor skeletal variations was established many years ago, we have since learned that the skeletal system possesses an extensive capacity to remodel during postnatal development, and current evidence indicates that many of the minor skeletal variations present in the term fetus are no longer evident postnatally. ... Thus, minor skeletal variations, particularly findings such as wavy ribs and minor delays in ossification are generally not considered adverse in and of themselves (Carney and Kimmel, 2007). ... The interpretation of fetal malformations can also be a challenge, particularly when faced with a low incidence of a particular malformation occurring in the high-dose group only. As highlighted by Palmer many years ago, 'because low rates of malformation are the rule, one faces the recurring nightmare of deciding whether one or two malformations are related to treatment or accidental' ... Currently there are few options for resolving these issues, which is of particular concern given the enormous impact on regulation of the chemical as well as the potential labeling of the compound as a teratogen. In some cases, the studies have been repeated using extremely large sample sizes, but this is obviously problematic in terms of animal use, costs, and time. Mechanistic studies are another option, although these may only be possible if higher doses can be used to increase the incidence. … statistics often are of limited help in resolving these uncertainties, as very large numbers of offspring are needed to achieve the statistical power needed to detect an increase in low incidence malformations. To overcome some of these statistical limitations, historical control data are considered in judging whether or not a low incidence finding seen in a treated group might have been a chance occurrence. However, historical control data should be used judiciously and within a reasonable time frame, as drift in the background incidence can occur over time, as can sudden spikes in the incidence of a particular effect."

The very extensive analysis by Holson et al. is based on experience with about 1,500 studies (Holson et al., 2006). It also is based on a 1984 analysis carried out by the National Center for Toxicological Research on behalf of FDA entitled *Reliability of Experimental Studies for Predicting Hazards to Human Development*, which was never published in the open literature. They show the background of "abnormal" reproductive outcome, for example the spontaneous resorption of small litter: 43% of rabbits with a single implant resorbed it and 10% terminated pregnancy prematurely via abortion. 3% and 5% abnormal outcomes were found for 2 and 3 implants, respectively. The authors also suggest: "*The slope of the dose-response curve (is) ... often steeper in developmental toxicity studies than in other toxicity studies*," which means that effects occur only close to maximum tolerated doses, which "*grossly overpredict risks*." Another problem they identify is the high background of spontaneous adverse developmental outcomes (Tab. 6.2).

Reproducibility

These screening protocols have been employed mainly in national and international programs to gather screening-level data for chemicals. However, this study design has limited sensitivity and produces a high level of equivocal results that often have to be further evaluated in more "definite studies," such as a prenatal developmental toxicity study and/or a two-generation study. Given that the screening requires 560 ani-

Tab. 6.2: Most commonly occurring developmental variations in control Hra:(NZW)SPF rabbits
(modified from Holson et al., 2006)

Total number examined (1992 – 2003)	Fetuses	Litters	% per litter
External	**10 278**	**1529**	–
Twinning	1	1	0.0 – 0.8
Visceral	**10 278**	**1529**	–
Accessory spleen(s)	1198	681	4.8 – 33.2
Major blood vessel variation	565	329	0.0 – 17.5
Gall bladder absent or small	150	115	0.0 – 7.8
Retrocaval ureter	142	110	0.0 – 5.4
Hemorrhagic ring around the iris	46	33	0.0 – 3.6
Spleen - small	6	6	0.0 – 1.0
Hemorrhagic iris	4	4	0.0 – 0.8
Liver - pale	2	2	0.0 – 0.6
Eye(s) - opacity	2	1	0.0 – 1.0
Accessory adrenal(s)	1	1	0.0 – 0.7
Renal papilla(e) not developed and/or distended ureter(s)	1	1	0.0 – 1.2
Skeletal	**10 278**	**1529**	–
13th full rib(s)	4082	1240	19.4 – 59.1
13th rudimentary rib(s)	1982	1042	8.1 – 32.5
27 presacral vertebrae	1724	766	4.5 – 32.1
Hyoid arch(es) bent	504	357	0.0 – 22.2
Sternebra(e) no. 5 and/or 6 unossified	448	274	0.0 – 11.4
Sternebra(e) with threadlike attachment	146	121	0.0 – 9.1
Sternebra(e) malaligned(slight or moderate)	117	108	0.0 – 5.0
Extra site of ossification anterior to sternebra no.1	106	84	0.0 – 7.4
Accessory skull bone(s)	80	69	0.0 – 5.0
7th cervical rib(s)	73	59	0.0 – 7.7
25 presacral vertebrae	35	31	0.0 – 7.4

The most commonly occurring manifestations of these findings are:
(1) right carotid and right subclavian arteries arising independently from the aortic arc (no brachiocephalic trunk),
(2) left carotid artery arising from the brachiocephalic trunk
(3) retroesophageal right subclavian artery.

Source: Data collected at WIL Research Laboratories, Inc.

mals per test, the application of this test in its present form as a screening tool should be reconsidered for large toxicological programs. The reasons for equivocal results can be several: One is that the data are simply inconclusive; another is that this is related to either variability or lack of reproducibility. Thus it is either reproducibility or robustness of the test that has an impact on reproducibility. An improvement of the test design to increase accuracy of the test by reducing the number of equivocal results is desirable. Notably, the "definitive" multi-generation studies also have a high rate of equivocal results: "*The number of equivocal results remained high across these six species at just under 25%*" (Bailey et al., 2005).

Hotchkiss et al. (2008) addressed the inherent variability of the litter-based endpoints: Power calculations were calculated for categorical effects based upon the numbers of malformed males versus males without malformations per dose group: "*If 20 animals per dose group are examined for malformations, then lesions occurring at an incidence of 25% or greater can be detected, whereas an incidence of 10% can be detected if all the pups are examined from 20 litters. If only ten males per group are examined, as recommended for histopathological analyses in some regulatory agency test guidelines, then effects are only detected statistically if about 50% or more of the tissues/organs are affected; a level of statistical power that many would consider inadequate.*"

Scientific Relevance

The relevance of studies raises a concern: "*However, if dosing was high enough to cause the above described 'maternal toxicity,' these doses often also cause some effects in offspring. So the crux is that, on one hand the experimenter must apply high doses in order to fulfill the guideline requirements, while on the other hand results achieved at such doses may lead to the classification of a compound.*" Holson et al. (2006) observe the problem of statistics applied without correction for the multiple endpoints assessed: "*Because, for example, a standard developmental toxicity study with ANOVA/Dunnett's and Kruskal-Wallis/Mann-Whitney statistical analyses performed on all parametric and nonparametric data, respectively, may involve as many as 100 to 300 individual hypothesis tests, the possibility exists for numerous spurious statistical findings.*" Another biasing effect is the "litter effect," i.e., the common observation that several fetuses of the same litter are affected, thereby "*... artificially inflating the apparent group response*" and leading to false-positive results.

Predictivity of point of reference (human reproductive toxicity)

The ability of animal models to predict the human response is a fundamental assumption in developmental toxicity and risk assessment, yet varying degrees of discordance among species are very common in actual practice. Pronounced interspecies variances have been described showing not more than 60% correlation between different laboratory mammalian species in the area of developmental toxicity. There is no reason to assume that any species predicts humans better than, e.g., mice predict rat developmental toxicity of a given chemical. Hurtt et al. (2003) have demonstrated by analyzing 91 veterinary drugs that no single species (rat, rabbit, or mouse) was capable of detecting more than 61% of the teratogens. However, this study should be interpreted with caution since Schardein (2000) has provided an extensive study in which several hundreds of chemicals have been assessed for their interspecies variations. Bailey (2005) examined the data for 11 groups of known human teratogens across 12 animal species and found huge variability in positive predictability, with a mean of 61% (Bailey et al., 2005): "*Of the 139 individual classifications across the species tested, a total of 78 (56%) were positive; the remaining 44% of results were almost entirely negative. The only encouraging aspect to come from these statistics appears to be the high positive predictability score for the hamster; however, the USFDA published a report detailing the responses of the mice, rats, rabbits, hamsters, and monkeys to 38 known human teratogens in which the high scoring hamster produced only a 45% rate of correct positives (USA FDA Federal Register 'Caffeine,' 1980). Furthermore, the mean percentage of correct positives from any one of these species was only 60%... The US FDA report also analyzed the rate of concordance between these species and humans for 165 compounds known to be non-teratogenic in the latter; the 'order of merit' for each species and its negative predictive value were completely different from that for the positive predictive values, ranging from 80% in monkeys to 35% in mice and hamsters. The mean negative predictive value for any of these species was 54%. Taken together, these predictive values of 60% and 54% for human teratogens and human non-teratogens, respectively, represent a poor return on the investment of animals, time, labor and money. The 57% mean value is little better than the 50% that would have been obtained by pure chance.*"

The "precautionary" response of regulatory toxicology was to test in more than one laboratory animal species in order to reduce the 40% missed potential developmental toxicants. However, this inevitably increases the already 40% false-positive classifications (Hartung, 2009a). Whether we can afford this substantial over-labeling, especially in high-production volume chemical evaluation programs, has been discussed elsewhere (Hartung and Rovida, 2009a).

Discordance in developmental toxicity testing certainly seems to conflict with the widely held dogma stating that the basic events in embryo development are highly conserved across species, even for species as disparate as fruit flies, frogs, mice, and humans. This degree of conservation mainly applies to the most fundamental processes in embryogenesis, such as establishment of the general body plan, pattern formation, cellular induction, and regulation of differentiation via signaling pathways. On the other hand, pharmacokinetics and, in particular, maternal metabolism, can vary widely between species and are likely to drive interspecies discordance. Placental anatomy and physiology also vary greatly between conventional test species and humans. In fact, rats, mice, and rabbits utilize two very different types of placentae – the inverted visceral yolk sac placenta which is extremely important in early pregnancy, as well as a chorioallantoic placenta which does not become functional until mid-pregnancy. In contrast, humans only utilize a chorioallantoic type of placenta throughout most of gestation (Georgiades et al., 2002).

Holson et al. (2006) list the following limitations for reproductive toxicity assessments for the most common species:

> *"Rat – susceptible to dopamine agonists (dependence on prolactin for maintenance of early pregnancy), prone to premature reproductive senescence following treatment with GABAnergic and other CNS-active agents, increased susceptibility to Leydig cell tumors, increased susceptibility to mammary tumors, inverted yolk sac placentation, limited fetal period.*
>
> *Rabbit – Consume diet inconsistently, prone to abortion and toxemia, induced ovulatory, sensitive to local gastrointestinal disturbances (e.g., antibiotics), not routinely used in repeated-doe toxicity studies, prone to resorption when few implantations are present, inverted yolk sac placentation."*

Specificity

There are many examples of positive results in the routine species that have little or no effect in humans ("false-positives"), especially at normal exposures and therapeutic dose levels. Notable examples include glucocorticoids, benzodiazepines, caffeine, carbon dioxide, dopamine, indomethacin, and aspirin (Bailey et al., 2005; Hartung, 2009c). A simple calculation shows that a prevalence of 2.5% reproductive toxicants in humans among industrial chemicals when tested in two species (correlating with each other and humans at 60%) will result in 65% of all substances labeled false-positive, while 2.1% real-positives (85% of all positives) of the toxicants are identified (Hartung and Rovida, 2009a; Hartung, 2009a).

In 1983, Brown and Fabro estimated, that "Of those agents thought not to be teratogenic in man, only 28% are negative in all species tested" (Brown and Fabro, 1983) (Tab. 6.3).

They also did not find a strong concordance of potency (Tab. 6.4).

Tab. 6.3: Concordance of human and animal teratogenicity data

(modified from Brown and Fabro, 1983)

Human teratogens[Ψ]		**Human non-teratogens**[🎗]	
Test species	*% with positive response (correct positives)*	Test species	*% with no positive response (correct negatives)*
Mouse	85%[‡]	Mouse	35%
Rat	80%	Rat	50%
Rabbit	60%	Rabbit	70%
Hamster	45%	Hamster	35%
Monkey	30%	Monkey	80%
Two or more species	80%	Two or more species	50%
Any one species	97%	All species	28%

From US FDA

[Ψ] 38 compounds: "reports of birth defects in humans associated with intake."

[🎗] 165 compounds: "for which human teratologic effects have not been reported."

[‡] From the published information, the exact meaning of an 85% response rate is not clear. It could mean, for example, 85% of the *agents* were positive in at least one mouse study, or of all *tests* of these agents in the mouse, 85% were positive.

Tab. 6.4: Comparison of teratogenic potency of chemicals in humans and animals

(modified from Brown and Fabro, 1983)

	Lowest effective dose (mg/kg/day)			
Chemical	***Humans***	***Animals***	**Species**	**Ratio of Animal dose:human dose**
Methyl mercury	0.005	0.2	Cat, rat	50
DES[Ψ]	0.02	0.2	Rhesus monkey	10
Methotrexate	0.042	0.2	Rat	4.8
Aminopterin	0.05	0.1	Rat	2.0
PCBs[🎗]	0.07	0.125	Rhesus Monkey	1.8
Thalidomide	0.5	2.5	Rabbit	5.0
Phenytoin	2.0	50	Mouse	25
Alcohol	400	1500	Rat	3.8

From the Council on Environmental Quality

[Ψ] Diethylstilbestrol

[🎗] Polychlorinated biphenyls

Similarly, Bailey and Knight (2005) summarized their collected data (Bailey et al., 2005): *"This means that of 1223 definite, probable, and possible animal teratogens, fewer than 2.3% were linked to human birth defects."*

The consequence of low specificity in order to boost sensitivity, which can be seen as "precautionary," creates concerns as to the societal costs (Durodie, 2003). A breakdown of embryotoxic effects of 74 industrial chemicals, which have been tested according to EU Directive 67/548/EEC B31 in the New Chemical Database, showed that 34 chemicals have demonstrated effects on the offspring, but only two chemicals have been classified as developmentally toxic according to the standards applied by the national competent authorities (Bremer and Hartung, 2004). This demonstrates the lack of confidence in the specificity of this "definitive" test.

Sensitivity

The same analysis by Bremer and Hartung (2004) showed that 55% of these chemical effects to the offspring could not be detected within multi-generation studies (Fig. 6.3), which suggests that either the developmental toxicity screening tests are over-predictive or that the multi-generation assays lack sensitivity (Bremer and Hartung, 2004).

This is in contrast to claims that "*... Every chemical or drug known to be teratogenic in humans, with possibly two exceptions, is also teratogenic in one or more laboratory species*" (Schardein, 2000). One such exception is the prostaglandin E_1 analogue misoprostol: Treatment of humans with this drug for peptic ulcer disease or to initiate labor has a strong association with fetal malformations, but is not teratogenic in the rat (Bailey et al., 2005). However, the identification of known human teratogens also was not necessarily in the routine species, and we have to keep in mind that this was often retrospective analysis, where the effect in humans was known and was looked for in the animal studies, creating considerable bias.

Applicability domain

An applicability domain, i.e., the part of the chemical universe where the animal tests give correct predictions, has not been established for the different animal studies.

Taken together, no comprehensive critical evaluation of current *in vivo* testing is available, with the exceptions of a book chapter by Holson et al. (2006) and two narrative reviews (Bailey et al., 2005; Carney et al., 2011). There is some concern, which warrants a systematic review. Evidence-based toxicology offers a toolbox for such evaluations. Hartung and Hoffmann (Hartung, 2010c; Hoffmann and Hartung, 2006) conclude: "*Thus, a crucial need remains for an organized and critical analysis of the primary literature in reproductive toxicology to evaluate the concordance of regulatory reproductive toxicity studies to human exposure outcomes*." A critical problem is that reference data from humans are difficult to obtain from epidemiology (Friedman, 2009). Our knowledge of human teratogens is very much limited to drugs. Furthermore, we are lacking a process such as IARC for carcinogenicity in the field to achieve consensus on the reproductive toxicity of substances. Validation of novel tests against the traditional animal models should be done with

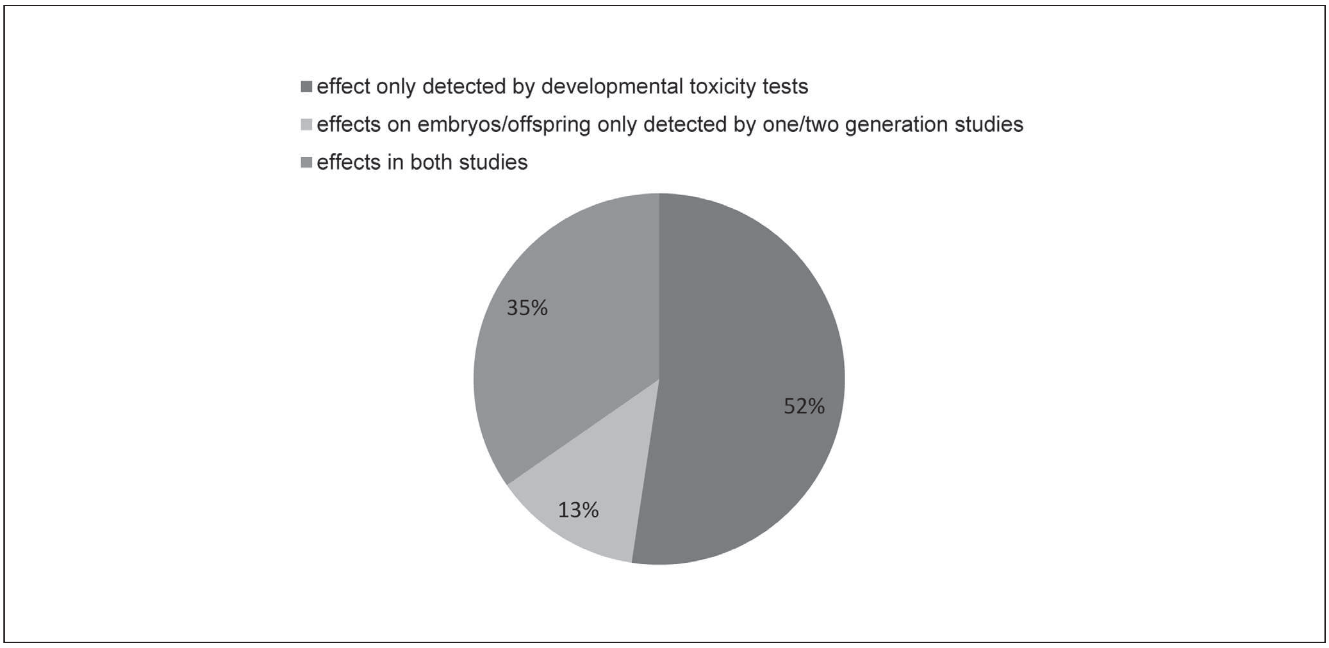

Fig. 6.3: Correspondence of developmental toxicity screening studies and (multi-)generation reproductive toxicity studies in the ECB database

(reprinted with permission from Bremer and Hartung, 2004)

The figure shows that not all embryotoxic effects will be picked up in one/two generation studies. Additional developmental toxicity tests are necessary. Further investigation is necessary to understand whether one/two generation studies can be combined with a set of *in vitro* methods for developmental toxicity in a conceptual framework that could perform reliable hazard identification of a chemical.

caution. A formal invalidation (Balls and Combes, 2005) is not likely to find major support in the absence of valid alternative approaches, given the importance of this subject. Nevertheless, it might be worthwhile to examine the *in vivo* reproductive toxicity tests applying the principles of evidence-based toxicology (EBT) (Hartung, 2009b). The development of *in vitro* methods might be furthered by evidence that the current test system is not providing the safety information we are looking for.

6.2.2 Reduction to key events

Most reproductive toxicity testing is done to exclude teratogenic effects. For this reason, an early focus in alternative method development was on tests for embryonic malformations (Augustine-Rauch et al., 2010). The most complete reflection of embryonic development apparently can be achieved with zebrafish embryos (Selderslaghs et al., 2011; Sukardi et al., 2011; Weigt et al., 2010, 2011; Yang et al., 2009), for example using dynamic cell imaging, or frog eggs (FETAX assay) (Hoke and Ankley, 2005), which has been evaluated more critically by ICCVAM[1]. It seems to be timely to evaluate available protocols and datasets and define a protocol for formal validation.

By 2002, three well-established tests had already been validated, i.e., the mouse embryonic stem cell test, the whole rat embryo culture, and the limb bud assay (Genschow et al., 2002, 2004; Piersma et al., 2004; Spielmann et al., 2004). They obviously cover only a small part of the reproductive cycle and only a small though critical part of embryonic development. Among them, the murine embryonic stem cell test (EST) has attracted most interest. Originally a counting of beating heart cells formed, it is now adapted to other endpoints and to human cells (Leist et al., 2008a). At present, the EST has its application primarily in in-house hazard identification. To reach regulatory implementation, further characterization is needed, such as definition of biological and chemical applicability domain, mechanistic studies to identify developmental pathways, comprehensive comparison of the developmental processes active in EST, differentiation with *in vivo* embryogenesis, and ultimately predictability of the EST for the developmental phase covered.

