AATEX 1, 127–141 (1992)

Validation of *in Vitro* Tests for General Toxicity

Björn Ekwall, MD, PhD

Dept. Toxicology, Univ. of Uppsala Biomedical Center
Box 594, S–75124 Uppsala, Sweden

INTRODUCTION

In the last twenty years, animal experimentation in toxicology has been supplemented and in some cases replaced with *in vitro* methods. The new *in vitro* toxicology (1–4) consists mainly of methods based on cell and organ cultures, but includes test tube experiments with bodily macromolecules, receptors, cell organelles and transmitters, as well. Quantitative Structure-Activity Relationship (QSAR) prediction of human toxicity based on physicochemical properties also belongs to *in vitro* toxicology.

The new methods have been introduced gradually over the years, with a different impact on the various areas of toxicology. The new methods were first introduced in mechanistic and biotransformation studies—within this field approximately 25% of toxicological experiments are now performed *in vitro*. In this field, the animal experiments point out lesions, the causes of which are further elucidated in test tubes, in which the molecular or cellular events can be studied without disturbance from the multitude of irrelevant processes which usually are present in a corresponding animal experiment.

The introduction of *in vitro* methods in testing began in the early seventies as a consequence of bacterial and cellular studies of mutagenesis and cancer mechanisms. Preceeded by Bruce Ames Salmonella test, several cellular short term tests on mutagenicity, transformation and promotion, i.e. components of carcinogenicity induced by chemicals, were developd, validated and accepted for general use. The mutagenesis tests were mechanistically based, which certainly facilitated introduction, compared with an hypothetical acceptance only based on empirical correlations from validation studies.

In the last decade, a host of *in vitro* methods have been developed for the purpose of testing various types of non-genetic, general toxicity. These methods are generally not mechanistically defined. Instead cell cultures and other biological systems in vitro are used as "mini-mice", *i.e.* used empirically to measure all types of toxicity on grounds of a supposed similarity to human targets. One of these developments is quite old, that is the prediction of local irritancy of bodily implants, such as plastic or metallic protheses, dental materials, etc. by simple test systems with cell lines (5). Most other proposed *in vitro* tests have been developed recently, such as tests for acute systemic toxicity, acute systemic target organ toxicity (nervous system, liver, heart, kidney, the immune system, etc.), local irritancy (skin, eye, lung, phototoxicity), teratogenicity, and various toxicokinetic tests. For each of these purposes some 10–50 *in vitro* test methods have been developed, totalling 400–500 different methods

Some of these methods have rapidly been adopted by some industries as preliminary sceening tools, to be evaluated by parallel use of the *in vitro* method and conventional methods, to be ultimately accepted or rejected at a later stage. Thus quite many methods for prediction of acute toxicity, eye irritancy, phototoxicity, and teratogenicity are just now in practical use on evaluation premises (5). Cellular tests for local and systemic toxicity of plastic devices, etc. are in fact

recommended as the test of first choice in the U.S. Pharmacopeia (6). However, the bulk of newly developed methods are not subject to industrial evaluation by practice, and in similarity to those which are, no method has gained a wide acceptance by industry in general, legal authorities, or scientific community.

Why are so many *in vitro* methods waiting for, but not getting general acceptance? Are they not good enough, that is are they not competitive with existing conventional animal tests? Nobody knows the exact answer to these questions. Probably the answer is rather complex. To finalize this section of the chapter, the author will try to give a tentative explanation.

In the main, it is easy for a laboratory to develop and propose an *in vitro* test for a certain purpose (for a certain type of general toxicity). The principle is to analyse som type of chemical toxicity, i.e. the effect of the chemical insult, into components such as injury to basal functions of cells, injury to organ-specific functions of cells, injury to specific noncellular structres in the tissue involved, the inflammatory reaction to injury, and the toxicokinetics of the chemical in the target tissue. Then, a test tube (cell culture) test is constructed, either to simulate all these factors compositely, or to simulate one factor to become a partial test of the discussed type of toxicity. Thus, *in vitro* eye irritancy tests are either of a composite type (the CAM test) or represent possible targets in the eye (tests with simple cell lines for basal cytotoxicity (19), tests with specialized cells such as corneal epithelium, tests with a protein matrix simulating corneal structure, tests of inflammatory messengers, etc).

On the contrary, it will be very difficult and expensive for a single laboratory to prove that the test works, i.e. is predictive of the human toxicity aimed at. In case of eye irritancy, the laboratory has to prove that the test (used in other laboratories) will predict eye irritancy in man better than the existing Draize test, or at least as good, for most classes of chemicals (because the Draize test is used to test unknown chemicals, for instance in nature-derived cosmetics). This will require validation (5, 7, 8), i.e. evaluation of the (technical) reliability as well as of the relevance of the method for the purpose aimed at, in this case human eye irritation. If the test is a composite method, this will probably require testing of hundreds of carefully selected chemicals with (ideally) a human eye irritancy to compare *in vitro* results to, plus Draize test results, to be compared to the human toxicities (in order to prove that the in vitro test is as good or better than the Draize test). The testing ought to be blind. If a test of one component of eye irritancy is going to be validated, the other components must be brought into the validation process, also, as *in vivo* data or as other *in vitro* test results. This is a "mission impossible" for most laboratories, having developed *in vitro* tests. So most method developers have performed a preliminary, small evaluation of their methods, adapted to the real world resources. Ten to thirty chemicals with animal test results found in literature have been tested and compared with the proposed test. This will not lead the test into practical, generally accepted use, but may lead to further "practical" evaluation in an industry or an entry to further evaluation by other researchers/organisations.

Thus, the bottle-neck to implementation of *in vitro* methods to test for general toxicities of various types is probably deficient validation, caused by lack of resources of individual laboratories. Another problem is that no good human toxicity data exist for certain types of general toxicity (eye irritation, teratogenicity, etc.) which on principle makes acceptance of non-mechanistic *in vitro* tests difficult. Finally, it is important to realize that deficient validation is not the only stumbling-block to future effective *in vitro* testing of general toxicity. For many types of toxicity and toxicokinetics there are simply no proposed *in vitro* tests, such as for subchronic and chronic systemic

toxicity, excretion of chemicals in the liver and kidney, and so forth. There is also a lack of *in vitro* tests for subtle and reversible toxicity of various types. Stimulatory toxicity is not represented well *in vitro*. As long as important areas of toxicity are not covered by *in vitro* tests, the success of validation procedures cannot lead to a general replacement of animal toxicity tests.

