Evidence-Based Toxicology – the Toolbox of Validation for the 21st Century?

Thomas Hartung

Johns Hopkins University, Bloomberg School of Public Health, Dept. Environmental Health Sciences, Center for Alternatives to Animal Testing (CAAT), Doerenkamp-Zbinden Chair for Evidence-based Toxicology, Baltimore, MD, USA, and Professor of Pharmacology and Toxicology, University of Konstanz, Germany

Summary

Validation has become a primary driver of the evolution of toxicological methods. Agreement at OECD level currently requires validation of new approaches for consideration in test guideline development. Several examples of this exist. However, the toxicology in the 21st century movement, prompted by the 2007 NRC/NAS vision document, might lead to a revolutionary change in the toxicological toolbox. The challenge is whether the validation process, as it has been formalized over the last two decades, meets the needs for this paradigm shift.

The concept of evidence-based medicine (EBM) has emerged from clinical medicine, which retrospectively assesses the evidence of adequacy of a given approach. This is not typically done in prospective studies – the equivalent of validation studies might be multicenter randomized trials. Evidently, where such unambiguous evidence is available, no other assessment is necessary. EBM, however, has developed procedures, including meta-analysis, to collect and evaluate all the available evidence where no such definitive study is available. The recent successful introduction of retrospective validation, i.e. the collection and evaluation of existing evidence from various sources, represents a step in this direction. Here, we will explore new toxicological approaches via evidence-based toxicology (EBT).

Keywords: *alternative methods, evidence-based medicine, quality assurance, cell culture, genomics*

1 Introduction

Toxicity testing is at a critical crossroad, prompted three years ago by the publication of the NRC report on Toxicology in the 21st Century (Tox-21c) (NRC, 2007) and the discussion it triggered (Collins et al., 2008; Gibb, 2008; Andersen and Krewski, 2009; Kavlock, 2009; Kavlock et al., 2009). The report addressed the adequacy of the current tools with regard to throughput, costs, predictability of human responses, and animal use. Similar thoughts were followed in the 2004 FDA Critical Path Initiative (Coons, 2009). Obviously, the current methods for hazard assessment are largely the same for industrial chemicals (Hartung, 2010b), nanoparticles (Hartung, 2010c), pesticides, and drug candidates. The assessment of toxicological risk relies primarily on *in vivo* animal experiments that were designed decades ago and cost about \$ 3 billion/year worldwide (Bottini and Hartung, 2009). Their low throughput has led to a backlog of substances whose potential toxicity remains to be adequately assessed (Grandjean and Landrigan, 2006; Judson et al., 2009) and hinders front loading of toxicity testing in the drug development process. Species differences in toxicity responses require the use of uncertainty factors for human risk assessment (NRC, 2000). Moreover, to minimize the risk of not identifying a human hazard, animal studies are usually performed at doses or-

ders of magnitude higher than realistic human exposure (Hartung, 2009a). In the field of drug development about 92% of substances fail during clinical trials, about 20% of which are due to toxic effects in humans not identified in pre-clinical animal testing (FDA, 2004). Thus, there are serious concerns regarding the efficiency and relevance of current toxicity testing methods for human health effects.

2 Validation – a blessing or a curse for a paradigm shift in toxicology?

Most test methods in regulatory toxicology have been established by fixing acute demands with reasonably available tools identified by consensus processes. Though scientific truth is established on the basis of irrefutable evidence, majority of opinion governs regulatory toxicology. The process is driven by historic requirements, contemporary scientific knowledge, individuals who happen to be in charge, political decisions, etc. and is thus circumstantial rather than strategic. No objective criteria other than reproducibility, some reference toxicants, and cost/effort considerations are available, creating what might be termed a status of “face validity” or apparent validity. Actual relevance is typically established by “learning from experience.” Heed-

ing the warning of Petr Skrabanek and James McCormick that “*Learning from experience might be nothing more than making the same mistakes with increasing confidence*” (Skrabanek and McCormick, 1989), we might ask ourselves, which mechanisms there are to challenge the validity of these methods?

The situation is different when consensus has been established for particular tools. Such tools offer a point of reference to validate new approaches. Validation is defined (Balls et al., 1990; Hartung et al., 2004; OECD, 2005) as the independent assessment of a method for a defined purpose as to its reproducibility, scientific basis, and reliability/relevance. A culture of prospective ring trials has developed which is capable of re-assuring the one-to-one replacement of a method by a better one, e.g., one limiting or refining animal use.

Tremendous problems arise (Balls et al., 2006; Hartung, 2007a) where no reference method exists, the reference method is flawed, or the purpose or applicability of both methods is overlapping but not identical. Unfortunately, most areas of toxicology are a mixture of those problems. This is why very few replacement methods have been accepted, and, when they are, they often outperform the reference method and demonstrate its flaws. One obvious solution would be the joint comparison with an independent point of reference, such as human data, but this is rarely available and even more rarely done. The typical validation is done against the reference method only.

The concept of a reference method is often perverted to the concept of a “gold standard”: this is most appropriate for reference substances and the definition of units of measure (where the concept originates). However, scientific methods arise from a competition of ideas and the point of reference must evolve. The term “traditional methods” is more adequate. However, we should be clear that this implies that we cannot get better than this reference method. The only alternative is to anchor our validation with other reference points such as clinical data, composite knowledge on certain substances, or understood mechanisms. We have suggested earlier (Hartung, 2007a; Hofmann et al., 2008; Ahr et al., 2008) that such points of reference are chosen, but this will require expert judgment and consensus processes (for example see Vinken et al., 2008).

The concept of formal validation and its expanding definition in detail has taught a lot about method evaluation: There is no doubt that quality assurance of methods must be based on method definition (including purpose and applicability domain) and reproducibility. This is actually also the easy part; it gets difficult when scientific basis and relevance are addressed. Traditional validation has not really embraced scientific relevance and the relevance assessments suffer from the limitations of their point of comparison.

Furthermore, formal validation has become a lengthy and costly exercise. Idealized time estimates are three years for validation plus two for peer review and costs of \$ 500,000 involving 3 or more laboratories. For a new battery of tests or a major revision of the toxicological toolbox the feasibility is quickly questionable. Similar to the problem of mixtures in toxicology, the problem of mixing methods in integrated testing strategies quickly exposes its limits, as far too many combinations are possible. So either we limit ourselves to the rigidity of fixed

combinations or we need to find faster and more flexible approaches. The feasibility of a validation study must consider the required study size and duration relative to the speed of technological progress.