The entire reproductive cycle with its vulnerabilities most probably cannot be broken down to one or few key events. For practical purposes, however, we might test for these, especially when certain alerts lead to these test needs, typically from findings in repeated-dose testing. Why study the entire reproductive cycle when an alert already hints at a certain problem? If these data are insufficient for regulatory decisions, but alerts have been identified, the existing data can be used as the basis for the development of a tailored testing scheme. Depending on the nature of the alerts, test batteries of specific validated *in vitro* tests could be triggered in order to confirm or refute observed concerns. For example a histopathology in testes observed in repeated-dose studies will be followed up by tests on spermatotoxicity models, not a two-generation study if the classification cannot be done based on the finding alone. This approach, which we termed "*alert-driven testing*" (Bremer et al., 2007a), seeks targeted testing that provides sufficient toxicological data for hazard identification, but also keeps *in vivo* testing to a minimum.

A frequent possible scenario for an alert-driven strategy could be unclear histopathological observations in the testes in subacute or chronic toxicity studies. These findings should not automatically trigger additional animal-intensive tests for reproductive toxicity. These effects should be further explored, however, by using *in vitro* testing batteries analyzing cytotoxic effects on specific cell populations of the reproductive organs and/or by analyzing relevant hormone production or by monitoring gametogenesis *in vitro*. The obtained data will identify if observed changes in the tissues of reproductive organs are reprotoxic effects or if the observed effects are related to general toxicity. The establishment of relevant databases (Judson, 2010) such as the Fraunhofer society's database (Bitsch et al., 2006) or ToxCast's ToxRefDB (Knudsen et al., 2009; Martin et al., 2009a,b) will support the development of such a scientific approach. A query of the former database (Bremer et al., 2007a) containing 329 chemicals tested in repeated dose studies (rats) and 203 chemicals (mice) has demonstrated that major targets of chemicals showing toxicological effects on the testes are target cells that can also be cultured *in vitro*. However, substantial research efforts are still necessary to maintain the functionality of target cells *in vitro* and to convert these *in vitro* models into predictive tests using specific functions as toxicological endpoints. Changes of the functionality of certain target cells will point to the relevant target mechanisms and will support the interpretation if the observed effects are relevant to humans.

6.2.3 Negative exclusion by lack of key property

Instead of positively identifying a key property, which would lead to classification, it is often more attractive to exclude a key property to come to no classification. This is especially the case when hazards are relatively rare and positive identification will not save many test efforts. Properties that especially come to mind are the barrier models suggesting limited bioavailability. This can be oral availability on the side of the mother or the placental barrier. Models are available for both (Bremer et al., 2007a; Mose and Knudsen, 2006; Myren et al., 2007; Poulsen et al., 2009) and oral uptake (see Chapter 2), but a key problem is whether they are sufficiently predictive to completely rule out a possible effect, especially as the placental barrier changes its properties over time during pregnancy. Bremer et al. note the differences of the placenta between rodents and humans (Bremer et al., 2007a): "*Rodents have an inverted yolk sac placenta, which is responsible for the histiotrophic nutrition of the embryo during the first few days of embryonic development. Interference with the function of this placenta, due to the accumulation of a chemical, can cause embryonic death or embryonic toxicity/malformations not occurring in species that lack this placental structure. This can result in the false classification of a chemical as a developmental toxicant ... This reasoning is, to some extent, also valid for the rabbit. Conversely, there may be examples where the yolk sac placenta may protect the embryo by*

[1] http://iccvam.niehs.nih.gov/methods/development/dev.htm

hindering the access of a chemical to it. ... The toxicant concentration reaching the embryo is a critical factor in developmental toxicity. Among the mechanisms regulating the disposition of toxicants from the maternal circulation to the embryo, drug efflux transporters play a key role, and are possibly responsible for interspecies variability." This argues very much for the use of human placentas, which are relatively easily available.

A very promising approach is to use thresholds of toxicological concern (Kroes et al., 2005), i.e., TTC, to define exposure limits below which an effect is sufficiently improbable, as reproductive toxicity is considered a threshold effect (Piersma et al., 2011). Van Ravenzwaay and coworkers (2011) determined such a TTC for reproductive toxicity at 8 µg/kg bw/d. These approaches can be further refined either by distinguishing classes of chemicals or using internal TTC, i.e., basing the threshold on plasma concentrations actually achieved. An interesting option would be to use experimental barrier model data to modify the TTC level.

6.2.4 Optimization of existing tests

Reproductive Toxicity is by its very nature characterized by "complexity layered on complexity," and the devil might be found in the details. Reproducibility, robustness, and reliability combined with a relevant, sound scientific base will be critical for an acceptable test going forward.

The three embryotoxicity tests validated in 2002 have received considerable interest for further optimization. In order to review and discuss the next steps of using the tests, an ECVAM workshop was held in January 2003 (Spielmann et al., 2006). A panel of 12 European and American experts from industry, academia, and governmental institutions analyzed the tests for chemical and pharmaceutical safety testing *in vitro*. The outcome of the workshop can be summarized as follows (Spielmann et al., 2006):

1. The tests are reliable and transferable to other laboratories.
2. The prediction models need to be revised in order to receive a better discrimination between non- and weak/moderate embryotoxic chemicals.
3. The tests should also be applied to industrial chemicals to demonstrate the reliability and relevance of the system, since within the formal validation study primarily pharmaceuticals have been tested.
4. The selected strong developmental toxicants represent a limited number of mechanisms of toxicity, mostly affecting cell proliferation. Strong embryotoxic chemicals with other toxicological mechanisms should be tested in order to enhance the reliability for a wider applicability of the tests for a broader range of chemicals.
5. A metabolic system to detect proteratogenic compounds has to be integrated in order to extend the applicability.
6. Other differentiation pathways have to be included in the tests. Additional major target tissues such as the nervous system and the skeletal system have to be included in order to get precise information about the teratogenic potential of chemicals.

A lot of work has been done to further optimize the standard murine EST, which was shown to distinguish the different embryotoxic potentials of even very closely related structures (de Jong et al., 2011). Optimizations of protocols were reported (De Smedt et al., 2008; Seiler et al., 2004; Seiler and Spielmann, 2011). In order to move towards transcriptomics read-outs, PCR was employed to monitor specific gene expression (Pellizzer et al., 2004). Knudsen and colleagues (Knudsen et al., 2011) demonstrated the mEST's ability to capture data on disruption of developmental signaling pathways as a potential alternative for assessing developmental toxicity. "*His example focused on the expression of genes for the 17 + 2 conserved signaling pathways critical to early development (National Research Council, 2000), taking the hypothesis that an abnormal activation or inhibition of signaling pathways can lead to developmental toxicity. The test system uses murine ESCs cultured 3 days as hanging drops that form 'embryoid bodies' with gene expression patterns for ectodermal, mesodermal, and endodermal lineages. Analysis of gene expression at 5 days revealed the top expressed signaling pathways as Cadherin, Wnt/β-catenin, Hedgehog, Integrin, ND, Nuclear Hormone, and Receptor Ser/Thr kinase*."

The EC report (Adler et al., 2011) (cited literature there) gives a comprehensive overview on variants of the embryonic stem cell tests with respect "*... to their readouts but also in the target cell differentiation (Peters et al., 2008; Zur Nieden et al., 2004). Depending on the area of application, effects on differentiating neural cells (Stummann et al., 2009b; Theunissen et al., 2010), cardiomyocytes (Buesen et al., 2009) and skeletal cells (Stummann et al., 2009b; Zur Nieden et al., 2004; Zur Nieden et al., 2010) have been investigated. Effects on the quantity of differentiated target cells have been assessed by using immunological methods such as flow cytometry (Buesen et al., 2009) or molecular biological methods such as RT-PCRs and omics (Chapin et al., 2007; Osman et al., 2010; van Dartel et al., 2009; van Dartel et al., 2010; West et al., 2010; Winkler et al., 2009; Zur Nieden et al., 2001; Zur Nieden et al., 2004). Several of the methodologies could also be automated in order to increase the throughput of substances and make the test available for screening purposes (Peters et al., 2008)*." A key development certainly is to translate the EST to human stem cells (Pal et al., 2011; Pellizzer et al., 2005; West et al., 2010). This promises, finally, to overcome species differences for the key health concern of reproductive toxicity.

A key limitation of many *in vitro* tests is the lack of metabolizing capacity (Coecke et al., 2006). Efforts to combine the EST with metabolizing systems have been described (Bremer et al., 2002; Hettwer et al., 2010) with kinetics modeling. Another improvement represented the combination with kinetic modeling (Verwei et al., 2006). Similar optimization work was also carried out for the whole embryo culture (Piersma et al., 2008) adding metabolizing systems (Luijten et al., 2008). The added value and validity of these variants should be assessed systematically. Notably, the modular approach (Hartung et al., 2004) would allow assessing only the aspects that have been changed, and, by establishing performance standards for the murine EST, validity could possibly be established with reasonable effort.

Other promising approaches include the use of, or combination with, computation models of development pathways and systems and, finally, high-throughput *in vitro* approaches as, for example, those being utilized by the EPA ToxCast program (Sipes et al., 2011a) (see below).

6.2.5 In silico approaches

The development of reliable QSARs for reproductive toxicity is currently suffering due to a lack of high quality *in vivo* data and the complexity of the reproductive toxicity endpoint, which involves several known and unknown toxicological mechanisms. It should be stressed that QSARs can be based on either *in vivo* or on *in vitro* data. The uncertainty of the origin of data should be taken into account when integrating these models into testing strategies.

Some commercially available toxicity prediction software packages are claiming to detect reproductive toxicants. Maslankiewicz et al. (Bremer et al., 2007a) have reported that the software program DEREKfW has been challenged with around 100 reproductive toxicants included in Annex I of Directive 67/548/EEC, and 90% of chemicals classified for "impaired fertility" and 81% of chemicals that cause harm to the unborn child were not detected. The TSCA chemical category list of the new chemical program of US-EPA failed in 77% to detect EU-classified chemicals causing adverse effects to mammalian fertility and 82% of developmental toxicants have not been correctly identified. This is in strong contrast to mere internal validations that show results of >80% correlation for reproductive toxicity (Matthews et al., 2007), illustrating the importance of objective assessments.

A working group of ILSI/HESI assessed structure/activity relationships (SAR) (Julien et al., 2004) and summarized: "*The Working Group's investigation of two statistically based SAR systems that have been applied to developmental toxicity elucidated the difficulties in predictive modeling of this toxicity. With a statistically based approach, the activity (or inactivity) of each training set compound must be captured in a way that can be correlated with the presence or absence of chemical structural features. This poses a number of methodological challenges. The particular 'activity' representing developmental toxicity must be defined. Also, an objective, rational, reproducible, and transparent process for scoring a training set compound for the activity must be developed. Additional methodological challenges derive from the dynamic nature of development and the general sparseness of published developmental toxicity data.*

To advance the potential of SAR for predictive modeling of developmental toxicity, it will be necessary to develop general scientific agreement on valid and transparent methodology for selecting, categorizing, and scoring developmental toxicity data. Such methodology should be developed by an interdisciplinary panel of developmental toxicologists and developmental biologists, working in consultation with SAR model developers and individuals with other relevant expertise (e.g., biostatisticians). The recommendations from this panel should undergo peer review.

The Working Group recommended three research efforts that will inform the development of improved methodology: 1) A systematic and holistic analysis of developmental toxicity data of adequate quality and quantity should be conducted. Toward this aim, a comprehensive, publicly available electronic database of developmental toxicity data is needed. Such a compilation would allow investigators to methodically examine a range of considerations when selecting and utilizing toxicological data in training sets (i.e., various experimental factors, various approaches to combining/separating categories of endpoints, and alternative scoring systems); 2) Training sets for discrete developmental endpoints should be developed. This would allow examination of the process used to assemble training sets, as well as the effect of alternative processes on the predictive performance of the model. The Working Group considers these first two recommended efforts, compiling/analyzing a comprehensive database and developing/investigating alternative training sets, as complementary and iterative exercises; 3) The combined use of multiple types of tools and approaches for screening should be investigated.

In conclusion, the Working Group recognizes there is a need for valid and efficient methods to screen large numbers of environmental contaminants for their potential to pose a developmental hazard. Whereas the use of SAR models for exploratory studies is encouraged, statistically based SAR models, in their current form, are not yet sufficiently developed or validated to yield confident predictions with which to identify potential developmental toxicants in a screening program. The Working Group believes that the efforts recommended in this report will contribute to improving the potential of statistically based models for this application."

Similarly, Cronin et al. (2002) summarize: "*There are a number of problems with applying QSARs to reproductive toxicology notably the complexity, subtlety, and sometimes ill-defined nature of the endpoints and lack of data available for modeling.*" Hewitt et al. (2010) conclude similarly: "*This study demonstrates the limited success of current modeling methods when used in isolation. However, the study also indicates that when used in combination, in a weight-of-evidence approach, better use may be made of the limited toxicity data available and predictivity improved.*" Recommendations (condensed here) are provided as to how this area could be further developed in the future:

- Availability of suitable toxicity data; almost exclusively data collected for pharmaceutical compounds, which may prevent the study of predictions made for industrial chemicals.
- Placental transfer can be useful as a modulating factor.
- Existing "global" (Q)SAR models for reproductive and developmental toxicity must be treated with caution. Given the plethora of different mechanisms (many of which are unknown) involved within reproductive and developmental toxicity, a single "catch all" (Q)SAR model is likely to show limited performance. If literature data are available, a number of structurally/mechanistically restricted "local" (Q) SARs would be more appropriate.
- At present, category formation approaches are promising but they are limited, both by available data from which to select category members and by the approaches available to define categories.
- Currently, the structural alert approach, as used in DEREKfW, requires more alerts to be developed for reproductive and developmental toxicity.
- The importance of time-dependent effects should also be considered.

- A weight-of-evidence prediction is dependent upon whether a valid chemical category can be formed for read-across, (Q) SAR models, and chemical profilers for specific reproductive toxicity effects.
- A need for collaboration between scientists with experience in computational modeling and those with experience in interpretation of developmental toxicity data has been highlighted.
- There is value in considering more than one *in silico* approach within a weight-of-evidence framework.

There is clearly a need for access to existing animal and human data to improve the situation. New technologies and bioinformatic methods can only be utilized if there is increased sharing of data. A number of research efforts already allow global access to information such as ACTOR, ToxRefDB, the ILSI-HESI toxicogenomics project, etc. This concept of data sharing has also been incorporated into REACH, which requires that toxicological data be made publicly available, but the summarizing data typically do not qualify for modeling approaches. The need to make data available extends to publicly recorded human clinical trials and pregnancy registries.

Complementary to this issue of globally available data is the need for consistent and universally accepted terminology for characterizing effects. Historically, the developmental toxicology community has embraced this concept, with international collaborative projects and publications on terminology used in the evaluation of fetal specimens (e.g., Makris et al., 2009; Wise et al., 1997). This same attention to consistency and precision in terminology must also be applied to new technologies for developmental toxicity testing.

Altogether, it is unlikely that *in silico* approaches as stand-alone methodologies will make a major contribution to reproductive toxicology in the near future. This is in line with some growing skepticism on (Q)SAR as stand-alone methods in regulatory safety assessments in general (Doweyko, 2004; Hartung, 2009b; Hawkins, 2004; Raunio, 2011).

In contrast to the above mentioned drawbacks of (Q)SARs, computational toxicology based on High-Throughput Screening (HTS) data, and cell agent-based models (ABMs) have been able to simulate prototype toxicity pathways that affect growth, morphogenesis, and development. *In vitro* profiling manages to screen for targets, pathways, and processes to build predictive signatures for discrete adverse outcomes from animal data or human epidemiology where available. Functional assays must extend these signatures to mechanistic relationships and pathway-based inferences for an integrated testing strategy. As we increase biological knowledge, it will be necessary to build and utilize biologically informed models that can simulate downstream consequences of perturbation. In this regard, computational systems biology is needed to reconstruct higher-order biological effects from the more fundamental *in vitro* data. These predictive models demonstrated the feasibility of predicting ToxRefDB animal toxicity solely from the results of HTS data. In the future, it will be necessary to perform forward validation of these models without dependence on animal data (for compounds lacking such data). This highlights the need for "virtual models" in which a toolbox of dynamic models can be used to interpret HTS data and pathway-based information. Latest studies from ToxCast have demonstrated the feasibility of predictive modeling of fertility, blood vessel development, and prenatal developmental toxicity (Sipes et al., 2011a). Angiogenesis can be considered an example, as cell-agent based models (ABMs) for angiogenesis have been developed that recapitulate HTS data at a histological scale (Kleinstreuer et al., 2011). In this regard, (Q)SARs may become more informed as we train these read-across methods with information from HTS data and cellular ABMs. Another approach emerging from the above mentioned data is called: "Towards a virtual embryo." The final goal is to apply HTS data, *in silico* tools, and models to look globally at developmental processes and toxicities in a new way. Predictive and mechanistic models would dynamically integrate data with relevant information about embryonic systems. Applying "Virtuomics" and running "what-if" scenarios to predict adverse outcomes from different perturbations might allow scientifically-based predictions of how development might be affected across a range of complex factors. A toolbox of virtual tissue models may someday comprise a modular virtual embryo for simulating important information as part of an integrated testing strategy[2].

6.2.6 Information-rich single tests

Complex phenomena such as elements of the reproductive cycle and their perturbations usually can be captured better by multiple endpoints than by a single biomarker. Functional endpoints such as formed beating heart cells in the EST already integrate many biological pathways, but new technologies allow assessing a multitude of measurements using high-content technologies. Both omics and image analysis can add new qualities to interpretation of the biological models (Hartung and Leist, 2008). However, it is important to keep in mind that whatever fancy analysis we add, it can hardly overcome the limitations of the underlying model (Hartung, 2010b, 2011). So the same considerations of the limits of both animal and cell models apply.

We also need to keep in mind that the novel technologies pose an enormous challenge to the validation process as exemplified for toxicogenomics approaches (Corvi et al., 2006). The number of parameters to control and document, the sometimes high cost per single measurement limiting replicates and numbers of substances tested, or the complex prediction models for information-rich methods, as well as the rapid turnover of technological change are only a few examples of challenges faced.

In vitro work so far has combined mainly whole embryo culture and transcriptomics (Luijten et al., 2010) or the EST with metabolomics (Kleinstreuer et al., 2011; West et al., 2010), proteomics (Groebe et al., 2010; Klemm et al., 2008; Klemm and Schrattenholz, 2004; Seiler and Spielmann, 2011) or transcriptomics, as summarized recently (van Dartel and Piersma, 2011). These approaches use patterns or biomarkers derived from a training set of substances to identify substances with similar mode of action. Their predictive value looks promising but awaits formal validation.

[2] http://www.epa.gov/ncct/v-Embryo/

6.2.7 Integrated testing strategies (ITS)

ITS are a consequence of REACH (van Leeuwen et al., 2007), which argues for the use of *all* available information and views use of the definitive animal experiment only as a last resort. However, the ITS suggested for reproductive toxicity (Fig. 6.4) is relatively simple, not really accommodating any alternative methods.

The idea of a comprehensive ITS would be to provide for as many substances as possible enough information to avoid the ultimate animal test. Ideally, all aspects of the human reproductive cycle would be mapped and translated into test components. At the same time, there are some dominant findings, which lead to classification as a reproductive toxicant (see Fig. 6.5).

