Report and Recommendations of the CAAT[1]/ERGATT[2] Workshop on the Validation of Toxicity Test Procedures[3]

Michael Balls[4], Bas Blaauboer[5], David Brusick[6], John Frazier[7], Denise Lamb[8], Mark Pemberton[9], Christoph Reinhardt[10], Marcel Roberfroid[11], Herbert Rosenkranz[12], Beat Schmid[13], Horst Spielmann[14], Anna-Laura Stammati[15] and Erik Walum[16]

[4]FRAME, Eastgate House, 34 Stoney Street, Nottingham NG1 1NB, UK; [5]RITOX, University of Utrecht, P.O. Box 80.176, 3508 TD Utrecht, The Netherlands; [6]Hazelton Laboratories America, 9200 Leesburg Turnpike, Vienna, VA 22182, USA; [7]CAAT, School of Hygiene and Public Health, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205, USA; [8]MAFF, Harpenden Laboratories, Hatching Green, Harpenden, Herts, AL5 2BD, UK; [9]ICI Central Toxicology Laboratory, Alderley Park, Macclesfield, Cheshire SK10 4TN, UK; [10]SIAT (Swiss Institute for Alternatives to Animal Testing) ETH-Zentrum, Turnerstrasse 1, CH-8092 Zurich, Switzerland; [11]Unité de Biochemie Toxicologique, Universite Catholique de Louvain 7369, Avenue de Mounier 73, B-1200 Bruxelles, Belgium; [12]Department of Environmental Health Sciences, School of Medicine, Case Western Reserve University, 2119 Abington Road, Cleveland, OH 44106, USA; [13]Zyma SA, CH-1260 Nyon, Switzerland; [14]ZEBET, Bundesgesundheitsamt, P.O. Box 33 00 13, D-1000 Berlin, FRG; [15]Istituto Superiore di Sanita, Viale Regina Elena 299, I-00199 Roma, Italy; [16]Unit of Neurochemistry and Neurotoxicology, University of Stockholm, S-106 91 Stockholm, Sweden

Preface

Developments in cell culture and in bioanalytical and computer technology have resulted in the rapid proliferation of new procedures for use in the hazard evaluation of chemicals (1). These procedures range from computer-based structure-activity relationship (SAR) methods to cell-based or tissue-based *in vitro* testing systems. Many of these procedures provide valuable toxicological data which have not previously been utilised in hazard evaluation and risk assessment. Data provided by these procedures are potentially useful to industry and regulatory agencies for decision-making. However, before any new procedures can be applied, it is essential to demonstrate that they provide reliable information which is relevant to the decision-making process. The process by which the reliability and relevance of a test are established is called validation.

Discussions which took place between members of the European Research Group for Alternatives in Toxicity Testing (ERGATT) and John Frazier, Associate Director of the Johns Hopkins Center for Alternatives to Animal Testing (CAAT), in Berlin in April 1989 and in Nottingham in July 1989, identified a shared concern that the scientific validation of toxicity test procedures, though of crucial significance, had not been fully or adequately addressed. These discussions led to a plan to assemble a small, international group of individuals with extensive and varied experience in relation

[1]CAAT — The Johns Hopkins Center for Alternatives to Animal Testing; [2]ERGATT — European Research Group for Alternatives to Animal Testing; [3]This document represents the agreed report and recommendations of the participants as individual scientists.

to the scientific validation of toxicity test procedures, to consider the question in depth and to make recommendations for the guidance of others. As a result, a Workshop was held on 8–12 January 1990 at Hotel Arvenbüel, Amden, Switzerland.

Although most scientists have an inherent concept of what constitutes the validation of test procedures, the scientific basis and the necessary components of this process have not been fully described in a formal exposition (2, 3). Therefore, the objectives of this international Workshop were:

To conduct extensive discussions on all matters related to the validation of toxicity test procedures and to produce an authoritative report, to be published in a peer-review journal for the guidance of researchers, regulators and others.

It is recognised that several approaches to validation may be scientifically acceptable. The plan described in this report is proposed as only one such approach. It is hoped that this report will stimulate and widen discussion within the scientific community, resulting in a better definition of the validation process. Individuals with comments on this report are encouraged to send them to Michael Balls (ERGATT, c/o FRAME, 34 Stoney Street, Nottingham NG1 1NB, UK) or John Frazier (CAAT, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205, USA).

The report which follows is the culmination of the efforts of the participants in this Workshop and represents a consensus of their opinions as individual scientists. It is our expectation that a better understanding of the validation process will facilitate technology transfer and the acceptance of new testing procedures for use in hazard evaluation of chemicals.

The Workshop was generously supported by a donation from the Proctor & Gamble Company (Miami Valley Laboratories, P.O. Box 398707, Cincinnati, OH 45239, USA). FRAME (Fund for the Replacement of Animals in Medical Experiments) kindly provided the secretariat. The organisers are particularly grateful to Susi Goll (Fonds für Versuchstierfreie Forschung, Biberlinstrasse 5, 8032 Zurich, Switzerland) for her help in making the local arrangements, and to Vivienne Hunter (FRAME) for secretarial assistance before, during and after the Workshop.

