Thomas Hartung, ECVAM

Food for Thought ... on Validation

Food for thought?

Publishing a scientific journal in the internet era is an enormous challenge: Articles are commonly retrieved by browsing databases, with the consequence that fewer and fewer people hold the journal in their hands and actually screen the contents of an issue. Many miss what makes a scientific journal different from a repository of articles – the active forum of exchange of a specific section of the scientific community. With my new series of "Food for Thought" articles *ALTEX* aims to enhance this form of communication by providing a very personal view on topics in the field of alternative methods that might need some more thought.

As a first topic I have chosen the core of ECVAM's business, validation itself. I want to share and discuss a couple of thoughts, identified problems and emerging solutions, but not give final answers. Comments and feedback are more than welcome as is active participation in driving the field toward solutions for these open problems.

What does validation of alternative methods mean?

According to the dictionary, the word "alternative" has only been used to mean "better than the established" since 1970. This is in fact an excellent description of what science is about in general: progress to something better than what is already established.

Therefore, it is remarkable that the term today is so closely associated with alternatives to animal experiments. In this field we are striving toward a more humane science, a science that avoids and reduces the suffering of animals. Remarkably, this area is also on the frontier of safeguarding the quality of science. Validation of alternative methods means proving the relevance of the scientific methods we use. This is an astonishingly new concept, which has only developed over the past 20 years. Indeed, it started out in this field of biomedicine, in the field of alternative methods.

There are two main reasons, why it is this field that has taken a certain lead: On the one hand, in the context of (regulatory) safety assessments very much is at stake (our health, possible scandals) and, on the other hand, at the same time we are in the scientific field with perhaps the most traditional (not to say "old-fashioned") approaches. Where else in science do we still use experimental set-ups that have hardly changed for 60 years?

These two aspects are interdependent: where consumer and patient safety are concerned, approaches are not changed easily; regulators have good reasons to be conservative. To meet the challenges of self-renewal and at the same time maintain safety standards, validation was introduced to ascertain the suitability of new methods and to allow change only for the better and only on solid ground. In this context, validation ensures the "survival of the fittest". But, in an area in which protection of some "fossils" hinders normal evolution, "un-natural selection" becomes necessary.

The principle of validation is depicted in Figure 1: The alternative test method consists of the test system and a data analysis procedure (DAP). The latter is an algorithm used to calculate the result from the raw data. The reference test serves as the point of comparison. If no such reference test exists or can be used for comparison, a consensus standard is used, i.e. experts agree on the reference, for example a number of positive and negative substances.

The test results of the alternative test are then compared to the reference results. For this purpose a prediction model, which converts the results of the alternative method into the categories or units of measurement of the reference method (for example a measure of cytotoxicity is converted into classes of toxicity according to EU classification), is typically required. There are three principle aspects of validity highlighted by grey boxes: (i) reproducibility of the test system, (ii) its

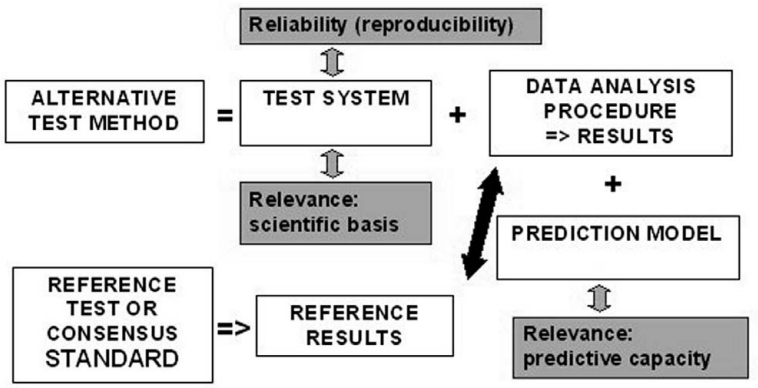

Fig. 1: Definition of validation

scientific mechanistic basis and (iii) its predictive capacity for the reference results. Furthermore, quality control of the test (standardisation, notably including the definition of its purpose) and the validation process itself (mainly its transparency and independence) are requirements and form the definition of the validation process.

