Zuerst ersch. in : ALTEX Proceedings ; 2 (2013), 1. - S. 27-37

Summary and Validation of New Animal-Free Toxicity Tests

Marcel Leist, Marina Hasiwa, and Mardas Daneshian

Doerenkamp-Zbinden Chair for In-vitro Toxicology and Biomedicine and Center for Alternatives to Animal Testing in Europe, University of Konstanz, Konstanz, Germany

Summary

Alternatives to animal testing have been developed mainly in the fields of toxicology and vaccine testing. Typical examples are the evaluation of phototoxicity, eye irritation, or skin corrosion resulting from cosmetics and industrial chemicals. Examples also can be found in other biomedical areas, however, including the control of the quality of drug preparations or for the control of the production process of biologics. For regulatory purposes, the quality, transferability, and predictivity of an alternative method need to be evaluated. This procedure often is called the "validation process" of a new method. It follows defined rules, and several governmental institutions have been established to perform, supervise, or advise on this process. As this often results in a delay of method implementation, different alternatives for the evaluation of a method's suitability and quality are subject to discussion. We describe here the principles of model development and quality control. We also give an overview of methods that have undergone validation. Strengths and shortcomings of traditional approaches are discussed, and new developments and challenges are outlined.

Keywords: validation process, OECD test guidelines, ICCVAM, ECVAM, JaCVAM

1 Introduction

Validation is a normal procedure in all fields of science once a test is developed (Hartung, 2007). The validation process is intended to provide confidence in the results, to define where the test may or may not be applied, and to give an account of test characteristics such as precision, accuracy, specificity, sensitivity, robustness, and transferability. The establishment, validation, and documentation of test methods in different areas of science all have been extensively covered in the specialist literature. This includes specific recommendations published by regulatory bodies. For instance, OECD GD 34 gives guidance on "*Development, Validation and Regulatory Acceptance of New and Updated Internationally Acceptable Test Methods in Hazard Assessment*." While the predictivity and biological relevance often are difficult to quantify, the quality of an assay system may be assessed by strictly quantitative methods.

2 Theoretical consideration on the setup of methods

What is essential for the setup of a method before validation? The setup and later validation of a toxicological test system requires some initial thought on model setup. The general rules for good scientific experiments may provide some initial guidance (Burgess, 2001). Three basic requirements must be fulfilled:

(a) *Reproducibility*: The experiment needs to be independent of the observer, the place, and the time when it is performed. That means that it should be repeatable by anyone (skilled in the art) and anywhere. The data should be quantifiable in order to establish reproducibility and comparability of the data.

(b) *Relevance*: The reason and rationale for the experiment should be clear and, most importantly, it should be embedded in a plausible biological context. This means, in a wider sense, that it is hypothesis-driven.

(c) *Hypothesis-generating*: The results of the experiment must point beyond the experiment itself and make predictions for other conditions. The predictions made must be testable and falsifiable.

These can be transferred to the requirements for model development, where three major criteria need to be fulfilled:

(aa) *Reliability*: This refers to the robustness and reproducibility of the model. Validation of this aspect should be mandatory for each model used, independent of the legal context or other implications. It is an evaluation of the technical quality of a model.

(bb) *Translation*: This refers to the scientific relevance of the model. Often judgments of this aspect will require that time and experience be gained through use of the model. Deep knowledge of biological processes involved in the model – and in reality – is required as well.

(cc) *Predictivity*: This aspect deals with the capacity of the model to yield results that correlate well with reality.

Specificity and sensitivity are among the parameters that describe this aspect. Notably, sensitivity and specificity are not only technical reliability parameters, as they change over time and with experience gained. Any given number is valid only in relation to the "gold standard" or the "reality" used as surrogate for reality. This point often is neglected or not recognized.

The model (i.e., the toxicological test used, e.g., as *in vitro* replacement method, or as animal model) itself is built from four elements. Each element can be validated and adapted individually:

(aaa) *Biological system*: This may be a dendritic cell or a guinea pig, or a differentiating stem cell.

(bbb) *Exposure scheme*: The guinea pig may be dosed orally or dermally, once or repeatedly, with a certain vehicle, for a certain time. The stem cell may be exposed to a chemical with or without medium change, during a certain time window and in a specified solvent.

(ccc) *Assay endpoint*: Death of a cell or of the guinea pig, measured by a specified viability assay, or using a specified humane endpoint; or skin reddening or altered differentiation, determined by PCR or immunocytochemistry. The type of endpoint chosen can completely change the outcome of an assay. In this context it is of utmost importance to distinguish endpoints that describe the biological system from endpoints that describe the behavior of the test in the presence of chemicals. These separate issues require independent optimization and characterization. For instance, a person's body weight can be measured well on scales (to give a good readout on general growth characteristics of a person = biological system), but this endpoint will hardly respond to acute poisoning of the person. Instead, blood pressure or vomiting activity may be good measures of human poisoning (toxicological test), but they in turn give little information on the growth activity over time.

(ddd) *Prediction model*: Translation of an endpoint outcome into toxicological information. For instance: is reversible light skin reddening interpreted as a sign of sensitization? Or, is a change of gene expression of marker x interpreted as a toxicological change? Is there a binary outcome (toxic/non-toxic)? Or, are there more than two classes (mild, moderate, severe irritants, and how are the boundaries defined); if there are two or more assay endpoints, how are they combined into a final toxicity statement? During validation, the prediction model also needs scrutiny, and the questions asked are as follows: Is there a threshold (different from the statistical threshold) for when an effect can be considered biologically relevant? How is the outcome interpreted when more than one endpoint is measured (e.g., general cytotoxicity and functional impairment or effects on two different cell types)? Is an increase compared to normal good, when a decrease is bad? How should data be interpreted when a compound alters the baseline values for the endpoint (e.g., colored compound in spectrophotometric assays, reducing agents in tetrazolium reduction assays)?

Before validation of a method can be initiated, all the various scriticized recently for its slow progress and potentially faulty outcomes, and alternatives are being considered. The evidence-based toxicology alternative has attempted to suggest alternative validation approaches (Hartung, 2010). It must be noted, however, that these require even more stringent definitions of the above criteria and of assay quality. Technical assay quality assessment is an indispensable step that should occur prior to any further validation steps addressing translation and predictivity.

3 Quality aspects of test systems

The description of a test system for regulatory purposes requires a standard operation procedure (Sesardic et al., 2004). This would, e.g., provide information on source and characterization of cells, a sufficient description of culture conditions for maintenance and experiment, and information on which parameters are critical and what affects them (Coecke et al., 2005). It also includes measurement methods, essential instrumentation, important manipulation steps, details on the determination of endpoints, and a description of the data processing.

Validation of model relevance needs to answer, for instance, the following questions: What human problem is modeled? What biological effect is it designed to measure? Which effects is the test designed to predict? Can it detect deviations from normal to both sides, or does the test work only for one side?