3 The needs of Tox-21c in validation

Tox-21c proposed a paradigm shift in toxicology. Instead of relying on traditional animal experiments, the report proposes the application of the latest advances in science and technology to develop more relevant test strategies. It has some similarities with alternative methods movement, but it is less animal welfare-driven and prompted more by limitations in throughput and predictivity (Hartung, 2008b; Hartung, 2010a). It also does not rely on a one-by-one replacement of methods with better (3Rs) ones. Instead, pathways of toxicity (PoT) shall be identified using *in vitro* cell systems (preferably human), high throughput testing, ‘omics’ approaches, systems biology, and computational modeling. A new battery of tests covering PoT will then form the basis for a new approach, which promises a truly revolutionary change (Hartung, 2008a). The so called “pathways of toxicity” are defined as changes in normal biological processes, e.g. cell function, communication, and adaptation to environmental changes which are expected to result in adverse health effects (NRC, 2007). Toxins can affect multiple PoT in different cells systems, leading to various adverse effects (cell phenotypes) (Fig. 1). One familiar PoT described in the Tox-21c report is the signaling by estrogens in which initial exposure results in en-

Fig. 1: Scheme of how toxins affect various pathways of toxicity (PoT) in diverse cell types leading to different or similar phenotypes.

hanced cell proliferation. The identification of these pathways would likely provide more accurate information for chemical or drug risk to humans than current animal tests (Gibb, 2008). Although a broad discussion has ensued on the design and feasibility of this new concept for toxicology, in practice we are still at the very beginning of the paradigm shift. This relates to the development of new technologies as well as to their quality assurance, acceptance, and integration (Hartung, 2009c). Tox-21c emphasizes the concept of PoT, i.e. a mechanistic foundation and a reconstruction of safety assessments acknowledging the limitations of traditional approaches. It is obvious that this will require a role of the assessment of the scientific basis and relevance different to traditional validation. The question to be addressed here is, whether EBT lends itself to this problem.

Tox-21c is envisaged as a toxicity testing approach, using primarily cellular and *in silico* methods as well as lower organisms and:

- Is based on identified and annotated PoT, probably requiring a type of Human Toxicology Project (Seidle and Stephens, 2009) for their comprehensive mapping
- Is based on information-rich technologies (Hartung and Leist, 2008; Leist et al., 2008b) such as omics, image analysis and high-throughput testing, at least to identify the PoT or signatures of toxic effects but likely also for the actual testing; expectations are high that systems biology approaches might be translated to form a systems toxicology, which means the complex integration of several omics by bioinformatics.
- Combines different methods in integrated testing strategies (ITS)
- Is combined with PBPK modeling
- Is predictive of human health risks (and environmental hazards; it should be noted that the concept of Tox-21c has not been largely discussed for this area)

Obviously, the new approach finally must meet similar standards as to quality assurance (validation including GLP and other good practices) and International harmonization (Bottini et al., 2008; Hartung, 2009c) as the current schemes. Thus we will need to explore what this means for validation.

It is assumed that several hundred PoT exist, unlikely many thousands. Figure 2 shows how the identification of more and more PoT by expanding the number of toxicants and the number of cell systems should come to a saturation, i.e. indicating that the mapping is comprehensive. However, the number of variants of PoT is a possible threat to a PoT-based toxicity testing with regard to validation needs. How can we possibly validate these? Actually, the question starts one step earlier, i.e. how can we identify and annotate them? In the end, we likely need a project similar to a Human Toxicology Project to map PoT. In contrast to the currently used phenomenological “black box” animal test, PoT need to be identified in human *in vitro* systems to provide more relevant, accurate and mechanistic information for the assessment of human toxicological risk. The future ultimate goal would be to map the entirety of human PoT, the human toxome. The concentration at which a substance triggers a PoT might be extrapolated to a relevant human blood or tissue concentration and finally a corresponding dose by (retro-) PBPK modeling (Yang et al., 1998; Barton et al., 2007; Bouvier d’Yvoire et al.,

2007; Chiu et al., 2007; Clewell and Clewell, 2008), informing human risk assessment. Retro-QSAR indicates that instead of current use, i.e. to describe the resulting tissue levels and kinetics for a given dose, the theoretical dose necessary to achieve this tissue level needs to be modeled. This type of new tool will also require validation and it remains to be explored, whether the principles of validation as adapted to (Q)SAR are applicable (Hartung and Hoffmann, 2009).

However, probably more important, if a substance does not trigger any of these PoT, for the first time it may be possible to establish the non-toxicity of a substance at a given concentration, which might be combined later with thresholds of toxicological concern (TTC) (Kroes et al., 2005). The major advance is that external doses need not be used for TTC, but threshold tissue levels can be used, bringing the concept one step closer to the effect. We might be able to establish the no-effect levels in tissues with the PoT-based methods and then estimate by retro-QSAR, which dose would be required to achieve such levels.

It is assumed that Tox-21c will also integrate further *in silico* approaches (Hartung and Hoffmann, 2009). However, the Tox-21c literature is relatively vague about Integrated Testing Strategies (ITS). Practical considerations, in contrast, have led to suggestions of tiered testing strategies and, at least, the vision of integrated or “intelligent” testing (van Leeuwen et al., 2007). The most advanced are found in the REACH test guidance (European Chemicals Agency, 2008). These represent special challenges to their validation (Kinsner-Ovaskainen et al., 2009); a working definition of ITS has been proposed in this EPAA/ECVAM workshop: “*In the context of safety assessment,*

Fig. 2: As more substances are tested in more cell systems an increasing number of PoT will be identified until the entirety of PoT is mapped.

an Integrated Testing Strategy is a methodology which integrates information for toxicological evaluation from more than one source, thus facilitating decision-making. This should be achieved whilst taking into consideration the principles of the Three Rs (reduction, refinement and replacement)." This definition is very much imprinted by the largely European alternative method concept (Hartung, 2010a) and has little specificity. The workshop defined components of ITS as building blocks if they are "single test or a battery of tests with associated prediction model(s)/data interpretation procedures (DIPs), and ending with a decision step." The definition is in so far confusing as the prediction models and DIPs have so far been interpreted as an algorithm to estimate the *in vivo* result, which means something only the overall ITS can deliver. It is further stated that components of ITS require assessment of reproducibility and transferability, while "relevance (predictive capacity) belongs to the entire test strategy." Arguably, the authors state that such validation is only required for ITS for "Replacement of Test Guideline used for regulatory purposes" while not for screening, hazard classification and labeling, and risk assessment. It is obvious that Tox-21c will have to address these questions.