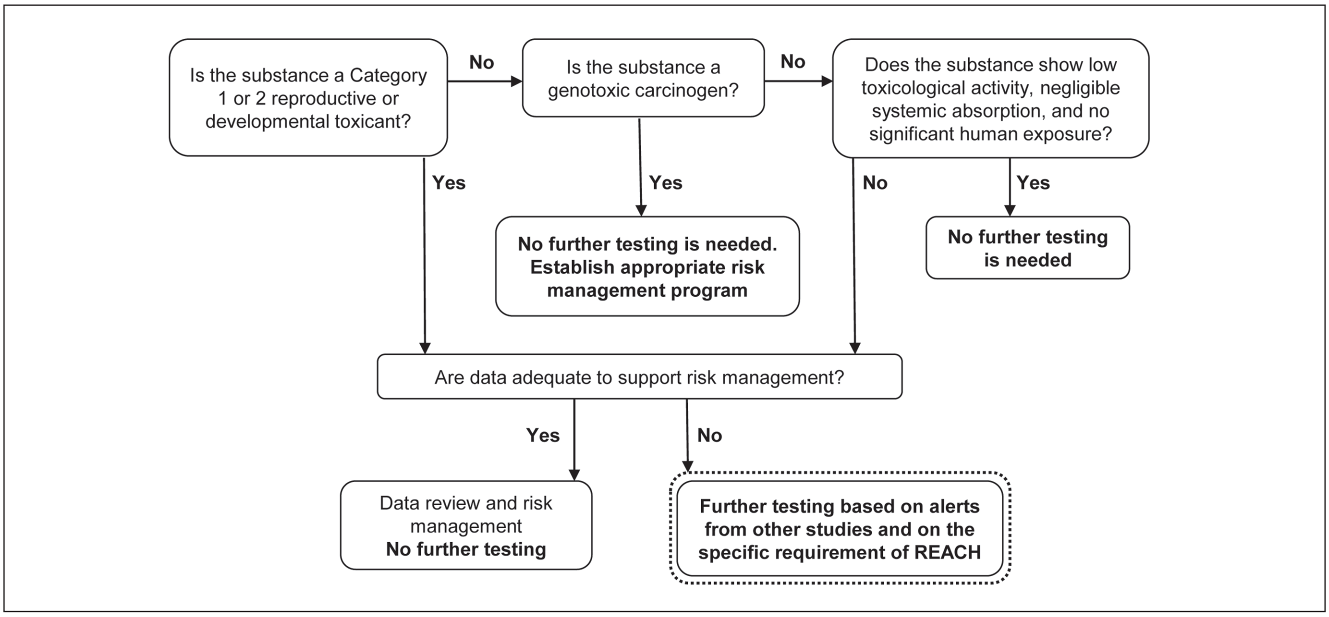
Is the substance a Category 1 or 2 reproductive or developmental toxicant?
No
Is the substance a genotoxic carcinogen?
No
Does the substance show low toxicological activity, negligible systemic absorption, and no significant human exposure?
Yes
Yes
No
Yes
No further testing is needed. Establish appropriate risk management program
No further testing is needed
Are data adequate to support risk management?
Yes
No
Data review and risk management No further testing
Further testing based on alerts from other studies and on the specific requirement of REACH

Fig. 6.4: Flow chart for considering whether a substance requires additional testing under REACH for reproductive or developmental toxicity

(modified from Scialli, 2008)

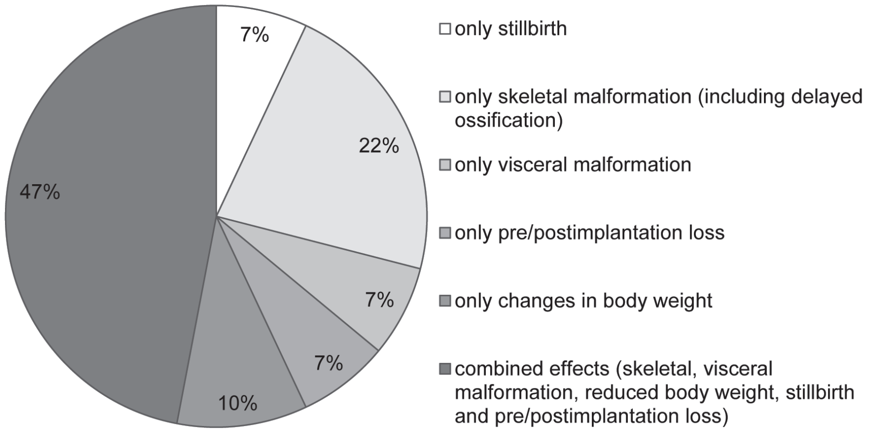

Fig. 6.5: Breakdown of embryotoxic effects of 74 industrial chemicals, which have been tested according to EU Directive 67/548/EEC B31

(reprinted with permission from Bremer and Hartung, 2004)

The figure presents a breakdown of embryotoxic effects of 74 industrial chemicals, which have been tested according to EU Directive 67/548/EEC B31. Even if 34 chemicals have demonstrated effects on the offspring only 2 chemicals have been classified as developmental toxic according to the standards applied by the national competent authorities. However, by analyzing all the developmental toxic effects the data demonstrate mainly that combined embryotoxic effects have been detected, but some chemicals also induce specific effects, such as delayed ossification and other skeletal effects. It is important that the experimental design of *in vitro* tests will be set up in a way that these effects can be detected.

We have earlier suggested (Bremer and Hartung, 2004) using this for a prevalence-based testing strategy, creating tests for an ITS specifically addressing these aspects. This reduces mapping the human reproductive cycle to those elements which really lead to classifications. We called this a “prevalence-driven approach” (Fig. 6.6).

This concept was further refined by Bremer et al. (2007b), who studied in more detail available information on endpoints leading to classifications in various databases. Despite a number of exhaustive database and literature searches, data satisfying the inclusion criteria for this analysis could not be located in the public domain for more than half (53%) of the substances classified by regulators as being toxic to reproduction. The analysis was limited to data on 71 classified reproductive toxicants. Statistically and biologically significant positive effects have been reported as absolute frequency (i.e., the total number of times a positive effect was detected in a particular sub-endpoint, irrespective of the dose at which the effect was seen). The most

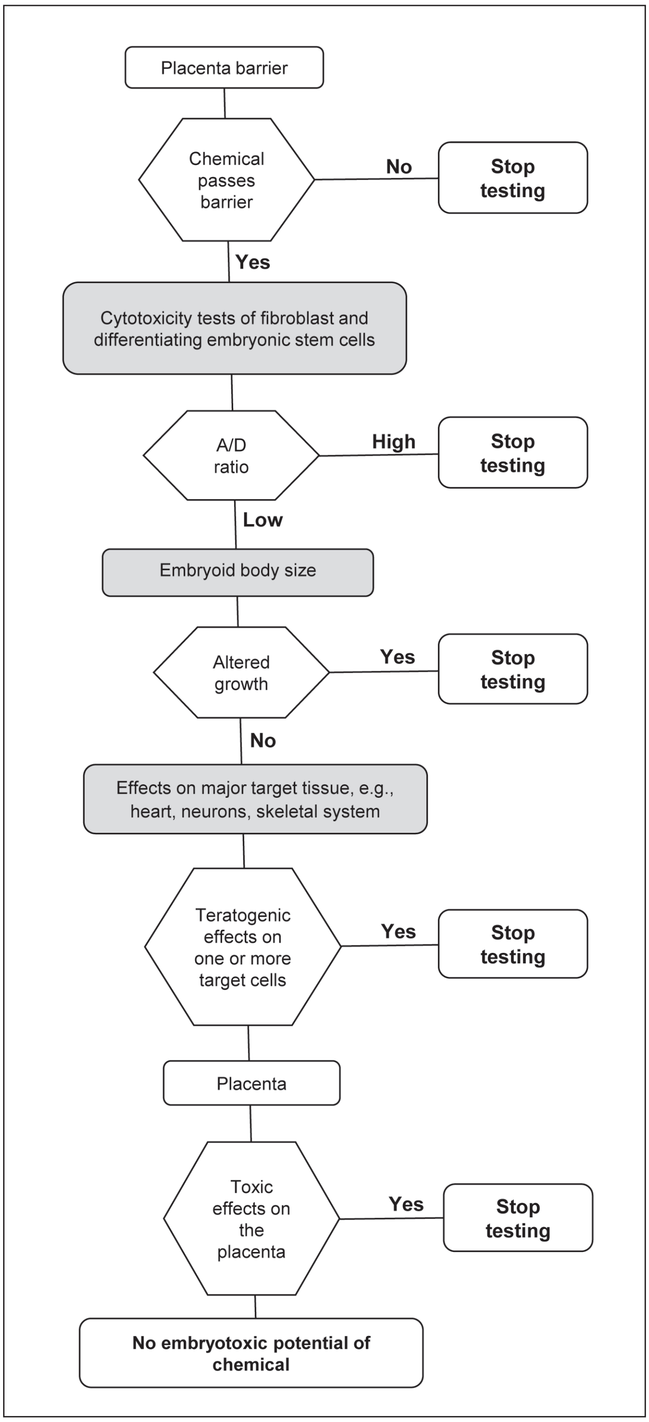
Placenta barrier
Chemical passes barrier
No
Stop testing
Yes
Cytotoxicity tests of fibroblast and differentiating embryonic stem cells
A/D ratio
High
Stop testing
Low
Embryoid body size
Altered growth
Yes
Stop testing
No
Effects on major target tissue, e.g., heart, neurons, skeletal system
Teratogenic effects on one or more target cells
Yes
Stop testing
Placenta
Toxic effects on the placenta
Yes
Stop testing
No embryotoxic potential of chemical

Fig. 6.6: Proposal for a test strategy in order to detect the embryotoxic hazard of chemicals
(modified from Bremer and Hartung, 2004)
The flow chart demonstrates a proposal for a testing strategy to detect the embryotoxic hazard of chemicals. Three tests based on embryonic stem cells and their differentiated counterparts have been combined. The reliability of the test strategy has to be tested by using selected chemicals with various toxicological pathways. It has to be proven that all toxicological mechanisms will be detected or if additional systems such as tests for receptor-mediated embryotoxicity must be included. It should be pointed out that such a test strategy should be part of a general testing scheme for toxicological profiling of chemicals. Chemicals with a known cytotoxic effect probably will not enter into this testing scheme. Chemicals that are known to be metabolized will be tested in combination with a biotransformation system. *In vitro* tests, in grey, have been developed, but further test optimization and validation is required.

frequent ones were 39 cases of body weight changes as a more general toxicity parameter, 30 cases of testicular weight/histopathology, 28 offspring body weight at birth, and 25 each for sperm morphology, sperm count, pregnancy rate, and live offspring. Interestingly, uterine weight/histopathology was on the lower end with only five cases. Most of the reported effects are not isolated, but also appear in combinations. It is, therefore, highly relevant for a further analysis, in particular for sub-endpoints occurring with a lower prevalence, to determine if they are associated with a more frequently occurring effect. Such analysis could allow focusing test development on the most relevant modes of action. These would further diminish the relevance to test for a sub-endpoint with a lower prevalence such as, e.g., parturition. Even if parturition is a sub-endpoint with a low prevalence of a health effect, which has per se a low prevalence in the universe of industrial chemicals, the competent authorities currently request testing for such an endpoint.

For developmental toxicity the search strategy described above identified reliable data for 202 of the classified substances. Given the extensive range of histopathological, functional, clinical, and other evaluations undertaken in the context of a developmental toxicity study, standardization is important not only in relation to the selection of study endpoints but also in the terminology used to communicate study results. For the purposes of this analysis, studies were analyzed and catalogued in a manner consistent with the recommendations of Chahoud and colleagues (Chahoud et al., 1999) using sub-endpoint definitions proposed by MacKenzie and Hoar (Derelanko and Hollinger, 2001). The frequency with which standardized sub-endpoints from guideline prenatal developmental toxicity and developmental toxicity studies were reported positive for the 202 substances in this database ranged from 78 for postimplantation and dead, 77 skeletal, 60 body weight, 55 external limbs and digits, etc. Offspring sex ratio (4) and parturition (2) were the least frequent. These preliminary analyses illustrate how we might be guided in developing an ITS of components most relevant for regulatory decision making.

The limited availability of full study records in the public domain impedes this approach, but the more recent data made available, for example via the ToxRefDB, might help here: Knudsen et al. (2009) characterized 283 chemicals (mainly pesticides) tested in both rats and rabbits; 53 chemicals (18.7%) had lowest effect levels on development that were either specific (no maternal toxicity) or more sensitive than the maternal animal in either species: "*The primary expressions of developmental toxicity in pregnant rats were fetal weight reduction, skeletal variations and abnormalities, and fetal urogenital defects. General pregnancy/fetal losses were over-represented in the rabbit, as were structural malformations to the visceral body wall and CNS. Based upon administered doses, there was a clear hierarchy to the sensitivity and specificity of [developmental lowest effect levels] dLELs in comparing species, with rat development being more sensitive with regards to the number of endpoints affected and the number of active chemicals. Many of these relationships are consistent with previous database studies of developmental toxicology, indicating that they are driven by the biology of the test species*."

A more detailed analysis of the same database was presented by Martin et al. (2009b): "*19 highly prevalent effects identified treatment-related changes to reproductive performance including fertility, mating, gestational interval, implantations, litter size, and live birth index, demonstrating effects at different stages of the reproductive cycle. ... The fairly restricted set of 19 effects characterized 151 of the 152 chemicals that demonstrated any reproductive toxicity. Additionally, these 19 effects identified 229 of the 269 chemicals that caused any offspring toxicity. The remaining 40 chemicals not identified were predominantly affecting pup weight only. This supports the hypothesis that we can extract a small finite set of key reproductive effects from this dataset for use in developing robust predictive signatures*." This strongly supports the idea that a rather limited set of critical endpoints might be mapped by either mode of action or PoT-based tests. It is hoped that with the expansion of the ToxCast program further chemical classes will be entered. The ontology of effects developed here represents on its own a very valuable tool for the field. Similarly, proprietary data could be analyzed, even in a blinded manner, to establish more robust frequencies, and companies should be encouraged to share these. The analyses of metabolites and biological pathways are likely to identify additional nodes that may be important to develop specific tests for predictive reproductive toxicity and the PoT approach is therefore important for ITS.

Ideally, not only hazard information is used for an ITS. ITS do not necessarily use only new (*in vitro*) test data but can incorporate *in silico* estimations and modeling. A promising integration of different information sources is the combination of *in vitro* studies with kinetic modeling (Andersen et al., 2005), which has been suggested as *Biologically Based Dose-Response Modeling* for developmental toxicity (Lau et al., 2000). Note also that existing data can be used, ranging from cell-based tests to animal and human data. To the extent that existing information shall be integrated into the ITS, it will be necessary to assess its quality. A tool was developed (Schneider et al., 2009) to objectively assign the so-called Klimisch scores to either *in vivo* or *in vitro* studies as a first step.

Principles for systematic ITS composition (Jaworska and Hoffmann, 2010) and validation (Kinsner-Ovaskainen et al., 2009) are only emerging. A certain consensus exists that the reproducibility of each ITS component needs to be assessed. However, it is not clear how the predictive value can be assessed without an enormous number of substances tested and in the absence of an animal model as point of reference for the components. An evaluation stressing more the scientific validity of the components appears to be a pragmatic solution. Earlier, we termed this a mechanistic validation (Coecke et al., 2007; Hartung, 2007b, 2010b), as it confirms that the model reflects a scientifically established relevant mechanism, differentiating it from an empiric reproduction of a reference test. The careful selection of reference compounds (Hoffmann et al., 2008) will become even more important.

6.2.8 Pathways of Toxicity (PoT) and systems toxicology

The area of reproductive toxicity testing appears to be very well suited for PoT-based approaches as currently pioneered by the

US-EPA: The ToxCast project has mapped a multitude of pathway assays to animal reference data: Sipes et al. (2011a) have delivered a very impressive proof of principle of the PoT concept across species. This has been started for zebrafish (Sipes et al., 2011b), demonstrating the common basis of PoT with mammals. The two most promising alternatives for hazard-based identification of developmental activity in the ToxCast battery are non-animal embryonic stem cells and zebrafish embryos; although the latter should only be seen as an interim approach until full replacement tests are available. These models, in tandem with >600 ToxCast assays, provide a unique resource for this prioritization. The performance of these test systems needs to be looked at closely within the context of ToxRefDB animal bioassay data. Early results comparing zebrafish with pregnant rat and rabbit have shown similar concordance (e.g., ~56-60%) between rat-zebrafish, rabbit-zebrafish, and rat-rabbit. As such, the need to develop *in vitro* extrapolation from concentration response to *in vivo* dosimetry, cross-species differences, and life-stage assessments is required. Although most HTS assays were based on human cells, they could distinguish PoT that are active in either rats or rabbits, explaining species differences. This shows how the change in resolution allows annotating PoT to different species and measuring them with PoT specific assays with high throughput.

Similar efforts should be extended for the (favorably human) EST (Kleinstreuer et al., 2011; West et al., 2010). The obvious potential of combing stem cell methods with Tox-21c approaches was stressed earlier (Chapin and Stedman, 2009). This extends from embryotoxicity to other areas such as male fertility (Krtolica and Giritharan, 2010) and to link toxicity related biomarkers uncovered using hES cells with the PoT concept. However, the vision of Tox-21c is not that a tremendous number of assays in a centralized facility are used for each and every substance. Once critical PoT are identified, they can be translated into rather simple assays. For developmental toxicity, for example, a total of 17 intracellular pathways have been identified as involved in organogenesis (for review see: Anon., 2000), cyto-differentiation, growth and tissue renewal, of which 5 appear to be the most relevant for early development (Fig. 6.7).

As part of a collaborative project linked to the ReProTect project, Michael Schwarz and his group in Tübingen, Germany, have established a system with ECVAM in which mouse embryonic stem cells were stably transfected with luciferase reporters specific for the Wnt/beta-Catenin and the TGF-β signaling pathways (the so-called ReproGlo assay, (Uibel et al., 2010)). The effects of several known human teratogens and non-teratogens, including thalidomide, have been investigated

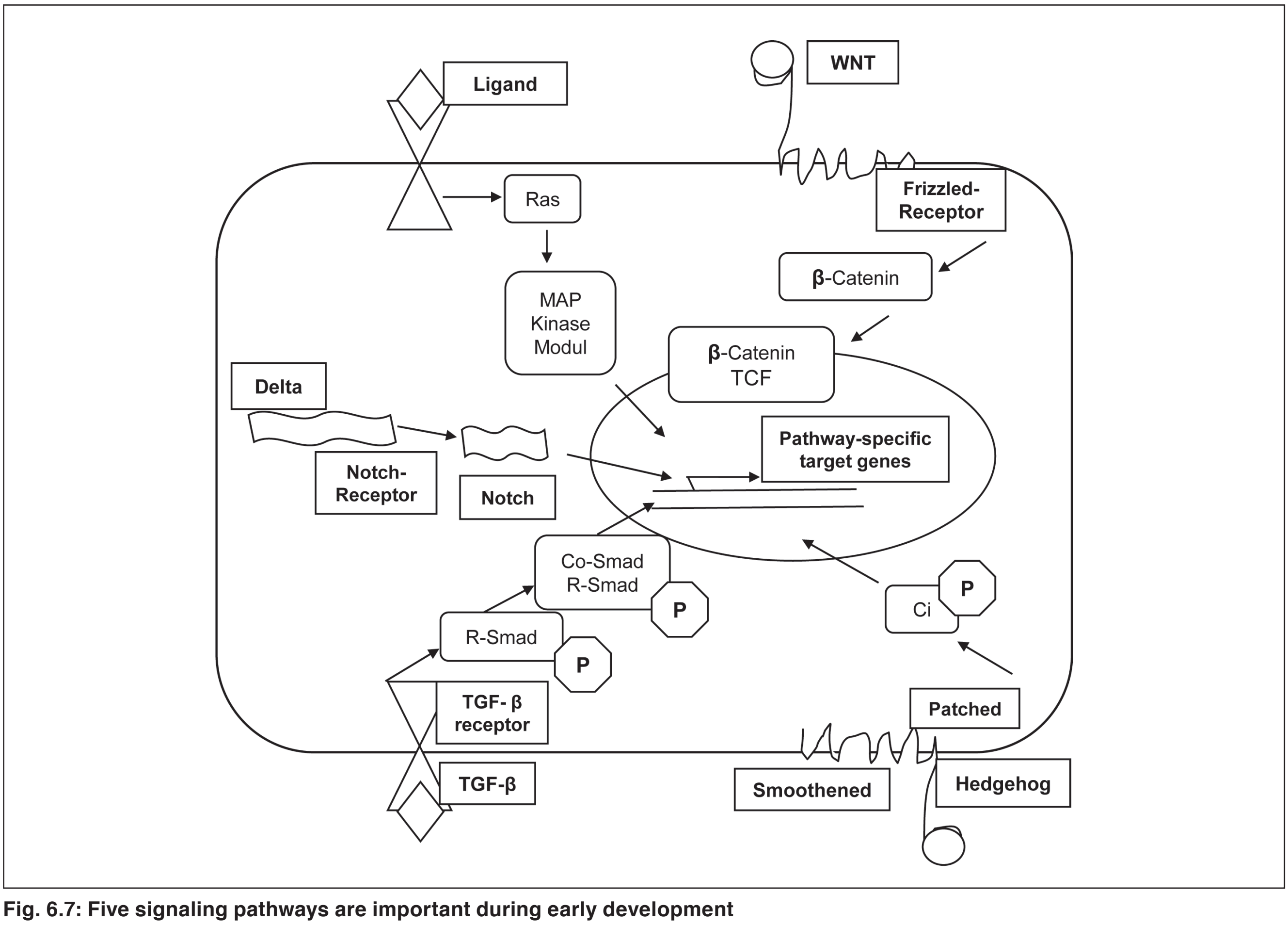

Fig. 6.7: Five signaling pathways are important during early development
(Redrawn with permission from Anon, 2000).

in this system. The undifferentiated cells are incubated for only 24 hours; the system is based on a multi-well format and thus is well suited for high-throughput analysis. It also allows the determination of non-specific toxicity (Alamar Blue assay) and the specific response (luciferase-reporter readout) on one and the same plate. The test correctly identified human reproductive toxicants such as lithium chloride, retinoic acid, the potency of different valproic acid derivatives and (with a metabolizing system) cyclophosphamide. This nicely illustrates that PoT-based assays, if representing nodes in the perturbed physiological networks, most likely can cover substantial parts of the universe of toxicants.