Important assay performance validation questions include: Does a compound that should change the endpoint do this – and by how much does it do this (=dynamics of the response, maximum possible deviation of endpoint); does a compound that is not expected to change the endpoint behave neutrally? It is frequently neglected, although scientifically important, that besides negative (NC) and positive (PC) controls (as above), many systems also require unspecific controls (UC). The response dynamics of a PC, and thus the performance of the test method, cannot be qualified without assessing the response to UC. It is important to re-challenge the test method with a new set of PC and NC (learning set, training set of chemicals) to assess its performance with respect to unknown compounds.

Frequently, test methods should assess specific adverse effects (SAE) independent of general cytotoxicity (GC). For instance, inhibition of neurite outgrowth can be measured meaningfully only in a concentration range that does not kill the cells (Kuegler et al., 2010). The toxicity range of test compounds may be determined as follows: a general cytotoxicity/viability test is run over a wide range of concentrations, initially with 10-fold dilutions. After identification of the relevant range, re-testing is performed in a more narrow range (3-fold dilutions) to identify the highest non-cytotoxic concentration (HNCC) within the conditions of the assay (e.g., a given time frame). For most practical purposes this may be done by using the mathematically defined IC_{10} value of the cytotoxicity concentration response curve and moving to the left by a certain factor (e.g., HNCC = EC_{10} x 0.02). Ideally, GC should be determined in parallel/simultaneously with SAE. Inability to measure GC does not

mean that it does not occur. This applies, in particular, to short-term assays (few hours), as most GC endpoints require several hours to become manifest.

Each experimental setup requires controls to indicate whether the experimental system reacts correctly, i.e., in the right direction, or in the right range. They give us an acceptance criterion for believing the other data obtained from unknown samples by the test method. The concept of acceptance criteria is highly important in all quantitative experimental sciences. Test systems, especially in *in vitro* toxicology, are usually so complex that they require known positive and negative controls to be measured along with the unknown samples (Leist et al., 2010). Only if these controls fulfill the acceptance criteria can the other experimental data be taken into consideration. Data from an experiment that did not fulfill the acceptance criteria cannot be used.

4 Controls and considerations required for the validation of assay predictivity

The predictive power usually is validated by examination of the correlation of assay results with a gold standard. However, correlation does not mean causality, even if the correlation is very good (Balls et al., 2006). On the one hand, the correlation may be real but exist only within a small range or under specific conditions or for a limited class of compounds. On the other hand, the correlation may not really exist but is suggested by the choice of compounds along the continuum of effects. This argument has an important practical implication for test compound selection. For instance, if the question is whether a simple 24 h fibroblast cytotoxicity assay correlates with a complex endpoint, such as chronic toxicity or carcinogenicity, it can be possible to find a good correlation if the 20 test compounds are comprised of 10 compounds of very low cytotoxicity and 10 compounds of high cytotoxicity. Some assays tend to agree when extremes are used, but the resulting (mathematically) good correlations may not hold true for test compounds in the intermediate range. Are such cases relevant and common? Yes, they are, particularly in studies using multiple endpoints. When dozens or hundreds of endpoints are used, such artificial correlations are likely to appear for at least some of them. Typical examples are -omics studies suggesting a correlation between some metabolite or protein modification with toxicity (Leist et al., 2008). For such studies, appropriate statistics use measures to counteract the effect of multiple endpoints on apparent significance of effects (false discovery rate corrections – FDR) (Benjamini et al., 2001).

The minimum information required on the response dynamics is the linear and dynamic range of the endpoint and the detection limit. Moreover, information should be provided regarding how stable (robust) a readout is. For instance, when neurite growth is measured, data are required on the length under optimal conditions (S) and on the variation of length under these conditions (V); in addition, the minimum length (N.B., this is not necessarily zero; it may, for example, be 50% of the maximum length measured in the presence of the strongest known growth inhibitor) that can be observed under the given assay conditions needs to be determined (B). Also, its variation (N) is an essential piece of information. From these data, the signal-noise ratio [S/N-ratio or (S-B)/N)] can be calculated. These data also can be used to define the detection limit [e.g.: B + (5 x N)]. Another quality parameter of the test system (independent of any test compound) is the z' factor, which, ideally, should be >0.5 and indicates the detection power of the system [z' = 1 - ((3 x (V + N)/(S - B)]. The procedures used to determine z' or S/N ratio also are well suited to detect systematic errors in the assay setup.

Toxicity curves do not necessarily follow a simple mathematical model, and they do not need to reach zero (viability) within the tested range of concentrations. For instance, only a subpopulation of cells may be affected. This means that EC_{50} values cannot be extrapolated. A meaningful EC_{50} requires that real data points (ideally ≥2) exist on both sides of the EC_{50}.

5 Validation criteria and the validation process

The validation process itself has evolved over time to allow higher throughput, flexibility, and efficiency. For this, it is important to recall the main elements of an alternative method. As is evident, a test system is involved. This needs to be coupled with analysis endpoints and a data analysis procedure. Sometimes the third component is neglected: the prediction model relating the results of the method to predictions for human safety. A modular approach (Hartung et al., 2004) has been useful to accelerate the validation procedure. First, the reliability of the test system needs to be validated. This includes testing the descriptive assay parameters (accuracy, precision, detection limit, linear range, robustness, specificity, sensitivity, response dynamics) at increasing levels of complexity, i.e., within a laboratory (different operators) and between different laboratories (transferability). In parallel, the mechanistic validity and scientific relevance can be evaluated. In a third line of validation, the predictive capacity is evaluated. Until now, this has been done by correlation of the test results with the results of animal experiments. This process may yield information on applicability domains (e.g., only certain types of chemicals, but not others).

6 Validation by comparison with animal data

The field of alternative methods has tended to focus on one particular aspect of validation: the comparison to animal data. In this sense, validation and the phrases "valid methods" and "validated methods" have been used in legal texts, such as the European regulation on chemicals, REACH, the seventh amendment of the European Cosmetics Directive, and the new directive on the use and protection of experimental animals (2010/63EU). One of the consequences was the creation of a European validation agency in the field of toxicology, the European Centre for Validation of Alternative Methods

(ECVAM) in Ispra (Italy). Comprehensive validation is a prerequisite for the adoption of a new method into a legal framework such as the OECD test guidelines or the European pharmacopoeia.

Validation in comparison with animal data has been criticized frequently. One argument is that animal experiments may not be suitable as a gold standard, as they do not correlate well enough with human data (Bahramsoltani et al., 2009). Another argument is that such a correlative process is not possible when test batteries are used that do not model a defined animal experiment (Hartung, 2008). Therefore, new ideas have been proposed to overcome this problem. The most extreme approaches suggest neglecting the correlation aspects initially and, instead, focusing much more on the first two domains of validation: high quality of the test system and high scientific relevance which may, by themselves, provide a good predictivity for human safety. Such concepts, at present, are being tested and further developed with great speed.