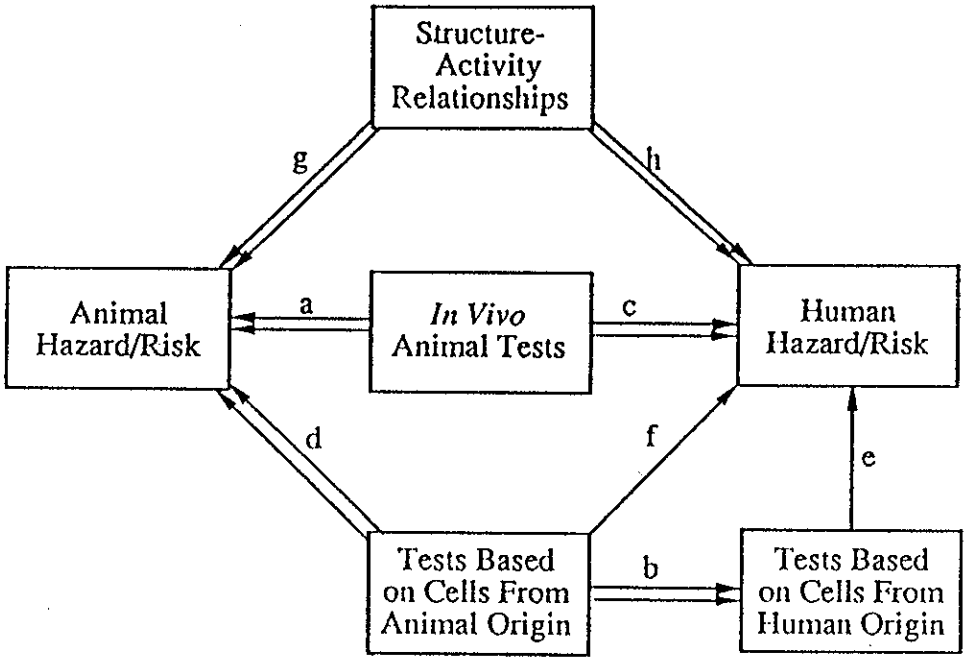
Introduction

The widespread use of chemicals in modern society requires that the risks associated with their use must be determined. Toxicology is the science which considers these societal issues. The two most important aims of toxicology are to determine the possible adverse effects of chemicals and to evaluate the particular risk they may pose for man, as well as for other organisms and ecosystems.

In most cases, hazard identification and risk evaluation are based on results derived from animal experiments. This reliance on animal data, however, results in a number of problems and concerns. Firstly, the results must be extrapolated from animal models to humans in order to assess their relevance, and this cannot be accomplished without introducing a significant degree of uncertainty. Secondly, the use of animals raises ethical questions, especially since many *in vivo* animal procedures in toxicology inevitably result in animal distress. These issues, together with advances in biotechnology, have led to the introduction of a number of new approaches that have less or no dependence on whole animal procedures. Many of these newer methods consist of the use of theoretical models or of cell and tissue cultures. These procedures also have a number of other advantages. The range of experimental possibilities is much greater than for *in vivo* animal methods, and the experimental conditions are often more readily controllable. Therefore, they can often provide the opportunity for a detailed analysis of the toxic mechanism(s) of action of a chemical at the molecular, cellular, tissue or organ levels. Furthermore, results from these procedures can be incorporated into chemical hazard evaluation. It is to be expected that this approach will lead to better use of all available toxicological information, to less dependence on animal testing, and to a more appropriate use of remaining animal tests. With respect to this last point, the performance of non-animal procedures can precede animal testing, leading to more-efficient experimental design and reduce animal distress. Moreover, extrapolation models from animals to man can be improved by the use of these methods (Figure 1). Eventually, non-animal procedures may come to replace all animal tests.

To derive the maximum value from new methods and to ensure their acceptance, they

Figure 1: The relationships between biological test systems to be used in toxicological hazard evaluation



a =animal to animal extrapolations; b & c = animal to human extrapolation; d & e = in vitro/in vivo extrapolations; f = a combination of in vitro to *in vivo* extrapolations; g & h = *computer-based predictions of animal and human hazard. These comparisons must also take into consideration many other biologically-important factors, including kinetics. Double arrows indicate extrapolations which are currently being made or are being investigated experimentally; single arrows, e & f = extrapolation procedures that may be of great value in overcoming current problems in extrapolation but require additional research effort (Adapted from M. Wooder, Stockholm, 1982).*

must be fully and properly validated. This document will discuss and describe the validation process, the main aim of which is to make available reliable and relevant methods that can be used for specific purposes in toxicology and toxicity testing.

Following significant discussion of what constitutes the proper boundaries of the validation process, the Workshop participants agreed that **test development**, the steps involved in establishing and defining a new procedure, and **acceptance**, the steps involved in taking the decision to use a particular procedure for a specified purpose, are not legitimate components of the validation process. New tests must be fully developed before they are admitted to the validation process. This does not mean that fine-tuning of the test cannot occur during certain preliminary stages of validation. However, once a procedure has begun validation, it must successfully proceed through four stages to qualify as a validated test namely: **intralaboratory** assessment, **interlaboratory** assessment, test **database development** and, finally, **evaluation**. At the end of this process, a new procedure may be described as scientifically validated for a

specific purpose, if it has performed satisfactorily according to defined performance criteria. At this time, industry or regulatory authorities will have adequate documentation for making a decision as to whether or not the scientifically validated test can be considered acceptable for their particular needs. These steps are summarised in Figure 2.

Purpose of Validation Studies and Test Selection

The main areas of toxicology on which various tests developed during the last decade have focused are: all aspects of acute toxicity; reproductive toxicology; mutagenicity/carcinogenicity; and target organ toxicity. In most of these areas, several individual tests have been actively evaluated, while many others are in the initial processes of development and validation.