6.2.9 General considerations

Machine learning and 'omics'

Considering the fact that we do not know everything, and our current knowledge base is growing quickly, the identification of pathways and nodes of biology that seem to be state-of-the-art today might be outdated tomorrow. If we develop specific assays focused only on what we know, we limit our ability to uncover new mechanisms based on new compounds, mixtures, or metabolites. An example of this, even within a platform, can be seen in metabolomics. When one uses only a targeted approach, the ability to learn about new biomarkers is limited. In contrast, an untargeted approach opens the number of possible endpoints to all of the measureable metabolites. Furthermore, tests that rely on vast amounts of biomarker data, such as those obtained from metabolomic fingerprints, allow a continual "machine learning" approach of the predictive models. The utility of omics-based approaches seems to be an especially efficient manner by which to examine many biomarkers simultaneously to create knowledge bases that will allow a machine-learning approach to insure inclusion of important information.

Compound-related issues

The number of compounds with clearly known human reproductive or developmental adverse effect is relatively small; therefore it becomes difficult to compile a relevant set of compounds on which to build predictive model *in vitro* systems. Furthermore, assuming the efforts to predict human reproductive or developmental effects are successful, compounds that cause human reproductive toxicity will not go forward, and the set of compounds to be used as reference molecules will become self-limited. Predictive training sets need to be standardized, and to really understand the utility of training sets it is recommended that similar structures are used, especially when they segregate differently (toxic vs nontoxic). Otherwise, efforts should be made to assure that structures are as diverse as possible to facilitate maximizing the "chemical and biological space" of the predictive models. Lastly, it would be interesting to select some compounds to be predicted from the PoT key metabolites and pathway regulators. In conclusion, it is essential to have a set of compounds compiled and recommended for use as a gold standard training set.

Dosing: Steady state versus C_{max}

It is not clear what would be the best approach to determining concentrations of compounds to be used in *in vitro* studies. Current expert opinion is that cytotoxicity information is not valuable, and that triggering of pathways in reproductive and developmental toxicity should be evaluated. Frequently dose-response analyses for toxicity are performed and subtoxic or low doses relative to cytotoxicity are chosen. Given that the *in vitro* models are very different from *in vivo*, comparisons between these two systems are useful to assure relevance. A related concern is whether steady state concentrations *in vivo* are the reference or whether C_{max} values might be more appropriate. Many metabolic specialists claim that high doses of compounds trigger later cellular events, which are not likely to be seen at longer-term steady state levels. These facts should be taken into consideration when developing or refining *in vitro* tests.

Short versus long-term effects

Evidence should be built into *in vitro* predictive models that allow for an understanding of the early identification of events that may take longer timeframes to be observed *in vivo*. If exposure to a compound *in vivo* takes weeks or months to produce an event, do the *in vitro* tests performed for shorter periods of time have the potential to identify these compounds?

Biological systems

The main interest lies in developing biological systems that model *in vivo* systems. Obviously, the key issue is that the systems must be relevant (and most likely based on human rather than animal materials). The more complicated systems typically move toward systems with cellular interactions and ultimately toward 3D cultures. The caution is that the compound doses delivered to the cells in these experiments need to be carefully considered. The more complex systems typically evolve to polarized cells, and it is becoming very clear that cells interact very differently with compounds delivered apically versus basolaterally (Benet et al., 2003). Therefore, the addition of compounds to such models needs to be evaluated with relevance to the *in vivo* situation. Both the advantages and disadvantages of 3D and complex models need to be considered.

In vivo factors

Many upstream risk factors are associated with human developmental defects as an interaction of multiple factors relating to genetics, environment, and socioeconomic status. The latter includes factors such as prenatal healthcare, maternal nutrition, anxiety, general health, and drug use/abuse. These may be difficult to unravel *in vivo* (adverse outcome pathways) and to quantify *in vitro* (toxicity pathways). As such, alternative methods need to address key molecular pathways and cellular processes that propagate information across multiple scales of biological organization in the developing embryo. Particularly important, but as yet under-represented in alternative models, is a systematic approach to characterize and analyze multicellular networks within the context of normal biological architecture. Assays that address 3D configuration and extracellular matrix biology are needed.

6.3 Conclusions and recommendations: reproductive toxicity

It is probably too simplistic just to break developmental and reproductive toxicity down into a series of hazards. Issues that should be considered or addressed in developmental toxicity testing were recently listed by Makris et al. (2011):

– Translational medicine, cross-species extrapolation
– Mode of action data
– Cumulative exposure issues
– Critical windows of exposure and effect
– Latency of response
– Structural vs. functional outcomes

This chapter very much agrees with the emphasis on mode of action, or even with finer resolution to PoT. This corresponds very well with an emphasis on functional instead of structural outcomes. It is hoped that the annotation of PoT to species will help the cross-species extrapolation. It is also hoped that the early events (points of chemical interaction) will also be predictive for the more latent manifestations. Exposure considerations have not been addressed here, with the exception of the TTC concept.

There are many considerations involved in non-animal testing for stages of the reproductive cycle, and an integrated strategy combining *in vitro* methods with high-throughput screening (HTS), predictive computational models, and computer simulation provides the foreseeable path forward.

Recommendations: reproductive toxicity

The following key recommendations are made:

1. The limitations of the pertinent animal test protocols for reproductive toxicity testing should be systematically reviewed in the spirit of evidence-based toxicology.
2. The TTC approaches can be further refined by distinguishing classes of chemicals or using internal TTC, i.e., basing the threshold on plasma concentrations actually achieved. An interesting option would be to use experimental barrier model data to modify the TTC level.
3. The zebrafish embryo teratogenicity assay should be evaluated for defining a protocol that will allow formal validation, although the test should be seen as an interim approach until a full animal-free replacement is available.
4. A human stem cell-based test employing either human embryonic stem cells or induced pluripotent stem cells should be validated. An evaluation of stem cell variants and prediction models should be carried out, especially since the assay is considered for possibly replacing the second species in reproductive toxicity testing by FDA. Similarly, the variants of whole embryo culture should be followed up almost a decade after validation of the original protocol.
5. The advantage of using human rather than animal derived biological test systems should be taken into account for every optimization or new development of a test system that is designed for human risk assessment.
6. For *in silico* approaches, the ILSI/HESI Working Group recommendations are reiterated: "*1) A systematic and holistic analysis of developmental toxicity data of adequate quality and quantity should be conducted. Toward this aim, a comprehensive, publicly available electronic database of developmental toxicity data is needed. Such a compilation would allow investigators to methodically examine a range of considerations when selecting and utilizing toxicological data in training sets (i.e., various experimental factors, various approaches to combining/separating categories of endpoints, and alternative scoring systems); 2) Training sets for discrete developmental endpoints should be developed. This would allow examination of the process used to assemble training sets, as well as the effect of alternative processes on the predictive performance of the model. The Working Group considers these first two recommended efforts, compiling/analyzing a comprehensive database and developing/investigating alternative training sets, as complementary and iterative exercises; 3) The combined use of multiple types of tools and approaches for screening should be investigated.*"
7. Typical alerts leading to reproductive toxicity testing from repeated dose studies or developmental toxicity testing studies should be identified in order to develop mechanistic *in vitro* tests to clarify the alert.
8. The analysis of findings in reproductive toxicity studies leading to classifications should be consolidated to identify modes of action to translate into test modules for an ITS.
9. ITS as Bayesian networks of mode of action tests should be formed and optimized by machine learning.
10. PoT from the most promising *in vitro* tests (stem cells, zebrafish, whole embryo culture) should be mapped to feed into a Human Toxome database. Similarly, analysis of samples from animal experiments might allow PoT identification using omics approaches.
11. Identified PoT should lead to specific test development, preferably HTS compatible.
12. A probabilistic risk assessment condensing the information from PoT-based tests and other sources needs to be developed.

7 Overall Conclusions

Author conclusions: Thomas Hartung
Discussants: David A. Basketter, Bas Blaauboer, Robert Burrier, Harvey Clewell, Mardas Daneshian, Chantra Eskes, Alan Goldberg, Nina Hasiwa, Sebastian Hoffmann, Joanna Jaworska, Ian Kimber, Tom Knudsen, Robert Landsiedel, Marcel Leist, Paul Locke, Gavin Maxwell, James McKim, Emily A. McVey, Gladys Ouédraogo, Grace Patlewicz, Olavi Pelkonen, Erwin Roggen, Annamaria Rossi, Costanza Rovida, Irmela Ruhdel, Michael Schwarz, Andreas Schepky, Greet Schoeters, Nigel Skinner, Kerstin Trentz, Marian Turner, Philippe Vanparys, James Yager, Joanne Zurlo

Alternative approaches as one-by-one replacements of animal tests have advanced over the last two decades, and formal validation has delivered the proof-of-principle that they do not lower safety standards (Westmoreland et al., 2010). Increasingly, international acceptance of these methods is being achieved. However, currently validated tests address mainly topical and acute toxicities. The advances in technologies and the gain of toxicological knowledge also appear to make novel approaches feasible for systemic toxicities in a not-so-distant future. Based on the recent analysis commissioned by the European Commission (Adler et al., 2011) and its independent review (Hartung et al., 2011), this expert group has started to set priorities and to identify a roadmap for such a transition. The five whitepapers prepared for this purpose differ in approach and style, even after discussion, revision, compilation, and editing.

The framework for a strategy to replace animal tests has only been developed during the writing of the whitepapers. It has been applied to carcinogenicity (Chapter 5) and reproductive toxicity testing (Chapter 6), but not to the other three fields.

One reason for the differences between the chapters is the different status of the areas. Reproductive toxicity testing has been pioneered by the ReProTect and the ToxCast projects, and the modes of action for genotoxic and non-genotoxic carcinogenicity appear to be more limited in number and better understood than for chronic organ toxicities. Those two areas form a group, together with repeated dose toxicity testing (Chapter 4), as all three areas are suitable for ITS and PoT-based approaches, which represent not only a departure from one-to-one replacement strategies but also a revolution of testing strategies brought about during the last decade. It appears that the large number of target tissues and modes of action will make an ITS approach difficult, requiring that these be broken down to PoT and Pot-based assays, which can then be combined in a HTS platform.

The situation for toxicokinetics (Chapter 2) and skin sensitization (Chapter 3) appears to be very different from the three areas above: Toxicokinetics has to be seen more as the necessary complement to all novel approaches. New assays in this field will be used to enable quantitative *in vitro*/*in vivo* extrapolation (QIVIVE); here, the main approaches are *in silico* modeling and the integration of input from *in vitro* barrier and metabolism models. With a targeted effort, especially broadening the database for modeling, and the necessary funding, an important contribution could be expected in a few years, in line with the earlier reports' judgment (Adler et al., 2011; Hartung et al., 2011). Skin sensitization has seen the development of about 20 *in vitro* and *in silico* models, several of which look very promising and are currently undergoing validation. We will have to see whether and how to combine these tests in the most meaningful way. Eventually, ITS will be set up that reflect the different modes of action and steps in the pathophysiology of skin sensitization.

The main conclusions and recommendations of the report can be summarized as follows:

1. Toxicokinetics

- Represents a necessary complement to all *in vitro* approaches to allow QIVIVE
- need for "*in vitro* kinetics" of chemicals in the experimental systems with the goal of producing proper kinetic parameters for QIVIVE
- *In silico* approaches need to be further optimized
- Need for more comprehensive data collections, especially *in vitro* data from barrier models
- Problems mainly in the fields of bioavailability and urinary excretion
- Achievable with reasonable investment

2. Sensitization

- Reasonably good animal model (LLNA) capable of generating potency and dose response information
- Multiple *in vitro* assays available but unclear which test methods provide potency information
- The need to build mechanistic understanding to enable data integration for potency determination for hazard characterization & risk assessment remains an important *in vitro* challenge

3. Repeated dose testing

- Tox-21c approaches based on PoT represent the key perspective; need to focus on defining levels that cause adverse effects rather than just hazard identification
- Need for data sharing from industry
- Need for models for PoT identification (e.g., stem cells)
- Need for co-cultures, 3D models, and long-term models
- Human disease knowledge and known toxicants must be exploited

- Increased focus on modeling of inflammatory/immunological damage

4. Carcinogenicity

- Possible abolition of current test via an objective assessment with tools of Evidence-based Toxicology (EBT)
- Important ongoing work to optimize genetic toxicity battery
- Further evaluation of cell transformation assay required
- ITS including non-genotoxic modes of action should be developed
- Tox-21c approaches based on PoT (including metabolomics) represent a key opportunity

5. Reproductive Toxicity

- Analysis of current animal tests by EBT approaches
- Validation of (human) embryonic stem cell test variants
- Validation of zebrafish egg test for teratogenicity
- Extension of ITS approaches, extending the approach of ReProTect
- Extension of the ToxCast program currently pioneering PoT-based assessments
- Tox-21c approaches based on PoT, especially mapping the PoT for reproductive toxicity for a Human Toxome database

References

Abadie-Viollon, C., Martin, H., Blanchard, N., et al. (2010). Follow-up to the pre-validation of a harmonised protocol for assessment of CYP induction responses in freshly isolated and cryopreserved human hepatocytes with respect to culture format, treatment, positive reference inducers and incubation conditions. *Toxicol In Vitro 24*, 346-356.

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

Aeby, P., Ashikaga, T., Bessou-Touya, S., et al. (2010). Identifying and characterizing chemical skin sensitizers without animal testing: Colipa's research and method development program. *Toxicol In Vitro 24*, 1465-1473.

Alden, C. L., Smith, P. F., Piper, C. E., et al. (1996). A critical appraisal of the value of the mouse cancer bioassay in safety assessment. *Toxicol Pathol 24*, 722-725.

Aldenberg, T. and Jaworska, J. (2010). Multiple test in silico Weight-of-Evidence for toxicological endpoints. In M. Cronin and J. Madden (eds.), *In Silico Toxicology: Principles and applications*. Cambridge: Royal Society of Chemistry.

Ameen, C., Strehl, R., Bjorquist, P., et al. (2008). Human embryonic stem cells: current technologies and emerging industrial applications. *Crit Rev Oncol Hematol 65*, 54-80.

Ames, B. N. and Gold, L. S. (1990). Chemical carcinogenesis: too many rodent carcinogens. *Proc Natl Acad Sci U S A 87*, 7772-7776.

Ames, B. N., Profet, M., and Gold, L. S. (1990). Nature's chemicals and synthetic chemicals: comparative toxicology. *Proc Natl Acad Sci U S A 87*, 7782-7786.

Ames, B. N. and Gold, L. S. (2000). Paracelsus to parascience: the environmental cancer distraction. *Mutat Res 447*, 3-13.

An, S., Kim, S., Huh, Y., et al. (2009). Expression of surface markers on the human monocytic leukaemia cell line, THP-1, as indicators for the sensitizing potential of chemicals. *Contact Dermatitis 60*, 185-192.

Andersen, M. E. (1991). Physiological modelling of organic compounds. *Ann Occup Hyg 35*, 309-321.

Andersen, M. E., Thomas, R. S., Gaido, K. W., et al. (2005). Dose-response modeling in reproductive toxicology in the systems biology era. *Reprod Toxicol 19*, 327-337.

Andersen, M. E., Clewell, H. J., Carmichael, P. L., et al. (2011). Can case study approaches speed implementation of the NRC report: "toxicity testing in the 21st century: a vision and a strategy?". *ALTEX 28*, 175-182.

Anisimov, V. N., Ukraintseva, S. V., and Yashin, A. I. (2005). Cancer in rodents: does it tell us about cancer in humans? *Nat Rev Cancer 5*, 807-819.

Anonymous (2000). *Scientific Frontiers in Developmental Toxicology and Risk Assessment*. Committee on Developmental Toxicology, Board on Environmental Studies and Toxicology, Commission on Life Sciences, National Research Council. Vol. Washington, DC: National Academy Press. http://www.nap.edu/catalog/9871.html

Anthony, A., Caldwell, J., Hutt, A. J., et al. (1987). Metabolism of estragole in rat and mouse and influence of dose size on excretion of the proximate carcinogen 1'-hydroxyestragole. *Food Chem Toxicol 25*, 799-806.

Api, A. M., Basketter, D. A., Cadby, P. A., et al. (2008). Dermal sensitization quantitative risk assessment (QRA) for fragrance ingredients. *Regul Toxicol Pharmacol 52*, 3-23.

Arkusz, J., Stepnik, M., Sobala, W., et al. (2010). Prediction of the contact sensitizing potential of chemicals using analysis of gene expression changes in human THP-1 monocytes. *Toxicol Lett 199*, 51-59.

Ashby, J. and Tennant, R. W. (1991). Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat Res 257*, 229-306.

Ashby, J. (1997). Cell transformation assays as predictors of carcinogenic potential. *Toxicol Pathol 25*, 334-335.

Ashikaga, T., Sakaguchi, H., Sono, S., et al. (2010). A comparative evaluation of in vitro skin sensitisation tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). *Altern Lab Anim 38*, 275-284.

Asp, J., Steel, D., Jonsson, M., et al. (2010). Cardiomyocyte clusters derived from human embryonic stem cells share similarities with human heart tissue. J Mol Cell Biol 2, 276-283.

Augustine-Rauch, K., Zhang, C. X., and Panzica-Kelly, J. M. (2010). In vitro developmental toxicology assays: A review of the state of the science of rodent and zebrafish whole embryo culture and embryonic stem cell assays. *Birth Defects Res C Embryo Today 90*, 87-98.

Bailey, J., Knight, A., and Balcombe, J. (2005). The future of teratology research is in vitro. *Biogenic Amines* 97–145.

Balls, M. and Combes, R. (2005). Validation at a crossroads. *Altern Lab Anim 33*, 187.

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

Balls, M. and Clothier, R. (2010). A FRAME response to the Draft Report on Alternative (Non-animal) Methods for Cosmetics Testing: Current Status and Future Prospects – 2010. *Altern Lab Anim 38*, 345-353.

Banas, A., Teratani, T., Yamamoto, Y., et al. (2007). Adipose tissue-derived mesenchymal stem cells as a source of human hepatocytes. *Hepatology 46*, 219-228.

Barlow, S. and Schlatter, J. (2010). Risk assessment of carcinogens in food. *Toxicol Appl Pharmacol 243*, 180-190.

Barratt, M. D. (2000). Prediction of toxicity from chemical structure. *Cell Biol Toxicol 16*, 1-13.

Basketter, D. and Maxwell, G. (2007). In vitro approaches to the identification and characterization of skin sensitizers. *Cutan Ocul Toxicol 26*, 359-373.

Basketter, D. A., Gerberick, F., and Kimber, I. (2007). The local lymph node assay and the assessment of relative potency: status of validation. *Contact Dermatitis 57*, 70-75.

Basketter, D. A. (2008). Nonanimal alternatives for skin sensitization: a step forward? *Toxicol Sci 102*, 1-2.

Basketter, D. A. and Kimber, I. (2009). Updating the skin sensitization in vitro data assessment paradigm in 2009. *J Appl Toxicol 29*, 545-550.

Basketter, D. A. and Kimber, I. (2011). Predictive tests for irritants and allergens and their use in quantitative risk assessment. In J. D. Johansen, P. F. Frosch and J. Lepoittevin (eds.), *Contact Dermatitis*. Berlin: Springer.

Bauch, C., Kolle, S. N., Fabian, E., et al. (2011). Intralaboratory validation of four in vitro assays for the prediction of the skin sensitizing potential of chemicals. *Toxicol In Vitro 25*, 1162-1168.

Baylin, S. B. and Ohm, J. E. (2006). Epigenetic gene silencing in cancer – a mechanism for early oncogenic pathway addiction? *Nat Rev Cancer 6*, 107-116.

Benet, L. Z., Cummins, C. L., and Wu, C. Y. (2003). Transporter-enzyme interactions: implications for predicting drug-drug interactions from in vitro data. *Curr Drug Metab 4*, 393-398.

Benfenati, E., Benigni, R., Demarini, D. M., et al. (2009). Predictive models for carcinogenicity and mutagenicity: frameworks, state-of-the-art, and perspectives. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev 27,* 57-90.

Benford, D., Bolger, P. M., Carthew, P., et al. (2010). Application of the Margin of Exposure (MOE) approach to substances in food that are genotoxic and carcinogenic. *Food Chem Toxicol 48, Suppl 1*, S2-24.

Benigni, R. and Giuliani, A. (2003). Putting the Predictive Toxicology Challenge into perspective: reflections on the results. *Bioinformatics 19,* 1194-1200.

Benigni, R. (2004). Chemical structure of mutagens and carcinogens and the relationship with biological activity. *J Exp Clin Cancer Res 23*, 5-8.

Benigni, R. and Zito, R. (2004). The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results. *Mutat Res 566*, 49-63.

Benigni, R. and Bossa, C. (2006). Structure-activity models of chemical carcinogens: state of the art, and new directions. *Ann Ist Super Sanita 42*, 118-126.

Benigni, R., Bossa, C., Tcheremenskaia, O., et al. (2010). Alternatives to the carcinogenicity bioassay: in silico methods, and the in vitro and in vivo mutagenicity assays. *Expert Opin Drug Metab Toxicol 6*, 809-819.

Benigni, R. and Bossa, C. (2011). Alternative strategies for carcinogenicity assessment: an efficient and simplified approach based on in vitro mutagenicity and cell transformation assays. *Mutagenesis 26*, 455-460.