The field of cosmetics is a good example for progress in the establishment and validation of alternative methods: replacement methods for some toxicological domains have been validated. These include phototoxicity, skin corrosion, skin irritation, eye corrosion, and eye irritation. Refinement/reduction methods also are available for acute oral toxicity (altered variants of the LD_{50} test) and skin sensitization (local lymph node assay) (ICCVAM, 2006). Many of these tests have been accepted by the OECD and, to a large extent, some have been substituted for the corresponding animal experiments.

According to current legislation, in 2013, animal testing for cosmetics has to stop in further toxicological domains. These domains include toxicokinetics, skin sensitization, repeated dose toxicity, carcinogenicity, and reproductive toxicity. A recent report published by the European Commission stated that sufficiently validated methods are not yet available in these domains. This opinion was confirmed by a large expert panel (Hartung et al., 2011) assembled by the Center for Alternatives to Animal Testing in Europe – CAAT-Europe (Daneshian et al., 2010). Thus, test development and validation are well under way, with high pressure in these domains.

7 Toxicological and other methods that have been validated

More than 80 methods have been validated or are in some more or less advanced state of validation. These include more than 50 *in vitro* tests, 10 using isolated organs, several refined *in vivo* tests, and testing strategies that combine *in vitro* and *in vivo* approaches. *In vitro* is defined as: "no animals are involved," and the test is based on cell systems or isolated organs. Refined *in vivo* methods often involve the use of anesthetics and analgesics, and humane endpoints are applied. Furthermore, the development of tiered testing or testing strategies reduces the number of animals involved.

Many alternative methods are anchored in OECD (Organization for Economic Cooperation and Development) Guidelines. The guidelines for the testing of chemicals, as stated on the OECD website[1] "*are a collection of about 100 of the most relevant internationally agreed testing methods used by government, industry, and independent laboratories to identify and characterize potential hazards of new and existing chemical substances, chemical preparations, and chemical mixtures. They are a basic set of tools used primarily in regulatory safety testing and subsequent chemical and chemical product notification and chemical registration. In addition, they also can be used for the selection and ranking of candidate chemicals during the development of new chemicals and products and in toxicology research*."

Another important source is the European Pharmacopeia, and their mission is stated on their website[2]: "*The texts of the European Pharmacopoeia (Ph. Eur.) concern the qualitative and quantitative composition of medicines, the tests to be carried out on medicines, on the raw materials used in the production of medicines and on the intermediates of synthesis. It contains texts covering substances, excipients and preparations for pharmaceutical use of chemical, animal, human or herbal origin, homoeopathic preparations and homoeopathic stocks, antibiotics, as well as dosage forms and containers. The texts also cover biologicals, blood and plasma derivatives, vaccines and radiopharmaceutical preparations. They are legally binding*."

In the US, the Office of Chemical Safety and Pollution Prevention (OCSPP)[3], under the umbrella of the Environmental Protection Agency (EPA), addresses the harmonization of chemical and pesticide testing.

Given that the area of test methods is under permanent development, it is rather challenging to keep track of the current situation. There are several sources available that try to document the status of 3R methods, but none can claim to be complete.

Information databases that may be consulted include:

- AltTox.org (update 27.09.2011, used for this survey). http://alttox.org/ttrc/validation-ra/validated-ra-methods.html
- The Canadian Council of Animal Care in Science (CCAC/CCPA) (most information of April 2009, used for this survey). http://3rs.ccac.ca/en/searches-and-animal-index/ccac-reference-database.html
- The European Commission through the responsible Institute for Health and Consumer Protection (IHCP) on the website of the European Centre for the Validation of Alternative Methods (ECVAM) (last update 30.06.2011, used for this survey). http://tsar.jrc.ec.europa.eu/documents/TSAR_public_ongoing_validation_studies_2011-06-30.pdf

A summary of the most prominent and widely accepted methods is provided below. The implementation of such assays in regular testing differs considerably between countries, institutions, and exact data requirements.

[1] http //www oecd ilibrary org/environment/oecd guidelines for the testing of chemicals section 4 health effects 20745788
[2] http //www edqm eu/en/European Pharmacopoeia 1401 html
[3] http //www epa gov/aboutepa/ocspp html

Acute aquatic toxicity

One validated test is anchored in the OECD TG 203[4], using an upper threshold concentration (UTC) step-down approach, which reduces the number of fish used by 65% (Hutchinson et al., 2003; Jeram et al., 2005; ECVAM, 2006b). Another test is under validation by ECVAM for an OECD Project to assess the transferability and reliability of the zebrafish embryo toxicity test for prediction of acute toxicity to fish. It is expected to be ready for implementation in 2012 (Selderslaghs et al., 2009, 2010).

Acute mammalian toxicity

Acute mammalian toxicity is divided into three subareas by their route of application. For the oral route, three tests have been validated and were implemented in the OECD TG 420[5], 423[6], 425[7]. All three methods reduce the number of animals used from 25 to 5-9 (van den Heuvel et al., 1990; Schlede et al., 1992, 1995; Diener and Schlede, 1999). The Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) recommended two *in vitro* tests to be implemented into a tiered testing strategy to reduce the number of animals further (ICCVAM, 2001). Another test is under validation by ECVAM and is considered to be a follow-up validation study on the predictive capacity of the 3T3/Neutral Red Uptake cytotoxicity test to identify non-toxic substances for acute oral toxicity and its potential inclusion into an in vitro testing strategy for acute oral toxicity, which is expected to be finalized in 2012 (Stokes et al., 2006).

With respect to inhalation exposure, the original OECD TG 403[8] is under revision to implement two validated tests that suggest humane endpoints and therefore are considered refinement methods.

One test is available for acute dermal toxicity, and it also applies humane endpoints. This may lead to a new OECD document (Draft TG 434).

Non-vaccine biologics

The Mouse LD_{50} Assay for Botox Potency Testing is a prominent example in the area of non-vaccine biologics. Eight alternative assays are available in different stages of regulatory acceptance. Some tests may have a large economic impact, as they are proprietary and implementation in a guideline would enforce their use by potential competitors. The Snap-25 test is listed as a method for replacement in the European Pharmacopeia for final batch testing (Ekong et al., 1997; Gaines Das et al., 1999; Sesardic et al., 2004), while two other assays are recognized by ICCVAM but further development is recommended (ICCVAM, 2008). There are three non-lethal mouse models, two listed in the European Pharmacopeia, and one is accepted only for BoNT type A (ICCVAM, 2008). Furthermore, there are two organ models; one is listed in the European Pharmacopeia (ICCVAM, 2008). The test for calcitonin bioactivity developed by Novartis is another example of an accepted alternative in the field of biologics (Hartung, 2001).