In the following paragraphs, performed and ongoing validation of *in vitro* tests for general toxicity will be discussed, with a focus on tests for acute systemic toxicity, eye irritancy and skin irritancy. Some single-laboratory validation will be discussed, since it had in some cases put a method into practical use, or provoked development of still better methods. Thereafter, ongoing multicenter validation studies will be described. Finally, proposed multicenter validation will be reviewed, followed by the author's criticism as well as constructive proposals to a future, effective multicenter validation.

Performed single-laboratory validation

As a prerequisite to later discussion the semantics of validation ought to be clarified. The word comes from the evaluation of genotoxic *in vitro* tests, and has been introduced into the area of non-genotoxic evaluation by others (7, 8) to denote establishing relevance and reliability of a test for a specific purpose. By reliability is meant reproducibility of a test and interlaboratory variation. While evaluation is a word for a process aiming at, but not necessarily leading to acceptance of a test, the word validation somehow suggests that an absolute goal (acceptance by regulatory authorities?) must be reached. In spite of that, validation is used by most researchers as the name of a process, specifically meaning evaluation of *in vitro* methods, but without the demand that the process must be complete in order to reach an absolute goal. Others (9) avoid the word, with pretty good reasons. The author of this chapter is not happy with the word, but will use it as a handy, short term for evaluation of the reliability and/or relevance of *in vitro* tests, meaning a process, which may include steps of investigations ultimately aimed at acceptance of a test, but which will not guarantee acceptance-instead, multicenter validation of 50 tests will guarantee that most of them will fail!

The interrelation and relative importance of the relevance and the reliability of tests are connected with the semantics of validation. A positive outcome of an evaluation of the relevance of a test, performed with a large number of chemicals, automatically implies considerable reliability. In contrast, a very reliable test may be useless for a purpose, due to irrelevance. Also, most modern cytotoxicity tests with objective end-point determinations are reproducible and results usually do not vary much between laboratories. On the other hand, some of those tests have been shown to lack relevance when scrutinized. To summarize, relevance and reliability are of equal importance for an acceptance of a method, while, at the same time, validation ought to be focused on the evaluation of relevance, the most critical part of performance of *in vitro* toxicity tests. This attitude has also been expressed in most single-laboratory validation studies-some of them have only evaluated relevance, having good reasons to believe that the reliability of the method would be sufficient, or could be made sufficient, in case of a positive evaluation for relevance.

Table 1 presents some examples of validation of methods to test local irritancy of muscular implants, acute systemic toxicity and eye irritancy, performed by the laboratories developing the tests. The validation of toxicity tests with cell lines performed against rat and mouse LD50 data with 200 chemicals actually represents five similar validation studies, each performed with some 40 chemicals. Each of these studies showed a relatively good correlation between the *in vitro* test results and the toxicity *in vivo* (around 70% correct predic-

Table 1. Examples of performed single-laboratory validation

Test systems	Type of toxicity	No. of chemicals	Species compared to	Type of comparison	Reference
Cell line toxicity (Agar overlay, etc.)	Irritation of muscular implants	1013	rabbits	qualitative (+/−)	10,11
Cell line toxicity (KB, MIT-24, etc.)	Acute systemic	200*	rat, mouse	linear regression	12–16
MIT-24	Acute systemic	50	human	quantitative	17,19
Cell line toxicity (NR, Comassie Blue)	Eye irritancy	150	rabbits	rank correlation	20
CAM and HET-CAM	Eye irritancy	250	rabbits	rank correlation	21,22
Eytem[tm]	Eye irritancy	1500	rabbits	rank correlation	23,24

KB=Kenacid Blue; MIT-24=Metabolic Inhibition Test, supplemented with microscopy after 24 h; NR=Neutral Red uptake; CAM=Chorioallantoic membrane Test
*the total number of compounds included in five separate studies

tion), which has served to introduce such tests for screening purposes in industry. Details of the tests and the outcome of validation has been described in a recent review (5).

However, this single-laboratory validation has not won general acceptance for any of the methods, except for the cytotoxicity tests of local irritancy of intramuscular implants. None of these methods has totally replaced any animal experiments. To a certain extent this could be due to short-comings of the validation made.

These short-comings may be analysed by a comparison of the studies with a set of ideal requirements for a convincing validation study. Such requirements have been listed in the right column of Table 2, which also presents less optimal conditions in the columns to the left. Before the single-laboratory validation is discussed, three important procedures, which will determine the results of validation studies will be outlined (upper 9 paragraphs, Table 2).

The *selection of chemicals* used in validation is critical to the outcome of the study (18, 25). A large number of chemicals, including all relevant classes ought to be used. The chemicals should not be selected-if chemicals which are easily tested *in vitro* or having a supposed cytotoxic action *in vivo* are over-represented the results will appear better than reality. If the test is going to substitute for an *in vivo* test used to test chemically unknown substances (undefined mixtures, natural products) all chemical classes must be represented. The best way to secure a non-biased selection is probably to make a random choice out of a chemical registry (e.g. NIOSH Registry), or similar measures. If a test is validated for testing a restricted number of classes of chemicals, it will probably not be of much practical value. Another prerequisite for a convincing validation study is to use defined chemicals, and not proprietary mixtures or coded industrial products. If studies are performed with chemicals, testing and validation will be possible to reproduce in other laboratories, and thus aquire a scientific status. Credibility of a study will be low if results from evaluation are published without presentation of tabulated base data. Validation studies without tabulated data from tests of chemicals actually prove nothing.