Bernstein, L., Gold, L. S., Ames, B. N., et al. (1985). Some tautologous aspects of the comparison of carcinogenic potency in rats and mice. *Fundam Appl Toxicol 5*, 79-86.

Berwald, Y. and Sachs, L. (1963). In Vitro Cell Transformation with Chemical Carcinogens. *Nature 200*, 1182-1184.

Berwald, Y. and Sachs, L. (1965). In vitro transformation of normal cells to tumor cells by carcinogenic hydrocarbons. *J Natl Cancer Inst 35*, 641-661.

Bessems, M., t Hart, N. A., Tolba, R., et al. (2006). The isolated perfused rat liver: standardization of a time-honoured model. *Lab Anim 40*, 236-246.

Bhogal, N., Grindon, C., Combes, R., et al. (2005). Toxicity testing: creating a revolution based on new technologies. *Trends Biotechnol 23*, 299-307.

Bitsch, A., Jacobi, S., Melber, C., et al. (2006). REPDOSE: A database on repeated dose toxicity studies of commercial chemicals – A multifunctional tool. *Regul Toxicol Pharmacol 46*, 202-210.

Blaauboer, B. J. (2001). Toxicodynamic modelling and the interpretation of in vitro toxicity data. *Toxicol Lett 120*, 111-123.

Blaauboer, B. J., Clewell, H. J., Clothier, R., et al. (2001). In vitro methods for assessing acute toxicity: biokinetic determinations. Report of the International Workshop on in vitro methods for assessing acute systemic toxicity. NIEHS. 01-4499. NIH, 47-60. http://www.epa.gov/hpv/pubs/general/nih2001a.pdf

Blaauboer, B. J. (2002). The necessity of biokinetic information in the interpretation of in vitro toxicity data. *Altern Lab Anim 30, Suppl 2*, 85-91.

Blaauboer, B. J. (2003). The integration of data on physico-chemical properties, in vitro-derived toxicity data and physiologically based kinetic and dynamic as modelling a tool in hazard and risk assessment. A commentary. *Toxicol Lett 138*, 161-171.

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

Blakey, D., Galloway, S. M., Kirkland, D. J., et al. (2008). Regulatory aspects of genotoxicity testing: from hazard identification to risk assessment. *Mutat Res 657*, 84-90.

Bobr, A., Olvera-Gomez, I., Igyarto, B. Z., et al. (2010). Acute ablation of Langerhans cells enhances skin immune responses. *J Immunol 185*, 4724-4728.

Boekelheide, K. and Andersen, M. E. (2010). A mechanistic redefinition of adverse effects – a key step in the toxicity testing paradigm shift. *ALTEX 27*, 243-252.

Boekelheide, K. and Campion, S. N. (2010). Toxicity testing in the 21st century: using the new toxicity testing paradigm to create a taxonomy of adverse effects. *Toxicol Sci 114*, 20-24.

Boraso, M. and Viviani, B. (2011). Glia-neuron sandwich cocultures: an in vitro approach to evaluate cell-to-cell communication in neuroinflammation and neurotoxicity. *Methods Mol Biol 758*, 135-152.

Bottini, A. A. and Hartung, T. (2009). Food for thought ... on the economics of animal testing. *ALTEX 26*, 3-16.

Bottini, A. A. and Hartung, T. (2010). The economics of animal testing. *ALTEX* 67-77.

Bouvier d'Yvoire, M., Prieto, P., Blaauboer, B. J., et al. (2007). Physiologically-based Kinetic Modelling (PBK Modelling): meeting the 3Rs agenda. The report and recommendations of ECVAM Workshop 63. *Altern Lab Anim 35*, 661-671.

Breithaupt, H. (2006). The costs of REACH. REACH is largely welcomed, but the requirement to test existing chemicals for adverse effects is not good news for all. *EMBO Rep 7*, 968-971.

Bremer, S., Pellizzer, C., Coecke, S., et al. (2002). Detection of the embryotoxic potential of cyclophosphamide by using a combined system of metabolic competent cells and embryonic stem cells. *Altern Lab Anim 30*, 77-85.

Bremer, S. and Hartung, T. (2004). The use of embryonic stem cells for regulatory developmental toxicity testing in vitro – the current status of test development. *Curr Pharm Des 10*, 2733-2747.

Bremer, S., Brittebo, E., Dencker, L., et al. (2007a). In vitro tests for detecting chemicals affecting the embryo implantation process. The report and recommendations of ECVAM workshop 62 – a strategic workshop of the EU ReProTect project. *Altern Lab Anim 35*, 421-439.

Bremer, S., Pellizzer, C., Hoffmann, S., et al. (2007b). The development of new concepts for assessing reproductive toxicity applicable to large scale toxicological programmes. *Curr Pharm Des 13*, 3047-3058.

Bridges, B. A. (1988). Genetic toxicology at the crossroads – a personal view on the deployment of short-term tests for predicting carcinogenicity. *Mutat Res 205*, 25-31.

Brown, N. A. and Fabro, S. (1983). The value of animal teratogenicity testing for predicting human risk. *Clin Obstet Gynecol 26*, 467-477.

Brown, R. P., Delp, M. D., Lindstedt, S. L., et al. (1997). Physiological parameter values for physiologically based pharmacokinetic models. *Toxicol Ind Health 13*, 407-484.

Brunner, D., Frank, J., Appl, H., et al. (2010). Serum-free cell culture: the serum-free media interactive online database. *ALTEX 27*, 53-62.

Bucher, J. R. (1998). Update on national toxicology program (NTP) assays with genetically altered or "transgenic" mice. *Environ Health Perspect 106*, 619-621.

Bucher, J. R. (2000). Doses in rodent cancer studies: sorting fact from fiction. *Drug Metab Rev 32*, 153-163.

Buehler, E. V. (1965). Delayed Contact Hypersensitivity in the Guinea Pig. *Arch Dermatol 91*, 171-177.

Calabrese, E. J. (2009). The road to linearity: why linearity at low doses became the basis for carcinogen risk assessment. *Arch Toxicol 83*, 203-225.

Calabrese, E. J. (2011). Muller's Nobel lecture on dose-response for ionizing radiation: ideology or science? *Arch Toxicol 85*, 1495-1498.

Carfi, M., Gennari, A., Malerba, I., et al. (2007). In vitro tests to evaluate immunotoxicity: a preliminary study. *Toxicology 229*, 11-22.

Carney, E. W. and Kimmel, C. A. (2007). Interpretation of skeletal variations for human risk assessment: delayed ossification and wavy ribs. *Birth Defects Res B Dev Reprod Toxicol 80*, 473-496.

Carney, E. W., Ellis, A. L., Tyl, R. W., et al. (2011). Critical evaluation of current developmental toxicity testing strategies: a case of babies and their bathwater. *Birth Defects Res B Dev Reprod Toxicol 92*, 395-403.

Chahoud, I., Buschmann, J., Clark, R., et al. (1999). Classification terms in developmental toxicology: need for harmonisation. Report of the Second Workshop on the Terminology in Developmental Toxicology Berlin, 27-28 August 1998. *Reprod Toxicol 13*, 77-82.

Chapin, R. E. and Stedman, D. B. (2009). Endless possibilities: stem cells and the vision for toxicology testing in the 21st century. *Toxicol Sci 112*, 17-22.

Chen, J. J., Moon, H., and Kodell, R. L. (2007). A probabilistic framework for non-cancer risk assessment. *Regul Toxicol Pharmacol 48*, 45-50.

Chiba, M., Ishii, Y., and Sugiyama, Y. (2009). Prediction of hepatic clearance in human from in vitro data for successful drug development. *AAPS J 11*, 262-276.

Choi, S. M., Yoo, S. D., and Lee, B. M. (2004). Toxicological characteristics of endocrine-disrupting chemicals: developmental toxicity, carcinogenicity, and mutagenicity. *J Toxicol Environ Health B Crit Rev 7*, 1-24.

Clausen, B. E. and Kel, J. M. (2010). Langerhans cells: critical regulators of skin immunity? *Immunol Cell Biol 88*, 351-360.

Clewell, H. J., 3rd (1993). Coupling of computer modeling with in vitro methodologies to reduce animal usage in toxicity testing. *Toxicol Lett 68*, 101-117.

Coecke, S., Balls, M., Bowe, G., et al. (2005). Guidance on good cell culture practice. a report of the second ECVAM task force on good cell culture practice. *Altern Lab Anim 33*, 261-287.

Coecke, S., Ahr, H., Blaauboer, B. J., et al. (2006). Metabolism: a bottleneck in in vitro toxicological test development. The report and recommendations of ECVAM workshop 54. *Altern Lab Anim 34*, 49-84.

Coecke, S., Goldberg, A. M., Allen, S., et al. (2007). Workgroup report: incorporating in vitro alternative methods for developmental neurotoxicity into international hazard and risk assessment strategies. *Environ Health Perspect 115*, 924-931.

Cohen, S. M., Robinson, D., and MacDonald, J. (2001). Alternative models for carcinogenicity testing. *Toxicol Sci 64*, 14-19.

Cohen, S. M. (2004). Human carcinogenic risk evaluation: an alternative approach to the two-year rodent bioassay. *Toxicol Sci 80*, 225-229.

Cohen, S. M. and Arnold, L. L. (2011). Chemical carcinogenesis. *Toxicol Sci 120, Suppl 1*, S76-92.

Colacci, A., Mascolo, M. G., Perdichizzi, S., et al. (2011). Different sensitivity of BALB/c 3T3 cell clones in the response to carcinogens. *Toxicol In Vitro 25*, 1183-1190.

Collins, F. S., Gray, G. M., and Bucher, J. R. (2008). Toxicology. Transforming environmental health protection. *Science 319*, 906-907.

Collins, T. F. (2006). History and evolution of reproductive and developmental toxicology guidelines. *Curr Pharm Des 12*, 1449-1465.

Combes, R., Balls, M., Curren, R., et al. (1999). Cell transformation assays as predictors of human carcinogenicity – the report of ECVAM workshop 39. *Altern Lab Anim*, 745-767.

Combes, R., Grindon, C., Cronin, M. T., et al. (2007). Proposed integrated decision-tree testing strategies for mutagenicity and carcinogenicity in relation to the EU REACH legislation. *Altern Lab Anim 35*, 267-287.

Corvi, R., Ahr, H. J., Albertini, S., et al. (2006). Meeting report: Validation of toxicogenomics-based test systems: ECVAM-ICCVAM/NICEATM considerations for regulatory use. *Environ Health Perspect 114*, 420-429.

Cox, J. L. and Rizzino, A. (2010). Induced pluripotent stem cells: what lies beyond the paradigm shift. *Exp Biol Med (Maywood) 235*, 148-158.

Crebelli, R. (2000). Threshold-mediated mechanisms in mutagenesis: implications in the classification and regulation of chemical mutagens. *Mutat Res 464*, 129-135.

Creton, S., Aardema, M. J., Carmichael, P. L., et al. (2011). Cell transformation assays for prediction of carcinogenic potential: state of the science and future research needs. *Mutagenesis*

Crofton, K. M., Mundy, W. R., Lein, P. J., et al. (2011). Developmental neurotoxicity testing: recommendations for developing alternative methods for the screening and prioritization of chemicals. *ALTEX 28*, 9-15.

Cronin, M. T., Dearden, J. C., Duffy, J. C. (2002). The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints. *SAR QSAR Environ. Res. 13*, 167-76.

Cronin, M. T., Enoch, S. J., Hewitt, M., et al. (2011). Formation of mechanistic categories and local models to facilitate the prediction of toxicity. *ALTEX 28*, 45-49.

Crow, J. A., Borazjani, A., Potter, P. M., et al. (2007). Hydrolysis of pyrethroids by human and rat tissues: examination of intestinal, liver and serum carboxylesterases. *Toxicol Appl Pharmacol 221*, 1-12.

Daneshian, M., Leist, M., and Hartung, T. (2010). The Center for Alternatives to Animal Testing – Europe (CAAT-EU): a transatlantic bridge for the paradigm shift in toxicology. *ALTEX 27*, 63-69.

Darnell, M., Schreiter, T., Zeilinger, K., et al. (2011). Cytochrome P450-dependent metabolism in HepaRG cells cultured in a dynamic three-dimensional bioreactor. *Drug Metab Dispos 39*, 1131-1138.

Dash, A., Inman, W., Hoffmaster, K., et al. (2009). Liver tissue engineering in the evaluation of drug safety. *Expert Opin Drug Metab Toxicol 5*, 1159-1174.

Daston, G. P. and Seed, J. (2007). Skeletal malformations and variations in developmental toxicity studies: interpretation issues for human risk assessment. *Birth Defects Res B Dev Reprod Toxicol 80*, 421-424.

Davies, T. and Monro, A. (1995). Marketed human pharmaceuticals reported to be tumorigenic in rodents. *Int J Toxicol 14*, 90-107.

Davies, T. S., Lynch, B. S., Monro, A. M., et al. (2000). Rodent carcinogenicity tests need be no longer than 18 months: an analysis based on 210 chemicals in the IARC monographs. *Food Chem Toxicol 38*, 219-235.

Davila, J., Stedman, D., Engle, S., et al. (2008). Stem cell technology for embryotoxicity, cardiotoxicity, and hepatotoxicity evaluation. In S. Ekins and J. J. Xu (eds.), *Drug Efficacy, Safety, and Biologics Discovery: Emerging Technologies and Tools*. Hoboken, USA: John Wiley & Sons.

De Buck, S. S. and Mackie, C. E. (2007). Physiologically based approaches towards the prediction of pharmacokinetics: in vitro-in vivo extrapolation. *Expert Opin Drug Metab Toxicol 3*, 865-878.

de Jong, E., Doedee, A. M., Reis-Fernandes, M. A., et al. (2011). Potency ranking of valproic acid analogues as to inhibition of cardiac differentiation of embryonic stem cells in comparison to their in vivo embryotoxicity. *Reprod Toxicol 31*, 375-382.

De Silva, O., Basketter, D. A., Barrat, M. D., et al. (1995). Alternative methods for skin sensitisation testing: the report and recommendations of ECVAM workshop 19. *Altern Lab Anim*, 683-705.

De Smedt, A., Steemans, M., De Boeck, M., et al. (2008). Optimisation of the cell cultivation methods in the embryonic stem cell test results in an increased differentiation potential of the cells into strong beating myocard cells. *Toxicol In Vitro 22*, 1789-1796.

DeJongh, J., Verhaar, H. J., and Hermens, J. L. (1997). A quantitative property-property relationship (QPPR) approach to estimate in vitro tissue-blood partition coefficients of organic chemicals in rats and humans. *Arch Toxicol 72*, 17-25.

Dejongh, J., Forsby, A., Houston, J. B., et al. (1999). an integrated approach to the prediction of systemic toxicity using computer-based biokinetic models and biological in vitro test methods: Overview of a prevalidation study based on the ECITTS project. *Toxicol In Vitro 13*, 549-554.

Derelanko, M. J. and Hollinger, M. A. (2001). *Handbook of Toxicology*. Vol. 2. London: Informa Healthcare.

Dietrich, D. R. (2010). Courage for simplification and imperfection in the 21st century assessment of "Endocrine disruption". *ALTEX 27*, 264-278.

Doll, R. and Peto, R. (1981). The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J Natl Cancer Inst 66*, 1191-1308.

dos Santos, G. G., Reinders, J., Ouwehand, K., et al. (2009). Progress on the development of human in vitro dendritic cell based assays for assessment of the sensitizing potential of a compound. *Toxicol Appl Pharmacol 236*, 372-382.

dos Santos, G. G., Spiekstra, S. W., Sampat-Sardjoepersad, S. C., et al. (2011). A potential in vitro epidermal equivalent assay to determine sensitizer potency. *Toxicol In Vitro 25*, 347-357.

Doweyko, A. M. (2004). 3D-QSAR illusions. *J Comput Aided Mol Des 18*, 587-596.

Duff, T., Carter, S., Feldman, G., et al. (2002). Transepithelial resistance and inulin permeability as endpoints in in vitro nephrotoxicity testing. *Altern Lab Anim 30, Suppl 2,* 53-59.

Durodie, B. (2003). The true cost of precautionary chemicals regulation. *Risk Anal 23*, 389-398.

ECHA (European Chemicals Agency) (2008). Guidance on information requirements and chemical safety assessment, Chapter R.7a: Endpoint specific guidance. 1-428. http://www.echa.europa.eu/documents/10162/17224/information_requirements_r7a_en.pdf

Elespuru, R. K., Agarwal, R., Atrakchi, A. H., et al. (2009). Current and future application of genetic toxicity assays: the role and value of in vitro mammalian assays. *Toxicol Sci 109*, 172-179.

Emmendoerffer, A., Hecht, M., Boeker, T., et al. (2000). Role of inflammation in chemical-induced lung cancer. *Toxicol Lett 112-113*, 185-191.

Emter, R., Ellis, G., and Natsch, A. (2010). Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. *Toxicol Appl Pharmacol 245*, 281-290.

Ennever, F. K., Noonan, T. J., and Rosenkranz, H. S. (1987). The predictivity of animal bioassays and short-term genotoxicity tests for carcinogenicity and non-carcinogenicity to humans. *Mutagenesis 2*, 73-78.

Ennever, F. K. and Lave, L. B. (2003). Implications of the lack of accuracy of the lifetime rodent bioassay for predicting human carcinogenicity. *Regul Toxicol Pharmacol 38*, 52-57.

Entine, J. (2011). *Scared to Death – How Chemophobia Threatens Public Health*. Vol. New York: American Council on Science & Health.

EPA – U.S. Environmental Protection Agency (1988). Reference physiological parameters in pharmacokinetic modeling 600/6-88/004. http://nepis.epa.gov → Search: "600688004"

EPA – U.S. Environmental Protection Agency (2009). The U.S. Environmental Protection Agency's Strategic Plan for evaluating the toxicity of chemicals. 1-38. http://nepis.epa.gov → Search: "100K09001"

Esch, M. B., King, T. L., and Shuler, M. L. (2011). The role of body-on-a-chip devices in drug and toxicity studies. *Annu Rev Biomed Eng 13*, 55-72.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet 8*, 286-298.

Fagerholm, U. (2007). Prediction of human pharmacokinetics – evaluation of methods for prediction of hepatic metabolic clearance. *J Pharm Pharmacol 59*, 803-828.

Falsig, J., Latta, M., and Leist, M. (2004a). Defined inflammatory states in astrocyte cultures: correlation with susceptibility towards CD95-driven apoptosis. *J Neurochem 88*, 181-193.

Falsig, J., Porzgen, P., Lotharius, J., et al. (2004b). Specific modulation of astrocyte inflammation by inhibition of mixed lineage kinases with CEP-1347. *J Immunol 173*, 2762-2770.

Farmer, P. B. (2002). Committee on Mutagenicity of Chemicals in Food, Consumer Products and the Environment ILSI/HESI research programme on alternative cancer models: results of Syrian hamster embryo cell transformation assay. International Life Sciences Institute/Health and Environmental Science Institute. *Toxicol Pathol 30*, 536-538.

Febriana, S. A., Jungbauer, F., Soebono, H., et al. (2011). Inventory of the chemicals and the exposure of the workers' skin to these at two leather factories in Indonesia. *Int Arch Occup Environ Health*

Felter, S., Lane, R. W., Latulippe, M. E., et al. (2009). Refining the threshold of toxicological concern (TTC) for risk prioritization of trace chemicals in food. *Food Chem Toxicol 47*, 2236-2245.

Fielden, M. R., Nie, A., McMillian, M., et al. (2008). Interlaboratory evaluation of genomic signatures for predicting carcinogenicity in the rat. *Toxicol Sci 103*, 28-34.

Fielden, M. R., Adai, A., Dunn, R. T., 2nd, et al. (2011). Development and evaluation of a genomic signature for the prediction and mechanistic assessment of nongenotoxic hepatocarcinogens in the rat. *Toxicol Sci 124*, 54-74.

Filser, J. G., Csanady, G. A., Kreuzer, P. E., et al. (1995). Toxicokinetic models for volatile industrial chemicals and reactive metabolites. *Toxicol Lett 82-83*, 357-366.

Fleischer, M. (2007). Testing costs and testing capacity according to the REACH requirements. Results of a survey of independent and corporate GLP laboratories in the EU and Switzerland. *Journal of Business Chemistry 4*, 96-114.

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

Freedman, D. A. and Zeisel, H. (1988). From mouse to man: The quantitative assessment of cancer risks. *Statist Sci* 3-56.

Friedman, J. M. (2009). Big risks in small groups: The difference between epidemiology and counselling. *Birth Defects Res A Clin Mol Teratol 85*, 720-724.

Fung, V. A., Barrett, J. C., and Huff, J. (1995). The carcinogenesis bioassay in perspective: application in identifying human cancer hazards. *Environ Health Perspect 103*, 680-683.

Galbiati, V., Mitjans, M., and Corsini, E. (2010). Present and future of in vitro immunotoxicology in drug development. *J Immunotoxicol 7*, 255-267.