Vaccines

The testing of vaccines depends on their intended use, human or veterinary, and the testing addresses the vaccine's potency or safety separately.

For vaccine potency in veterinary use, the lethal challenge test was replaced by an enzyme-linked immunosorbent assay (ELISA), a biochemical analytical approach. It is implemented in the European Pharmacopeia, e.g., swine erysipelas vaccine (Pastoret et al., 2002; Rosskopf-Streicher et al., 1999, 2001).

In testing the vaccine potency for human use, seven tests are implemented in the European Pharmacopeia. The lethal paralytic challenge test for batch potency of tetanus toxoid vaccines may be replaced by an ELISA measurement (Balls and Hellsten, 2000b) and a toxin binding inhibition method (Balls and Hellsten, 2000a); the diphtheria vaccine may be tested via a cell-based assay and an ELISA (Council of Europe, 2008). Hepatitis B and Poliomyelitis vaccines are tested via serological antigen quantification (Council of Europe, 2008), and Rabies potency testing is done by using only one dilution, and humane endpoints are applied (Council of Europe, 2008).

The formerly used target animal vaccine safety test for veterinary use could be dropped as a result of a retrospective study conducted by ECVAM.

In the area of vaccine safety for human use, the following four tests are available: i) the abnormal toxicity test can be deleted from the testing scheme when batch consistency can be demonstrated (Schwanig et al., 1997); ii) the oral polio neurovirulence test conducted in monkeys may be replaced by an *in vitro* test called "MAPREC," but only for type 3 oral polio virus vaccines (WHO, 2005); iii) the use of transgenic mice instead of monkeys (TgPVR21) was validated by the World Health Organization (WHO) for type 1,2,3 oral polio vaccines (Dragunsky et al., 2003); and iv) the residual toxicity of diphtheria may be replaced by the Vero Cell Test (Council of Europe, 2008).

Chronic toxicity

In the area of chronic toxicity for pesticides, the 1-year dog study was found to be unnecessary based on a statement of the ECVAM Scientific Advisory Committee (ESAC) and the US Environmental Protection Agency (EPA). It was found that the 1-year study does not provide more information than the 90 day study, but some countries still require these data (OECD, 1998).

Eye corrosion and irritation

For eye corrosion and irritation studies ICCVAM will implement the routine use of anesthetics, systemic analgesics and humane

[4] http //www oecd org/dataoecd/17/20/1948241 pdf
[5] http //iccvam niehs nih gov/SuppDocs/FedDocs/OECD/OECD GL420 pdf
[6] http //iccvam niehs nih gov/SuppDocs/FedDocs/OECD/OECD GL423 pdf
[7] http //www oecd org/dataoecd/17/51/1948378 pdf
[8] http //www oecd org/dataoecd/54/55/41761261 pdf

endpoints, as well as several validated tests, laid down in the OECD TG 437[9] and 438[10], and an OECD GD supplement.

Furthermore, there is one test under validation by ECVAM and the European Cosmetics Association (COLIPA) to assess the transferability, reliability, and predictive capacity of two in vitro test methods based on reconstructed human tissue models, to be used as stand-alone test methods to identify chemicals not classified as eye irritant (non-irritant) (McNamee et al., 2007).

Food safety

In the area of food safety, two tests have been validated to replace the Mouse Bioassay for shellfish toxins (PSP). One screening method and a high performance liquid chromatography (HPLC) approach were accepted in the EU in 2010 (Jellett et al., 2002; Mackintosh et al., 2002; FAO/IOC/WHO, 2004).

Carcinogenicity

There are two tests under validation for carcinogenicity to assess protocol standardization, transferability, and reproducibility (but not performance) of three protocols of cell transformation assays: the Syrian hamster embryo (SHE) pH 6.7, the Syrian hamster embryo (SHE) pH 7.0, and the BALB/c 3T3 assays. Furthermore, a validation study is underway to verify if the Bhas 42 cells-based cell transformation assay might be an equivalent (Combes et al., 1999; Maurici et al., 2005; OECD, 2007). Results are expected in 2012.

Genotoxicity

The area of genotoxicity is covered by eight validated *in vitro* tests, which are part of a tiered testing strategy to reduce the number of animals. They are reflected in several OECD documents, and two *in vitro* comet assays are under validation (OECD, 1986a,b,c,d; OECD, 1997a,b,c).

Hematotoxicity

One hematotoxicity test for acute neutropenia (CFU-GM) has been validated by ECVAM. The test can be applied instead of a second animal species. Therefore, it is not considered a replacement, though a reduction in the number of animals is achieved (Pessina et al., 2003).

Phototoxicity

To determine phototoxicity, the European Commission accepted the in vitro neutral red uptake (NRU) phototoxicity test (OECD, 2004a) as method B.41 in Annex V of the EU Council Directive 67/548/EEC[11]. Animal methods to detect phototoxic effects of chemicals are prohibited in all Member States.

Pyrogenicity

To replace the rabbit pyrogen test, five *in vitro* tests based on human cell models have been validated by ECVAM (ECVAM, 2006c). They can be used to detect gram-negative mediated pyrogenicity. The official European Pharmacopeia listed test, the Limulus amoebocyte lysate assay (LAL), lacks the capability of detecting gram-positive stimuli. The cell-based assays also may be useful for gram-positive mediated pyrogenicity (ECVAM, 2006c; NICEATM-ICCVAM, 2007, 2008, 2009; Poole et al., 2003; Hoffmann et al., 2005a,b). This might lead to a full replacement of the rabbit test in the near future.

Reproductive and developmental toxicity

Due to the complexity of the reproductive cycle and the importance of the developmental process, not many alternatives are available in these areas. The OECD only recently accepted the extended one-generation study (OECD, 2008), which replaces the two-generation study (OECD, 1983). Furthermore, as stated in OECD TG 415[12]: "*For reproductive endpoints, it is envisaged that, as a first step and when available, information from repeat-dose studies (including screening reproductive toxicity studies, e.g., TG 422), or short term endocrine disrupter screening assays (e.g., Uterotrophic assay – TG 440; and Hershberger assay – TG 441) are used to detect effects on reproductive organs for males and females. This might include spermatogenesis (testicular histopathology) for males and estrous cycles, follicle counts/oocyte maturation, and ovarian integrity (histopathology) for females. The Extended One-Generation Reproductive Toxicity Study then serves as a test for reproductive endpoints that require the interaction of males with females, females with conceptus, and females with offspring and the F1 generation until after sexual maturity.*"

There are also two ECVAM-validated *in vitro* tests using embryos from animals (ECVAM, 2006a; Spielmann et al., 2006). In addition, there is one stem cell-based test (EST) available (ECVAM, 2002), which ECVAM recommended to be part of a tiered testing strategy.