Before constructing an ITS based on PoT, it will be necessary to identify a substantial number of PoT, ideally the entire human toxome. To achieve the mapping of the human toxome, it will be necessary to combine several of the latest emerging technologies in life sciences. Human embryonic stem cells represent one of the most promising human cell systems (Davila et al., 2004; Bremer and Hartung, 2004; Pellizzer et al., 2005; Adler et al., 2008; Leist et al., 2008a; Stummann and Bremer, 2008; Chapin and Stedman, 2009). Key advantages over other cell systems (Hartung, 2007b) are that they can generate an unlimited source of differentiated cells, which allows the testing of substances at a broad range of concentrations in a time and cost efficient manner. Secondly, they are pluripotent and can be differentiated into any tissue of the human body. However, standardized protocols are only emerging and it is not clear whether quality controls of Good Cell Culture Practice (Coecke et al., 2005) are sufficient for this new field. The use of a human cell system will avoid species differences in toxicity responses. In principal, however, any cellular system should be usable, but either human primary cells or organotypic cultures would be recommended.

The toxicity responses of the chosen cellular system towards reference substances will need to be studied especially by metabolic and genetic profiling methods. Metabolic profiling can be achieved based on a mass spectrometry based metabolomics approach, which has advantages over NMR-based metabolomics approaches in terms of sensitivity, number of detectable metabolites, and metabolite quantification. Since the metabolome is defined by gene, transcript, and protein changes, it is the omics science closest to the phenotype. The measured metabolic perturbations represent the final outcome of all physiological processes in the biological system making the approach most interesting for toxicology (Robertson, 2005). Until now, metabolomics has been mainly applied for *in vivo* toxicity studies based on metabolic profiling of non-invasive blood and urine samples. Previous work (van Vliet et al., 2008) has already demonstrated the value of metabolomics for *in vitro* toxicity studies.

In addition to metabolomic analysis, gene array and real-time PCR transcriptome or proteomic profiling will likely be applied. Gene array technologies allow measuring the expression of thousands of genes at the same time, including the entire human genome. In contrast, real-time PCR provides a more specific and accurate tool to further confirm the expression level of a particular gene. Transcriptomics provides a simple and sensitive endpoint to study the toxic responses in biological systems – in fact, gene expression changes are normally observed before any other changes occur. In addition, gene expression analysis in primary cultures is a promising tool to detect chemicals with toxic potential (Hogberg et al., 2009, 2010). In an ECVAM/IC-CVAM/NICEATM workshop (Corvi et al., 2006) we have pioneered the discussion of validation of toxicogenomics tools. In this area in contrast to other omics technologies at least standard for reporting studies exist (Brazma et al., 2001), but other omics such as metabolomics are following (Griffin et al., 2007; Sansone et al., 2007; van der Werf et al., 2007). Though it is not said that the PoT-based methods or method batteries will necessarily be omics technologies (omics might be used to identify PoT only and then reporter gene assays etc. are constructed to cover the PoT), it demonstrates the complexity of the problem to standardize and QA such methods. Running this under GLP will be extremely difficult, starting with the problem of GLP for cell culture methods in general. However, a major opportunity lies in the commercialization of new *in vitro* methods, as they lead to standardization and broad availability as kits and integrated machines. The ensuing dialogue of the *in vitro*/alternatives field with technology providers is most important. The recent concentration process forming a few very large technology providers (ThermoFischer, LifeTechnologies, Agilent, etc.) promises long-term availability of high-quality tools to base regulatory testing on. The chances for contract research organization (CRO) are enormous (Goldberg and Hartung, 2008) and again this quality-assures the methods for regulatory use.

Combining these omic approaches can establish a method for the identification of human PoT. The metabolic and genetic profiling data obtained need to be analyzed using the latest statistical data mining methodologies. First patterns in the datasets will be recognized, followed by the establishment of relationships between genes and metabolites. With the integration of existing biochemical knowledge, the biochemical pathways altered by substances to PoT can be identified in a systems biology approach. The validation of these PoT will among others rely on the use of gene silencing or knock-out technologies.

A key goal needs to be the definition and annotation of PoT and making them available in an open access database, which allows integration of PoT identified from other groups and systems. Pathway identification from omics approaches is approached in several areas including toxicology (Fischer, 2005; Cho et al., 2006; Bosl, 2007; Kwoh and Ng, 2007). CAAT is working with EPA ToxCast (Judson et al., 2010) toward this key effort. An important element will therefore be a series of consensus workshops defining the database, its content, its validation, and its governance. CAAT and its Transatlantic Think Tank for Toxicology (t⁴) (Daneshian et al., 2010) aim to organize these.

In order to confirm PoT, it will be necessary to inhibit either

the development of the respective signature by blocking a toxicants effect or enhancing/reproducing it with an activation of the respective PoT. The former can either be achieved (where available) with specific inhibitors or more generally with siRNA. Almost every cell including human embryonic stem cells can nowadays be transfected with siRNA (Martinez di Montemuros and Parise, 2008), most efficiently with lentiviruses (Zaehres et al., 2005), but also with lipofection techniques (Matin et al., 2004; Vallier et al., 2004). The suppression of the respective gene in the candidate PoT can be controlled by quantitative real-time PCR of its mRNA. The goal must be to reproduce the metabolic changes as assessed by metabolomics with a block of a key element of the PoT. This means, that the derangement of downstream metabolites in the PoT is assessed and compared to those observed, when the PoT is deranged by toxicants. Such experiments can be extended to co-exposure with toxin after transfection to analyze a possible aggravation of the metabolic derangement, especially on basis of metabolomics assessing the metabolites relevant for the PoT.

Obviously such an approach represents rather a validation of the scientific mechanism or mode of action than of the model. Quality assurance means here mainly to link PoT with effects of well-established toxicants. Only at a later stage the presence of a PoT in a given test system can serve as a validation of the test (battery), e.g. by showing that the very same reference substances interfere via the PoT with the test. This scenario stresses the opportunity lying in validation based on PoT and the concept of scientific relevance for validation. We have earlier suggested this as “mechanistic validation” (Coecke et al., 2007; Hartung, 2007a).

A key need for any type of validation is the definition of a prediction model, i.e. the translation of test results into a prediction of adversity. This is a special challenge for the information-rich methods envisaged for Tox-21c (Boekelheide and Campion, 2010). There is discussion about whether this needs to be done before a validation exercise, which is certainly more convincing. It is most important that such definition of adversity is not Gaussian, i.e. defining “not normal” by percentiles of the controls: First, values do not typically follow normal distributions, but, more importantly, this sets a prevalence of hazard at the given percentage of significance. With other words, a definition of positive test results by “significant change” is not appropriate. Only a definition of normal (non-toxic) and toxic by relating results to reference compounds with known classification can define adversity.