Some common misunderstandings about validation

Misunderstanding 1:
Validation is an animal welfare activity

Validation of alternative methods is primarily a quality control process. Its aim is to prevent premature or unsuitable methods from being used in sensitive contexts, such as safety assessments and product development. It has a gate-keeping function and is not there to promote alternatives. Indeed, only a tiny number of all scientific methods in use can be considered to be suitable substitutes for well established animal experiments; and again only some of these are sufficiently well standardised to allow them to enter the validation process. Of these few methods about one third will fail the prevalidation and roughly one third will fail the validation phase. Validation means a rigorous sorting out of many valuable *in vitro* and more recently *in silico* approaches (Worth et al., 2004) in scientific use to identify the very few which can then be considered validated.

Why does validation still promote alternatives and does not just represent an obstacle to their use? Optimistic answer: Because alternative methods simply are better and can stand the comparison with conventional animal-based methods for toxicity assessments. Notions like "humane science is the best science" or "animal welfare and good science are just two sides of the same coin" express this feeling.

There is actually quite a basis to support this view (Goldberg and Hartung 2006):

- animal tests reflect the scientific approach of the time they were developed, not necessarily today's scientific approach and understanding
- they never underwent proper quality control/validation, but are rather based on convention; a precise description of method performance is mostly missing
- they are "under-powered", which means that from the view point of a statistician (sorry to the animal-loving statisticians – this only refers to the professional view) far too few animals are used per experiment to allow conclusive results: costs, work and animal welfare limit the use of animals with regard to group sizes and numbers of repetitions.

The latter deficit is too often overcome by inappropriate use of statistics (the all-time favourite being the one-sided t-test) or by skipping statistics altogether. We happily add new endpoints to animal tests without compensating for multiple testing: A significance level of 5% implies that one out of 20 endpoints for a negative substance can be false-positive. Regulatory test SOPs often foresee 40 endpoints. That makes it very difficult to find a negative substance at all.

Another major shortcut used to squeeze results out of small groups is to use in-bred animals for the tests: testing identical twins eliminates variability of the assay – fine, if the endpoint under study is not affected by this variability, but does anyone control this properly?

Misunderstanding 2:
Validation is the calibration of a method

The term "validation" is commonly used in many contexts. For chemical and physical methods or in general in the quality assurance processes of ISO/GLP, it refers mainly to the assessment of reproducibility and the definition of controls to be applied. Scientists operating in analytical fields often do not understand the extent of work and time it takes to carry out a validation study for toxicological safety assessment purposes. The calibration of methods ascertains that we measure things right. The focus of the validation of alternative methods, however, is to assess whether we are measuring the right parameter (i.e. its relevance) and to ascertain to which test materials it is applicable (negatively said, the test limitations). It would probably have been better to coin a completely new word for the process of validation of alternative methods, such as "relevantification". Using the ambiguous term "validation" leads too many people to believe too soon that they understand the process.

Misunderstanding 3:
Animal tests have been validated by their long-term successful use

The best answer to this I have encountered so far (although it was originally meant in a different context) is the following: "Learning from experience may be nothing more than learning to make the same mistakes with increasing confidence." (Petr Skrabanek and James McCormick, Follies and Fallacies in Medicine, Tarragon Press, Glasgow, 1989).

The ideal evidence that could be used to evaluate animal-based approaches would be epidemiological studies in humans. However, such studies are usually not available or, even worse, cannot be done with only reasonable effort. Why should experience not collected systematically work here? It took us 50 years to prove that smoking causes cancer. How should we deal more effectively with all other potentially dangerous substances of less clearly assessable exposure? The 10 to 15 years latency between exposure and diagnosis of cancer cannot be circumvented in epidemiological studies. Other chronic health effects are no better – perhaps less time elapses between exposure and the development of symptoms of chronic systemic toxicity or reproductive toxicity, but the possible manifestations are also much more diverse (Prieto et al., 2006).

We could ask whether "predictive chronic toxicity" exists at all. A world with about 140.000 man-made chemicals that permits us to become almost 100 years old, cannot be that dangerous to our health. To attribute this to the successful sorting out of certain chemicals appears overbearing, given that systematic risk assessment of new chemicals has (in Europe since 1981) been done on the last 4.700 only.

The claim that animal experiments protect us from chemicals' adverse effects is difficult to disprove: most substances are safe anyway, others have not been on the market long enough to cause problems and in most cases the causal link between exposure and effect cannot

be scientifically demonstrated. Frustratingly, only the acute and topical toxic effects permit comparison of animal results with human data, and of course it is these which are relatively easy to replace by alternative methods anyway.