Test selection for inclusion in any validation study can only be considered in the context of the specific purpose for the study. This purpose must be carefully defined and specified by a series of descriptors (Table I). The first descriptor is related to the level of toxicological assessment, i.e. to:

a) toxic potential of chemicals
b) toxic potency and the classification of chemicals

Figure 2: The major steps in the pathway by which new procedures are developed, validated and accepted

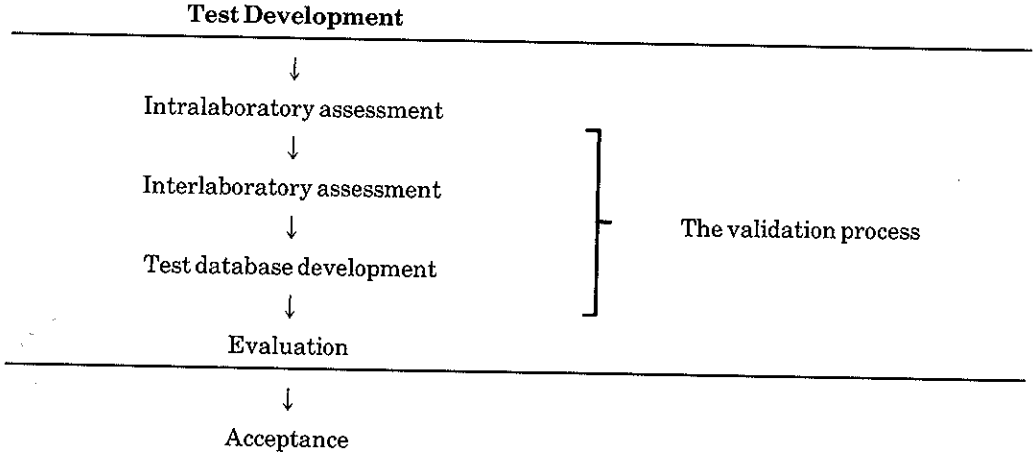


Table I: Descriptors required to define fully the purpose of a validation exercise

I.	**Levels of Toxicity Testing: Potential/Potency[1]/Hazard/Risk**
II.	**Type of Testing: Screening/Adjunct/Replacement**
III.	**Type of Toxicity: Ocular Irritation/Neurotoxicity/Hypersensitivity/Phytotoxicity etc.**
IV.	**Chemical Spectrum: Universe of all chemicals/particular chemical classes**

[1] *When testing for potency, it should be made clear whether the purpose is for ranking, priority setting or classification.*

c) hazard associated with the chemical
d) risk to specific populations represented by the chemical

In this context, **toxic potential** is an inherent property of a chemical. Potential to produce a toxic effect can be determined experimentally or theoretically. Many tests may be considered relevant for this testing purpose, even though they may not take into consideration such exposure of cells to the chemical as would occur *in vivo*. Thus, any properly-developed correlative or mechanistic test can be considered for inclusion in a validation study for the assessment of potential.

Since **toxic potency** is a relative measure of the toxicity of a chemical, it can be determined experimentally (dose-response relationships). The scale which is used to quantitate potency varies from one test to another, so potency is not an absolute value. However, within the confines of a given test, ranking and classification of chemicals as to their toxicity can be accomplished.

Hazard relates to the expression of toxic potential under specified conditions of exposure. It can be influenced by numerous factors, including toxicokinetics. Consequently, criteria for test selection for this purpose become more stringent, since influences other than the assessment of toxic potential are involved.

Risk assessment involves consideration of data gathered from *in vivo* and *in vitro* studies, together with information and experience obtained from other sources, e.g. exposure, and makes use of extrapolation procedures to predict toxicity in the target organism. Although the ultimate goal for *in vitro* systems is that they are validated for use in risk assessment, no single *in vitro* test has yet attained this goal. It is likely that a battery of tests will be required for this particular purpose in relation to any given potential hazard.

One of these four levels of assessment must be selected when defining the purpose of a validation study. In addition, the purpose of the study must include a descriptor relating to the type of testing activity required:

Screening tests: simple, rapid and inexpensive tests for use in making preliminary decisions or in setting priorities among large groups of chemicals for selection for further testing, either with intact animals or more-sophisticated *in vitro* tests;

Adjunct tests: tests conducted in conjunction with animal tests to evaluate toxicity, to elucidate mechanisms of toxicity or to further investigate specific observations without the need to perform additional animal studies;

Replacement tests: tests developed with the intent of replacing existing *in vivo* or *in vitro* tests, in all respects, including regulatory acceptance.

Other descriptors which must be included in defining the purpose of a validation study are:

a) the type of toxicity to be evaluated (e.g. ocular irritation, hepatotoxicity, teratogenicity); and
b) any restrictions on the chemical classes to be evaluated, which could reduce the breadth of applicability of the validated test (e.g. applicable only to chlorinated hydrocarbons, surfactants or aromatic amines).