4 The concept of evidence-based toxicology

Evidence-based medicine (EBM) is a largely accepted process in medicine, which aims to summarize and make available the best available evidence for a medical question in the most transparent and objective manner (Mayer, 2004). Certainly, many aspects of medicine depend on individual factors, values and choices, which are only partially subject to scientific methods. EBM seeks to clarify those aspects of medical practice that are in principle subject to scientific methods. In order to achieve

this, the Cochrane collaboration, a group of over 27,000 volunteering physicians and scientists in more than 90 countries, carries out systematic reviews of relevant medical questions, especially therapeutic interventions but also diagnostic strategies. Specific methods, especially statistical tools such as those related to meta-analysis, are developed and results made available via the Cochrane library. Already in 1993, Neugebauer and Holaday applied the principles to animal and *in vitro* data.

Some similarities to the problems of toxicology and especially the analogy of toxicology and diagnosis setting (Hoffmann and Hartung, 2005) prompted us and others to call for an evidence-based toxicology – EBT – (Guzelian et al., 2005, Hoffman and Hartung, 2006). These include:

- The mixture of traditional and scientific approaches
- The lack of quality assurance for many methods and the difficulty to retrieve such objective assessments lacking a central depository
- The information flood of scientific articles and the expansion of toxicology (larger programs addressing for example untested old chemicals, more legislations in more economically relevant countries, increasing risk avoidance, new products such as nanoparticles, biologicals, and cell therapies etc., concerns about mixtures, new hazards such as endocrine disruption, respiratory sensitization, immunotoxicity, neurodevelopmental toxicity, etc.)

Obviously, toxicology has a similar problem of information flooding and coexistence of traditional and modern methodologies, as well as various biases (Wandall et al., 2007). It is most difficult to find and summarize the relevant information for any given major question. Rudén (2001a,b) showed the divergence in judgment and limitations of analysis for the example of 29 cancer risk assessments carried out for trichloroethylene – 4 assessments concluded that the substance is carcinogenic, 6 said it is not, and 19 were equivocal. The main reason for this divergence was a selection bias in the materials considered, i.e. an average reference coverage of only 18%, an average citation coverage of most relevant studies of 80%, as well as an interpretation difference of most relevant studies in 27%, and the lack of study/data quality assessment not documented in 65% of the assessments. This indicates a tremendous problem for compiling evidence in toxicology.

- EBM has some key tools to further its purpose, which are little used in toxicology:
- Systematic reviews (to be distinguished from narrative reviews, which are the common form in science)
- Weighing of evidence by quality scores
- Meta-analysis and other statistical tools such as likelihood ratios (an application of the Bayes’ theorem especially for diagnostics, here basically the pretest odds multiplied by the likelihood ratio gives the post-test odds), AUC-ROC, i.e., the area under the receiver operating characteristic curve (reflects the relationship between sensitivity and specificity for a given test) or number needed to treat/harm (the effectiveness and safety of an intervention expressed in a clinically meaningful way)
- Formalized group decision taking processes, such as the Delphi method or the nominal group technique

A number of workshops took up this idea (Griesinger et al., 2009), a first tool for quality scoring toxicological studies was developed (Schneider et al., 2009), and the first chair for EBT was created at Johns Hopkins, notably also the site of the US Cochrane center. The EBT concept, however, is still in its infancy. This article is meant to explore whether there is a case of need to develop EBT from the Tox-21c movement.

5 Experiences with retrospective validation as a kind of EBM-like validation

Retrospective validation is an EBT-like approach, but remains anchored in the validation paradigm of reference methods. This means it is based on an evaluation of available information, but the comparison is done against a reference method. An ECVAM workshop (Hoffmann et al., 2008) aimed to open the concept of reference methods by suggesting the use of reference results, i.e., not to reproduce the results of a given method, but establish by weighing evidence the best possible results to be obtained for a panel of test compounds. This means that every substance chosen is evaluated with the relevant information available to assign the test result, which should be achieved for the substance. In other words, the goal would not be to reproduce the results of one animal test, but to identify correctly the positive and negative substances according to an overall assessment especially including human experiences.

Retrospective validation was introduced into the modular approach to validation (Hartung et al., 2004). It has been used successfully to validate the micronucleus test (Corvi et al., 2008) and some variants of *in vitro* eye irritation tests (Hartung et al., 2010). The cell transformation assays currently under peer review notably used a combination of retrospective data collection and prospective studies. What are the differences between a prospective and a retrospective validation? Obviously, a prospective study allows challenging a defined test according to the current understanding of its optimal test protocol, applicability, and test demands. Transfer of method to all laboratories in the round-robin can be controlled and a power analysis can establish the number of substances likely required. At least in principle the result is open; this means that different, more powerful statistics can be used than in a post-hoc analysis.

In contrast, a retrospective analysis depends entirely on what is available. There is an obvious danger of a selection bias here. Regulatory toxicology represents a special challenge, because of the proprietary nature of many data, the lack of incentive for publication (especially of negative data), and the prohibition/avoidance of repeated testing. Documentation of existing data often is heterogeneous and/or incomplete, impairing data analysis. In general, only when relatively large data sets are available from several sources it will make sense to collect and jointly analyze them. Ideally this should take place as a meta-analysis but this has not been applied to toxicology (Hartung, 2009b).

The EBM movement has its strength, however, in the rigor and transparency of the decision-making process. Certainly, the organization of available evidence according to the modular ap-

proach is an important step, but selection bias has not really been addressed and the decision process is very traditional. Systematic reviews of available evidence could serve as a role model (Hartung, 2009b).

In conclusion, retrospective analysis moves validation toward EBM with regard to its tools but leaves it chained to its point of comparison – the traditional method.