Misunderstanding 4:
Prevalidation is what happens before validation

Including a prevalidation step in the validation scheme was meant to introduce an important check-point before embarking on large ring trials (Hartung und Spielmann, 1995). However, the standardisation of methods and the reproducibility assessment are integral parts of the validation process themselves. Actually, in other fields, these alone are often considered as the validation (see misunderstanding 2). However, the work done in this phase, which may often last even longer and be more laborious than the actual final validation phase, is perceived as preparatory only. We have thus largely abandoned this term in the definition of the Modular Approach (Hartung et al., 2004), although it will no doubt continue to constitute an integral step when organising a prospective validation study.

Misunderstanding 5:
Omics technologies will quickly give us the means to test, but they are difficult to validate

Hopes raised for toxicogenomics, -proteomics and -metabonomics, etc. as novel alternative methods are high. Typically, whenever a technology is difficult to understand, it creates unjustified hopes or fears. The truth about omics as alternatives is just the other way around: Validation of "omic" technologies can be done quickly, but they are difficult to standardise and whether they are a suitable means to provide the answers we need, must still be shown. Nevertheless, efforts are ongoing to allow the applied use of the technology in regulatory frameworks (Corvi et al., 2006). The principal problems are:

- Measuring a lot (data-rich endpoints) does not improve the quality of the test system you start with ("trash in, trash out").
- The highly complex procedures are difficult to standardise and quality control is difficult.
- The technologies are still too demanding for them to be carried out in routine laboratories.

There is little doubt about how to validate them once they have reached a sufficient stage of maturation (Corvi et al., 2006): challenge them with a couple of substances and check reproducibility and predictive capacity.

Nevertheless, these techniques promise to screen more effectively for new biomarkers to measure the desired effect and to gain a deeper mechanistic understanding. Thereby, they will help to develop new alternative methods but, at this moment, they do not represent ready-to-use alternatives.

Some problems of the validation process

Problem 1: The point of reference

Figure 2 shows, what I call the "validation dilemma": We typically do not compare the results of the test to be validated with what we are actually interested in, i.e. in vivo human data. We simply do not usually have such data. Instead, we generally consider the animal experiment as the (one and only) gold standard. However, this means that from the beginning we can only approximate the test's performance; it is not possible to go further and improve the approach.

If the animal test is considered correct, we can only achieve a sensitivity or specificity below 100% for the new test and a correlation below 1. This instantly gives the impression of lower precision and lower safety levels. This means that (accepting that the animal test in reality is not perfect) everything that in fact represents an improvement will indeed appear to weaken the rigor of the assessment.

This is not a very scientific approach, where "striving for the better" should be the main principle. Imagine if religious reformers had from the beginning accepted the infallibility of the pope… (To be historically correct, the infallibility dogma was introduced only some 140 years ago, i.e. long after reformation). In future we will aim to use the term "traditional test" rather than "gold standard" to indicate the uncertainty of the point of reference.

Some options on how to overcome the dilemma come to mind:

- We can validate the animal experiment against human data (certainly limited).
- We can estimate how well effects in humans may be predicted by performing species comparisons: why should the rat predict human responses better than the mouse or guinea pig, etc. Noteworthy, where such species comparisons exist, correlations range typically around only 70%.
- We can estimate the reproducibility of animal tests from retrospective assessments of the variability within the test and between different tests. The differ-

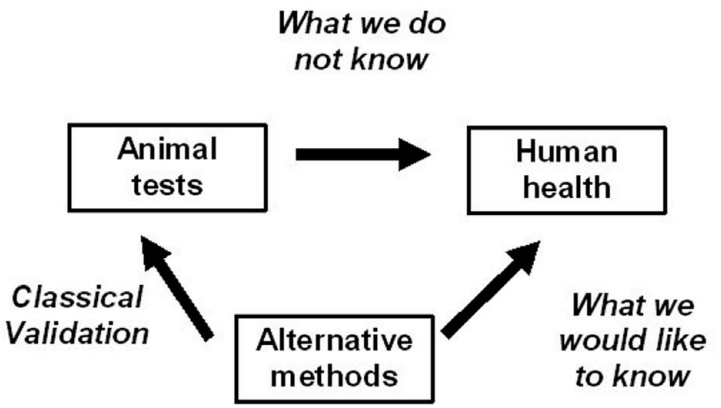

Fig. 2: The validation dilemma

ence in the response of animals in the same treatment group will certainly be smaller than that between experiments on different days or done in different laboratories.
- We can create a combined reference from all available data instead of comparing only with the animal experiment. By expert judgment, for example, the classification of a chemical is done on the basis of all available data, which might include non-guideline studies, structure/activity relationships, mechanistic information and human data. This compiled information should outperform what a single test can offer.