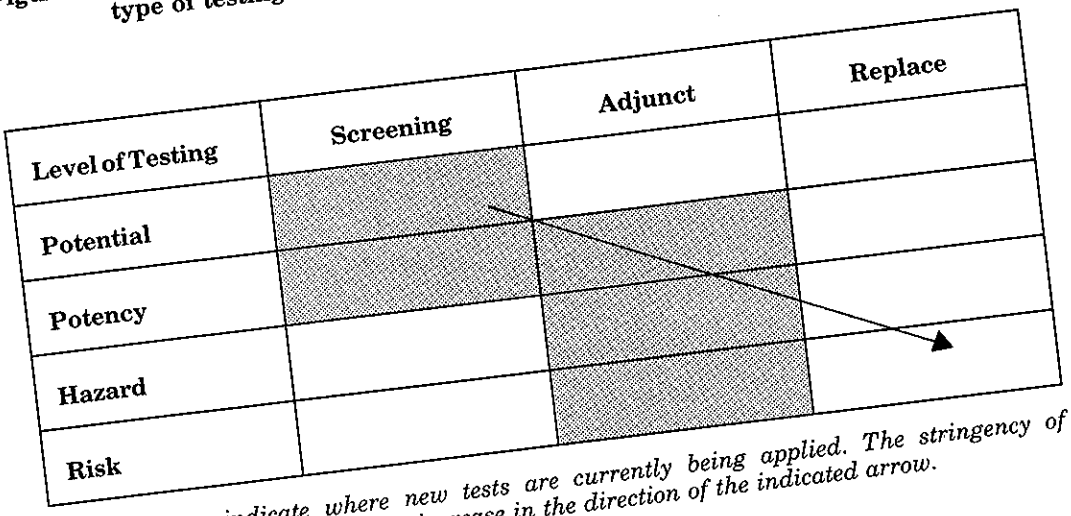
This set of descriptors will fully define the purpose of any validation study (Table I and Figure 3). Tests should only be considered for inclusion in a study if the specific purposes for which they have been developed are defined and consistent with the needs of the study. Consideration *must* be given to the degree of complexity of the individual tests proposed. In addition, tests *must* have been properly developed, and a need for them in relation to the availability of other tests *must* have been demonstrated.

In summary, the usefulness of a test to a validation study depends strongly on whether it is appropriate for solving a given problem. There is at present no single test which can be universally applied to identify all the chemicals which interfere (or could interfere) with biological systems or parts of them. There are many instances, however, in which tests are already developed and can be used for specific purposes, such as for selecting materials with lower toxic potential at an early stage of product development, or to provide information on the toxicity of chemicals which is not otherwise available from *in vivo* testing.

Chemicals Selection

The selection of chemicals in conjunction with

Figure 3: Classification scheme for tests based on level of toxicity assessment and type of testing activity



Shaded areas indicate where new tests are currently being applied. The stringency of performance criteria will always increase in the direction of the indicated arrow.

their appropriate reference classification is a critical aspect in evaluating the relevance of a test during the validation process. In an ideal validation programme, a test would be validated with a diverse selection of chemicals representing all the foreseeable areas of chemistry to which the test might be applied. Such a selection would include chemicals of different structural groupings, physical forms, physicochemical properties, mechanisms of action and spectra of broad biological activities. In statistical terms, the validation process would be most effective if these chemical types were randomly distributed among the classification groups that were to be subjected to the test.

Common sense dictates that it is not practicable to include in one validation programme, the large numbers of chemicals that would be required to represent all possible combinations of chemical properties. Consequently, the validation process must acknowledge the limited diversity of the chemicals selected and this must, in turn, dictate the structural areas within which the test is credited as being valid. It therefore follows that, for a test to be considered to have been validated for chemical groups not included in this validation process, further validation studies would be required.

Considerable debate has been directed at defining the numbers of chemicals that are required at each level of the validation process. It is generally accepted that increasing the numbers of chemicals used will increase the statistical confidence in the estimates of test performance measures, i.e., sensitivity and specificity. Nevertheless, it also increases costs and the time taken. Therefore, for the various stages identified in the development and validation of a method, the following numbers of chemicals are proposed for consideration (Table II).

Test development

An adequate number of chemicals should be used for the development of the method in the laboratory of origin. No fixed number of chemicals can be said to satisfy all the requirements for test development. However, a set of calibration chemicals (**Reference Set 1**, 5–10 chemicals with a range of potencies with respect to the biological endpoint of concern) should be available to investigators at this stage.

Intralaboratory assessment

Two sets of chemicals are used at this stage. **Reference Set 1**, (also referred to as the Calibration Chemical Set) should be used to calibrate the endpoint measurement of the method.

A second set of between 20 and 50

Table II: Sets of reference chemicals required for test validation

Reference Set 1:	Consisting of 5–10 chemicals, which can be used to calibrate test systems — also referred to as the **Calibration Chemical Set**.
Reference Set 2:	Consisting of 20–50 chemicals, which are employed in the extended evaluations of a test in the laboratory of origin — also referred to as the **Intralaboratory Reference Chemical Set**.
Reference Set 3:	Consisting of 10–20 chemicals which are used to evaluate interlaboratory reproducibility — also referred to as the **Interlaboratory Reference Chemical Set**.
Reference Set 4:	Consisting of 200–250 chemicals from widely distributed chemical classes, which represent a subset of the universe of chemicals and are used to fully develop the test database — also referred to as the **Database Reference Chemical Set**.

intralaboratory reference chemicals (**Reference Set 2**) should be used to evaluate the performance of the method in the laboratory of origin. This set of chemicals should be coded and tested blind by the researcher, to prevent the introduction of bias.

Interlaboratory assessment

As with the intralaboratory assessment, the calibration set of chemicals (Reference Set 1) can be used in a Preliminary Phase (see later section on Validation Programme Design and Practical Considerations) to standardise the test response in the individual laboratories taking part in the exercise. These chemicals can be used during the establishment of the method in each participating laboratory and in training personnel.

In order to determine reproducibility of the method among the participant laboratories, an interlaboratory reference set of chemicals (**Reference Set 3**) is required for the Definitive Phase. This set should overlap with the chemicals contained in Reference Set 2, in order to give some assurance that the performance observed in the intralaboratory assessment will be predictive of the potential to succeed in the subsequent validation exercise.

Assessment of reproducibility must involve the testing of chemicals under blind conditions. The precise numbers of chemicals required for this stage will depend upon the number of laboratories involved and should be established on the basis of statistical considerations.