6 The similarity of diagnosis setting and toxicology as a bridge between EBM and EBT

Toxicologists familiar with EBM often have problems translating these approaches to their field because most EBM evaluates therapeutic measures and studies outcome. In toxicology, we have few interventions (restrictions of use), and their outcome can usually not be assessed, as the substance is then simply not used and no experience is gained. However, a growing branch of EBM addresses diagnostic measures (Knotterus and Buntinx, 2009). Intellectually, running a couple of tests and setting a diagnosis or assigning toxicity is not different (Hoffmann and Hartung, 2005). In the end, the process is about detecting or excluding a disorder/hazard by increasing diagnostic certainty as to their presence or absence. We might learn from EBM of diagnostic measures. The first issue is its “discriminative power,” typically assessed as a table and deducing sensitivity and specificity. We have several times alluded to the problem of the prevalence of hazards impacting on the usefulness of tests (Hoffmann and Hartung, 2005; Hartung, 2009a), i.e., a sensitivity assessed by studying 25 toxic and 25 non-toxic substances does not reflect the sensitivity of the test in the real world, where only 5% of substances might be toxic. However, we must also ask what the discriminatory power is after cheaper and faster tests have been carried out, for instance, its added value in a test strategy. The value of a cancer bioassay for example will further diminish if we have sorted out mutagens. In testing strategies, each step might make only a rather small contribution; however, to evaluate the importance of such small steps it takes a relatively large study population (number of test substances). It is also of key importance to ask, what is the contribution to the decision making process. Data, which are collected but do not add to decisions (risk management) are, at minimum, a waste of money.

The evaluation of diagnostics may be flawed by several biases, most importantly spectrum bias and selection bias. “Spectrum bias” occurs if the diagnostic is assessed in a study population with a different clinical spectrum. Translated to toxicology, this might be the case if we validate a method for chemicals but base the validation mainly on drugs because good data are available. Selection bias refers to cases where there is a relation between the test result and the probability of being chosen. This would be the case, for example, if tested in parallel to the traditional test applied to suspicious substances. Observer bias would refer to a situation where (unconsciously) more effort is put into one method than the others or where skill and experience levels differ.

For diagnostic research, a number of different questions need to be addressed (Haynes and You, 2009):

- Phase I questions: Do patients with the target disorder have different test results from normal individuals? Translated to toxicology, this is the typical validation paradigm, testing known toxicants and non-toxicants. We typically derive sensitivity and specificity; we might learn from diagnostic tests that these have confidence intervals, which are not always calculated in toxicology.
- Phase II questions: Are patients with certain test results more likely to have the target disorder than patients with other test results? Translated to toxicology, this is translating to the test situation, i.e. taking into consideration at least prevalence of the hazard (and converting to predictive values) but might extend to practical aspects such as selection and spectrum and observer biases. Thus, it is moving the test evaluation from ideal test conditions to the routine situation. We have to admit that this is not typically done in validation in toxicology. There is even some resistance to applying the concept of prevalence.
- Phase III questions: Among patients in whom it is clinically sensible to suspect the target disorder, does the test result distinguish those with and without the target disorder? Translated to toxicology, this relates to testing strategies. “Clinically sensible to suspect” might relate to the presence of alerts and suspicious substances as well as results from other tests; if not everything is tested with the test in question, it is important to understand how this selection impacts on the prevalence of hazard in the tested subgroup. It translates pre-test likelihood of a hazard into post-test likelihood, i.e., whether the test moves us ahead in confirming the hazard. Noteworthy, here all cases impact where either the traditional method or the method under evaluation is “lost, not performed or intermediate”; it is most important that these are not simply excluded from the analysis of accuracy. It is also most important that results cannot be interpreted by the examiner (e.g. by post-hoc changing of thresholds) or are biased because the results of the traditional method are known. A special case is when the selection of substances through suspicions or test results is very effective; then the additional value of carrying out the test in question might become very small, i.e., its discriminatory value has been used up along the way of the testing strategy.
- Phase IV questions: Do patients who undergo the diagnostic test fare better (in their ultimate health outcomes) than similar patients who do not? Ultimate health outcomes in toxicology are difficult to obtain; this might be translated to the avoidance of market withdrawals, for instance.
- Phase V questions: Does the ultimate diagnostic test lead to better health outcomes at acceptable costs? Cost benefit analyses in toxicology are rare and are difficult if we cannot answer phase IV questions. Still estimates might be helpful to assess the economics of what we are doing (Bottini and Hartung, 2009).

We might learn a lot from the literature on diagnostic tests as how to derive the confidence intervals for our sensitivity and

specificity calculations, the pre- and post-test likelihoods of a diagnosis (hazard) and from these the informative/discriminatory value, etc. (Habbema et al., 2009). Noteworthy, alternative methods usually work by threshold setting making the continuous results of the test (e.g. % cytotoxicity) a dichotomous result (toxic or not). However, the absolute value often has informative value, i.e. a borderline value will usually have less confidence than an extreme value. For diagnostic test strategies, methods for assessment of the predictive value of combinations and what an individual test adds to it are available, especially multiple logistic regression formulations of Bayes’ theorem (Habbema et al., 2009); other approaches are tree-building methods and neural networks (Buntinx et al., 2009). Importantly, in the diagnostic field standards for reporting on diagnostic accuracy studies have also been developed (Bossuyt and Smidt, 2009), which could serve as models for toxicology. Even guidance for the systematic review of such studies is available (Horvath and Pewsner, 2004; Buntinx et al., 2009). These examples demonstrate the availability of a rich literature on the evaluation of diagnostic tests, which can be translated to the evaluation of (new) toxicological methods applying EBT concepts.

7 The need for probabilistic risk assessment

A key feature of EBM is the deduction of probabilities, odd ratios, and confidence intervals. This reflects a scientific reality in the life sciences, where few cases are black and white, at least without tremendous restrictions of the parameters under which the judgment is valid. In toxicology, we have maintained the presumption of a black and white world for a very long time. Substances are either carcinogenic or not, they are irritant or not, corrosive or not... though we sometimes introduce classes like weak or strong. However, all these represent only measures with uncertainties in the test system, uncertainty in threshold setting, and uncertainty in the extrapolation to humans as well as extrapolation to actual usage scenarios. Assigning a probability of a certain hazard appears to be much more adequate than the distinct categories we are using. Even better, these probabilities should come with confidence intervals. The goal of our quality assurance/validation must then be to establish such probabilities, which means a completely different type of prediction models. This is much more similar to the reasoning of pre- and post-test probabilities of a certain diagnosis in EBM.

It might be conceptually much easier for regulatory toxicologists to understand that each test shifts only the probability of hazard and the confidence intervals than to think in misclassifications. However, unlike EBM, where a patient is a person with a certain probability of a diagnosis, we are confronted with dose, which means a substance shows a hazard only at a certain dose. Thus, a dimension of complexity is added to our assessments, which complicates such test evaluations. It will have to be shown whether we can handle this complexity and especially whether enough data can be made available to enable such assessments.

8 Conclusions

Table 1 summarizes the key characteristics of traditional validation studies, retrospective validations, EBT-based assessments of methods and EBM of diagnostics.