Galbiati, V., Mitjans, M., Lucchi, L., et al. (2011). Further development of the NCTC 2544 IL-18 assay to identify in vitro contact allergens. *Toxicol In Vitro 25*, 724-732.

Gaylor, D. W. (2005). Are tumor incidence rates from chronic bioassays telling us what we need to know about carcinogens? *Regul Toxicol Pharmacol 41*, 128-133.

Gennari, A., Ban, M., Braun, A., et al. (2005). The use of in vitro systems for evaluating immunotoxicity: The report and recommendations of an ECVAM Workshop. *J Immunotoxicol 2*, 61-83.

Genschow, E., Spielmann, H., Scholz, G., et al. (2002). The ECVAM international validation study on in vitro embryotoxicity tests: results of the definitive phase and evaluation of prediction models. European Centre for the Validation of Alternative Methods. *Altern Lab Anim 30*, 151-176.

Genschow, E., Spielmann, H., Scholz, G., et al. (2004). Validation of the embryonic stem cell test in the international EC-

VAM validation study on three in vitro embryotoxicity tests. *Altern Lab Anim 32*, 209-244.

Georgiades, P., Ferguson-Smith, A. C., and Burton, G. J. (2002). Comparative developmental anatomy of the murine and human definitive placentae. *Placenta 23*, 3-19.

Gerberick, G. F., Vassallo, J. D., Bailey, R. E., et al. (2004). Development of a peptide reactivity assay for screening contact allergens. *Toxicol Sci 81*, 332-343.

Gerberick, G. F., Troutman, J. A., Foertsch, L. M., et al. (2009). Investigation of peptide reactivity of pro-hapten skin sensitizers using a peroxidase-peroxide oxidation system. *Toxicol Sci 112*, 164-174.

Gimble, J. and Guilak, F. (2003). Adipose-derived adult stem cells: isolation, characterization, and differentiation potential. *Cytotherapy 5*, 362-369.

Gold, L. S., Slone, T. H., Manley, N. B., et al. (1991). Target organs in chronic bioassays of 533 chemical carcinogens. *Environ Health Perspect 93*, 233-246.

Gold, L. S., Slone, T. H., and Ames, B. N. (1998). What do animal cancer tests tell us about human cancer risk?: Overview of analyses of the carcinogenic potency database. *Drug Metab Rev 30*, 359-404.

Gold, L. S., Manley, N. B., Slone, T. H., et al. (2005). Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature through 1997 and by the National Toxicology Program in 1997-1998. *Toxicol Sci 85*, 747-808.

Gomez-Lechon, M. J., Castell, J. V., and Donato, M. T. (2007). Hepatocytes – the choice to investigate drug metabolism and toxicity in man: in vitro variability as a reflection of in vivo. *Chem Biol Interact 168*, 30-50.

Goodman, J. I., Gollapudi, B., and Lehman-McKeeman, L. D. (2007). Genetic toxicity assessment: employing the best science for human safety evaluation. *Toxicol Sci 1*, 1.

Gottmann, E., Kramer, S., Pfahringer, B., et al. (2001). Data quality in predictive toxicology: reproducibility of rodent carcinogenicity experiments. *Environ Health Perspect 109*, 509-514.

Gray, G. M., Li, P., Shlyakhter, I., et al. (1995). An empirical examination of factors influencing prediction of carcinogenic hazard across species. *Regul Toxicol Pharmacol 22*, 283-291.

Greenhough, S., Medine, C. N., and Hay, D. C. (2010). Pluripotent stem cell derived hepatocyte like cells and their potential in toxicity screening. *Toxicology 278*, 250-255.

Greenland, S. (1998). Probability logic and probabilistic induction. *Epidemiology 9*, 322-332.

Griesinger, C., Hoffmann, S., Kinsner-Ovaskainen, A., et al. (2007). Proceedings of the First International Forum Towards Evidence-Based Toxicology. First International Forum Towards Evidence-Based Toxicology, Como, Italy. *Human Exp. Toxicol 28, Spec. Issue*, 83-176.

Grindon, C., Combes, R., Cronin, M. T., et al. (2008). An integrated decision-tree testing strategy for repeat dose toxicity with respect to the requirements of the EU REACH legislation. *Altern Lab Anim 36, Suppl 1*, 139-147.

Groebe, K., Hayess, K., Klemm-Manns, M., et al. (2010). Protein biomarkers for in vitro testing of embryotoxicity. *J Proteome Res 9*, 5727-5738.

Guguen-Guillouzoa, C., Corlua, A., Guillouzo, A. (2010) Stem cell-derived hepatocytes and their use in toxicology. *Toxicology 270*, 3-9.

Gulden, M. and Seibert, H. (1997). Influence of protein binding and lipophilicity on the distribution of chemical compounds in in vitro systems. *Toxicol In Vitro 11*, 479-483.

Gulden, M., Morchel, S., and Seibert, H. (2001). Factors influencing nominal effective concentrations of chemical compounds in vitro: cell concentration. *Toxicol In Vitro 15*, 233-243.

Gulden, M., Morchel, S., Tahan, S., et al. (2002). Impact of protein binding on the availability and cytotoxic potency of organochlorine pesticides and chlorophenols in vitro. *Toxicology 175*, 201-213.

Gulden, M. and Seibert, H. (2003). In vitro-in vivo extrapolation: estimation of human serum concentrations of chemicals equivalent to cytotoxic concentrations in vitro. *Toxicology 189*, 211-222.

Gulden, M., Dierickx, P., and Seibert, H. (2006). Validation of a prediction model for estimating serum concentrations of chemicals which are equivalent to toxic concentrations in vitro. *Toxicol In Vitro 20*, 1114-1124.

Guyton, K. Z., Kyle, A. D., Aubrecht, J., et al. (2009). Improving prediction of chemical carcinogenicity by considering multiple mechanisms and applying toxicogenomic approaches. *Mutat Res 681*, 230-240.

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

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

Hartung, T. (2007a). Food for thought ... on cell culture. *ALTEX 24*, 143-152.

Hartung, T. (2007b). Food for thought ... on validation. *ALTEX 24*, 67-80.

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

Hartung, T. (2008b). Food for thought ... on animal tests. *ALTEX 25*, 3-16.

Hartung, T. (2008c). Toward a new toxicology – evolution or revolution? *Altern Lab Anim 36*, 635-639.

Hartung, T. and Leist, M. (2008). Food for thought ... on the evolution of toxicology and the phasing out of animal testing. *ALTEX 25*, 91-102.

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

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

Hartung, T. (2009c). Per aspirin ad astra. *Altern Lab Anim 37, Suppl 2*, 45-47.

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

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

Hartung, T. and Rovida, C. (2009a). Chemical regulators have overreached. *Nature 460*, 1080-1081.

Hartung, T. and Rovida, C. (2009b). That which must not, can not be... A reply to the EChA and EDF responses to the REACH analysis of animal use and costs. *ALTEX 26*, 307-311.

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

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

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

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

Hartung, T. (2010e). Food for thought ... on alternative methods for nanoparticle safety testing. *ALTEX 27*, 87-95.

Hartung, T., Bruner, L., Curren, R., et al. (2010). First alternative method validated by a retrospective weight-of-evidence approach to replace the Draize eye test for the identification of non-irritant substances for a defined applicability domain. *ALTEX 27*, 43-51.

Hartung, T. (2011). From alternative methods to a new toxicology. *Eur J Pharm Biopharm 77*, 338-349.

Hartung, T., Blaauboer, B. J., Bosgra, S., et al. (2011). An expert consortium review of the EC-commissioned report "alternative (Non-Animal) methods for cosmetics testing: current status and future prospects – 2010". *ALTEX 28*, 183-209.

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

Hartung, T. and Sabbioni, E. (2011). Alternative in vitro assays in nanomaterial toxicology. *Wiley Interdiscip Rev Nanomed Nanobiotechnol*

Haseman, J., Melnick, R., Tomatis, L., et al. (2001). Carcinogenesis bioassays: study duration and biological relevance. *Food Chem Toxicol 39*, 739-744.

Haseman, J. K., Boorman, G. A., and Huff, J. (1997). Value of historical control data and other issues related to the evaluation of long-term rodent carcinogenicity studies. *Toxicol Pathol 25*, 524-527.

Hawkins, D. M. (2004). The problem of overfitting. *J Chem Inf Comput Sci 44*, 1-12.

Helma, C. and Kramer, S. (2003). A survey of the Predictive Toxicology Challenge 2000-2001. *Bioinformatics 19*, 1179-1182.

Henn, A., Lund, S., Hedtjarn, M., et al. (2009). The suitability of BV2 cells as alternative model system for primary microglia cultures or for animal experiments examining brain inflammation. *ALTEX 26*, 83-94.

Hestermann, E. V., Stegeman, J. J., and Hahn, M. E. (2000). Serum alters the uptake and relative potencies of halogenated aromatic hydrocarbons in cell culture bioassays. *Toxicol Sci 53*, 316-325.

Hettwer, M., Reis-Fernandes, M. A., Iken, M., et al. (2010). Metabolic activation capacity by primary hepatocytes expands the applicability of the embryonic stem cell test as alternative to experimental animal testing. *Reprod Toxicol 30*, 113-120.

Hoffmann, S. and Hartung, T. (2005). Diagnosis: toxic! – trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations. *Toxicol Sci 85*, 422-428.

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

Hoffmann, S., Edler, L., Gardner, I., et al. (2008). Points of reference in the validation process: the report and recommendations of ECVAM Workshop 66. *Altern Lab Anim 36*, 343-352.

Hoke, R. A. and Ankley, G. T. (2005). Application of frog embryo teratogenesis assay-Xenopus to ecological risk assessment. *Environ Toxicol Chem 24*, 2677-2690.

Holson, J., Nemec, M., Stump, D. G., et al. (2006). Significance, reliability, and interpretation of developmental and reproductive toxicity study findings. In R. Hood (eds.), *Developmental and Reproductive Toxicology. A Practical Approach*. London, UK: Informa Healthcare.

Hotchkiss, A. K., Rider, C. V., Blystone, C. R., et al. (2008). Fifteen years after "Wingspread" – environmental endocrine disrupters and human and wildlife health: where we are today and where we need to go. *Toxicol Sci 105*, 235-259.

Houston, J. B. and Carlile, D. J. (1997). Prediction of hepatic clearance from microsomes, hepatocytes, and liver slices. *Drug Metab Rev 29*, 891-922.

Houston, J. B. and Galetin, A. (2008). Methods for predicting in vivo pharmacokinetics using data from in vitro assays. *Curr Drug Metab 9*, 940-951.

Huang, S., Wiszniewski, L., and Derouette, J. (2009). Disease Models: In vitro organ culture models of asthma. *Drug Discovery Today 6*, 137-144.

Huff, J. (1999). Long-term chemical carcinogenesis bioassays predict human cancer hazards. Issues, controversies, and uncertainties. *Ann N Y Acad Sci 895*, 56-79.

Hurtt, M. E., Cappon, G. D., and Browning, A. (2003). Proposal for a tiered approach to developmental toxicity testing for veterinary pharmaceutical products for food-producing animals. *Food Chem Toxicol 41*, 611-619.

Imai, T., Imoto, M., Sakamoto, H., et al. (2005). Identification of esterases expressed in Caco-2 cells and effects of their hydrolyzing activity in predicting human intestinal absorption. *Drug Metab Dispos 33*, 1185-1190.

Imai, T. (2006). Human carboxylesterase isozymes: catalytic properties and rational drug design. *Drug Metab Pharmacokinet 21*, 173-185.

Inoue, T., Tanaka, K., Mishima, M., et al. (2007). Predictive in vitro cardiotoxicity and hepatotoxicity screening system using neonatal rat heart cells and rat hepatocytes 6th World Congress on Alternatives & Animal Use in the Life Sciences, Tokyo. *AATEX 14, Spec. Issue*, 457-462.

Jacobs, A. (2009). An FDA perspective on the nonclinical use of the X-Omics technologies and the safety of new drugs. *Toxicol Lett 186*, 32-35.

Jaenisch, R. (2009). Stem cells, pluripotency and nuclear reprogramming. *J Thromb Haemost 7, Suppl 1*, 21-23.

Jager, T., Vermeire, T. G., Rikken, M. G., et al. (2001). Opportunities for a probabilistic risk assessment of chemicals in the European Union. *Chemosphere 43*, 257-264.

Jamei, M., Marciniak, S., Feng, K., et al. (2009). The Simcyp population-based ADME simulator. *Expert Opin Drug Metab Toxicol 5*, 211-223.

Janer, G., Hakkert, B. C., Slob, W., et al. (2007). A retrospective analysis of the two-generation study: what is the added value of the second generation? *Reprod Toxicol 24*, 97-102.

Jaworska, J., McDowell, R., and Aardema, M. (2005). Application of decision theory to interpretation of in vitro tests battery results. 5th World Congress on Alternatives and Animal Use in the Life Sciences, Berlin. *ALTEX 22, Spec. Issue*, 132.

Jaworska, J., Gabbert, S., and Aldenberg, T. (2010). Towards optimization of chemical testing under REACH: a Bayesian network approach to Integrated Testing Strategies. *Regul Toxicol Pharmacol 57*, 157-167.

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

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

Jemal, A., Siegel, R., Ward, E., et al. (2009). Cancer statistics, 2009. *CA Cancer J Clin 59*, 225-249.

Jemal, A., Siegel, R., Xu, J., et al. (2010). Cancer statistics, 2010. *CA Cancer J Clin 60*, 277-300.

Jensen, J., Hyllner, J., and Bjorquist, P. (2009). Human embryonic stem cell technologies and drug discovery. *J Cell Physiol 219*, 513-519.

Johansson, H., Lindstedt, M., Albrekt, A. S., et al. (2011). A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. *BMC Genomics 12*, 399.

Johnson, F. M. (1999). Carcinogenic chemical-response "fingerprint" for male F344 rats exposed to a series of 195 chemicals: implications for predicting carcinogens with transgenic models. *Environ Mol Mutagen 34*, 234-245.

Johnson, F. M. (2001). Response to Tennant et al.: Attempts to replace the NTP rodent bioassay with transgenic alternatives are unlikely to succeed. *Environ Mol Mutagen 37*, 89-92.

Johnson, F. M. and Huff, J. (2002). Bioassay bashing is bad science. *Environ Health Perspect 110*, A736-737; author reply A737-739.

Johnson, F. M. (2003). How many high production chemicals are rodent carcinogens? Why should we care? What do we need to do about it? *Mutat Res 543*, 201-215.

Jowsey, I. R., Basketter, D. A., Westmoreland, C., et al. (2006). A future approach to measuring relative skin sensitising potency: a proposal. *J Appl Toxicol 26*, 341-350.

Judson, R., Richard, A., Dix, D. J., et al. (2009). The toxicity data landscape for environmental chemicals. *Environ Health Perspect 117*, 685-695.

Judson, R. (2010). Public databases supporting computational toxicology. *J Toxicol Environ Health B Crit Rev 13*, 218-231.

Judson, R. S., Martin, M. T., Reif, D. M., et al. (2010). Analysis of eight oil spill dispersants using rapid, in vitro tests for endocrine and other biological activity. *Environ Sci Technol 44*, 5979-5985.

Judson, R. S., Kavlock, R. J., Setzer, R. W., et al. (2011). Estimating toxicity-related biological pathway altering doses for high-throughput chemical risk assessment. *Chem Res Toxicol 24*, 451-462.

Julien, E., Willhite, C. C., Richard, A. M., et al. (2004). Challenges in constructing statistically based structure-activity relationship models for developmental toxicity. *Birth Defects Res A Clin Mol Teratol 70*, 902-911.

Kaderali, L., Dazert, E., Zeuge, U., et al. (2009). Reconstructing signaling pathways from RNAi data using probabilistic Boolean threshold networks. *Bioinformatics 25*, 2229-2235.

Kaplan, D. H., Kissenpfennig, A., and Clausen, B. E. (2008). Insights into Langerhans cell function from Langerhans cell ablation models. *Eur J Immunol 38*, 2369-2376.

Kaplan, D. H. (2010). In vivo function of Langerhans cells and dermal dendritic cells. *Trends Immunol 31*, 446-451.

Karleta, V., Andrlik, I., Braunmuller, S., et al. (2010). Poloxamer 188 supplemented culture medium increases the vitality of Caco-2 cells after subcultivation and freeze/thaw cycles. *ALTEX 27*, 191-197.

Kavlock, R. and Dix, D. (2010). Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. *J Toxicol Environ Health B Crit Rev 13*, 197-217.

Kedderis, G. L., Carfagna, M. A., Held, S. D., et al. (1993). Kinetic analysis of furan biotransformation by F-344 rats in vivo and in vitro. *Toxicol Appl Pharmacol 123*, 274-282.

Kedderis, G. L. and Held, S. D. (1996). Prediction of furan pharmacokinetics from hepatocyte studies: comparison of bioactivation and hepatic dosimetry in rats, mice, and humans. *Toxicol Appl Pharmacol 140*, 124-130.

Kedderis, G. L. (1997). Extrapolation of in vitro enzyme induction data to humans in vivo. *Chem Biol Interact 107*, 109-121.

Kim, J. H. and Scialli, A. R. (2011). Thalidomide: the tragedy of birth defects and the effective treatment of disease. *Toxicol Sci 122*, 1-6.

Kimber, I. and Dearman, R. J. (1991). Investigation of lymph node cell proliferation as a possible immunological correlate of contact sensitizing potential. *Food Chem Toxicol 29*, 125-129.

Kimber, I. and Basketter, D. A. (1997). Contact sensitization: a new approach to risk assessment. *Hum Ecol Risk Assess 3*, 385-395.

Kimber, I., Gerberick, G. F., and Basketter, D. A. (1999a). Thresholds in contact sensitization: theoretical and practical considerations. *Food Chem Toxicol 37*, 553-560.

Kimber, I., Pichowski, J. S., Basketter, D. A., et al. (1999b). Immune responses to contact allergens: novel approaches to hazard evaluation. *Toxicol Lett 106*, 237-246.

Kimber, I., Pichowski, J. S., Betts, C. J., et al. (2001). Alternative approaches to the identification and characterization of chemical allergens. *Toxicol In Vitro 15*, 307-312.

Kimber, I., Dearman, R. J., Basketter, D. A., et al. (2002). The local lymph node assay: past, present and future. *Contact Dermatitis 47*, 315-328.

Kimber, I., Cumberbatch, M., and Dearman, R. J. (2009). Langerhans cell migration: not necessarily always at the center of the skin sensitization universe. *J Invest Dermatol 129*, 1852-1853.

Kimber, I., Basketter, D. A., and Dearman, R. J. (2010). Chemical allergens – what are the issues? *Toxicology 268*, 139-142.

Kimber, I., Basketter, D. A., Gerberick, G. F., et al. (2011). Chemical allergy: translating biology into hazard characterization. *Toxicol Sci 120, Suppl 1,* S238-268.

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

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

Kirkland, D., Pfuhler, S., Tweats, D., et al. (2007). How to reduce false positive results when undertaking in vitro genotoxicity testing and thus avoid unnecessary follow-up animal tests: Report of an ECVAM Workshop. *Mutat Res 628*, 31-55.

Kirkwood, T. B. (2008). A systematic look at an old problem. *Nature 451*, 644-647.

Kirsch-Volders, M., Aardema, M., and Elhajouji, A. (2000). Concepts of threshold in mutagenesis and carcinogenesis. *Mutat Res 464*, 3-11.

Kleinstreuer, N. C., Smith, A. M., West, P. R., et al. (2011). Identifying developmental toxicity pathways for a subset of ToxCast chemicals using human embryonic stem cells and metabolomics. *Toxicol Appl Pharmacol 257*, 111-121.

Klemm, M. and Schrattenholz, A. (2004). Neurotoxicity of active compounds – establishment of hESC-lines and proteomics technologies for human embryo- and neurotoxicity screening and biomarker identification. *ALTEX 21, Suppl 3*, 41-48.

Klemm, M., Groebe, K., Soskic, V., et al. (2008). (Stem cell-based in vitro models as alternative methods for toxicity and efficacy tests in animals). *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 51*, 1033-1038.

Kluwe, W. M. (1982). Overview of phthalate ester pharmacokinetics in mammalian species. *Environ Health Perspect 45*, 3-9.

Knight, A., Bailey, J., and Balcombe, J. (2006a). Animal carcinogenicity studies: 2. Obstacles to extrapolation of data to humans. *Altern Lab Anim 34*, 29-38.

Knight, A., Bailey, J., and Balcombe, J. (2006b). Animal carcinogenicity studies: 1. Poor human predictivity. *Altern Lab Anim 34*, 19-27.

Knight, A., Bailey, J., and Balcombe, J. (2006c). Animal carcinogenicity studies: implications for the REACH system. *Altern Lab Anim 34, Suppl 1*, 139-147.

Knight, A. W., Little, S., Houck, K., et al. (2009). Evaluation of high-throughput genotoxicity assays used in profiling the US EPA ToxCast chemicals. *Regul Toxicol Pharmacol 55*, 188-199.

Knudsen, T. B., Martin, M. T., Kavlock, R. J., et al. (2009). Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB. *Reprod Toxicol 28*, 209-219.