Test database development

The development of the test database will provide the ultimate data set upon which the performance of the test can be measured. Sufficient chemicals should be selected to ensure that there is representation of all the foreseeable areas of application. For widespread application, as many as 250 diverse chemicals (**Reference Set 4**) may be required, but they can overlap with the chemicals in sets 2 and 3. It is acknowledged that there will be exceptions, e.g. where the proposed area of application is restricted to defined chemical groups. For such exceptions, fewer chemicals will be required. Creative approaches to chemical selection are needed to reduce the magnitude of this task (see, for example, 4).

As with both intralaboratory and interlaboratory assessment, this part of validation should be conducted under blind conditions. The process of database compilation should have appropriate quality control, which may involve the use of specific chemicals of known activity, also tested under blind conditions, which do not compromise the integrity of the overall blind study.

Chemicals selected for the validation of one biological endpoint will not necessarily be appropriate for the validation of another biological endpoint. Consequently, several different sets of chemicals will need to be compiled for the wide range of biological endpoints for which tests may be developed. However, some chemicals will be acceptable for the validation of more than one biological endpoint.

The definition and provision of sets of reference chemicals will permit the direct comparison of different tests developed for the prediction of the same biological endpoint, thereby permitting the selection of tests on the basis of common performance criteria. It is proposed that these sets of chemicals should form the basis of a chemical reference bank in order to facilitate the provision of chemicals and reference data for the validation of tests internationally.

The Use of *In Vivo* Toxicological Data for Classification of Reference Chemicals

The objective of collecting data derived from humans and animals is to provide a basis for classifying chemicals with respect to specific biological endpoints, which will serve as a reference against which the performance of tests developed to predict these endpoints can be judged.

Although ecotoxicological concerns are important, most toxicological data are generated with the aim of predicting effects in humans. In many situations, animal species are used as surrogates for humans. Thus, in practice, *in vivo* toxicological studies provide three forms of reference data, i.e., data generated in experimental animals, in other animals, and in humans.

Species differences in responses to chemicals can lead to differences in reference classification, so it is critical that tests are validated against the appropriate reference data in order to comply with the objective of the prediction. Ideally, reference data generated in appropriate animals should be used if effects in animals are to be predicted, while reference data generated in humans should be used if effects in humans are to be predicted. In practice, this latter goal can rarely be achieved at present.

Toxicological data experimentally generated in animals under laboratory conditions

Over the past 50 years, a considerable volume of toxicological data has been generated in laboratory animals on a wide range of chemicals and mixtures. Nevertheless, recent attempts at assessing the completeness of these toxicological profiles on existing chemicals, have shown that the data are not comprehensive for most chemicals, and that, for many chemicals, little or no data exist.

Many of the *in vivo* test methods used for producing these data are complex, since many factors influence both exposure to the chemical and the expression of toxicity (Table III). These factors, together with differences in assessment and interpretation of the toxicity, will lead to the generation of diverse data.. Finally, not all the data generated by toxicologists are available for review, as much of this information has not been considered "suitable" for publication, e.g. it comprises negative data or concerns proprietary chemicals. Consequently, even the most

Table III: Factors which contribute to variability in *in vivo* data

Methods:	— test method — interpretation of method — exposure route, vehicles, doses, duration, etc.
Animals:	— species, strains — diet, maintenance conditions — health status — age, sex
Data:	— incomplete data — unavailability of raw data — data handling, manipulation — lack of GLP compliance — inadequate number of test animals
Classification:	— undefined ambiguous criteria — absence of confidence limits for accuracy of classification

comprehensive literature search performed by a competent information scientist will probably have to be supplemented by direct requests to industry that previously-unpublished data be made available.

Industry should play a stronger role in releasing and publishing data which are not currently available, specifically for validation purposes. In those circumstances where insufficient data are available and/or the data are not of adequate quality to ensure the reliable classification of a chemical, then the only options available are either to substitute an alternative chemical or to test the chemical *in vivo*. The testing of chemicals *in vivo* solely for test validation purposes will inevitably raise ethical questions.

Human toxicological data

Worldwide, much effort has been invested in collecting human toxicological data, and the available sources of such data can be categorised under the following six headings:

1. The medical and scientific literature — reports of individual cases and reviews.
2. Regulatory authorities — raw data for risk evaluation, as submitted by industrial and pharmaceutical companies for specific regulatory purposes.
3. Industrial companies — data from clinical trials, human volunteer studies, health and safety monitoring of workers, reports of adverse reactions, post-marketing surveillance, and poisoning incidents.
4. Poison Information Centres (PICs) and Drug Information Centres (DICs) — which collect data for a number of purposes, including diagnosis, risk assessment and management of poisoning, together with surveillance of outcome for the assessment of treatment and prevention (toxico-vigilance). Data collected by such routine surveillance schemes is often subject to bias and may be incomplete for the purpose of studying a particular chemical or chemical group. However, it can be used as a basis for the design of more-complete studies.
5. Other national/international organisations — which may conduct epidemiological surveys (e.g. adverse drug reaction monitoring schemes, the UN Environment Programme).
6. Epidemiological studies — including those developed within the toxicology monitoring services of PICs.

The nature of data obtained from these sources is highly variable, primarily because they are collected and collated for a variety of purposes. Consequently, it is not easy to use them in a uniform fashion in a reference database.

Problems in the use of human toxicological data in a reference data base include the following:

a) human data are diverse in nature and quality;
b) data frequently relate to formulated products rather than to single chemical compounds;
c) conditions of exposure are not controlled;
d) members of human populations are frequently exposed to more than one chemical or product;
e) methods of data collection vary;
f) data frequently describe symptoms (e.g. headache), rather than representing objective measures of toxic effects.