This will still leave us in most cases with a reference that is based largely on animal data, since these are the data sets which are required by regulators and are carried out for these purposes under Good Laboratory Practice (GLP) principles. The drawback of a lot of these so-called high quality animal datasets is that the data are often not public, since they are only provided confidentially to regulatory institutions and are typically not published in the public domain. This is where we crucially need industry to support the development of alternative methods. This type of information is also of critical importance for the development of structure/activity relationships. Unfortunately, the provision of (raw) data is often not enough: the substance needs to be made available at the same time to challenge the alternative test system.

In practice, high quality data exists (although confidential and without raw data) for chemicals notified over the last 25 years under the Dangerous Substance Directive in the New Chemicals Database (to date about 4.700 compounds). However, few of these substances are commercially available in laboratory quantities and in sufficient purity; for many new chemicals production has been discontinued already. Again, the call is to industry to support validation by providing suitable substances with their datasets.

Expectations of the European Partnership for Alternative Approaches to Animal Testing (EPAA, http://ec.europa.eu/ enterprise/epaa/index_en.htm), a partnership between the European Commission and industry comprising an increasing number of individual companies (to date 27) and trade associations (to date 7), are high.

Problem 2: Precautionary toxicology

The idea of precautionary toxicology is to opt for the worst case assumption as to the toxic properties of a substance in the absence of proof against this. The concept evolved out of the German socio-legal tradition of the 1930s (*"Vorsorgeprinzip"*). In 2000, the European Commission issued a Communication on the precautionary principle, in which it adopted a procedure for its application, as done earlier in the Maastricht Treaty. The principle has been translated to many EU policies, including areas beyond environmental policy, such as the EU food law and policies relating to consumer protection, trade and research, and technological development. A working definition and implementation strategy for the EU context has been proposed (Fisher et al., 2006):

"Where, following an assessment of available scientific information, there are reasonable grounds for concern for the possibility of adverse effects but scientific uncertainty persists, provisional risk management measures based on a broad cost/benefit analysis whereby priority will be given to human health and the environment, necessary to ensure the chosen high level of protection in the Community and proportionate to this level of protection, may be adopted, pending further scientific information for a more comprehensive risk assessment, without having to wait until the reality and seriousness of those adverse effects become fully apparent".

The idea is intriguing: better sacrifice a couple of innocent chemicals than suffer from surprises of products on the market. But, what type of reference for validation is this, where we have an inflation of false-positives? For some areas like cancer and reproductive toxicology it has been shown that we are facing most likely 10-times more false than real positives (Kirkland et al., 2005; Hoffmann and Hartung, 2005; Bremer et al., 2007; Kirkland et al., 2007). Whether we can afford this for REACH, i.e. applying this to the most valuable chemicals we have, is outside the scope of this analysis. But how could this possibly serve as the standard, the reference point, for validation? It will be most difficult to devise any test to identify such false-positives. Thus, a precautionary approach is not only suicidal to many newly developed substances (and possibly if applied within REACH to our most valuable chemical products). It also closes the door on its own successful replacement by creating an obstacle for any substitute, which would have to find the same false-positives (Hoffmann and Hartung, 2005).

Problem 3: Test guidelines instead of standard protocols for animal tests

The most effective measure adopted to reduce unnecessary animal tests was the introduction of "mutual acceptance of data" (MAD) by the OECD. This means that substances are tested only once and not repeatedly in every country of notification. However, this usually means that a compromise must be made in creating test guidelines, which are ill-defined and allow for many variants covering the various member state versions of the test. This means that the jointly accepted data originate from quite different test versions, notably, without any proof of their equivalence. How can such a diverse set of data serve as a point of reference? How can a substitute predict such a combination of different tests with regard to their outcome? Hardly any of these substances will be tested in more than one variant. So, who can tell which of the results can serve as a point of reference for validation? Thus, what initially served as a measure to reduce unnecessary animal testing, has turned into an obstacle to its real, final resolution (Hoffmann and Hartung, 2006a).