The *in vivo data used* is a second important factor in validation studies. Very few studies have been published concerning the prediction of acute human toxicity by animal lethality tests, or the prediction by Draize's tests of human ocular and skin irritancy (referenced in 18). But toxicologists generally think that these correlations are not especially good (70%?), due to species differences. An

Table 2. Steps of refinement of validation studies

Early approaches		Advanced approaches
Few chemicals (10–50)	Larger number of chemicals, various classes	Large number (200), all relevant classes
Easily tested chemicals	Some water-insoluble, volatile, etc. chemicals	Non-selected chemicals
Propriety formulations	Defined chemicals	High purity chemicals
Known in vivo data	Blind in vivo data	Coded substances
In vivo data from various protocols/literature, not suitable for comparisons	In vivo data from various protocols, suited to comparisons	Suitable in vivo data from the same protocol-designed experiments
Comparisons to animal toxicity	Comparisons to human toxicity	Comparisons to human toxicity, plus parallel comparisons of animal and human toxicity
Qualitative comparisons (+/−) of in vitro and in vivo data	Semiquantitative comparisons (Rank correlation)	Quantitative linear comparisons
Correlative, empirical comparisons, only	Correlative approach, and ad hoc analysis of outliers	Correlation, plus predetermined qualitative analysis (mechanisms, toxicokinetics, etc.)
Direct comparison between in vitro and gross in vivo toxicity	In vitro results, plus known data on other toxicity or toxicokinetics compared to gross in vivo toxicity	In vitro results, plus supplementary data compared as complete models to gross toxicity
One method validated in developing laboratory	Multilaboratory reliability studies of the test, also (Round Robin)	Multilaboratory relevance studies of the test, also
Centrally operated multicenter validation of a few methods	Such validation of many methods for a certain type of toxicity	Such validation for an unlimited number of methods (contest) for many purposes
Multicenter validation to select the best test for a purpose	Multicenter validation to select the best combination of tests for a purpose (multivariate analysis)	Multicenter validation to select a minimal number of tests to be differently combined to predict various types of toxicity

ideal validation trying to lead *in vitro* tests into practical use as predictors of human toxicity, replacing animal tests, would therefore use human toxicity data as the sole reference. Every effort should be done to make such data available or standardized from scattered data of uneven quality. Otherwise every *in vitro* method would be less good than the corresponding animal experiment, going to be replaced. If animal results are used as reference, the lack of precision of the animal method will just be added to the inevitable lack of precision (in predicting animal toxicity) of the *in vitro* method. The species gap would thus be unnecessarily introduced in the new *in vitro* tests. Animal data should instead by used in validation to be compared to the human data of the chemicals used in the study, to establish a baseline for judging if *in vitro* tests are better or worse than the animal results (5). This is indeed validation of animal tests, paradoxically enforced on *in vitro* cytotoxicologists!

Good human *in vivo* data may be generated by testing on volunteers of some types of toxicity, such as skin and eye toxicity. The tests should produce dose-response curves for fixed end-points. In the case of eye irritancy, testing can only include a minimal response.

The third important procedure in validation

is *the method to compare in vitro and in vivo data.* In validation of the mechanistic mutagenicity tests a qualitative *in vitro/in vivo* comparison (+/−) was considered to be sufficient, or rather, it was difficult to establish a quantitative *in vitro/in vivo* relationship. However, all *in vitro* tests of general toxicity are quantitative, and above all, are thought to directly simulate quantitative interactions between chemicals and cells (receptors, macromolecules, etc.) in the body. The tests therefore ought to be quantitatively validated by linear regression analysis or similar methods (5). This analysis must take into consideration different mechanisms, targets and toxicokinetics of outliers. In this analysis of outliers, ad hoc explanations should be avoided-if possible, cathegorisation of outliers should be included in the design of the validation programme. Technically, linear analysis requires exact toxicity data without cut-off results.

Most validation studies performed up to date compare the results of *in vitro* toxicity tests directly with a type of human toxicity such as acute systemic toxicity and skin or eye irritancy. At the same time, most tests have been developed to account of only one of many interactions involved in the gross toxic effect in man, including toxicity to targets not measured by the test, and toxicokinetics (absorption, metabolism, distribution and excretion for systemic toxicity or elimination from the site in case of local toxicity). Comparison of the results from a single test with gross toxicity in man may only determine the degree of influence or part taken by the tested toxicity in gross toxicity, if this determination is not obscured by toxicokinetics. This type of validation will discard tests measuring a relatively rare effect, and, further, cannot discriminate between important supplementary tests and such tests with the same information as other tests. Comparisons of test results to doses to whole animals, whether local or systemically, instead of comparisons to tissue concentrations must result in errors, which can be corrected if toxicokinetic data are interpolated in the comparison.

It is therfore desirable to compare *in vitro* toxicity test results, plus toxicity and toxicokinetics not measured by the test, with *in vivo* toxicity. A historical example may illustrate this: Some early studies did compare cell toxicity and animal LD50 and found a relatively good correlation, indicating influence of cell toxicity on rodent lethality (referenced in 15). Simply by introducing two factors not accounted for by the cytotoxicity tests, i.e. A. Known non-cytotoxic lethal action of some compounds, and B. The passage of the blood-brain barrier of other compounds, into the same type of cytotoxicity tests/LD50 comparisons, the author could preliminary show a relevance of the cytotoxicity test results for most ocmpounds (15). Maybe, that the different in vitro toxicity/eye irritancy correlations for different classes of chemicals will turn out to be a toxicokinetic phenomenon, possible to eliminate by addition of toxicokinetic information to test batteries, such as tests for absorption and elimination rates?

Let us now look at the validation studies examplified in Table 1. It is clear that these pioneer studies do not fulfill all the discussed criteria of optimal validation. In case of selection of chemicals, most studies, except the various acute systemic toxicity validation studies (12–17) have used many comounds. However, in some of the studies the numbers represent products and materials, that are impossible to request for retesting, and for some studies all raw data have not been presented (22, 24). Concerning the *in vivo* data, most studies used animal data. Implantation of materials in rabbits will only give a semi-quantitative response, which explains the qualitative comparison. The Draize test scores validated against in the eye irritancy studies are composites of subjective semi-quantitative scores, not suited for validation purposes. Blind testing has been performed only in some of the studies (14, 20, 21, 24). Concerning the methods of comparison, the Draize test data, in similarity to the local

implantation data, will only permit rank correlations. Some of the *in vitro* tests are also scored similarily (10, 21, 22) which also makes the linear correlation approach difficult. All studies except one (17) compared the *in vitro* test result directly with dose-defined, gross animal toxicity.

How was reliability of tests validated in the discussed studies? In some of the more ambitious studies reliability was studied, with use of coded substances (22, 24). Other studies evaluated relevance of methods, which in validation by the same laboratory had been subject to multilaboratory evaluation of reliability, such as the kenacid Blue (KB) method (14) and Agar overlay test (10). Remaining studies did not include reliability evaluation. This is not surprising, since reliability and relevance very well may be studied separately, without an obligatory coupling to the same study.