PICs are probably the most useful and accessible source of data on acute human poisoning (5). It is the aim of recent initiatives undertaken by the Commission of the European Communities (CEC) and the International Programme on Chemical Safety (IPCS-WHO/ILO/UNEP) to promote standardisation of data collection by PICs and to make such information more readily available on a worldwide basis. Ironically, the least accessible raw data are held by industrial companies, although they have the greatest resources and experience in the storage and handling of such information.

Because human data are available from such a diverse range of sources, it is not easy to incorporate them into a uniform reference data base. Although data sets can be readily assessed individually, only in a few cases is it possible to compare them. The available human data are being used to identify areas for further research, e.g. where there are gaps in knowledge of a particular chemical or exposed population. In a very few cases, human data are used to assess the relevance to man of data generated *in vitro* and *in vivo* in laboratory studies (e.g. 6).

In order that better use can be made of all available data by toxicologists, epidemiologists, regulatory authorities, international organisations, governments, and others, it is essential that:

1. methods of data collection, collation,

evaluation and expression be standardised, wherever this is desirable and practicable;
2. the data and data sets themselves are validated in standardised ways which have been agreed with, and are acceptable to, the toxicological community as a whole;
3. the data be made freely available to the scientific community.

It is also clear that most data on human toxicology differ widely from data derived from laboratory experiments on animals. This is mainly because of differences in the purpose for collecting the data. Experimental toxicologists generally measure specific biological endpoints after exposure of animals to particular compounds at particular doses. Human or clinical toxicologists tend to be concerned with the incidence of poisoning caused by particular chemicals or products in a certain population. They may also need to evaluate the use of a specific treatment regime in particular cases of poisoning. As a result of these different scientific objectives in experimental and human toxicology, the data obtained are presented and evaluated in very different ways. In practice, this means that information obtained in one discipline is rarely used in the other. Because of this, two separate bodies of literature have evolved. It is recommended that information toxicologists be encouraged to investigate ways of integrating these different types of information and making their results available for general use.

Data derived from veterinary and wildlife toxicology

As in the case of human toxicology, veterinary toxicology involves the collection of data from a variety of sources and for a variety of purposes, which are not always compatible with each other. For example, data may be received on the acute effects of household products on domestic pets following accidental exposure, or on acute and chronic data on the poisoning of farm animals and wild animals by industrial chemicals and agrochemicals.

It is highly desirable that more use be made of such information. For example, experience with pets and wildlife could be used, more frequently than at present, to warn of hazard within the home or in the environment in general. Also, veterinary clinical data could contribute to assessment of the hazards represented by particular chemicals or products, by assisting in the extrapolation of experimental results with laboratory animals to estimating likely hazard in man.

Assessment of toxicological data and reference classification

Despite dissimilarities between toxicological data generated in animals and humans, there exists a common approach to the assessment of such data. Before the assessment process can be started, all available data should be collected and, wherever possible, this should include both original and raw data. The assessment process should be conducted by a number of experts, commonly referred to as a peer-review panel. Prior to review of the data, the panel should generate a profile of those factors that influence exposure to the chemical and the expression of toxicity (Table III) and should estimate the likely influence of each factor on the available database.

The process of assessment includes the individual review of each piece of information for compliance with the objective determinants for the biological endpoint, and the reliability of the data, both in terms of the reputation of the source and in relation to other data reported on standard or known chemicals.

Once individual data have been accepted, the weight of evidence must be established, and an assessment of the confidence or precision of the final classification must be made. In this process, some data subsets may be found to be inconsistent with the reference classification decided upon on the weight of evidence. In such cases, data subsets should be fully evaluated in order to eliminate the introduction of bias into the judgement process.

Validation Programme Design and Practical Considerations

When a validation study is being planned, its objectives must be clearly defined, as must the ways in which they will be achieved. Experienced toxicologists and statisticians are essential for the study design and for the estimation of time and costs. Sufficient funding should be available to provide for an optimal programme design and to ensure completion of the study. The programme

design itself should meet recognised international standards (e.g. ISO 5725 1981). To avoid unnecessary duplication of effort, national and international agencies (e.g. CEC, OECD, WHO), as well as associations of the chemical, cosmetic and pharmaceutical industries, should be informed, at an early stage of planning, of the objectives and scope of all validation programmes, and regular updates on progress should be provided. It is also essential that the results of all validation programmes be published in the peer-review literature, made available to all interested parties and deposited in specialised data banks, such as *INVITTOX* (7), and that full descriptions of the techniques and statistical methods involved should be included.

In general, a validation study can be conducted for a single test. However, as a practical consideration, it would be more efficient to evaluate several tests simultaneously. A particular advantage of conducting multiple-test validation studies is the ability to select the best test system (or combination of test systems) for a particular purpose from among those validated.

Intralaboratory assessment

Reproducibility and performance should be tested on a broad spectrum of chemicals or a particular group for which the test is designed. The technical problems involved in testing toxic chemicals in *in vitro* systems have been extensively discussed by Frazier & Bradlaw (8).

Interlaboratory assessment

Determination of reproducibility within and among laboratories and their comparability is the aim of this stage of validation. Since time and money may be limiting factors, the numbers of laboratories and of test chemicals may be less than optimal, but in no case should this compromise the statistical validity of the study. Criteria for the choice of chemicals have been described above.

1. *Preliminary phase*

A final decision on the test protocol, including a range-finding procedure to establish appropriate dosages, selection of test chemicals, selection of participating laboratories, training of personnel, and, in particular, planning of scientific coordination, is essential at this stage.