Knudsen, T. B., Kavlock, R. J., Daston, G. P., et al. (2011). Developmental toxicity testing for safety assessment: new approaches and technologies. *Birth Defects Res B Dev Reprod Toxicol 92*, 413-420.

Kodell, R. L., Chen, J. J., Delongchamp, R. R., et al. (2006). Hierarchical models for probabilistic dose-response assessment. *Regul Toxicol Pharmacol 45*, 265-272.

Kola, I. and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov 3*, 711-715.

Krewski, D., Acosta, D., Jr., Andersen, M., et al. (2010). Toxicity testing in the 21st century: a vision and a strategy. *J Toxicol Environ Health B Crit Rev 13*, 51-138.

Kroes, R., Renwick, A. G., Cheeseman, M., et al. (2004). Structure-based thresholds of toxicological concern (TTC): guidance for application to substances present at low levels in the diet. *Food Chem Toxicol 42*, 65-83.

Kroes, R., Kleiner, J., and Renwick, A. (2005). The threshold of toxicological concern concept in risk assessment. *Toxicol Sci 86*, 226-230.

Kroes, R., Renwick, A. G., Feron, V., et al. (2007). Application of the threshold of toxicological concern (TTC) to the safety evaluation of cosmetic ingredients. *Food Chem Toxicol 45*, 2533-2562.

Krtolica, A. and Giritharan, G. (2010). Use of human embryonic stem cell-based models for male reproductive toxicity screening. *Syst Biol Reprod Med 56*, 213-221.

Kuegler, P. B., Zimmer, B., Waldmann, T., et al. (2010). Markers of murine embryonic and neural stem cells, neurons and astrocytes: reference points for developmental neurotoxicity testing. *ALTEX 27*, 17-42.

Kusama, M., Toshimoto, K., Maeda, K., et al. (2010). In silico classification of major clearance pathways of drugs with their physiochemical parameters. *Drug Metab Dispos 38*, 1362-1370.

Kusuhara, H. and Sugiyama, Y. (2009). In vitro-in vivo extrapolation of transporter-mediated clearance in the liver and kidney. *Drug Metab Pharmacokinet 24*, 37-52.

Lan, S. F. and Starly, B. (2011). Alginate based 3D hydrogels as an in vitro co-culture model platform for the toxicity screening of new chemical entities. *Toxicol Appl Pharmacol 256*, 62-72.

Langezaal, I., Coecke, S., and Hartung, T. (2001). Whole blood cytokine response as a measure of immunotoxicity. *Toxicol In Vitro 15*, 313-318.

Langezaal, I., Hoffmann, S., Hartung, T., et al. (2002). Evaluation and prevalidation of an immunotoxicity test based on human whole-blood cytokine release. *Altern Lab Anim 30*, 581-595.

Lankveld, D. P., Van Loveren, H., Baken, K. A., et al. (2010). In vitro testing for direct immunotoxicity: state of the art. *Methods Mol Biol 598*, 401-423.

Lass, C., Merfort, I., and Martin, S. F. (2010). In vitro and in vivo analysis of pro- and anti-inflammatory effects of weak and strong contact allergens. *Exp Dermatol 19*, 1007-1013.

Latta, M., Kunstle, G., Leist, M., et al. (2000). Metabolic depletion of ATP by fructose inversely controls CD95- and tumor necrosis factor receptor 1-mediated hepatic apoptosis. *J Exp Med 191*, 1975-1985.

Lau, C., Andersen, M. E., Crawford-Brown, D. J., et al. (2000). Evaluation of biologically based dose-response modeling for developmental toxicity: a workshop report. *Regul Toxicol Pharmacol 31*, 190-199.

Lave, L. B., Ennever, F. K., Rosenkranz, H. S., et al. (1988). Information value of the rodent bioassay. *Nature 336*, 631-633.

Leist, M., Bremer, S., Brundin, P., et al. (2008a). The biological and ethical basis of the use of human embryonic stem cells for in vitro test systems or cell therapy. *ALTEX 25*, 163-190.

Leist, M., Hartung, T., and Nicotera, P. (2008b). The dawning of a new age of toxicology. *ALTEX 25*, 103-114.

Leist, M., Efremova, L., and Karreman, C. (2010). Food for thought ... considerations and guidelines for basic test method descriptions in toxicology. *ALTEX 27*, 309-317.

Liao, K. H., Tan, Y. M., and Clewell, H. J., 3rd (2007). Development of a screening approach to interpret human biomonitoring data on volatile organic compounds: reverse dosimetry on biomonitoring data for trichloroethylene. *Risk Anal 27*, 1223-1236.

Liebsch, M., Grune, B., Seiler, A., et al. (2011). Alternatives to animal testing: current status and future perspectives. *Arch Toxicol 85*, 841-858.

Loeb, L. A. and Harris, C. C. (2008). Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Res 68*, 6863-6872.

Loizou, G. and Hogg, A. (2011). MEGen: A Physiologically Based Pharmacokinetic Model Generator. *Front Pharmacol 2*, 56.

Long, M. E. (2007). Predicting carcinogenicity in humans: The need to supplement animal-based toxicology. 6th World Congress on Alternatives & Animal Use in the Life Sciences, Tokyo. *AATEX 14, Spec. Issue*, 553-559.

Lorge, E. (2009). Genetic Toxicology Testing and its Relevance to Human Risk and Safety Evaluation. In B. Ballantyne, T. C. Marrs and T. Syversen (eds.), *General and Applied Toxicology*. Hoboken: Wiley.

Lotharius, J., Falsig, J., van Beek, J., et al. (2005). Progressive degeneration of human mesencephalic neuron-derived cells triggered by dopamine-dependent oxidative stress is dependent on the mixed-lineage kinase pathway. *J Neurosci 25*, 6329-6342.

Luch, A. (2005). Nature and nurture – lessons from chemical carcinogenesis. *Nat Rev Cancer 5*, 113-125.

Luijten, M., Verhoef, A., Westerman, A., et al. (2008). Application of a metabolizing system as an adjunct to the rat whole embryo culture. *Toxicol In Vitro 22*, 1332-1336.

Luijten, M., van Beelen, V. A., Verhoef, A., et al. (2010). Transcriptomics analysis of retinoic acid embryotoxicity in rat postimplantation whole embryo culture. *Reprod Toxicol 30*, 333-340.

Lund, S., Porzgen, P., Mortensen, A. L., et al. (2005). Inhibition of microglial inflammation by the MLK inhibitor CEP-1347. *J Neurochem 92,* 1439-1451.

Lund, S., Christensen, K. V., Hedtjarn, M., et al. (2006). The dynamics of the LPS triggered inflammatory response of murine microglia under different culture and in vivo conditions. *J Neuroimmunol 180*, 71-87.

Lutz, W. K. (2000). A true threshold dose in chemical carcinogenesis cannot be defined for a population, irrespective of the mode of action. *Hum Exp Toxicol 19*, 566-568; discussion 571-562.

MacDonald, J. S. (2004). Human carcinogenic risk evaluation, part IV: assessment of human risk of cancer from chemical exposure using a global weight-of-evidence approach. *Toxicol Sci 82*, 3-8.

MacKeigan, J. P., Murphy, L. O., and Blenis, J. (2005). Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. *Nat Cell Biol 7*, 591-600.

Magkoufopoulou, C., Claessen, S. M., Jennen, D. G., et al. (2011). Comparison of phenotypic and transcriptomic effects of false-positive genotoxins, true genotoxins and non-genotoxins using HepG2 cells. *Mutagenesis 26*, 593-604.

Magnusson, B. and Kligman, A. M. (1969). The identification of contact allergens by animal assay. The guinea pig maximization test. *J Invest Dermatol 52*, 268-276.

Maguire, T. J., Novik, E., Chao, P., et al. (2009). Design and application of microfluidic systems for in vitro pharmacokinetic evaluation of drug candidates. *Curr Drug Metab 10*, 1192-1199.

Makris, S. L., Solomon, H. M., Clark, R., et al. (2009). Terminology of developmental abnormalities in common laboratory mammals (version 2). *Reprod Toxicol 28*, 371-434.

Makris, S. L., Kim, J. H., Ellis, A., et al. (2011). Current and future needs for developmental toxicity testing. *Birth Defects Res B Dev Reprod Toxicol 92*, 384-394.

Martin, M. T., Judson, R. S., Reif, D. M., et al. (2009a). Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database. *Environ Health Perspect 117*, 392-399.

Martin, M. T., Mendez, E., Corum, D. G., et al. (2009b). Profiling the reproductive toxicity of chemicals from multigeneration studies in the toxicity reference database. *Toxicol Sci 110*, 181-190.

Martin, S. F., Dudda, J. C., Bachtanian, E., et al. (2008). Toll-like receptor and IL-12 signaling control susceptibility to contact hypersensitivity. *J Exp Med 205*, 2151-2162.

Martin, S. F. and Jakob, T. (2008). From innate to adaptive immune responses in contact hypersensitivity. *Curr Opin Allergy Clin Immunol 8*, 289-293.

Martin, S. F., Esser, P. R., Schmucker, S., et al. (2010). T-cell recognition of chemicals, protein allergens and drugs: towards the development of in vitro assays. *Cell Mol Life Sci 67*, 4171-4184.

Martin, S. F., Esser, P. R., Weber, F. C., et al. (2011). Mechanisms of chemical-induced innate immunity in allergic contact dermatitis. *Allergy 66*, 1152-1163.

Mascolo, M. G., Perdichizzi, S., Rotondo, F., et al. (2010). BALB/c 3T3 cell transformation assay for the prediction of carcinogenic potential of chemicals and environmental mixtures. *Toxicol In Vitro 24*, 1292-1300.

Matthews, E. J., Kruhlak, N. L., Cimino, M. C., et al. (2006). An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: II. Identification of genotoxicants, reprotoxicants, and carcinogens using in silico methods. *Regul Toxicol Pharmacol 44*, 97-110.

Matthews, E. J., Kruhlak, N. L., Daniel Benz, R., et al. (2007). A comprehensive model for reproductive and developmental toxicity hazard identification: II. Construction of QSAR models to predict activities of untested chemicals. *Regul Toxicol Pharmacol 47*, 136-155.

Mattison, D. R. (2010). Environmental exposures and development. *Curr Opin Pediatr 22*, 208-218.

Maxwell, G. and Mackay, C. (2008). Application of a systems biology approach to skin allergy risk assessment. *Altern Lab Anim 36*, 521-556.

Maxwell, G., Aeby, P., Ashikaga, T., et al. (2011). Skin sensitisation: the Colipa strategy for developing and evaluating nonanimal test methods for risk assessment. *ALTEX 28*, 50-55.

Mazur, C. S., Kenneke, J. F., Hess-Wilson, J. K., et al. (2010). Differences between human and rat intestinal and hepatic bisphenol A glucuronidation and the influence of alamethicin on in vitro kinetic measurements. *Drug Metab Dispos 38*, 2232-2238.

Mazzotti, F., Sabbioni, E., Ponti, J., et al. (2002). In vitro setting of dose-effect relationships of 32 metal compounds in the Balb/3T3 cell line, as a basis for predicting their carcinogenic potential. *Altern Lab Anim 30*, 209-217.

McKim, J. M., Jr. (2010). Building a tiered approach to in vitro predictive toxicity screening: a focus on assays with in vivo relevance. *Comb Chem High Throughput Screen 13*, 188-206.

McKim, J. M., Jr., Keller, D. J., 3rd, and Gorski, J. R. (2010). A new in vitro method for identifying chemical sensitizers combining peptide binding with ARE/EpRE-mediated gene expression in human skin cells. *Cutan Ocul Toxicol 29*, 171-192.

McNeish, J. (2004). Embryonic stem cells in drug discovery. *Nat Rev Drug Discov 3*, 70-80.

McNeish, J. D. (2007). Stem cells as screening tools in drug discovery. *Curr Opin Pharmacol 7*, 515-520.

Mendel, C. M. (1992). The free hormone hypothesis. Distinction from the free hormone transport hypothesis. *J Androl 13*, 107-116.

Meng, Q. (2010). Three-dimensional culture of hepatocytes for prediction of drug-induced hepatotoxicity. *Expert Opin Drug Metab Toxicol 6*, 733-746.

Miranda, J. P., Leite, S. B., Muller-Vieira, U., et al. (2009). Towards an extended functional hepatocyte in vitro culture. *Tissue Eng Part C Methods 15*, 157-167.

Moffat, J. and Sabatini, D. M. (2006). Building mammalian signalling pathways with RNAi screens. *Nat Rev Mol Cell Biol 7*, 177-187.

Morelli, M. A. (2000). Industry viewpoint on thresholds for genotoxic carcinogens. *Toxicol Pathol 28*, 396-404.

Mose, T. and Knudsen, L. E. (2006). Placental perfusion – a human alternative. *ALTEX 23, Suppl*, 358-363.

Munro, I. C., Renwick, A. G., and Danielewska-Nikiel, B. (2008). The Threshold of Toxicological Concern (TTC) in risk assessment. *Toxicol Lett 180*, 151-156.

Myren, M., Mose, T., Mathiesen, L., et al. (2007). The human placenta – an alternative for studying foetal exposure. *Toxicol In Vitro 21*, 1332-1340.

Nakamura, K., Mizutani, R., Sanbe, A., et al. (2011). Evaluation of drug toxicity with hepatocytes cultured in a micro-space cell culture system. *J Biosci Bioeng 111*, 78-84.

Nassim, N. T. (2010). The *Black Swan: The Impact of the Highly Improbable*. Vol. 2. New York, USA: Random House Trade Paperbacks.

Natsch, A., Emter, R., and Ellis, G. (2009). Filling the concept with data: integrating data from different in vitro and in silico assays on skin sensitizers to explore the battery approach for animal-free skin sensitization testing. *Toxicol Sci 107*, 106-121.

Neumann, H. G. (2009). Risk assessment of chemical carcinogens and thresholds. *Crit Rev Toxicol 39*, 449-461.

Noordegraaf, M., Flacher, V., Stoitzner, P., et al. (2010). Functional redundancy of Langerhans cells and Langerin+ dermal dendritic cells in contact hypersensitivity. *J Invest Dermatol 130*, 2752-2759.

NRC (2007). *Toxicity Testing in the Twenty-first Century: A Vision and a Strategy*. Washington DC, USA: National Academies Press. http://www.nap.edu/catalog.php?record_id=11970

O'Brien, J., Renwick, A. G., Constable, A., et al. (2006). Approaches to the risk assessment of genotoxic carcinogens in food: a critical appraisal. *Food Chem Toxicol 44*, 1613-1635.

OECD (1987). Acute Oral Toxicity. OECD Guideline for Testing of Chemical, TG No. 401. http://iccvam.niehs.nih.gov/docs/acutetox_docs/udpProc/udpfin01/append/AppI.pdf

OECD (2001). Prenatal Developmental Toxicity Study. OECD Guidelines for Chemical Testing, TG No. 414. http://www.oecd.org/dataoecd/18/15/1948482.pdf

OECD (2004). Skin Absorption: in vitro Method. OECD Guidelines for Chemical Testing, TG No. 428. http://iccvam.niehs.nih.gov/SuppDocs/FedDocs/OECD/OECDtg428.pdf

OECD (2007). Detailed review paper on cell transformation assays for detection of chemical carcinogens. Series on Testing and Assessment No. 31. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO%282007%2918&docLanguage=En

OECD (2009). Carcinogenicity Studies. OECD Guidelines for Chemical Testing, TG No. 451. http://www.oecd.org/dataoecd/30/46/41753121.pdf

OECD (2011). Extended one-generation reproductive toxicity study .OECD Guidelines for Chemical Testing, TG No. 443.

http://www.oecd-ilibrary.org → Search: "Extended one-generation reproductive toxicity study"

Ohshima, H., Tatemichi, M., and Sawa, T. (2003). Chemical basis of inflammation-induced carcinogenesis. *Arch Biochem Biophys 417*, 3-11.

Oliveira, P. A., Colaco, A., Chaves, R., et al. (2007). Chemical carcinogenesis. *An Acad Bras Cienc 79*, 593-616.

Olson, H., Betton, G., Robinson, D., et al. (2000). Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol 32*, 56-67.

Ouwehand, K., Spiekstra, S. W., Reinders, J., et al. (2010). Comparison of a novel CXCL12/CCL5 dependent migration assay with CXCL8 secretion and CD86 expression for distinguishing sensitizers from non-sensitizers using MUTZ-3 Langerhans cells. *Toxicol In Vitro 24*, 578-585.

Paine, M. F., Khalighi, M., Fisher, J. M., et al. (1997). Characterization of interintestinal and intraintestinal variations in human CYP3A-dependent metabolism. *J Pharmacol Exp Ther 283*, 1552-1562.

Pal, R., Mamidi, M. K., Das, A. K., et al. (2011). Human embryonic stem cell proliferation and differentiation as parameters to evaluate developmental toxicity. *J Cell Physiol 226*, 1583-1595.

Paterson, S. and Mackay, D. (1989). Correlation of tissue, blood, and air partition coefficients of volatile organic chemicals. *Br J Ind Med 46*, 321-328.

Patlewicz, G., Rodford, R., and Walker, J. D. (2003). Quantitative structure-activity relationships for predicting mutagenicity and carcinogenicity. *Environ Toxicol Chem 22*, 1885-1893.

Patlewicz, G., Aptula, A. O., Uriarte, E., et al. (2007). An evaluation of selected global (Q)SARs/expert systems for the prediction of skin sensitisation potential. *SAR QSAR Environ Res 18*, 515-541.

Pelkonen, O. and Turpeinen, M. (2007). In vitro-in vivo extrapolation of hepatic clearance: biological tools, scaling factors, model assumptions and correct concentrations. *Xenobiotica 37*, 1066-1089.

Pelkonen, O., Kapitulnik, J., Gundert-Remy, U., et al. (2008). Local kinetics and dynamics of xenobiotics. *Crit Rev Toxicol 38*, 697-720.

Pelkonen, O., Tolonen, A., Korjamo, T., et al. (2009). From known knowns to known unknowns: predicting in vivo drug metabolites. *Bioanalysis 1*, 393-414.

Pellizzer, C., Adler, S., Corvi, R., et al. (2004). Monitoring of teratogenic effects in vitro by analysing a selected gene expression pattern. *Toxicol In Vitro 18*, 325-335.

Pellizzer, C., Bremer, S., and Hartung, T. (2005). Developmental toxicity testing from animal towards embryonic stem cells. *ALTEX 22*, 47-57.

Pepe, M. S. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford Univ Pr.

Pessina, A., Albella, B., Bueren, J., et al. (2001). Prevalidation of a model for predicting acute neutropenia by colony forming unit granulocyte/macrophage (CFU-GM) assay. *Toxicol In Vitro 15*, 729-740.

Peyret, T. and Krishnan, K. (2011). QSARs for PBPK modelling of environmental contaminants. *SAR QSAR Environ Res 22*, 129-169.

Pfuhler, S., Kirkland, D., Kasper, P., et al. (2009). Reduction of use of animals in regulatory genotoxicity testing: Identification and implementation opportunities-Report from an ECVAM workshop. *Mutat Res 680*, 31-42.

Pfuhler, S., Carmichael, P., Fowler, P., et al. (2010a). Animal-free genotoxicity testing: The COLIPA program. *Toxicology Letters* S248.

Pfuhler, S., Kirst, A., Aardema, M., et al. (2010b). A tiered approach to the use of alternatives to animal testing for the safety assessment of cosmetics: genotoxicity. A COLIPA analysis. *Regul Toxicol Pharmacol 57*, 315-324.

Piersma, A. H., Genschow, E., Verhoef, A., et al. (2004). Validation of the postimplantation rat whole-embryo culture test in the international ECVAM validation study on three in vitro embryotoxicity tests. *Altern Lab Anim 32*, 275-307.

Piersma, A. H., Janer, G., Wolterink, G., et al. (2008). Quantitative extrapolation of in vitro whole embryo culture embryotoxicity data to developmental toxicity in vivo using the benchmark dose approach. *Toxicol Sci 101*, 91-100.

Piersma, A. H., Hernandez, L. G., van Benthem, J., et al. (2011). Reproductive toxicants have a threshold of adversity. *Crit Rev Toxicol 41*, 545-554.

Plant, N. (2008). Can systems toxicology identify common biomarkers of non-genotoxic carcinogenesis? *Toxicology 254*, 164-169.

Ponti, J., Munaro, B., Fischbach, M., et al. (2007). An optimised data analysis for the Balb/c 3T3 cell transformation assay and its application to metal compounds. *Int J Immunopathol Pharmacol 20*, 673-684.

Poulin, P. and Krishnan, K. (1995). An algorithm for predicting tissue: blood partition coefficients of organic chemicals from n-octanol: water partition coefficient data. *J Toxicol Environ Health 46*, 117-129.

Poulsen, M. S., Rytting, E., Mose, T., et al. (2009). Modeling placental transport: correlation of in vitro BeWo cell permeability and ex vivo human placental perfusion. *Toxicol In Vitro 23*, 1380-1386.