Endocrine active substances

There are two OECD accepted methods, anchored in the OECD TG 455[13] and 456[14]; they may be used for screening purposes.

The US EPA accepted a Tier 1 screening battery including, besides several *in vivo* assays, five *in vitro* tests that have been accepted by the Office of Prevention, Pesticides and Toxic Substances (OPPTS) and laid down as legally binding guidelines for the US as Series 890 OPPTS[15].

There are two methods under validation by ICCVAM, ECVAM, and the Japanese Centre for the Validation of Alternative Methods (JaCVAM) to assess the transferability and reliability of the assays to rank chemicals according to their potency for estrogen

9 http //iccvam niehs nih gov/SuppDocs/FedDocs/OECD/OECD TG437 pdf
10 http //ecvam jrc it/ft doc/OECD%20TG%20438 pdf
11 http //eur lex europa eu/LexUriServ/LexUriServ
12 http //www oecd org/dataoecd/23/10/46466062 pdf
13 http //iccvam niehs nih gov/SuppDocs/FedDocs/OECD/OECD TG455 pdf
14 http //bit ly/W7igDK
15 http //www epa gov/ocspp/pubs/frs/publications/Test Guidelines/series890 htm

receptor activation or suppression for use as a building block in future testing strategies to detect endocrine active compounds. Evaluation is expected to be finished in 2012.

Skin

There are many tests available for hazard assessment regarding the human skin. They are divided into four different areas: absorption, penetration, corrosion, and irritation.

For skin absorption and dermal penetration, a regulatory accepted dermal percutaneous test is available, which may replace the animal test when human skin is used (OECD, 2004b). More information is available in the OECD guidance document[16].

For skin corrosion, three different methods are available, all integrated in OECD guidelines (OECD, 2004c, d, 2006), but their use is specified. For example, the test "Corrositex" can be used to identify acids and bases, and substances that are identified as corrosives will not proceed further to the animal test. The "TER" test can distinguish between corrosives and non-corrosives, but non-corrosives will require further confirmation by an animal test. The human skin models (EpiSkin™, EpiDerm™, SkinEthic™) are accepted in the EU as full replacements for corrosivity testing anchored in Regulation 440/2008/EC. In the US these tests can be used to exclude corrosives, while negative results lead to an animal test.

Skin irritation can be detected via the above mentioned human skin models. Tests may be adopted in a new OECD draft guideline and may be used to identify irritants.

Dermal sensitization

Sensitization is detected by local lymph node *in vivo* assay (LLNA) which is recommended by ICCVAM as a stand-alone substitute for the guinea-pig sensitization test (OECD, 2002; ICCVAM, 1999; ISO, 2002; EPA, 2003; ECVAM, 2000). The reduced rLLNA is able to distinguish between sensitizers and non-sensitizers, and if a chemical is negative in the rLLNA, it will not proceed to the full LLNA, which results in the use of fewer animals (ECVAM, 2007). There are three methods under validation by ECVAM, ICCVAM, and JaCVAM to assess the assay's transferability and reliability in view of future incorporation into a testing strategy for full replacement of current regulatory animal tests. COLIPA, the EU project *Sens-it-iv*, and many others have developed a whole battery of pure *in vitro* assays that potentially could lead to a full animal replacement within a few years.

Toxicokinetics and metabolism

The area of toxicokinetics and metabolism is considered to be very complex and difficult to model. Nevertheless, two assays are under validation by ECVAM, ICCVAM, and JaCVAM to assess the transferability and reliability of measuring liver enzyme (Cytochrome P450) induction using the human cryoHepaRG® cell line and cryopreserved human hepatocytes to provide a human metabolically-competent model for use in future testing.

For the field of kinetics testing, as for sensitization, carcinogenicity, toxicokinetics, repeat dose toxicity, and reproductive toxicity, the status of replacement methods has been reviewed extensively (Adler et al., 2011; Hartung et al., 2011).

8 Old versus new approaches to validation

Validation approaches are closely linked to the concept initially used for model development. In this context, it is important to recall that there are two fundamentally different ways of constructing test systems, which we call here (a) "correlative approach," and (b) "re-constructive approach."

(a) Correlative approach

The correlative approach has been used most frequently for the establishment of alternative methods for animal experimentation and, therefore, the whole theoretical concept of "validation" has been adapted to this approach. In brief, this approach uses the test method as input-output system. Validation in this case is largely concerned with an evaluation of how well input correlates with output. The model itself often is a kind of black box, with only limited information available on the relevant processes and reactions inside. This has the key advantage that reasonable correlations can be obtained without the need for knowledge on mechanisms of toxicity, or of regulatory mechanisms within the model. The disadvantage, obviously, is that the relevant processes often are not known.

Examples illustrate this situation best. The first is the mouse cancer bioassay. When it was established, it was a true black box model. The rationale was only that some correlation was expected between compounds known to be carcinogenic in man and the ones triggering tumorigenesis in mice. Mechanisms of carcinogenesis were largely unknown and did not necessarily need to be known for this model. Very powerful carcinogens and clear non-carcinogens correlated nicely between this model and the situation in man. Problems became obvious when a lack of correlation was observed for several classes of compounds. For instance, the so-called peroxisome proliferators triggered hepatocarcinogenesis in the mouse but not in man.

Another example from the same field of toxicology is the Ames bacterial mutagenesis assay, which was introduced to detect mutagens, at that time believed to be carcinogens as well. Back-mutation of errors in the bacterial genes coding for histidine synthesis obviously have no resemblance or biological relevance with respect to human carcinogenesis, but the model achieved a reasonably good overall correlation and therefore was widely accepted. It only became evident later that about half of the human carcinogens are non-mutagens and, therefore, cannot be detected in this assay. A correction or adaptation of the assay is not possible, as it does not reflect human biology. It was established simply as a correlative black-box model.

A third example also comes from the field of carcinogenesis. The human cell transformation assay predicts mutagenic and

[16] http //www oecd org/dataoecd/0/1/46257610 pdf

epigenetic carcinogens with an astonishingly high specificity. It is still unclear why the assay works and what the underlying biological principle is. On the basis of correlations with human or animal data, however, the validation of the assay is far advanced.

The strength of this correlative model setup is proven by the assays that have been developed on this basis and have been successfully validated and used. Model development was possible without the requirement for in-depth biological knowledge. The weaknesses are demonstrated on the example of the embryonic stem cell test (EST). The EST uses murine embryonic stem cells (EST). They are differentiated with a very rough protocol to mixed cultures containing cardiomyocytes and pacemaker cells, resulting in patches of cells that beat spontaneously. Compounds are being tested for their potential to inhibit the development of these beating cell clumps. In initial validations, the assay was found to predict teratogens with high specificity and sensitivity, and it was recommended by ECVAM and the ECVAM Scientific Advisory Committee (ESAC) for regulatory use. A biological characterization had never taken place, and the mechanisms and regulations underlying this assay were never characterized. The use of a small number of validation compounds and the absence of biological knowledge harbors some dangers, as demonstrated by the history of this assay. In a broader validation with compounds chosen within the context of the ReProTect study, the assay failed (Marx-Stoelting et al., 2009).