Pro*blem 4: Standardisation versus flexibility of alternative methods*

The same problem applies to the alternative methods, where a very specific protocol has been validated, while before and afterwards many variants are in use. When mining existing data for retrospective analysis, the question arises which data originate from sufficiently similar test protocols? A tool might be borrowed here from clinical medicine, i.e. meta-analysis. This refers to an approach to

combine various clinical studies and assess their overall outcome. However, no such meta-analysis has so far been undertaken in our field. Problems to be clarified include how to identify the data to be included, since often relevant data are proprietary, and how to control the quality of the input data. With regard to the latter aspect, we together with a contractor are currently developing the necessary quality score for toxicological studies. How might such a quality score look? For example, there could be a number of different categories ranging from "case report in non peer-reviewed literature" to "multi-laboratory, blinded ring trial under GLP with independent management and assessment".

Another way to control the variants that will be used for regulatory purposes is to establish for each of the validated in vitro assays an adequate set of performance standards for each application area. This would allow assessing the equivalence of a test variant to the validated method.

Problem 5: "One by one replacement" versus "one by many replacements"

Unlike for the acute and topical toxicities, which have dominated the area of development and validation of alternatives so far, the more complex endpoints will not be replaced by single tests (except for some filter tests sorting out certain substances from the start). Instead, combinations of tests will be required. This concept is referred to as the "intelligent testing strategy" (ITS), "integrated testing", "test strategy" or "test battery", etc.

At this moment we lack most of the tools to compose and to validate such testing strategies. They certainly require more substances to be tested, both to determine the proper composition of the tests and to validate the strategy. The main question from my current point of view is whether we have to and can validate all the individual tests separately or only the overall strategy. The first approach is challenging with regard to the point of reference for each partial test, the latter approach is challenging with regard to the complexity of the analysis. A solution might lie in the middle, where each building block of the ITS needs to be evaluated for standardisation and reproducibility (modules 1-4 of the modular approach), and relevance (modules 5 and 6) is only assessed for the overall ITS. Discussions have only begun and also the ongoing large EU Integrated Projects (A-Cute-Tox, ReProTect, Sens-it-iv, OSIRIS and CarcinoGenomics) are challenged with these questions.

However, perhaps we are overestimating the challenge: Some validated tests are in fact already testing strategies – the embryonic stem cell test for example tests two cell types (embryonic stem cells and fibroblasts) with different endpoints (cardiac differentiation and cytotoxicity); it might well be that composing this test (strategy) could have been done in a more analytical manner, but it did not pose a real problem for validation.

Problem 6: Lack of post-validation surveillance

Science is continuously progressing and alternative methods will still face challenges from new results and technological developments after validation. We need to stay open to such insights and changes; otherwise we are in danger of creating another rigid, traditional approach.

A first step would be to follow up the fate of validated alternatives in practical use. A user forum, feed-back mechanisms (e.g. where the method failed) or workshops collecting experience from practice (experimenters and regulators) represent opportunities. We must be open to both a restriction and an extension of the applicability domain of tests based on new experiences. This can include also the withdrawal of the validity status of a method after a new peer review. Some alternative methods were validated a decade ago: time to ask where we are with regard to their implementation and experience from use.

Problem 7: Lacking implementation of the relevant mechanism as a criterion for validation

While reproducibility/reliability and relevance assessments are well structured parts of the validation process, the assessment of the scientific basis of a test method has not really been formalised. It is usually seen as part of the test definition and the absence of obvious concerns is taken as evidence of a justified mechanistic basis.