Also during this stage, unforeseen problems in the handling and transfer of data, distributing and handling of chemicals, including safety measures at the work place, and, last but not least, time and costs required for testing of each chemical in the interlaboratory validation, are identified and, where necessary, resolved. Since many technical problems that are part of test refinement may be recognized at this stage, scientific coordination is especially important. To reduce costs and to avoid problems during transfer of data, the development of test-specific software systems for data management to provide for efficient collection and analysis of all data, is recommended.

The aim of the preliminary phase, which should end with a short test run under the conditions of the Definitive Phase which is to follow, is to ensure that the interlaboratory assessment will not fail because of technical or logistical reasons. Before initiating the Definitive Phase, agreement must be reached on the final form of the standard test protocols.

2. *Definitive phase*

Quality control and a blind trial with coded chemicals are essential parts of this stage. Quality control should ensure that every laboratory is working strictly according to the test protocol. Although not all laboratories can conform to the exact requirements of Good Laboratory Practices, all participating laboratories should strive to attain maximum standards of performance. A basic set of information on the physicochemical properties of the chemicals should be provided for each of the coded chemicals of the blind trial. The specific information provided should be consistent with that usually available for routine toxicological testing, such as pH, density, solubility and stability in solvents/vehicles. All test chemicals should be treated as if they were potentially harmful, and safety instructions for use in any emergency should be deposited in sealed envelopes with a responsible authority for each participating laboratory.

Test database development

At the end of the interlaboratory assessment, it should be possible to decide whether it is worthwhile to proceed to the next stage of validation, namely, test database development. At this stage, a group of 200–250 carefully selected and coded reference

chemicals should be tested in a small number of laboratories, according to the previously-agreed standard procedures. The 200–250 chemicals need not all be tested in all the participating laboratories or even in a single laboratory (e.g, the overlapping of data sets from several laboratories is sufficient). The Database Reference Chemicals can be distributed among several testing laboratories by means of statistically defined overlapping subsets. Periodic inspection of testing activities should be conducted for quality control. Calibration chemicals (positive controls) should be periodically included in the coded chemical set, to confirm test performance. All coded data should be transferred to the programme data bank for final evaluation.

Evaluation

The codes should be broken by the project manager only when all the data for all the chemicals have been collected. The analysis of the data should be conducted in collaboration with investigators involved in the experimental studies, using procedures established during the planning stage of the study. Investigators who have a vested interest in the outcome of the validation study, e.g. individuals who developed specific tests under evaluation, should not participate in the evaluation process, since a conflict of interests could arise. The results should be evaluated by comparison with the validation criteria and final decisions made. The bases for these decisions are described in the next section.

Criteria for Evaluation of Validation Activities

For many reasons, it is difficult to establish a set of criteria that would be applicable to the validation of all types of procedures. The bases of the procedures are diverse, even including computer-based SAR analyses. However, several general criteria appear to be relevant to the validation of most tests or procedures, and these will be described in this section.

Tests that are selected for validation, often have specific, rather than general, applications in chemical hazard assessment. Some may be designed to evaluate specific groups of chemicals (e.g. irritants) and others may be intended for the replacement of existing animal tests. One test classification scheme can be based on the type of testing involved, e.g. screening, adjunct or replacement. It is generally accepted that the validation criteria for a test developed as a screening test would be less rigorous than those required for an adjunct test, which in turn would be less than those for a procedure intended to replace an existing animal test (Figure 3).

The criteria described below are applicable both to specific stages of the validation process and also to assessment of whether or not the complete validation process for a procedure has been successful. The criteria identified for assessing the outcome of a validation project are restricted by the status of all relevant factors at the time when each stage of the validation process was completed and the results were analysed. Validation is an on-going process and the evaluation of test performance could change with changes in technology or with changes in the reference classification.

Test performance

1. Reproducibility

Definitions of reproducibility differ across the range of procedures. For tests with binary or response classification results, reproducibility can be determined by the frequency of concordant responses over N number of trials for the same chemical. Methods for assessing reproducibility for other tests, for example, cytotoxicity tests which express their results in a single number (e.g. ID50), are more difficult to specify, and for computer-based tests (e.g. SAR procedures), reproducibility is not applicable. For the **intralaboratory** phase of validation, the degree of reproducibility should be determined and reported, but no specific decision criteria will be offered, since this forms part of the judgement as to whether the test will proceed to the next phase. It is expected that low intralaboratory reproducibility (e.g. 70%) will raise serious doubts concerning the likelihood that the test performance will ultimately satisfy the minimum validation criteria.

Reproducibility in the **interlaboratory** assessment phase is considered to be of major importance in judging the transferability of a test and the success of the validation process. Interlaboratory assessments should be designed to involve several (**preferably**

four) laboratories and blind evaluation of the reference compounds. A calibration set of chemicals should be established to provide an initial frame of reference for responses. An estimate of the number of chemicals required for the validation of a procedure ranges from 10 to 20, but may be dependent upon the characteristics of the reference database. Interlaboratory assessment can be designed in such a way that all chemicals do not have to be evaluated by each of the laboratories involved in the process. The following **lower limits** for interlaboratory reproducibility criteria are proposed:

Screening tests: c. 70%

Adjunct tests: c. 70%

Replacement tests: c. 80% (or better than the reproducibility of the particular animal test concerned).

Test performances below these lower limits are clearly unacceptable. In addition, each participating laboratory should be able to attain an intralaboratory reproducibility equivalent to that obtained by the originating laboratory in the earlier intralaboratory assessment.

These requirements will need to be modified on a case-by-case basis.