Recently, the practical and economical advantages of centrally organized multilaboratory validation programs has diminished the interest in performing single-laboratory studies. However, both types of validation are based on the same discussed principles. Future single-laboratory validation will certainly be competitive to multi-laboratory validation, due to its comparatively higher flexicility. Well done single-laboratory validation should be as acceptable to regulatory authorities as the corresponding multi-laboratory validation. Single-laboratory validation may also be developed by simple measures such as literature data or collaboration of two or a few laboratories with an interest in the same specialized area.

Organized multi-laboratory validation

As suggested by some researches in the early 1980s (19, 26), there are several advantages with organized collaborative validation efforts, compared to single-laboratory studies. Such multicenter validation is a further development of the multilaboratory evaluation of the reliability of isolated methods, performed already in the 1960s, and will in its present form consist of centrally directed testing of the same reference chemicals in many test systems/laboratories, with the purpose to evaluate both relevance and reliability of methods. There are several types of such studies, which may be cathegorized as stages between single-laboratory validation and multi-laboratory validation of all possible methods for all types of replacement purposes (illustrated in the lower 3 paragraphs in Table 2). Thus the development of multicenter validation is concieved as analogous to development of other important aspects of validation, such as improvement of selection of chemicals, reference *in vivo* data and the methods of correlation.

Some advantages with multicenter programs over single-laboratory validation may be listed.

1. Evaluation of reliability of methods can be performed in direct connection with the evaluation of relevance.
2. The credibility of results will increase, compared with single-laboratory validation, if an independent group, not involved in assay development, has selected the chemicals.
3. Multilaboratory validation is considerably more economical (feasible) than single-laboratory studies because efforts (expenses) to select chemicals and *in vivo* data, including the planning and execution of comparisons, can be made once for all, i.e. are "shared" by laboratories (18).
4. Once the procedures for validation of relevance have been determined, an unlimited number of methods may be validated, by use of computer techniques. Thus, no potentially valuable method has to be excluded from validation.
5. Only a multicenter study may directly validate predictivity of combined methods, for instance by computerized multivariate analysis.

6. Only a multi-laboratory program can effectively use models of various types of human toxicity (based on knowledge of the most critical targets for toxicity, mechanisms of toxicity and key toxicokinetic data) and thus evaluate various incoming data as possible parts of these models.
7. If incoming *in vitro* data are numerous and varied, several types of toxicity (acute systemic, eye and skin irritancy, teratogenicity, etc.) could be tried to be modeled in the program. As a result, some of the *in vitro* results may be shown to be predictive in several models. Thus, a minimal number of different *in vitro* tests able to predict various types of human toxicity could be defined, which would make future *in vitro* toxicity testing economical. Such validation can neither be performed by single-laboratory studies, nor by multi-laboratory studies focusing on one type of toxicity.
8. The testing of the same set of reference chemicals in many *in vitro* toxicity tests (different targets, different toxicity criteria), included in a multi-laboratory program, will result in comparisons of *in vitro* results. Differential cytotoxicity studies (38, 39) may be performed on a large scale, for instance. Such comparisons are not *in vitro/in vivo* validation, but will probably contribute much to the understanding and development of *in vitro* toxicology.

In recent years, several centers and organisations have begun to organize multi-laboratory validation studies. This is a result of many factors: A. Recognition of validation as a bottle-neck to the practical introduction of ethical, economical and target-related *in vitro* methods in toxicity testing; B. Recognition of the advantages of multi-laboratory validation; and C. In some cases socio-political pressure to introduce non-animal alternatives to specific tests, such as Draize eye irritancy test. Table 3 presents ongoing multilboratory validation studies. Most of them have been described briefly in a recent review (5).

As one would expect from newly started, centrally organized, second-generation validation studies, the methods to select chemicals, *in vivo* data, and correlation strategies (Table 2) are generally more efined in the multi-laboratory programs (Table 3), compared to the single-laboratory studies discussed previously. Since most of these studies, except for the ZEBET program (35), are intended to be pre-evaluation studies rather than full-fledged validation, the number of chemicals included at this stage is usually small. however, generally thhemical classes tested have been considered more, and the tests are performed with coded chemicals (the MEIC study is an exception; 25). The animal data compared with have generally been better defined, in some cases newly generated for this purpose (9, 27). human data are used by several studies. The correlative approach is limited by the Draize test data ompared to in all eye irritancy studies and consists of rank correlations, in the CTFA study replaced by a refined "concordance analysis" and linear regression (9). In FRAME and MEIC studies linear regression was used, in the latter study supplemented by predetermined analysis of outliers, plus the use of simple toxicokinetic modelling (25, 40). Two studies have been more centered around reliability evaluation and comparisons of *in vitro* data, than on evaluation against *in vivo* data for relevance (26, 34). Of the 150 FRAME chemicals selected only 56 were evaluated for relevance against rodent Ld50 (14). Instead. these studies have reached interesting conclusions from the *in vitro/in vitro* comparisons.

A look at the number of methods in Table 3 reveals that there are indeed two types of multi-labooratory validation studies.

The first type of studies focuses on the evaluation of reliability of a few (often two) methods, which then are used in several

Table 3. Ongoing multi-laboratory validation programmes

Organisation		Type of toxicity	Start, year	No. of chemicals (labs.)*	No. of methods	Species compared to	Reference
FRAME	U.K. Fund for Replacement of Animals in Medical Experiments	acute syst., eye irritancy	1982	150/56	6	mouse, rat	14,26–31
MEIC	Scandinavian Society of Cell Toxicology	acute syst., skin, etc.	1987	50	200**	man	5,25
MRES	French Ministry of Research	liver	1989***	30	2(6)	rodent, man	32
SDA	U.S. Soap and Detergent Association	eye irritancy	1988	23	14	rabbit	33
CFTA	U.S. Cosmetics, Fragances and Toiletries Association	"	1990	10+10	23**	rabbit	9
EC	Commission of European communities	"	1989	21	5	rabbit	34
ZEBET	German Ministry of Health	"	1989	35+200	2(35)	rabbit, man	35
OPAL	A French organisation for Assistence to Laboratory Animals	"	1990***	40	2	rabbit	36
JSAAE	Japan Society for Alternatives to Animal experiments	"	1990	52	2	rabbit	37

*number of laboratories used to evaluate inter-laboratory reproducibility
**Programs still open for new methods
***Completed 1991

laboratories to test the reference chemicals (26, 27, 32, 34–37). All such studies with few methods have also evaluated relevance of tests by *in vitro/in vivo* comparisons, but to a various extent, depending on variation of the quality of *in vivo* data and the number of chemicals with *in vivo* data relative to the total number of chemicals used in the reliability study. In general, these *reliability-focused multi-laboratory studies of a few methods* represent a relatively primitive approach to evaluate relevance, i.e. in similarity to single-laboratory validation these methods can only evaluate tests as single predictors of *in vivo* toxicity. Prediction of a test as measured in one study will thus be difficult to compare with prediction of other tests, as measured in other similar studies, previously. Performed single-laboratory validation has been criticized for being able to lead to acceptance of methods too early, before other, potentially more valuable methods have been pointed out by contest-type validation (5). The same criticism holds also for multi-laboratory validation of a few methods. Since the relevance of a method is in no way predicted by the realiability, the focusing of the above described studies on reliability evaluation does not seem to be rational and economic, either. Resources are used to sort out technically reliable tests, which at a later stage, in competition with other tests, may turn out to have low relevance.