However, it is exactly this mechanistic basis which links alternative methods to modern toxicology, which has in large parts become mechanism-based. It might in some cases be more important that a new, alternative method reflects certain key mechanisms of a health or environmental effect shown by some prototypical chemicals than to show that an outdated traditional test can be reproduced with the new test. Especially for newly emerging areas (neurodevelopmental toxicology, endocrine disruption, im-

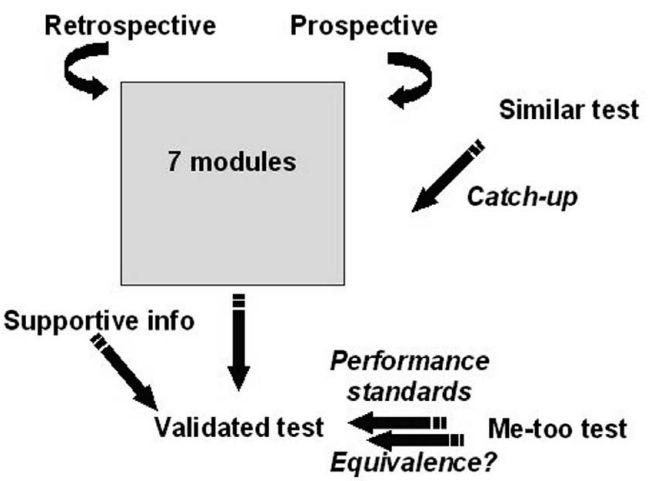

Fig. 3: Validation and Weight-of evidence

munotoxicity, respiratory irritation, respiratory sensitisation, etc.), for which no traditional tests exist, such an approach appears to be most promising. At the same time, this represents an enormous challenge to the transparency and scientific rigor of the process.

Where are we with regard to the validation of alternatives?

Validation is not a process set in stone, but one that can continuously be optimised and tailored. The definition of the Modular Approach (Hartung et al., 2004) represents only one landmark, which has already been pushed further by discussions accommodating different types of test validations as summarised in Figure 3 (see also: Balls et al., 2006; Hoffmann and Hartung, 2006a).

We have to distinguish first of all whether the data considered for validation already exist and are to be analysed "retrospectively" or whether a "prospective" study is to be carried out. These two approaches are not mutually exclusive, as data lacking in a retrospective analysis might be complemented in a prospective manner. During a running validation study a similar test might "catch up" and can still be included in the evaluation. Later, when a test has already been validated, new variants of the same test might emerge from different originators. The question of "equivalence" of these "me-too" developments (a term borrowed from the area of generics for pharmaceuticals) will arise. In order to avoid re-entering large trials, performance standards must be defined already in the context of the original validation and peer review to guide what will be required to attain equivalence. ECVAM is in the process of setting up the reference laboratory CORRELATE to carry out such assessments of equivalence among other things. Many of these processes require "weighing of evidence" (Balls et al., 2006).

Figure 4 summarises some of the open questions relative to the seven modules (left side) defined in 2004. They have been tackled in this article: How do we accommodate the mechanistic basis? Which variants of a test are equivalent? How can

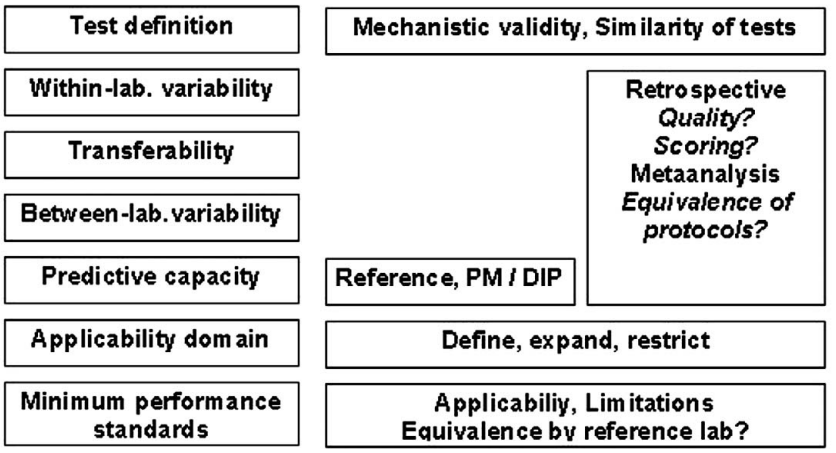

Fig. 4: Open questions modular approach

we quality score existing data? How does one perform a meta-analysis of data and what data must be included? What is our point of reference if not a traditional test? How can we derive a prediction model (PM, or as OECD prefers to call it a Data Interpretation Procedure, DIP)? How can we define and later change the applicability domain of a test? How can we assess the equivalence of a method that is similar to a validated one?