A priori, it may not seem necessary to understand an assay as long as it delivers good (= predictive and reproducible) results. Toxicological testing has largely adopted this approach, not just *in vivo*, but also *in vitro*. However, there are strong reasons to move ahead to mechanism-based *in vitro* tests to attribute a scientific rationale to the correlations found in new test systems. Paradoxically, modern technologies are especially likely to settle for black box approaches and blind correlations. Such approaches bear the risk of measuring trivialities if they are not based on a mechanistic rationale. For example: new metabolomic or transcriptomic fingerprints to predict complex forms of toxicity (e.g., developmental toxicity) may indeed only be expensive and sensitive measures of classical cytotoxicity. Results only gain scientific validity when they are controlled by various approaches and when falsification attempts of their predictions have failed.

(b) Re-constructive approach

The second type of modeling was termed the "re-constructive approach." This name was chosen because such models try to reconstruct reality using biological information and mathematical relationships between model parameters as building blocks. This approach requires an understanding of the biological process to be modeled, not only in qualitative, but also in quantitative terms. A biological process needs to be dissected into all its components. Each component needs to be understood. Moreover, the relationships between the components need to be understood, and mathematical approaches need to be developed to describe the relationship between all components and parameters. Finally, these elements can be used for "re-constructing" reality as closely as possible. An example is physiology-based pharmacokinetic (PBPK) modeling. The corresponding black box model is the injection of compounds into animals and the evaluation of their pharmacokinetic behavior (time course of plasma concentrations, urinary excretion …) and the correlation of this information with the expected behavior in man. PBPK modeling would use information on hepatic metabolism, solubility, lipophilicity, and renal excretion to model the behavior of the drug in a human body, using a set of differential equations. The validation of such models would refer not only to the input-output correlation but also to the construct of the model. This is a difficult task, and firm guidelines for this have not yet been established.

It is noteworthy that, in reality, the two extremes of black box modeling and pure reconstruction barely exist. Often, the approaches are combined to some extent. For instance, PBPK models would use information obtained from rodent models. Then information would be used from human and rodent liver metabolism, and this information then would be used to translate rodent information better into human information in an optimized PBPK model using the so-called parallelogram approach. Other examples, below, illustrate the incorporation of biological and mechanistic information into correlative models. For instance, in the case of skin irritation, originally the damage to skin was measured by classical viability assays. Attempts to account for inflammatory processes and active reactions of cells in the skin by measurement of chemokine release are ongoing. Also, in the field of sensitization, biological information is incorporated into available models. One approach, for instance, tested the effect of keratinocyte addition in a dendritic cell activation model to reflect their normal biological presence and role in co-stimulation and haptenization.

The validation of integrated testing strategies provides new dimensions of challenges. This will somehow require the validation not only of individual components but also of the relationships established between them and used for the overall modeling. Thus, this form of validation combines issues from the two types of model validation, correlative and re-constructive, discussed above. The challenges of such an approach may be illustrated by the example of dermal sensitization. An integrated test battery may involve a haptenization assay, measuring the covalent binding of the chemical to a peptide. It also may involve some physicochemical characterization to be used to predict skin penetration. A dendritic cell activation assay, in the absence or presence of keratinocytes, would be added. Eventually, T cell stimulation may be probed as well. Then the test strategy parts will have to be linked and weighted. One simple rule may be: if a compound is positive in one of the assays, it is considered a sensitizer. More complex sets of rules would use a hierarchical decision setup. For instance, compounds unlikely to penetrate the skin or without chemical reactivity may be considered of low hazard, even in the presence of positive dendritic cell activation. Such integrated testing strategies (ITS) may again

be validated by a correlative approach of the overall ITS versus reality. This most likely will be the first and most immediate solution in the near future. Re-constructive validations of such approaches will require huge amounts of data and experience, but they may become necessary in cases in which little "reality" information is available, such as the area of developmental neurotoxicity.

9 Outlook on validation in the field of developmental neurotoxicity (DNT)

Developmental neurotoxicity is an area that requires such new validation concepts, as not enough animal data are available. A recent review revealed that just over 100 compounds have been tested in studies using the OECD 426 draft guideline on developmental neurotoxicity. Most of these compounds were pesticides (66%); only eight industrial chemicals were included. Another review identified 174 compounds for which neurobehavioral risk assessment had been performed, in many cases on the offspring of the exposed animals (F1 generation) as well. Only 1% of these compounds were industrial chemicals, and thus the available data regarding the developmental neurotoxicity of industrial chemicals is rather limited. For some compounds, developmental neurotoxicity is the most sensitive of all toxicity endpoints evaluated in a broad safety evaluation battery. Thus, although developmental neurotoxicity appears to be an important domain of safety evaluation, test capacity is limited and test costs are extremely high. This puts pressure on the development of faster and cheaper *in vitro* systems that can predict developmental neurotoxicity, give information comparable of behavioral readouts, and facilitate screening, or at least prioritization of relevant drugs and chemicals for further testing. We envision that future *in vitro* test systems for developmental neurotoxicity will combine the above validation approaches with exposure information, and we suggest a strategy for test system development and cell-based risk assessment.

We propose that the emerging knowledge from molecular and cellular neuroscience and mechanistic neurotoxicology can be exploited to design *in vitro* tests that read out cellular and molecular endpoints that are predictive of behavioral signs of neurotoxic exposures in humans. For acute or more chronic neurotoxic effects, the onset of these effects is temporally associated with the onset of the chemical exposure and usually follows a dose-response relationship. The discipline of developmental/neurodevelopmental toxicology faces an additional problem, however. It is difficult to provide evidence for cause-effect relationships for processes with a long time lag between exposure (e.g., gestational) and effect (e.g., adult life). Suitable test systems for delayed effects need to be identified. Research in this particular area is motivated by increasing incidence of neurodevelopmental disorders such as autism, ADHD, and schizophrenia, as well as the growing awareness that environmental factors influence susceptibility to and/or severity of these diseases.

References

Adler, S., Basketter, D., Creton, S., et al. (2011). Alternative (non-animal) methods for cosmetics testing: current status and future prospects – 2010. *Arch Toxicol 85*, 367-385.

Bahramsoltani, M., Plendl, J., Janczyk, P., et al. (2009). Quantitation of angiogenesis and antiangiogenesis in vivo, ex vivo and in vitro – an overview. *ALTEX 26*, 95-107.

Balls, M. and Hellsten, E. (2000a). Statement on the application of the ELISA procedure for batch potency testing of tetanus vaccines for human use, I.f.H.C.P. Joint Research Centre, ECVAM Unit (ed.). Ispra, Italy: European Commission.