Some of the studies in Table 3 validate a higher number of methods, however, and share other similar characteristics (9, 25, 28, 33). Although these studies do not neglect the reliability evaluation, they are primarily concerned with evaluation of relevance of many methods at the same time, and may be called *relevance-focused multi-laboratory studies of multiple methods*, always introducing an element of contest or competition between methods. Within the group, the multiple-method approaches vary. Such studies could be wholly dependent on testing by laborator-

ies that develop or use the test, and just consist of listing chemicals to be tested by the laboratories (MEIC; 25) or a distribution of coded chemicals to such laboratories (FRAME; 28). On the other hand such studies can be done in contract laboratories, without connection with developing laboratories (32–36). The studies may be classified as open contests, if any relevant method has access to the competition (9, 25), or closed studies, if the organisation directing the study does nothing to invite developers/uses of tests. Also some actual programmes seem to have traits of both reliability-focused/few methods and relevance-focused/multiple-method validation (28, 34).

The CFTA and MEIC programmes are the most typical multiple-method programmes examplified in Table 3. Therefore, basic features of these programs will be decribed.

The CTFA program is a preliminary evaluation of local eye irritancy prediction of 23 methods. Evaluation is made with Draize test data from the same protocol, determined directly for the purpose. Different classes of chemicals and formulations of interest to the cosmetic industry are used for a step-wise evaluation. The first phase evaluates a series of hydro-alcoholic formulations, a second phase a series of oil-water emulsions, and so on. Tests not predicting the first type of substances well are not excluded from subsequent testing (9).

The MEIC program is a validation of around 200 methods with 50 chemicals. The chemicals all have human data on acute systemic toxicity and corresponding animal data, and will with time also get data from human volunteer testing of local irritancy. All *in vitro* methods with a possible relevance to various types of human general toxicity, *as judged by participating laboratories*, will be evaluated. Further, all incoming results from the volunteer laboratories will be evaluated for their combined relevance to all types of toxicity studied (acute systemic, chronic systemic, and local skin toxicities, plus various systemic target organ toxicities, as a minimum). Prediction of the human data by animal test results will be a baseline for judging the efficiency of *in vitro* prediction of the human toxicity. MEIC uses a stepwise evaluation process of 3 phases (I-chemicals nos. 1–10, II- nos. 11–30, III- nos. 31–50) (5, 25).

Only the multiple-method, multi-laboratory programs, in sharp contrast to both single-laboratory and reliability-focused multi-laboratory programs, will benefit from all previously listed advantages (page 7–8). The multiple-method studies may also incorporate most of the advanced aspects of validation, listed in the right column in Table 2. Before the discussion of multicenter validation is ended, one important improvement of the multimethod validation strategy ought to be discussed, with the CTFA and MEIC studies taken as examples.

The CTFA study validates specific tests for a certain purpose, i.e. tests developed to predict eye irritancy. Being multilaboratory, the program may evaluate directly the best combination of tests (whch is not the same as the best combination of the well correlated tests). It may also evaluate tests as partial predictors, along with other partial toxicity and toxicokinetic data, including tests which fit toxicokinetic modelling of eye toxicity. But, since its was focused on one type of toxicty only, the study may not incorporate *in vitro* tests developed for other purposes in the models (acute systemic, skin, and teratogenic toxicity, for instance) since they have not been included in the programme (discarded initially due to presumed irrelevance). Thus, the study cannot contribute to a future *in vitro* toxicity testing which is employing a minimal number of tests for modelling various types of human toxicity. Instead, the MEIC study will be able to investigate if this futuristic approach may be possible, by the use of a variety of *in vitro* tests as well as trials to model several types of toxicity from these tests. This advantage of the MEIC study is not just theoretical. In fact,

many *in vitro* tests developed today can be imagined to be used in models of most types of general toxicity, together with tests of toxicity to specific targets. Such tests include basal cytotoxicity tests, protein-denaturating tests, several physico-chemical data, tests on metabolism and distribution of chemicals, such as epithelial passage, etc. —It could even be speculated, that modeling of diverse types of toxicity (skin and eye irritancy, for instance) is perhaps more depending on the relative tuning of the same set of tests, than radically different tests in the sets.

It is very important that validation in the future shall not be hampered or frustrated by rigid theoretical considerations of the best way to validate. The tabulation of steps of refinement in Table 2, as well as the comments have been made to stimulate thinking, and not to make the reader believe in generally appliciable rules. Probably effective future validation will need studies on all levels discussed (single-laboratory, reliability-orientated multi-laboratory, and multiple-method multi-laboratory) used in concert. Methods could, for instance be evaluated for reliability in some programs, and at the same time be evaluated for specific or general relevance in other programs (5).—Just to discourage a strict interpretation of Table 2, it could for instance be pointed out that the listed advances in selection of chemicals, choice of *in vivo* data and demands on correlative methods are contradictory, and cannot be combined in most cases. Thus human data on systemic toxicity are not compatible with a refined selection of chemicals. Neither are certain human eye irritancy data (splash in the eye) compatible with linear evaluation.