Continued discussion must further shape the details. Practical work and perhaps also procedures and thinking are at this moment very much driven by the areas of chemicals and cosmetics; the potentials for pharmaceuticals and basic research for example have been discussed elsewhere (Hartung, 2002; Gruber and Hartung, 2004). These areas will also require specific considerations such as:

- parallel testing with the established tests for batch release/product control tests
- product-specific validations
- quality assurance of methods used in research (only example so far is the *in vitro* production of monoclonal antibodies) instead of defined tests
- general quality control issues such as Good Cell Culture Practice (Coecke et al., 2005), which is currently adapted to specific areas of interest such as stem cells
- opportunities to control pre-clinical safety assessments with the results of volunteer studies in drug development (noteworthy up to 30% of the drug candidates which make it into volunteer testing after having passed the toxicological tool box must be abandoned owing to toxic side-effects)
- methods for biologicals (especially human proteins and antibodies against human structures), which will be difficult to validate owing to the absence of relevant animal data

The changing political environment (7th amendment of the cosmetics directive 2003 and REACH 2006) has evoked an unprecedented validation programme: At this moment, 187 test methods are under validation. It is important to note that these are at very different stages of the process (between reproducibility assessments after test standardisation and final peer-review by ESAC, ECVAM's scientific advisory committee). As important, however, is that an approach towards an evidence-based toxicology is emerging (Hoffmann and Hartung, 2006b), in which validation shall form one pillar of the self-renewal of toxicology.

References

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 58a. *ATLA – Altern. Lab. Anim. 34,* 603-620.

Bremer, S., Pellizzer, C., Hoffmann, S. et al. (2007). The development of new concepts for assessing reproductive toxicity applicable to large scale toxicological programmes. *Current Pharmaceutical Design*, invited.

Coecke, S., Balls, M., Bowe, G., et al. (2005). Guidance on good cell culture practice. *ATLA – Altern. Lab. Anim. 33*, 261-287.

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

Fisher, E., Jones, J. and von Schomberg, R. (eds) (2006). *Implementing the precautionary principle: perspectives and prospects*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar

Goldberg, A. and Hartung, T. (2006). Not just for the rabbits. *Scientific American 294,* 84-91.

Gruber, F. P. and Hartung, T. (2004). Alternatives to animal experimentation in basic research. *ALTEX 21 Suppl. 1,* 3-31.

Hartung, T. (2002). Three Rs potential in the development and quality control of pharmaceuticals, *ALTEX 18, (Suppl. 1,),* 3-11.

Hartung, T. und Spielmann, H. (1995). Der lange Weg zur validierten Ersatzmethode. *ALTEX 12,* 98-103.

Hartung, T., Bremer, S., Casati, S., et al. (2004). Modular approach to the ECVAM principles on test validity. *ATLA – Altern. Lab. Anim. 32*, 467-72.

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

Hoffmann, S. and Hartung, T. (2006a). Designing validation studies more efficiently according to the modular approach: retrospective analysis of the EPISKIN test for skin corrosion. *ATLA – Altern. Lab. Anim. 34*, 177-191.

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

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

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

Prieto, P., Baird, A. W., Blaauboer, B. J. et al. (2006). The assessment of repeated dose toxicity in vitro: a proposed approach. *ATLA – Altern. Lab. Anim. 34,* 315-341.

Worth, A. P., Hartung, T. and Van Leeuwen, C. J. (2004). The role of the European centre for the validation of alternative methods (ECVAM) in the validation of (Q)SARs. *SAR QSAR Environ. Res. 15,* 345-358.

Worth, A. P., Van Leeuwen, C. J. and Hartung, T. (2004). The prospects for using (Q)SARs in a changing political environment – high expectations and a key role for the European Commission's joint research centre. *SAR QSAR Environ. Res. 15,* 331-343.

Acknowledgements

The continuous discussion and input of my co-workers in and outside ECVAM is gratefully appreciated. Especially, I would like to thank Sandra Coecke, Christoph Klein and Andrew Burke for critically reading this manuscript.

Correspondence to

Prof. Dr. med. Dr. rer. nat. Thomas Hartung
IHCP-ECVAM
Via E. Fermi 1
21020 Ispra
Italy
e-mail: thomas.hartung@ec.europa.eu