Balls, M. and Hellsten, E. (2000b). Statement on the application of the tobi test for batch potency testing of tetanus vaccines for human use, I.f.H.C.P. Joint Research Centre, ECVAM Unit (ed.). Ispra, Italy: European Commission.

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.

Benjamini, Y., Drai, D., Elmer, G., et al. (2001). Controlling the false discovery rate in behavior genetics research. *Behav Brain Res 125*, 279-284.

Burgess, C. (2001). *Valid Analytical Methods and Procedures*. 1st edition. Cambridge, UK: Royal Society of Chemistry.

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.

Combes, R., Balls, M., Curren, R., et al. (1999). Cell transformation assays as predictors of human carcinogenicity – the report and recommendation of ECVAM workshop 39. *ATLA 27*, 745-767.

Council of Europe (2008). 2.7.6 Assay of Diphtheria vaccine (adsorbed). In *European Pharmacopoeia*. 6th edition. Strasbourg, France.

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.

Diener, W. and Schlede, E. (1999). Acute toxic class methods: Alternatives to LD/LC_{50} tests. *ALTEX 16*, 129-134.

Dragunsky, E., Nomura, T., Karpinski, K., et al. (2003). Transgenic mice as an alternative to monkeys for neurovirulence testing of live oral poliovirus vaccine: validation by a WHO collaborative study. *Bull World Health Organ 81*, 251-260.

ECVAM (2000). ESAC statement on the validity of the Local Lymph Node Assay for skin sensitisation testing. Ispra, Italy: ECVAM.

ECVAM (2002). Embryonic Stem Cell Test (EST): INVITTOX no. 113. Embryotoxicity/Teratogenicity. Ispra, Italy: ECVAM.

ECVAM (2006a). ESAC Statement on the scientific validity of the upper threshold concentration (UTC) step-down approach – A new testing strategy to reduce the use of fish in acute aquatic toxicity testing. Ispra, Italy: ECVAM.

ECVAM (2006b). ESAC statement on the validity of in-vitro pyrogen tests. Ispra, Italy: ECVAM.

ECVAM (2006c). The Micromass Test – Method of Brown:

INVITTOX no122. Embryotoxicity/Teratogenicity. Ispra, Italy: ECVAM.

ECVAM (2007). ESAC Statement on the reduced local lymph node assay (rLLNA). Ispra, Italy: ECVAM.

Ekong, T. A., Feavers, I. M., and Sesardic, D. (1997). Recombinant SNAP-25 is an effective substrate for Clostridium botulinum type A toxin endopeptidase activity in vitro. *Microbiology 143*, 3337-3347.

EPA (2003). Health effects test guideline: OPPTS 870.2600 skin sensitization. EPA 712-C-03-197. Washington DC, USA.

FAO/IOC/WHO (2004). Report of the joint FAO/IOC/WHO ad hoc expert consultation on biotoxins in bivalve molluscs. Oslo, Norway.

Gaines Das, R. E., Heath, A. B., Martin, H., and Sesardic, D. (1999). Validation of in vitro assays for botulinum toxin: a case study. *Dev Biol Stand 101*, 267-276.

Hartung, T. (2001). Three Rs potential in the development and quality control of pharmaceuticals. *ALTEX 18, Suppl 1*, 3-13.

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

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

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

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

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 29*, 183-209.

Hoffmann, S., Luderitz-Puchel, U., Montag, T., and Hartung, T. (2005a). Optimisation of pyrogen testing in parenterals according to different pharmacopoeias by probabilistic modelling. *J Endotoxin Res 11*, 25-31.

Hoffmann, S., Peterbauer, A., Schindler, S., et al. (2005b). International validation of novel pyrogen tests based on human monocytoid cells. *J Immunol Methods 298*, 161-173.

Hutchinson, T. H., Barrett, S., Buzby, M., et al. (2003). A strategy to reduce the numbers of fish used in acute ecotoxicity testing of pharmaceuticals. *Environ Toxicol Chem 22*, 3031-3036.

ICCVAM (1999). The Murine Local Lymph Node Assay: A test method for assessing the allergic contact dermatitis potential of chemicals/compounds. NIH publication No. 99-4494. National Institute of Environmental Health Research Triangle Park, North Carolina, USA.

ICCVAM (2001). Guidance document on using in vitro data to estimate in vivo starting doses for acute toxicity. NIH Publication 01-4500. National Institute of Environmental Health Research Triangle Park, North Carolina, USA. http://www.epa.gov/HPV/pubs/general/nih2001b.pdf

ICCVAM (2006). ICCVAM test method evaluation report: In vitro cytotoxicity test methods for estimating starting doses for acute oral systemic toxicity tests, NIH Publication No. 07-4519. National Institute of Environmental Health Research Triangle Park, North Carolina. http://iccvam.niehs.nih.gov/docs/acutetox_docs/BRD_TMER/TMERmain_Nov2006.pdf

ICCVAM (2008). Report on the ICCVAM-NICEATM/ECVAM Scientific Workshop on Alternative Methods to Refine, Reduce or Replace the Mouse LD_{50} Assay for Botulinum Toxin Testing. NIH Publication 08-6416. National Institute of Environmental Health Research Triangle Park, North Carolina, USA. http://iccvam.niehs.nih.gov/docs/biologics-docs/BoNTwkshprept.pdf

ISO (2002). Biological evaluation of medical devices. Tests for irritation and delayed type hypersensitivity. BS EN ISO 10993-10:2009.

Jellett, J. F., Roberts, R. L., Laycock, M. V., et al. (2002). Detection of paralytic shellfish poisoning (PSP) toxins in shellfish tissue using MIST Alert, a new rapid test, in parallel with the regulatory AOAC mouse bioassay. *Toxicon 40*, 1407-1425.

Jeram, S., Sintes, J. M., Halder, M., et al. (2005). A strategy to reduce the use of fish in acute ecotoxicity testing of new chemical substances notified in the European Union. *Regul Toxicol Pharmacol 42*, 218-224.

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.

Leist, M., Hartung, T., and Nicotera, P. (2008). 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.

Mackintosh, F. H., Gallacher, S., Shanks, A. M., and Smith, E. A. (2002). Assessment of MIST Alert, a commercial qualitative assay for detection of paralytic shellfish poisoning toxins in bivalve molluscs. *J AOAC Int 85*, 632-641.

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

Maurici, D., Aardema, M., Corvi, R., et al. (2005). Carcinogenicity. *Altern Lab Anim 33, Suppl 1*, 177-182.

McNamee, P., Declercq, L., Smedt, A. D., et al. (2007). Update on the Colipa research programme for development of in vitro alternative methods for eye irritation. in 6th World Congress on Alternatives & Animal Use in the Life Sciences. Tokyo, Japan. *AATEX 14, Spec Issue*, 337-341.