The following hypotheses are put forward:

1. Traditional validation allows only substituting a method with something similar and does not accommodate paradigm shifts due to its comparison to the traditional test. The predictivity of a test approach can thus not be changed.
2. The pressing need to renovate methods for regulatory toxicology calls for information-rich, complex methods. These represent a challenge for quality assurance (such as GLP) and validation. While test definition and reproducibility can be handled similarly, scientific relevance will need to be stressed to compensate for the difficult predictive relevance in the absence of a reference test.
3. EBM for diagnostics can serve as a role model for test evaluation in toxicology. It offers the advantages of stronger emphasis on scientific validity, more transparent processes, and more extensive biometrical evaluation. It also appears to be more flexible and possibly faster to accommodate new evidence as well as test variants, especially compared to prospective validation studies.
4. We need to develop meta-analysis tools for retrospective validation and EBT. They will be most useful for the risk assessment process for individual compounds, where several studies are available. For this purpose, the quality score concept for toxicological studies needs to be further developed.
5. The concept of pathways of toxicity (PoT) needs to mature (definition, identification, annotation) to develop validation strategies.
6. Integrated Testing Strategies (ITS) pose an unresolved problem for validation; Tox-21c will need to rely to large extent on ITS. The key problems are how to avoid the rigidity of fixed combinations of methods, how to handle the multiple

Tab. 1: Comparison of different quality assurance approaches to toxicological and diagnostic tests

	Traditional validation	Retrospective validation	EBT	EBM for diagnostics
Data base	- Prospective studies	- Collected data	- Collected data - Literature review	- Literature review
Point of reference	- Animal test result	- Animal test result	- Scientific state of the art - Expert consensus on reference	- Scientific state of the art - Clinical diagnosis and outcome
Assessment parameters	- Reproducibility - Transferability - Reliability (to predict animal)	- Reproducibility - Transferability - Reliability (to predict animal)	- Reproducibility - Transferability - Reliability (to predict human) - Post-test probability of hazard	- Post-test probabilities of diagnosis - Various performance measures
Process owners	- Validation Management Group - Trial centers	- Validation Management Group	- Expert working group	- Expert working group
Style	- Actual testing - Compilation of dossier - Narrative	- Compilation of dossier - Narrative	- Systematic review - Meta-analysis to be developed - Transparent - Objective	- Systematic review - Meta-analysis - Transparent - Objective
Peer-review	- Final dossier	- Final dossier	- Strategy before assessment - Result	- Strategy before assessment - Result
Publication	- Validity statement – Scientific article - sometimes Background Review Document	- Validity statement – Scientific article - sometimes Background Review Document	- Guidance and documentation of process in EBT portal to be established	- Guidance and documentation of process in Cochrane Library

testing fallacy, and the high number of reference compounds needed for evaluation of decision trees.

7. The discriminatory value of a test changes in an ITS by changing pre-test to post-test probability of hazard. Our validations need to accommodate this by including prevalence and moving to predictive values.
8. Tox-21c relies first on PoT, their combinations, and the signatures they leave in information-rich systems. A key problem is the definition of adversity and its interplay with threshold setting, and thus the outcome of validation. Tox-21c relies secondly on retro-PBPK, for which no quality assurance and validation experience exists. The validation of the combined results represents the next level of complexity.
9. Since a key goal of regulatory toxicology is the identification of non-toxic substances or non-toxic doses of toxicants, Tox-21c requires a comprehensive mapping of relevant PoT. It will be critical to identify areas for exploration of the concept with likely limited numbers of PoT such as endocrine disruption to test the concept. For complex hazards, however, validation for identification of non-toxic substances will only be possible after a type of Human Toxicology Project mapping the human toxome.
10. The possibility to expand To-21c to ecotoxicology needs to be explored. The conservation of PoT across species is critical here.
11. When moving to probabilistic risk assessment, new measures of performance such as post-test hazard probability need to be developed, which operate with probability of hazard and its confidence intervals. This is further complicated as this is a dose-dependent estimate.

The challenge to Tox-21c will be to steer toward quality control without the creation of obstacles by formal validation. A balance between precaution and innovation is necessary, and this requires informed decisions by the actors in the regulatory arena. EBM has shown how the informed decision process in clinical medicine can be served. EBT promises to be its translation for an informed decision process in risk assessment.

References

- Adler, S., Pellizzer, C., Hareng, L. et al. (2008). First steps in establishing a developmental toxicity test method based on human embryonic stem cells. *Toxicol. In Vitro* 22, 200-211.
- Ahr, H.-J., Alepee, N., Breier, S. et al. (2008). Barriers to validation. A report by European Partnership for Alternative Approaches to animal testing (EPAA) working group 5. *ATLA* 36, 459-464.
- Andersen, M. E. and Krewski, D. (2009). Toxicity testing in the 21st century: bringing the vision to life. *Toxicol. Sci.* 107, 324-330.
- 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. *ATLA* 34, 603-620.
- Balls, M., Blaauboer, B., Brusick, D. et al. (1990). Report and recommendations of the CAAT/ERGATT workshop on the validation of toxicity test procedures. *ATLA* 18, 313-337.
- Barton, H. A., Chiu, W. A., Woodrow Setzer, R. et al. (2007). Characterizing uncertainty and variability in physiologically based pharmacokinetic models: state of the science and needs for research and implementation. *Toxicol. Sci.* 99, 395-402.
- 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.
- Bosl, W. J. (2007). Systems biology by the rules: hybrid intelligent systems for pathway modeling and discovery. *BMC Syst. Biol.* 1, 13.
- Bossuyt, P. M. M. and Smidt, N. (2009). Standards for reporting on diagnostic accuracy studies. In J. A. Knotterus and F. Buntinx, *The evidence base of clinical diagnosis – theory and methods of diagnostic research* (167-179, 2nd edition). Chichester, UK: Wiley Blackwell.
- Bottini, A. A. and Hartung, T. (2009). Food for thought... on the economics of animal testing. *ALTEX* 26, 3-16.
- Bottini, A. A., Alepee, N., de Silva, O. et al. (2008). Optimization of the post-validation process. The report and recommendations of ECVAM workshop 67. *ATLA* 36, 353-366.
- Bouvier d'Yvoire, M., Prieto, P., Blaauboer, B. J. (2007). Physiologically-based kinetic modelling (PBK modelling): Meeting the 3Rs Agenda. The report and recommendations of ECVAM workshop 63. *ATLA* 35, 661-671.
- Brazma, A., Hingamp, P., Quackenbush, J. (2001). Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* 29, 365-371.
- 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.
- Buntinx, F., Aergeerts, B. and Macaskill, P. (2009). Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests. In J. A. Knotterus and F. Buntinx, *The evidence base of clinical diagnosis – theory and methods of diagnostic research* (180-212, 2nd edition). Chichester, UK: Wiley Blackwell.
- 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.
- Chiu, W. A., Barton, H. A., DeWoskin, R. S. et al. (2007). Evaluation of physiologically based pharmacokinetic models for use in risk assessment. *J. Appl. Toxicol.* 27, 218-237.
- Cho, C. R., Labow, M., Reinhardt, M. et al. (2006). The application of systems biology to drug discovery. *Curr. Opin. Chem. Biol.* 10, 294-302.
- Clewell, R. A. and Clewell, H. J. 3rd (2008). Development and specification of physiologically based pharmacokinetic models for use in risk assessment. *Regul. Toxicol. Pharmacol.* 50, 129-143.
- Coecke, S., Goldberg, A. M., Allen, S. et al. (2007). Incorporating in vitro alternative methods for developmental neurotoxicity into international hazard and risk assessment strategies. *Environ. Health Persp.* 115, 924-931.
- 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. *ATLA* 33, 261-287.
- Collins, F. S., Gray, G. M. and Bucher, J. R. (2008). Toxicology:

- Transforming environmental health protection. *Science* 319, 906-907.
- Coons, S. J. (2009). The FDA's critical path initiative: a brief introduction. *Clin. Ther.* 31, 2572-2573.
- Corvi, R., Albertini, S., Hartung, T. et al. (2008). ECVAM retrospective validation of in vitro micronucleus test (MNT). *Mutagenesis* 23, 271-283.
- Corvi, R., Ahr, H. J., Albertini, S. et al. (2006). Meeting report: Validation of toxicogenomics-based test systems: ECVAM-ICVAM/NICEATM considerations for regulatory use. *Environ. Health Perspect.* 114, 420-429.
- 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.
- Davila, J. C., Cezar, G. G., Thiede, M. et al. (2004). Use and application of stem cells in toxicology. *Toxicol. Sci.* 79, 214-223.
- European Chemicals Agency (2008). Guidance on information requirements and chemical safety assessment (2178). Helsinki, Finland: European Chemicals Agency. Available at: http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_en.htm (accessed 17 April 2009).
- FDA (2004). FDA: Challenge and opportunity on the critical path to new medical products. <http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunities-Reports/ucm077262.htm> (accessed 20 April 2010).
- Fischer, H. P. (2005). Towards quantitative biology: integration of biological information to elucidate disease pathways and to guide drug discovery. *Biotechnol. Annu. Rev.* 11, 1-68.
- Gibb, S. (2008). Toxicity testing in the 21st century: a vision and a strategy. *Reprod. Toxicol.* 25, 136-138.
- Goldberg, A. and Hartung, T. (2008). The emerging new toxicology – an opportunity for contract research. *Eur. Pharmaceut. Contractor*, 42-46.
- Grandjean, P. and Landrigan, P. J. (2006). Developmental neurotoxicity of industrial chemicals. *Lancet* 368, 2167-2178.
- Griesinger, C., Hoffmann, S., Kinsner-Ovaskainen, A. (2009). Proceedings of the first international forum towards evidence-based toxicology. Conference Centre Spazio Villa Erba, Como, Italy. 15-18 October 2007. *Hum. Exp. Toxicol., Spec. Issue: Evidence-Based Toxicology (EBT)* 28, 81-163.
- Griffin, J. L., Nicholls, A. W., Daykin, C. A. et al. (2007). Standard reporting requirements for biological samples in metabolomics experiments: mammalian/in vivo experiments. *Metabolomics* 3, 179-188.
- Guzelian, P. S., Victoroff, M. S., Halmes, N. C. et al. (2005). Evidence-based toxicology: a comprehensive framework for causation. *Hum. Exp. Toxicol.* 24, 161-201.
- Habbema, J. D. F., Eijkemans, R., Krijnen, P. and Knotterus, J. A. (2009). Analysis of data on the accuracy of diagnostic tests. In J. A. Knotterus and F. Buntinx, *The evidence base of clinical diagnosis – theory and methods of diagnostic research* (118-145, 2nd edition). Chichester, UK: Wiley Blackwell.
- Hartung T. (2010a). Lessons learned from alternative methods and their validation for a new toxicology in the 21st century. *J. Toxicol. Env. Health* 13, 277-290.
- Hartung, T. (2010b). Food for thought ... on alternative methods for chemical safety testing. *ALTEX* 27, 3-14.
- Hartung, T. (2010c). 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. and Hoffmann, S. (2009). Food for thought on ... in silico methods in toxicology. *ALTEX* 26, 155-166.
- 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). A toxicology for the 21st century: Mapping the road ahead. *Toxicol. Sci.* 109, 18-23.
- Hartung, T. (2008a). Towards a new toxicology – evolution or revolution? *ATLA* 36, 635-639.
- Hartung, T. (2008b). Food for thought ... on animal tests. *ALTEX* 25, 3-9.
- Hartung, T. and Leist, M. (2008). Food for thought ... on the evolution of toxicology and phasing out of animal testing. *ALTEX* 25, 91-96.
- Hartung, T. (2007a). Food for thought ... on validation. *ALTEX* 24, 67-72.
- Hartung, T. (2007b). Food for thought ... on cell culture. *ALTEX* 24, 143-147.
- Hartung, T., Bremer, S., Casati, S. et al. (2004). A modular approach to the ECVAM principles on test validity. *ATLA* 32, 467-472.
- Haynes, R. B. and You, J. J. (2009). The architecture of diagnostic research. In J. A. Knotterus and F. Buntinx, *The evidence base of clinical diagnosis – theory and methods of diagnostic research* (20-41, 2nd edition). Chichester, UK: Wiley Blackwell.
- Hoffmann, S., Edler, L., Gardner, I. et al. (2008). Points of reference in validation – the report and recommendations of ECVAM Workshop. *ATLA* 36, 343-352.
- Hoffmann, S. and Hartung, T. (2006). Towards an evidence-based toxicology. *Hum. Exp. Toxicol.* 25, 497-513.
- Hoffmann, S. and Hartung, T. (2005). Diagnosis: Toxic! – Trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations. *Toxicol. Sci.* 85, 422-428.
- Hogberg, H. T., Kinsner-Ovaskainen, A., Coecke, S. et al. (2010). mRNA expression is a relevant tool to identify developmental neurotoxicants using an in vitro approach. *Toxicol. Sci.* 113, 95-115.
- Hogberg, H. T., Kinsner-Ovaskainen, A., Hartung, T. et al. (2009). Gene expression as a sensitive endpoint to evaluate cell differentiation and maturation of the developing central nervous system in primary cultures of rat cerebellar granule cells (CGCs) exposed to pesticides. *Toxicol. Appl. Pharmacol.* 235, 268-286.
- Horvath, A. R. and Pewsner, D. (2004). Systematic reviews in laboratory medicine: principles, processes and practical considerations. *Clin. Chim. Acta* 342, 23-39.
- Judson, R. S., Houck, K. A., Kavlock, R. J. (2010). In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ. Health Perspect.* 118, 485-492.
- Judson, R., Richard, A., Dix, D. J. et al. (2009). The toxicity data