Proposed validation

With time, so many proposed *in vitro* toxicity tests have heaped up without reaching general acceptance, due mainly to the "mission impossible" characteristics of single-laboratory validation. This has stimulated *in vitro* toxicologists to discuss possible ways to cope with the problems, i.e. to organize multi-laboratory validation programs. This has not only been the concern for *in vitro* toxicologists. Also animal rightists, industry and governmental agencies, including bureaus of OECD and EC are engaged in the possibilities to promote *in vitro* toxicity tests by organized, well-planned studies. Some organisations, such as FRAME (Fund for Replacement of Animals in Medical Experiments, England), CAAT (Center for Alternatives to Animal Testing, Baltimore, U.S.A.) and ERGATT (European Research Group for Alternatives to Animal Testing) have relayed impulses from the different sources mentioned into proposals for future multicentre validation. Such proposed programmes has been discussed in a recent review (5). The first proposed multicenter programmes were outlined by FRAME (26, 28). Then CAAT arranged special conferences on validation, including another proposal for organized validation (41, 42 and 7). Later, CAAT and the ERGATT group, in collaboration with FRAME, arranged a consensus conference on validation, the result of which was guidelines for validation of toxicity tests, presented as "scientific validation" (8). The CAAT/ERGATT conference was somewhat later followed up by a ERGATT/CEC (Commission of the European Communities) conference on acceptance of validated tests (43).

Except for a proposal for a relevance-focused, straightforward multi-method approach to validate eye irritancy (44), referred to in a recent review (5), no other than the above-mentioned proposals have been made. Since the CAAT/ERGATT report represents the most recent consensus reached by the researchers involved in earlier proposals, it will be discussed here, to supplement the discussion of on-going multi-laboratory programmes.

The CAAT/ERGATT document is stated to be an authoritative report or guidelines for validating any type of *in vitro* tests, including

genetic and non-genetic tests (8).

The "scientific validation" process was outlined by the workshop to consist of four stages: 1. Intralaboratory assessment to evaluate reproducibility and preliminary relevance with use of 5–10+20–50 compounds, in the laboratory of origin. 2. Interlaboratory assessment to evaluate inter-laboratory variation and reproducibility, with use of 5–10+10–20 substances, in at least but probably not more than 4 laboratories. 3. Test database development, to evaluate relevance, with use of 200–250 chemicals, tested in a small number of laboratories. 4. Evaluation. Both evaluation of performance of single tests and tests as parts of batteries are made. All substances are tested blind (coded). Tests are selected to be included in the process by judgement of their appropriateness for solving given problems-also the need for a test in relation to the availability of other tests must have been demonstrated. Tests are rejected from the process 1. if they fail in the intralaboratory assessment (<70% reproducibility) and 2. if they fail in the interlaboratory assessment (<70–80% interlaboratory reproducibility).

Further, "scientific validation" studies should have descriptors: 1. Level of toxicological assessment (four levels: A. Toxic potential; B. Toxic potency, C. Hazard; and D. Risk to specific populations); and 2. The required testing activity (three levels: A. Screening tests; B: Adjunct tests; and C. Replacement tests). These descriptors (of studies as well as tests) were thought to be decisive for selection of tests to a study. Also, these descriptors would influence the stringency of the criteria of test performance. Thus, screening of potential toxicity might be made with less relevant tests, than, for example, the replacement of an animal test for risk evaluation (A/A and D/C levels, respectively). The proposal discussed the three vital parts of any validation study, i.e. selection of chemicals, the choice of *in vivo* reference data, and the methods of comparisons. Not much was said about chemical selection, other than the recommendation to define structures of the chemicals used in the study, so that validated tests, when used, will not be presumed to be valid for chemicals structures not included in the validation.

Concerning *in vivo* data, the ideal use of human data was acknowledged. This having been said, use of animal data was considered fully compatible with the proposal.

Concerning methods for validating reliability, qualitative, stochastic methods were discussed, without any attempts to define methods capable of evaluation cytotoxic dose-response. Likewise, the methods to evaluate relevance were based on binary classitication schemes, without any discussion of how to deal with continous responses. Scrutiny of a close mechanistic relationship between *in vitro* tests and *in vivo* toxicity was recommended, but no methods to perform such a scrutiny were defined. Battery selection was described, based on pure statistical methods, such as Bayesian approaches based on already validated isolated tests, as well as multivariate regression analysis. Modelling of *in vivo* toxicity by use of combined toxicity and toxicokinetic data or other hard modelling with combined *in vitro* data (40) were not included in the proposed process.

A unique feature of the proposed validation was the claim that the process itself (the described measures taken) would be able to "certify" tests as valid, in an absolute sense. This was thought to be attained by use of the discussed descriptors. Another typical feature is that the process is not an open competition type of validation, but rather reminds of a panel controlled, closed machinery to produce "certified" tests.

The "scientific validation" proposed seems to be a reliability-focused multicenter validation, quite similar to the on-going programs of that type described in Table 3 (34–37). As such, it may reject potentially relevant tests at an early stage, due to a low reliability of the test at the stage of development validated. Also, reliable, but potentially irrelevant and

redundant methods (in retrospect) will be included in a complete data base development, at high costs. Potentially relevant methods may even by prohibited from entering the process, by prejudgement based on questionable hypotheses.

When it comes to chemical selection, the choice of *in vivo* data and the evaluation methods, the process is non-sophisticated. Validation of tests for certain classes of chemicals will probably lead to tests with a low appliciability and acceptance. Also, the use of animal reference data and the binary comparisons of data will lead to a very limited usefulness of validated tests. As discussed before, such tests are almost guaranteed to be less good than the corresponding conventional antimal test, when it comes to predict human toxicity (addition of the lack of precision of 1. prediction of animal toxicity by the test *in vitro*, and 2. prediction of human toxicity by the animal test, validated against).

The presentation of a relatively commonplace validation programme as "scientific" validation is disputable. The term pretends a special, authoritative status for the CAAT/ERGATT proposal, compared to other validation efforts made or going to be made. In fact, some ongoing programs are probably scientifically more advanced than the CAAT/ERGATT proposal, such as the relevance-focused multicenter programs. Equally disputable is the claim that the CAAT/ERGATT process is able to certify tests as validated, and ready for legal acceptance. No such claims can be made for the results of any validation program, or at least will not be taken seriously by prudent legislators, who themselves have to certify tests, based on their own expertise and judgement of the results of the outcome of all available validation efforts, including many other pieces of information.

It is important for future progress of *in vitro* toxicity testing that the validation process is not defined, but considered, in similarity to method development, to be a continous progress towards of better strategies to evaluate new tests. As in the case of method development, no organisation should be able to control or monopolize validation, even if organized action is needed in the field. It is thus important that the much needed, planned chemical data banks for validation purposes (18) as well as the functioning and planned *in vitro* toxicity data banks are administered by independent (governmental ?) bodies, free from vested interests, and that their services are tailored to suit all parties, including individual scientists.