NICEATM-ICCVAM (2007). Independent peer review panel report: Five in vitro test methods proposed for assessing potential pyrogenicity of pharmaceuticals and other products.

NICEATM-ICCVAM (2008). ICCVAM background review document: Validation status of five in vitro test methods proposed for assessing potential pyrogenicity of pharmaceuticals and other products.

NICEATM-ICCVAM (2009). In vitro pyrogen test methods: Overview and recommendations.

OECD (1983). One-generation reproduction toxicity study. OECD Guidelines for Testing of Chemicals. Test Guideline No. 415. OECD, Paris.

OECD (1986a). Genetic Toxicology: In Vitro Sister Chromatid Exchange Assay in Mammalian Cells. OECD Guidelines for Testing of Chemicals. Test Guideline 479. OECD, Paris.

OECD (1986b). Genetic Toxicology Saccharomyces cerevisiae,

Gene Mutation Assay. OECD Guidelines for Testing of Chemicals.Test Guideline No. 480. OECD, Paris.
OECD (1986c). Genetic Toxicology: Saacharomyces cerevisiae, Mitotic Recombination Assay. OECD Guidelines for Testing of Chemicals. Test Guideline 481. OECD, Paris.
OECD (1986d). Genetic Toxicology: DNA Damage and Repair, Unscheduled DNA synthesis in Mammalian Cells in vitro. OECD Guidelines for Testing of Chemicals. Test Guideline 482. OECD, Paris.
OECD (1997a). Bacterial Reverse Mutation Test. OECD Guidelines for Testing of Chemicals. Test Guideline 471. OECD, Paris.
OECD (1997b). In Vitro Mammalian Chromosome Aberration Test. OECD Guidelines for Testing of Chemicals. Test Guideline 473. OECD, Paris.
OECD (1997c). In Vitro Mammalian Cell Gene Mutation Test. OECD Guidelines for Testing of Chemicals. Test Guideline 476. OECD, Paris.
OECD (1998). Repeated Dose 90-Day Oral Toxicity Study in Non-Rodents. OECD Guidelines for Testing of Chemicals. Test Guideline 409. OECD, Paris.
OECD (2002). Skin Sensitisation: Local Lymph Node Assay. OECD Guidelines for Testing of Chemicals. Test Guideline 429. OECD, Paris.
OECD (2004a). In Vitro 3T3 NRU Phototoxicity Test. OECD Guidelines for Testing of Chemicals. Test Guideline 432. OECD, Paris.
OECD (2004b). Skin Absorption: In Vitro Method. OECD Guidelines for Testing of Chemicals. Test Guideline 428. OECD, Paris.
OECD (2004c). In Vitro Skin Corrosion: Transcutaneous Electrical Resistance Test (TER). OECD Guidelines for Testing of Chemicals. Test Guideline 430. OECD, Paris.
OECD (2004d). In Vitro Skin Corrosion: Human Skin Model Test. OECD Guidelines for Testing of Chemicals. Test Guideline 431. OECD, Paris.
OECD (2006). In Vitro Membrane Barrier Test Method for Skin Corrosion. OECD Guidelines for Testing of Chemicals. Test Guideline 435. OECD, Paris.
OECD (2007). Detailed review paper on cell transformation assays for detection of chemical carcinogens (Draft 4th version). OECD Environment, Health and Safety Publications Series on Testing and Assessment No. 31.
OECD (2008). Extended One-Generation 15 Reproductive Toxicity. OECD Guidelines for Testing of Chemicals. Test Guideline 416. OECD, Paris.
Pastoret, P. P., Sitrin, R., Merten, O. W., et al. (2002). New vaccine production technologies and their impact on the use of animals. *Dev Biol (Basel) 111*, 219-220.
Pessina, A., Albella, B., Bayo, M., et al. (2003). Application of the CFU-GM assay to predict acute drug induced neutropenia: an international blind trial to validate a prediction model for the maximum tolerated dose (MTD) of myelosuppressive xenobiotics. *Toxicol Sci 75*, 355-367.
Poole, S., Mistry, Y., Ball, C., et al. (2003). A rapid "one-plate" in vitro test for pyrogens. *J Immunol Methods 274*, 209-220.
Rosskopf-Streicher, U., Johannes, S., Wilhelm, M., et al. (1999). Potency testing of swine erysipelas vaccines by serology results of a pre-validation study. *ALTEX 16*,123-128.
Rosskopf-Streicher, U., Johannes, S., Wilhelm, M., and Cussler, K. (2001). Quality control of inactivated erysipelas vaccines: results of an international collaborative study to establish a new regulatory test. *Vaccine 19*, 1477-1483.
Sesardic, D. Jones, R. G., Leung, T., Alsop, T., et al. (2004). Detection of antibodies against botulinum toxins. *Mov Disord 19, Suppl 8*, S85-S91.
Selderslaghs, I. W., Van Rompay, A. R., De Coen, W., and Witters, H. E. (2009). Development of a screening assay to identify teratogenic and embryotoxic chemicals using the zebrafish embryo. *ReprodToxicol 28*, 308-320.
Selderslaghs, I. W., Hooyberghs, J., De Coen, W., and Witters, H. E. (2010). Locomotor activity in zebrafish embryos: a new method to assess developmental neurotoxicity. *Neurotoxicol Teratol 32*, 460-471.
Schlede, E., Mischke, U., Roll. R., and Kayser, D. (1992). A national validation study of the acute-toxic class method – an alternative to the LD_{50} test. *Arch Toxicol 66*, 455-470.
Schlede, E., Mischke, U., Diener, W., and Kayser, D. (1995). The international validation study of the acute toxic class method (oral). *Arch Toxicol 69*, 659-670.
Schwanig, M., Nagel, M., Duchow, K., and Krämer, B. (1997). Elimination of abnormal toxicity test for sera and certain vaccines in the European Pharmacopoeia. *Vaccine 15*, 1047-1048.
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.
Stokes, W. S., Schechtman, L. M., Rispin, A., et al. (2006). The use of test method performance standards to streamline the validation process. *ALTEX 23, Suppl*, 342–345.
van den Heuvel, M. J., Clark, D. G., Fielder, R. J., et al. (1990). The international validation of a fixed dose procedure as an alternative to the classical LD_{50} test. *Food Chem Toxicol 28*, 469-482.
WHO (2005). Final report: IABS scientific workshop on neurovirulence tests of live virus vaccines (7pp.). Geneva, Switzerland.

Correspondence to

Marcel Leist, PhD
Doerenkamp-Zbinden Chair for *In-vitro*
Toxicology and Biomedicine and Center for
Alternatives to Animal Testing in Europe
University of Konstanz
78467 Konstanz
Germany
e-mail: marcel.leist@uni-konstanz.de