Pratt, I., Barlow, S., Kleiner, J., et al. (2009). The influence of thresholds on the risk assessment of carcinogens in food. *Mutat Res 678*, 113-117.

Pritchard, J. B., French, J. E., Davis, B. J., et al. (2003). The role of transgenic mouse models in carcinogen identification. *Environ Health Perspect 111*, 444-454.

Prusakiewicz, J. J., Ackermann, C., and Voorman, R. (2006). Comparison of skin esterase activities from different species. *Pharm Res 23*, 1517-1524.

Punt, A., Freidig, A. P., Delatour, T., et al. (2008). A physiologically based biokinetic (PBBK) model for estragole bioactivation and detoxification in rat. *Toxicol Appl Pharmacol 231*, 248-259.

Punt, A., Paini, A., Boersma, M. G., et al. (2009). Use of physiologically based biokinetic (PBBK) modeling to study estragole bioactivation and detoxification in humans as compared with male rats. *Toxicol Sci 110*, 255-269.

Python, F., Goebel, C., and Aeby, P. (2007). Assessment of the U937 cell line for the detection of contact allergens. *Toxicol Appl Pharmacol 220*, 113-124.

Rall, D. P. (2000). Laboratory animal tests and human cancer. *Drug Metab Rev 32*, 119-128.

Raunio, H. (2011). In silico toxicology – non-testing methods. *Front Pharmacol 2*, 33.

Reuben, S. (2010). Reducing Environmental Cancer Risk: What We Can Do Now: 2008–2009 Annual Report, President's Cancer Panel. National Cancer Institute. *NIH*, 1-240. http://deainfo.nci.nih.gov/advisory/pcp/annualReports/pcp08-09rpt/PCP_Report_08-09_508.pdf

Reuter, H., Spieker, J., Gerlach, S., et al. (2011). In vitro detection of contact allergens: development of an optimized protocol using human peripheral blood monocyte-derived dendritic cells. *Toxicol In Vitro 25*, 315-323.

Rhomberg, L. E. (2011). Interpreting dose-response information on intermediate stages of causal cascades in toxicity mode of action. Society of Toxicology 50th Annual Meeting, Washington, DC.

Richert, L., Abadie, C., Bonet, A., et al. (2010). Inter-laboratory evaluation of the response of primary human hepatocyte cultures to model CYP inducers – a European Centre for Validation of Alternative Methods (ECVAM) – funded pre-validation study. *Toxicol In Vitro 24*, 335-345.

Rostami-Hodjegan, A. and Tucker, G. T. (2007). Simulation and prediction of in vivo drug metabolism in human populations from in vitro data. *Nat Rev Drug Discov 6*, 140-148.

Rotroff, D. M., Wetmore, B. A., Dix, D. J., et al. (2010). Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. *Toxicol Sci 117*, 348-358.

Rovida, C. and Hartung, T. (2009). Re-evaluation of animal numbers and costs for in vivo tests to accomplish REACH legislation requirements for chemicals – a report by the transatlantic think tank for toxicology (t^4). *ALTEX 26*, 187-208.

Rovida, C. (2010). Food for thought ... why no new in vitro tests will be done for REACH by registrants. *ALTEX 27*, 175-183.

Rovida, C. (2011). Local lymph node assay: how testing laboratories apply OECD TG 429 for REACH purposes. *ALTEX 28*, 117-129.

Rovida, C., Longo, F., and Rabbit, R. R. (2011). How are reproductive toxicity and developmental toxicity addressed in REACH dossiers? *ALTEX 28*, 273-294.

Rule, A. D., Gussak, H. M., Pond, G. R., et al. (2004). Measured and estimated GFR in healthy potential kidney donors. *Am J Kidney Dis 43*, 112-119.

Ryan, C. A., Gerberick, G. F., Gildea, L. A., et al. (2005). Interactions of contact allergens with dendritic cells: opportunities and challenges for the development of novel approaches to hazard assessment. *Toxicol Sci 88*, 4-11.

Sakaguchi, H., Ashikaga, T., Miyazawa, M., et al. (2006). Development of an in vitro skin sensitization test using human cell lines; human Cell Line Activation Test (h-CLAT). II. An inter-laboratory study of the h-CLAT. *Toxicol In Vitro 20*, 774-784.

Sakai, A., Sasaki, K., Hayashi, K., et al. (2011). An international validation study of a Bhas 42 cell transformation assay for the prediction of chemical carcinogenicity. *Mutat Res 725*, 57-77.

Salsburg, D. (1983). The lifetime feeding study in mice and rats – an examination of its validity as a bioassay for human carcinogens. *Fundam Appl Toxicol 3*, 63-67.

Sambuy, Y., De Angelis, I., Ranaldi, G., et al. (2005). The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on Caco-2 cell functional characteristics. *Cell Biol Toxicol 21*, 1-26.

Schaafsma, G., Kroese, E. D., Tielemans, E. L., et al. (2009). REACH, non-testing approaches and the urgent need for a change in mind set. *Regul Toxicol Pharmacol 53*, 70-80.

Schardein, J. (2000). *Chemically Induced Birth Defects*. Vol. 3. London, UK: Informa Healthcare.

Scharf, J., Ramadori, G., Braulke, T., et al. (1996). Synthesis of insulinlike growth factor binding proteins and of the acid-labile subunit in primary cultures of rat hepatocytes, of Kupffer cells, and in cocultures: regulation by insulin, insulinlike growth factor, and growth hormone. *Hepatology 23*, 818-827.

Schmidt, C. W. (2002). Assessing assays. *Environ Health Perspect 110*, A248-251.

Schmidt, M., Raghavan, B., Muller, V., et al. (2010). Crucial role for human Toll-like receptor 4 in the development of contact allergy to nickel. *Nat Immunol 11*, 814-819.

Schneider, K., Schwarz, M., Burkholder, I., et al. (2009). "ToxR-Tool", a new tool to assess the reliability of toxicological data. *Toxicol Lett 189*, 138-144.

Scholz, D., Poltl, D., Genewsky, A., et al. (2011). Rapid, complete and large-scale generation of post-mitotic neurons from the human LUHMES cell line. *J Neurochem 119*, 957-971.

Scialli, A. R. (2008). The challenge of reproductive and developmental toxicology under REACH. *Regul Toxicol Pharmacol 51*, 244-250.

Seagle, C., Christie, M. A., Winnike, J. H., et al. (2008). High-throughput nuclear magnetic resonance metabolomic footprinting for tissue engineering. *Tissue Eng Part C Methods 14,* 107-118.

Seibert, H., Morchel, S., and Gulden, M. (2002). Factors influencing nominal effective concentrations of chemical compounds in vitro: medium protein concentration. *Toxicol In Vitro 16*, 289-297.

Seidle, T. (2006). Chemicals and Cancer: What the Regulators Won't Tell You. PETA Europe Ltd. http://www.peta.nl/pdf/PetaCancerReport.pdf

Seidle, T., Prieto, P., and Bulgheroni, A. (2011). Examining the regulatory value of multi-route mammalian acute systemic toxicity studies. *ALTEX 28*, 95-102.

Seiler, A., Visan, A., Buesen, R., et al. (2004). Improvement of an in vitro stem cell assay for developmental toxicity: the use of molecular endpoints in the embryonic stem cell test. *Reprod Toxicol 18*, 231-240.

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

Selderslaghs, I. W., Blust, R., and Witters, H. E. (2011). Feasibility study of the zebrafish assay as an alternative method to screen for developmental toxicity and embryotoxicity using a training set of 27 compounds. *Reprod Toxicol*

Shanks, N., Greek, R., and Greek, J. (2009). Are animal models predictive for humans? *Philos Ethics Humanit Med 4*, 2.

Silbergeld, E. K. (2004). Commentary: the role of toxicology in prevention and precaution. *Int J Occup Med Environ Health 17*, 91-102.

Silva Lima, B. and Van der Laan, J. W. (2000). Mechanisms of nongenotoxic carcinogenesis and assessment of the human hazard. *Regul Toxicol Pharmacol 32*, 135-143.

Sipes, N. S., Martin, M. T., Reif, D. M., et al. (2011a). Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data. *Toxicol Sci 124*, 109-127.

Sipes, N. S., Padilla, S., and Knudsen, T. B. (2011b). Zebrafish: as an integrative model for twenty-first century toxicity testing. *Birth Defects Res C Embryo Today 93*, 256-267.

Skelin, M., Rupnik, M., and Cencic, A. (2010). Pancreatic beta cell lines and their applications in diabetes mellitus research. *ALTEX 27*, 105-113.

Snyder, R. D. and Green, J. W. (2001). A review of the genotoxicity of marketed pharmaceuticals. *Mutat Res 488*, 151-169.

Sobels, F. H. (1987). Environmental mutagenesis in retrospect. *Mutat Res 181*, 299-310.

Speit, G. (2009). How to assess the mutagenic potential of cosmetic products without animal tests? *Mutat Res 678*, 108-112.

Spielmann, H. and Gerbracht, U. (2001). The use of dogs as second species in regulatory testing of pesticides. Part II: Subacute, subchronic and chronic studies in the dog. *Arch Toxicol 75*, 1-21.

Spielmann, H., Genschow, E., Brown, N. A., et al. (2004). Validation of the rat limb bud micromass test in the international ECVAM validation study on three in vitro embryotoxicity tests. *Altern Lab Anim 32*, 245-274.

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

Spielmann, H. (2010). The EU Commission's Draft Report on Alternative (Non-animal) Methods for Cosmetics Testing: Current Status and Future Prospects – 2010: a missed opportunity. *Altern Lab Anim 38*, 339-343.

Stokes, W. S. and Wind, M. (2010). Validation of innovative technologies and strategies for regulatory safety assessment methods: challenges and opportunities. *ALTEX 27*, 87-95.

Storm, J. E., Collier, S. W., Stewart, R. F., et al. (1990). Metabolism of xenobiotics during percutaneous penetration: role of absorption rate and cutaneous enzyme activity. *Fundam Appl Toxicol 15*, 132-141.

Sukardi, H., Chng, H. T., Chan, E. C., et al. (2011). Zebrafish for drug toxicity screening: bridging the in vitro cell-based models and in vivo mammalian models. *Expert Opin Drug Metab Toxicol 7*, 579-589.

Sung, J. H., Esch, M. B., and Shuler, M. L. (2010). Integration of in silico and in vitro platforms for pharmacokinetic-pharmacodynamic modeling. *Expert Opin Drug Metab Toxicol 6*, 1063-1081.

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell 126*, 663-676.

Takahashi, K., Tanabe, K., Ohnuki, M., et al. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell 131*, 861-872.

Takayama, S., Thorgeirsson, U. P., and Adamson, R. H. (2008). Chemical carcinogenesis studies in nonhuman primates. *Proc Jpn Acad Ser B Phys Biol Sci 84,* 176-188.

Taylor, K., Casalegno, C., and Stengel, W. (2011). A critique of the EC's expert (draft) reports on the status of alternatives for cosmetics testing to meet the 2013 deadline. *ALTEX 28*, 131-148.

Teeguarden, J. G. and Barton, H. A. (2004). Computational modeling of serum-binding proteins and clearance in extrapolations across life stages and species for endocrine active compounds. *Risk Anal 24*, 751-770.

Thilly, W. G. (2003). Have environmental mutagens caused oncomutations in people? *Nat Genet 34*, 255-259.

Toh, Y. C., Lim, T. C., Tai, D., et al. (2009). A microfluidic 3D hepatocyte chip for drug toxicity testing. *Lab Chip 9*, 2026-2035.

Toivonen, H., Srinivasan, A., King, R. D., et al. (2003). Statistical evaluation of the Predictive Toxicology Challenge 2000-2001. *Bioinformatics 19*, 1183-1193.

Toth, B. (2002). Species susceptibilities to chemical carcinogens: a critical appraisal of the roles of sex hormones (endocrine status) and nutritional influences. *In Vivo 16*, 161-166.

Troutman, J. A., Foertsch, L. M., Kern, P. S., et al. (2011). The incorporation of lysine into the peroxidase peptide reactivity assay for skin sensitization assessments. *Toxicol Sci 122*, 422-436.

Tukov, F. F., Maddox, J. F., Amacher, D. E., et al. (2006). Modeling inflammation-drug interactions in vitro: a rat Kupffer cell-hepatocyte coculture system. *Toxicol In Vitro 20*, 1488-1499.

Tweats, D. J., Scott, A. D., Westmoreland, C., et al. (2007). Determination of genetic toxicity and potential carcinogenicity in vitro – challenges post the Seventh Amendment to the European Cosmetics Directive. *Mutagenesis 22*, 5-13.

Uibel, F., Muhleisen, A., Kohle, C., et al. (2010). ReProGlo: a new stem cell-based reporter assay aimed to predict embryotoxic potential of drugs and chemicals. *Reprod Toxicol 30*, 103-112.

van Dartel, D. A. and Piersma, A. H. (2011). The embryonic stem cell test combined with toxicogenomics as an alternative testing model for the assessment of developmental toxicity. *Reprod Toxicol 32*, 235-244.

van de Kerkhof, E. G., de Graaf, I. A., and Groothuis, G. M. (2007). In vitro methods to study intestinal drug metabolism. *Curr Drug Metab 8*, 658-675.

van der Jagt, K., Munn, S., Torslov, J., et al. (2004). Alternative approaches can reduce the use of test animals under REACH. Addendum to the report "Assessment of additional testing needs under REACH. Effects of (Q)SARs, risk based testing and voluntary industry initiatives". European Commission. *Joint Research Centre*. 1-31. http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/8790/1/EUR%2021405%20EN.pdf

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

van Kesteren, P. C., Zwart, P. E., Pennings, J. L., et al. (2011). Deregulation of cancer-related pathways in primary hepatocytes derived from DNA repair-deficient Xpa-/-p53+/- mice upon exposure to benzo[a]pyrene. Toxicol Sci 123, 123-132.

van Leeuwen, C. J., Patlewicz, G. Y., and Worth, A. P. (2007). Intelligent Testing Strategies, in Risk Assessment of Chemicals. In C. J. van Leeuwen and T. G. Vermeire (eds.), *Risk Assessment of Chemicals. An Introduction*. Heidelberg, Germany: Springer.

van Leeuwen, I. M. and Zonneveld, C. (2001). From exposure to effect: a comparison of modeling approaches to chemical carcinogenesis. *Mutat Res 489*, 17-45.

Van Oosterhout, J. P., Van der Laan, J. W., De Waal, E. J., et al. (1997). The utility of two rodent species in carcinogenic risk assessment of pharmaceuticals in Europe. *Regul Toxicol Pharmacol 25*, 6-17.

van Ravenzwaay, B. (2010). Initiatives to decrease redundancy in animal testing of pesticides. *ALTEX 27*, 159-161.

van Ravenzwaay, B., Dammann, M., Buesen, R., et al. (2011). The threshold of toxicological concern for prenatal developmental toxicity. *Regul Toxicol Pharmacol 59*, 81-90.

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

Vandebriel, R. J. and van Loveren, H. (2010). Non-animal sensitization testing: state-of-the-art. *Crit Rev Toxicol 40*, 389-404.

Vanhaecke, T., Pauwels, M., Vinken, M., et al. (2011). Towards an integrated in vitro strategy for repeated dose toxicity testing. *Arch Toxicol 85*, 365-366.

Vanparys, P., Corvi, R., Aardema, M., et al. (2011). ECVAM prevalidation of three cell transformation assays. *ALTEX 28*, 56-59.

Verdonck, F. A., Van Sprang, P. A., and Vanrolleghem, P. A. (2005). Uncertainty and precaution in European environmental risk assessment of chemicals. *Water Sci Technol 52*, 227-234.

Verwei, M., van Burgsteden, J. A., Krul, C. A., et al. (2006). Prediction of in vivo embryotoxic effect levels with a combination of in vitro studies and PBPK modelling. *Toxicol Lett 165*, 79-87.

Vinken, M., Doktorova, T., Ellinger-Ziegelbauer, H., et al. (2008). The carcinoGENOMICS project: critical selection of model compounds for the development of omics-based in vitro carcinogenicity screening assays. *Mutat Res 659*, 202-210.

Virshup, D. M. and Shenolikar, S. (2009). From promiscuity to precision: protein phosphatases get a makeover. *Mol Cell 33*, 537-545.

Viviani, B. (2006). Preparation and coculture of neurons and glial cells. *Curr Protoc Cell Biol Chapter 2*, Unit 2 7.

Volbracht, C., Leist, M., and Nicotera, P. (1999). ATP controls neuronal apoptosis triggered by microtubule breakdown or potassium deprivation. *Mol Med 5,* 477-489.

Volz, A., Piper, H. M., Siegmund, B., et al. (1991). Longevity of adult ventricular rat heart muscle cells in serum-free primary culture. *J Mol Cell Cardiol 23*, 161-173.

Ward, J. M. (2007). The two-year rodent carcinogenesis bioassay – Will it survive? *J Toxicol Pathol 1*, 13.

Waters, M. D., Jackson, M., and Lea, I. (2010). Characterizing and predicting carcinogenicity and mode of action using conventional and toxicogenomics methods. *Mutat Res 705*, 184-200.

Weber, F. C., Esser, P. R., Muller, T., et al. (2010). Lack of the purinergic receptor P2X(7) results in resistance to contact hypersensitivity. *J Exp Med 207*, 2609-2619.

Weigt, S., Huebler, N., Braunbeck, T., et al. (2010). Zebrafish teratogenicity test with metabolic activation (mDarT): effects of phase I activation of acetaminophen on zebrafish Danio rerio embryos. *Toxicology 275*, 36-49.

Weigt, S., Huebler, N., Strecker, R., et al. (2011). Zebrafish (Danio rerio) embryos as a model for testing proteratogens. *Toxicology 281*, 25-36.

West, P. R., Weir, A. M., Smith, A. M., et al. (2010). Predicting human developmental toxicity of pharmaceuticals using human embryonic stem cells and metabolomics. *Toxicol Appl Pharmacol 247*, 18-27.

Westmoreland, C., Carmichael, P., Dent, M., et al. (2010). Assuring safety without animal testing: Unilever's ongoing research programme to deliver novel ways to assure consumer safety. *ALTEX 27*, 61-65.

Wetmore, B. A., Wambaugh, J. F., Ferguson, S. S., et al. (2011). Integration of dosimetry, exposure and high-throughput screening data in chemical toxicity assessment. *Toxicol Sci 125*, 157-174.

Widschwendter, M. and Jones, P. A. (2002). DNA methylation and breast carcinogenesis. *Oncogene 21*, 5462-5482.

Wilcox, N. and Goldberg, A. (2011). Food for thought ... on validation. A puzzle or a mystery: an approach founded on new science. *ALTEX 28*, 3-8.

Wilkinson, G. R. and Schenker, S. (1975). Drug disposition and liver disease. *Drug Metab Rev 4*, 139-175.

Williams, G. M. (2001). Mechanisms of chemical carcinogenesis and application to human cancer risk assessment. *Toxicology 166*, 3-10.

Wise, L. D., Beck, S. L., Beltrame, D., et al. (1997). Terminology of developmental abnormalities in common laboratory mammals (version 1). *Teratology 55*, 249-292.

Wogan, G. N., Hecht, S. S., Felton, J. S., et al. (2004). Environmental and chemical carcinogenesis. *Semin Cancer Biol 14*, 473-486.

Woodcock, J. and Woosley, R. (2008). The FDA critical path initiative and its influence on new drug development. *Annu Rev Med 59*, 1-12.

Woosley, R. L. and Cossman, J. (2007). Drug development and the FDA's Critical Path Initiative. *Clin Pharmacol Ther 81*, 129-133.

Yang, C. H., Glover, K. P., and Han, X. (2010). Characterization of cellular uptake of perfluorooctanoate via organic anion-transporting polypeptide 1A2, organic anion transporter 4, and urate transporter 1 for their potential roles in mediating human renal reabsorption of perfluorocarboxylates. *Toxicol Sci 117*, 294-302.

Yang, J., Jamei, M., Yeo, K. R., et al. (2007). Prediction of intestinal first-pass drug metabolism. *Curr Drug Metab 8*, 676-684.

Yang, L., Ho, N. Y., Alshut, R., et al. (2009). Zebrafish embryos as models for embryotoxic and teratological effects of chemicals. *Reprod Toxicol 28*, 245-253.

Zimmer, B., Schildknecht, S., Kuegler, P. B., et al. (2011). Sensitivity of dopaminergic neuron differentiation from stem cells to chronic low-dose methylmercury exposure. *Toxicol Sci 121*, 357-367.

Acknowledgements

The workshop and the preparation of the five whitepapers were supported by CAAT-Europe (Daneshian et al., 2010) and the Doerenkamp-Zbinden Foundation. T. Hartung would like to thank Andre Kleensang and Robert Borotkanics for input on the text used in the introduction. H. Clewell would like to thank his colleagues at The Hamner – Miyoung Yoon, Mel Andersen, and Ed Lecluyse – for their invaluable contributions to the preparation of an initial version of this manuscript, as well as the American Chemistry Council for funding his research on *in vitro* to *in vivo* extrapolation.

Correspondence to

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