Future perspectives

Based on practical use of results of *in vitro* cytotoxicity tests, the author believes that toxicity testing in a foreseeable future (10–20 years) will be regularily performed in cell cultures and test tubes, rather than in animals. This will not exclude that some future testing for specific purposes will have to be performed in animals. Furthermore, such test tube testing will be more predictive of human toxicity than the animal tests. If this belief is going to be realized, it is not the first time in history when ethical and scientific achievements have joined-think of the replacement of human and animal slavery by machines in the 19th century.

If the prophecy is true, this development would be the result of many new types of *in vitro* toxicity and toxicokinetic tests. Also, these, and some of the present tests, would have been validated in refined programs. Probably many of the validation studies would have been relevance-focused multicenter studies using human data (which can be obtained if we try hard enough). These studies would probably have included multifactorial modelling of many *in vitro* data to account of human toxicity, including computerized physiological kinetic modelling. The computer, fed by the results from test tubes, would thus parallel the experimental animal of today! At the same time, the gradual acceptance of new *in vitro* methods would not only depend on formal validation programmes. It may very well be

that the parallel experience in various industrial laboratories of *in vitro* and *in vivo* methods, with time will convert most toxicologists to *in vitro* toxicologists. Finally, let us hope that such a development, if ever achieved, would not have been dependent on purely political or societal attempts to stop animal tests at a time when they still offered better hazard protection than the developing test tube toxicology.

References

1) Bourdeau, P. et. al. (Eds) (1990) Short-term Toxicity Tests for Non-genotoxic Effects. John Wiley & Sons, London.
2) Atterwill, C.K. & Steele, C.E. (1987) In vitro Methods in Toxicology. Cambridge University Press, Cambridge.
3) Mehlman, M.A. (Ed.) (1989) Bechmarks: Alternative Methods in Toxicology, Princeton Scientific Publishing, Princeton.
4) Walum, E., Stenberg, K. & Jensen, D. (1990) Understanding Cell Toxicology-Principles and Practice, Ellis Horwood, Chichester.
5) Ekwall, B., Bondesson, I., Hellberg, S., Högberg, J., Romert, L., Stenberg, K. & Walum, E. (1991) Validation of in vitro cytotoxicity tests-past and present strategies. *ATLA*, 19: 226–233.
6) U.S. Pharmacopoeia, 1990 Edition.
7) Frazier, J.M. (1990) Scientific Criteria for Validation of In Vitro Toxicity Tests. Paris: OECD
8) Balls, M., Blaauboer, B., Brusick, D., Frazier, J., Lamb, D., Pemberton, M., Reinhardt, C., Roberfroid, M., Rosenkranz, H., Schmid, B., Spielman, H., Stammati, A. & Walum, E. (1990) Report and recommendations of the CAAT/ERGATT workshop on the validation of toxicity test procedures. *ATLA*, 18: 313–337.
9) Gettings, S.D., Dipasquale, L.C, Bagely, D.M., Chudkowski, M., Demetrulias, J.L., Feder, P.I., Hintze, K.L., Marenus, K.D., Pape, W., Roddy, M., Schnetzinger, R., Silber, P., Teal, J.J. & Weise, S.L. (1990) The CTFA evaluation of alternatives program: An evaluation of in vitro alternatives to the Draize primary eye irritation test, (Phase I) Hydroalcoholic formulations; A preliminary communication. *In Vitro Toxicol.*, 3: 293–302.
10) Autian, J. (1973) The new field of plastics toxicology: Methods and results. In CRC Critical Reviews In Toxicology, Vol. 2 (Ed. L. Golberg), 1–40, Cleveland, CRC Press.
11) Wilsnack, R.E. (1976) Quantitative cell culture biocompatibility testing of medical devices and correlation to animal tests. Biomaterials. *Medical Devices & Artificial Organs*, 4: 235–261.
12) Walum, E. & Peterson, A. (1984) On the application of cultured neuroblastoma cells in chemical toxicity screening. *Journal of Toxicology and Environmental Health*, 13: 511–520.
13) Halle, W. & Göres, E. (1987) Vorhersage von LD50-werten mit der Zellkultur (Quantitative prediction of LD50 with use of cell cultures), (German). *Pharmazie*, 42: 245–248.
14) Clothier, R.H. & Hulme, L.M. (1989) Comparison of the in vitro cytotoxicities and acute in vivo toxicities of 59 chemicals. *Molecular Toxicology*, 1: 571–577.
15) Ekwall, B. (1983) Correlation between cytotoxicity in vitro and LD50 values. Pharmacology and Toxicology (Acta Pharmacol. et Toxicol.) 52, Suppl. II, 80–98.
16) Fry, J.R., Garle, M.J., Hammond, A.H. & Hatfield, A. (1990) Correlation of acute lethal potency with in vitro cytotoxicity. *Toxicology in Vitro*, 4: 175–178.
17) Ekwall, B. (1980) Preliminary studies on the validity of in vitro measurement of drug toxicity using HeLa cells: III. Lethal action to man of 43 drugs relted to the HeLa cell toxicity of the lethal drug concentrations. *Toxicol. Letters*, 5: 319–331.
18) Purchase, I. (1990) An international reference chemical data bank would accelerate the developement, validation and regulatory acceptance of alternative toxicology tests. *ATLA*, 18: 345–348.
19) Ekwall, B. (1983) Screening of toxic compounds in mammalian cell cultures. Annals of the New York Academy of Sciences No. 407, 64–77.
20) Shopsis, C., Borenfreund, E. & Stark, D.M. (1987) Validation studies on a battery of potential in vitro alternatives to the Draize test. In Alternative methods in Toxicology. Vol. 5, In vitro Toxicology Approaches to Validation (Ed. A.M. Goldberg), pp. 31–44, Mary Ann Liebert, New York.
21) Bagley, D.M., Rizvi, P.Y., Kong, B.M. & DeSalva, S.J. (1988) An improved CAM assay for predicting ocular irritation potential. In Alternative Methods in Toxicology, Vol. 6, Progress in In Vitro Toxicology (Ed. A.M. Goldberg), pp 131–138, Mary Ann Liebert, New York.
22) Luepke, N-P. & Wallat, S. (1987) HET/CAM: reproducibility studies. In Alternative Methods in Toxicology, Vol. 5, In Vitro Toxicology: Approaches to validation (Ed. A.M. Goldberg), pp. 353–363, Mary Ann Liebert, New York.
23) Soto, R.J., Servi, M.J. & Gordon, V.C. (1989) Evaluation of an alternative method for ocular irritation in in vitro toxicology. In Alternative Methods in Toxicology, Vol. 7, In Vitro Toxicology: New Directions (Ed. A.M. Goldberg), Mary Ann liebert, New York.
24) Gordon, V.C., Kelly, C.P. & Bergman, H.C. (1990) Application of the Eytex^{tm} method. *Toxicology In vitro*, 4: 314–317.
25) Bondesson, I., Ekwall, B., Hellberg, S., Romert, L., Stenberg, K. & Walum, E. (1989) MEIC-a new international multicenter project to evaluate the relevance to human toxicity of in vitro cytotoxicity tests. *Cell Biology & Toxicology*, 5: 331–348.
26) Balls, M. & Bridges, J.W. (1984) The FRAME research program on *in vitro* cytotoxicology. In Alternative Methods in Toxicology, Vol. 4, Acute