- landscape for environmental chemicals. *Environ. Health Perspect.* 117, 685-695.
- Kavlock, R. J. (2009). The future of toxicity testing – the NRC vision and the EPA's ToxCast program national center for computational toxicology. *Neurotoxicol. Teratol.* 31, 237.
- Kavlock, R. J., Austin, C. P. and Tice, R. R. (2009). Commentary – toxicity testing in the 21st century: Implications for Human Health Risk Assessment. *Risk Analysis* 29, 485-487.
- 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. *ATLA* 37, 1-8.
- Knotterus, J. A. and Buntinx, F. (2009). *The evidence base of clinical diagnosis – theory and methods of diagnostic research* (2nd edition, pp.302). Chichester, UK: Wiley Blackwell.
- Kroes, R., Kleiner, J. and Renwick, A. (2005). The threshold of toxicological concern concept in risk assessment. *Toxicol. Sci.* 86, 226-230.
- Kwoh, C. K. and Ng, P. Y. (2007). Network analysis approach for biology. *Cell Mol. Life Sci.* 64, 1739-1751.
- 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.
- Mayer, D. (2004). *Essential evidence-based medicine* (1-381). Cambridge: Cambridge University Press.
- Martinez di Montemuros, F. and Parise, P. (2008). New technologies from siRNA world. *Minerva Biotech.* 20, 3-11.
- Matin, M. M., Walsh, J. R., Gokhale, P. J. et al. (2004). Specific knockdown of Oct4 and beta2-microglobulin expression by RNA interference in human embryonic stem cells and embryonic carcinoma cells. *Stem Cells* 22, 659-668.
- Neugebauer, E. A. and Holaday, J. W. (1993). *Handbook of Mediators in Septic Shock* (1-608, 1st edition). Florida, USA: CRC-Press, Boca Raton.
- NRC (2007). Committee on toxicity testing and assessment of environmental agents, national research council: toxicity testing in the 21st century: a vision and a strategy. The National Academies Press, 196.
- NRC (2000). NRC: Scientific frontiers in developmental toxicology and risk assessment. National Academy Press, Washington DC.
- OECD (2005). *Guidance document on the validation and International acceptance of new or updated test methods for hazard assessment. ENV/JM/MONO(2005)14*. OECD Series on Testing and Assessment 34, 96pp. Paris, France: Organisation for Economic Co-operation and Development. Available at: [http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=env/jm/mono\(2005\)14&doclanguage=en](http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=env/jm/mono(2005)14&doclanguage=en)
- Pellizzer, C., Bremer, S. and Hartung, T. (2005). Developmental toxicity testing from animal towards embryonic stem cells. *ALTEX* 22, 47-57.
- Robertson, D. G. (2005). Metabonomics in toxicology: a review. *Toxicol. Sci.* 85, 809-822.
- Rudén, C. (2001a). Interpretations of primary carcinogenicity data in 29 trichloroethylene risk assessments. *Toxicol.*, 169-225.
- Rudén, C. (2001b). The use and evaluation of primary data in 29 trichloroethylene carcinogen risk assessments. *Regulat. Toxicol. Pharmacol.* 34, 3-16.
- Sansone, S. A., Schober, D., Atherton, H. J. et al. (2007). Metabolomics standards initiative: ontology working group work in progress. *Metabolomics* 3, 249-256.
- 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.
- Seidle, T. and Stephens, M. L. (2009). Bringing toxicology into the 21st century: A global call to action. *Toxicol. In Vitro* 23, 1576-1579.
- Skrabanek, P. and McCormick, J. (1989). *Follies and fallacies in medicine*. Glasgow: Tarragon Press.
- Stummann, T. C. and Bremer, S. (2008). The possible impact of human embryonic stem cells on safety pharmacological and toxicological assessments in drug discovery and drug development. *Curr. Stem Cell Res. Ther.* 3, 117-130.
- Vallier, L., Rugg-Gunn, P. J., Bouhon, I. A. et al. (2004). Enhancing and diminishing gene function in human embryonic stem cells. *Stem Cells* 22, 2-11.
- van der Werf, M. J., Takors, R., Smedsgaard, J. et al. (2007). Standard reporting requirements for biological samples in metabolomics experiments: microbial and in vitro biology experiments. *Metabolomics* 3, 189-194.
- van Leeuwen, C. J., Patlewicz, G. Y. and Worth, A. P. (2007). Intelligent testing strategies. In C.J. Van Leeuwen and T.G. Vermeire (eds.), *Risk Assessment of Chemicals. An Introduction* (467-509, 2nd edition). Dordrecht, The Netherlands: Springer.
- van Vliet, E., Morath, S., Eskes, C. et al. (2008). A novel in vitro metabolomics approach for neurotoxicity testing, proof of principle for methyl mercury chloride and caffeine. *Neurotoxicol.* 29, 1-12.
- Vinken, M., Doktorova, T., Ellinger-Ziegelbauer, H. et al. (2008). The carinoGENOMICS project: critical selection of model compounds for the development of omics-based in vitro carcinogenicity screening assays. *Mutat. Res.* 659, 202-210.
- Wandall, B., Hansson, S. O. and Rudén, C. (2007). Bias in toxicology. *Arch. Toxicol.* 81, 605-617.
- Yang, R. S., Thomas, R. S., Gustafson, D. L. et al. (1998). Approaches to developing alternative and predictive toxicology based on PBPK/PD and QSAR modeling. *Environ. Health Perspect.* 106, Suppl. 6, 1385-1393.
- Zaehres, H., Lensch, M. W., Daheron, L. et al. (2005). High-efficiency RNA interference in human embryonic stem cells. *Stem Cells* 23, 299-305.

Correspondence to

Prof. Thomas Hartung, MD, PhD
Johns Hopkins Bloomberg School of Public Health
Dept. Environmental Health Sciences
Center for Alternatives to Animal Testing
615 N. Wolfe St.
Baltimore, MD, 21205, USA
e-mail: THartung@jhsph.edu