Toxicity Testing: Alternatice approaches (Ed. A.M. Goldberg. pp. 62–79. Mary Ann Liebert, New York.

27) Balls, M & Clothier, R. H. (1991) Comments on the scientific validation and regulatory acceptance to *in vitro* toxicity tests. *Toxicology In Vitro*, 5: 535–538.

28) Balls, M., Riddell, R.J., horner, S.A. & Clothier, R.H. (1987) The FRAME approach to the development, validation, and evaluation of in vitro alternative methods. In alternative Methods in Toxicology, Vol. 5, In Vitro Toxicology: Approaches to Validation/Ed. A.M. Goldberg) pp. 45–58, Mary Ann Liebert, New York.

29) Clothier, R.H. & Balls, M. (1990) Validation of alternative toxicity tests: principles, practices and cases. *Toxicology In Vitro*, 4: 692–693.

30) Reader, S.J., Blackwell, V., O'Hara, R., Clothier, R.H., Griffin, G. & Balls, M. (1989) A vital dye release method for assessing the short-term cytotoxic effects of chemicals and formulations. *ATLA*, 17: 28–37.

31) Shaw, A.J., Clothier, R.H. & Balls, M., (1990) Loss of trans-epithelial impermeability of a confluent monolayer of Madin-Darby Canine Kidney (MDCK) cells as a determinant of ocular irritancy potential. *ATLA*, 18: 145–151.

32) Fautrel, A., Chesné, C., Guillouzo., Braut, F., Rahmani, R., De Sousa, G., Placidi, M., Pichon, J., Hoellinger, H., Vintézou, P., Diarte, I., Melchion, C., Cordier, A., Lorenzon, G., Penicourt, M, Vannier, B., Fournex, R., Peloux, A.F., Bichet, N., Gouy, D., Cono, J.P. & Lounes, R. (1991) A multicenter study of acute in vitro hepatic cytotoxicity, *Toxicology in Vitro*, 5: 543–547

33) Booman, K.A., DeProspo, J., Demetrulias, J., Driedger, A., Griffith, J.F., Grochoski, G., Kong, B., McCormick, W.C. 3rd, North-Root, H., Rozen, M.G. & Sedlak R.I. (1989) The SDA alternatives programme: comparison of in vitro data with Draize test data. *Journal of Toxicology-Cutaneous & Ocular Toxicology*, 8: 35–49.

34) Zanielli, P. & H Pirovano, R. (1991) Interlaboratory validation exercise on various in vitro cytotoxicity tests to be used for regulatory purposes for irritation labelling. *ATLA*, In Press.

35) Spielman, H., Gerner, I., Kalweit, S. & Wirnsberger, T. (1991) Validation project of alternatives to the Draize test in Grmany. *Toxicology in Vitro*, 5: 539–542.

36) Blein, O. & Adolphe, M. (1991) A French project of correlation and validation of alternatives to the Draize eye irritation test, Toxicology in vitro, 5, 555–557.

37) Sasaki, K., Tanaka, N., Watanabe, M. & Yamada, M. (1991) Comparison of Cytotoxic effects to chemicals in four different cell types. *Toxicocogy In vitro*, 5: 403–406.

38) Ekwall, B. & Acosta, D. (1982) *In vitro* comparative toxicity of selected drugs and chemicals in HeLa cells, chang liver cells, and rat hepatocytes. *Drug & Chemical Toxicology*, 5: 219–231.

39) Ekwall, B. & Ekwall, K. (1988) Comments on the use of diverse cell systems in toxicity testing. *ATLA*, 15: 193–200.

40) Ekwall, B., Bondesson, I., Castell, J.V., Gómez-Lechón, M.J., Hellberg, S., Högberg, J., Jover, R., Ponsoda, X. Romert, L., Stenberg, K. & Walum, E. (1989) Cytotoxicity evaluation of the first ten MEIC chemicals: acute lethal toxicity in man predicted by cytotoxicity in five cellular assays and by oral LD50 in rodents. *ATLA*, 17: 83–100.

41) Frazier, J.M., Gad, S.C., Goldberg, A.M. & McCulley, J.P. (1986) Validation. In Alternative Methods in Toxicology, Vol. 4, A Critical Evaluation of Alternatives to Acute Ocular Irritation Testing (Ed. A.M. Goldberg) pp. 113–117. Mary Ann liebert, New York.

42) Frazier, J.M., (1987) The validation approach of the Johns Hopkins Center for Alternatives to Animal testing. In Alternative Methods in Toxicology, Vol. 5, In Vitro Toxicology: Approaches to Validation/Ed. A.M. Goldberg) pp 87–90, Mary Ann Liebert, New York

43) Balls, M., Botham, P., Cordier, A., Fumero, S., Kayser, D., Koëter, H., Koundakjian, P., Lindquist, N-G., Meyer, O., Pioda, L., Reinhardt, C., Rozemond, H., Smyroniotis, T., Spielman, H., Van Looy, H., Van der Venne, M-T. & Walum, E. (1990) Report and recommendations of an international workshop on promotion of the regulatory acceptance of validated non-animal toxicity test procedures. *ATLA*, 18: 339–344.

44) Gad, S.C. (1989) Acute ocular irritation evaluation: in vivo and in vitro alternatives and making them the standard for testing. In Benchmarks: Alternative Methods in Toxicology (Ed. M.A. Mehlman), pp. 137–193, Princeton University Press, Princeton.