2. *Relevance*

(i) Correlation with reference classification
Since most, if not all, of the methods subjected to formal validation are predictive, the degree of predictivity that will satisfy the intended application of the procedure must be established. For most tests (those with binary responses or which produce response classifications), levels of sensitivity and specificity can be calculated from the data generated in the database development, and from these data, a predictivity value can be determined. Test performance will be determined in the evaluation phase. Approximately 250 chemicals will be needed for the general validation of a procedure. A description of the methods involved in the calculation of predictive value for a binary classification scheme is given in Appendix B. Some procedures which involve continuous measures of response, such as tests for cytotoxicity, are not amenable to sensitivity and specificity calculations as defined, and thus cannot meet this criterion. New methods to evaluate performance of these types of tests are needed. The minimum specifications proposed for predictivity are:

Screening tests: predictive value >0.50 (e.g. better than random classification)

Adjunct tests: predictive value ≥0.75

Replacement tests: predictive value ≥0.90

The actual predictive value of a screening test will depend on the purpose for which it is used, but should be at least better than tossing a coin!

It should be noted that, when individual tests are combined into a battery, the predictivity of the individual tests may be less demanding than if the tests were used in isolation. Thus, a test with a relatively low predictive value in general, i.e., for a wide spectrum of chemical, may have a high predictive value for a specific chemical class or for a group of chemicals which act through a specific mechanism. In this case, the test may have considerable value as a component of a test battery.

ii) Mechanistic similarity
In moving from a screening activity to the replacement of an existing animal procedure, the degree of similarity between what is being measured in the predictive test and in the animal model, becomes more important. While examples of real and hypothetical exceptions to this requirement can be given, the ability to demonstrate a close mechanistic relationship will enhance the acceptance of the data and provide confidence that a previously contested chemical will be properly assessed and/or classified. The recommended criteria for this attribute of tests are:

Screening tests: mechanistic similarity is not necessary if the empirical correlation/ predictivity criterion is met;

Adjunct tests: mechanistic relatedness is desired, but not required, for tests used for this purpose;

Replacement tests: mechanistic similarity for replacement tests is generally required, but exceptions could exist.

(iii) Logistical considerations
This refers to a group of attributes of a procedure or test, which determine its acceptability for routine application in toxicology laboratories (Table IV). The attributes included are:

1. Cost per chemical evaluated in the test;

Table IV: Criteria for logistical considerations

Criteria	Screening Test	Adjunct Test	Replacement Test
Cost per test	Minimal	Moderate	Probably not a limiting factor; cost of the animal test may be a guide
Performance time	Rapid	Rapid	Not a limiting maintenance factor
Installation cost	Minimal	Moderate	Probably not a limiting factor, but important in comparing competing tests
Complexity	Must be simple to perform/evaluate/ interpret	Moderate complexity can be acceptable	High complexity is not necessarily a limiting factor

2. Time required to evaluate a chemical in the test and for maintenance of test components (e.g. growing stocks of cells, maintaining computer equipment);
3. Installation cost of the test in a laboratory (including training of personnel, equipment costs);
4. Complexity of performing the procedure and interpreting the results obtained.

Tests which are inappropriate for their intended use, because it is not feasible to apply them in the context of that use, are of little value. Again, in moving from screening tests to those intended for the replacement of animal procedures, the need for low performance costs, for example, becomes less limiting in assessing the acceptability of the test.

(iv) Evaluation

Determination of the success of the validation process for a given procedure will to some extent depend upon the purpose underlying its validation. For general validation purposes, however, it would be expected that a procedure intended as a *replacement* for an existing animal procedure would have to meet all the criteria associated with that intended application. For a *screening* procedure, the minimum would be to meet all the criteria and, hopefully, exceed them in as many areas as possible. For *adjunct* tests, the criteria are less well defined, because of the varied applications of procedures in an adjunct capacity. Therefore, conformity with all the general criteria is of less importance than for the other two applications. Again, it should be emphasised that it is recognised that the criteria will be modified as technology and scientific information change and advance. Improvements in the size and quality of the reference chemicals set used in the validation process may also occur, which would necessitate further testing and validation. Thus, validation of a test or procedure is an ongoing process, and the assessment of its success or failure in the context of this document can only be based on the conditions and the state-of-the-art which exist at the time.

Battery Selection

Once individual tests have been validated, it is possible to construct batteries of tests which can combine the high sensitivity of one test with the high specificity of another. Similarly, tests which reflect different mechanisms, as well as tests which respond to different classes of chemicals, can be combined. One major requirement of batteries is that individual tests should generally be independent of one another. This can be determined by cluster analysis and/or statistical methods (9–12). One convenient method for predicting the probability of a toxicological endpoint based upon the results of individual tests in a battery of *in vitro* tests is by application of

Bayes' formula (11, 13). Advantages of using this formula include the fact that it can handle combinations of positive and negative results which are then weighed in accordance with the documented response of the individual test, so it can also be applied to *in vitro* results that are expressed in dose-response curves. Moreover, computer programs are available to facilitate the calculations. Modelling of databases indicates that including 3–4 individual tests in a battery is generally sufficient for making accurate predictions.

In addition to the construction of batteries based upon validated *in vitro* assays, it is possible to approach the problem empirically by including the results of individual tests, including non-validated ones, in a multivariate regression analysis. While this retrofit may explain retrospectively the behaviour of a data set, its predictive ability will depend upon the quality of the data used to generate the regression equation. The predictivity will increase as the database is expanded, i.e., as individual tests are, in effect, validated (14).

The Bayesian approach can be similarly broadened by the inclusion of dynamic programming (11), which greatly reduces the mainframe computer time required for analysis.

The evaluation of the predictive performance of a battery must be subjected to the same criteria of evaluation with respect to the reference chemicals as are applied to individual tests.

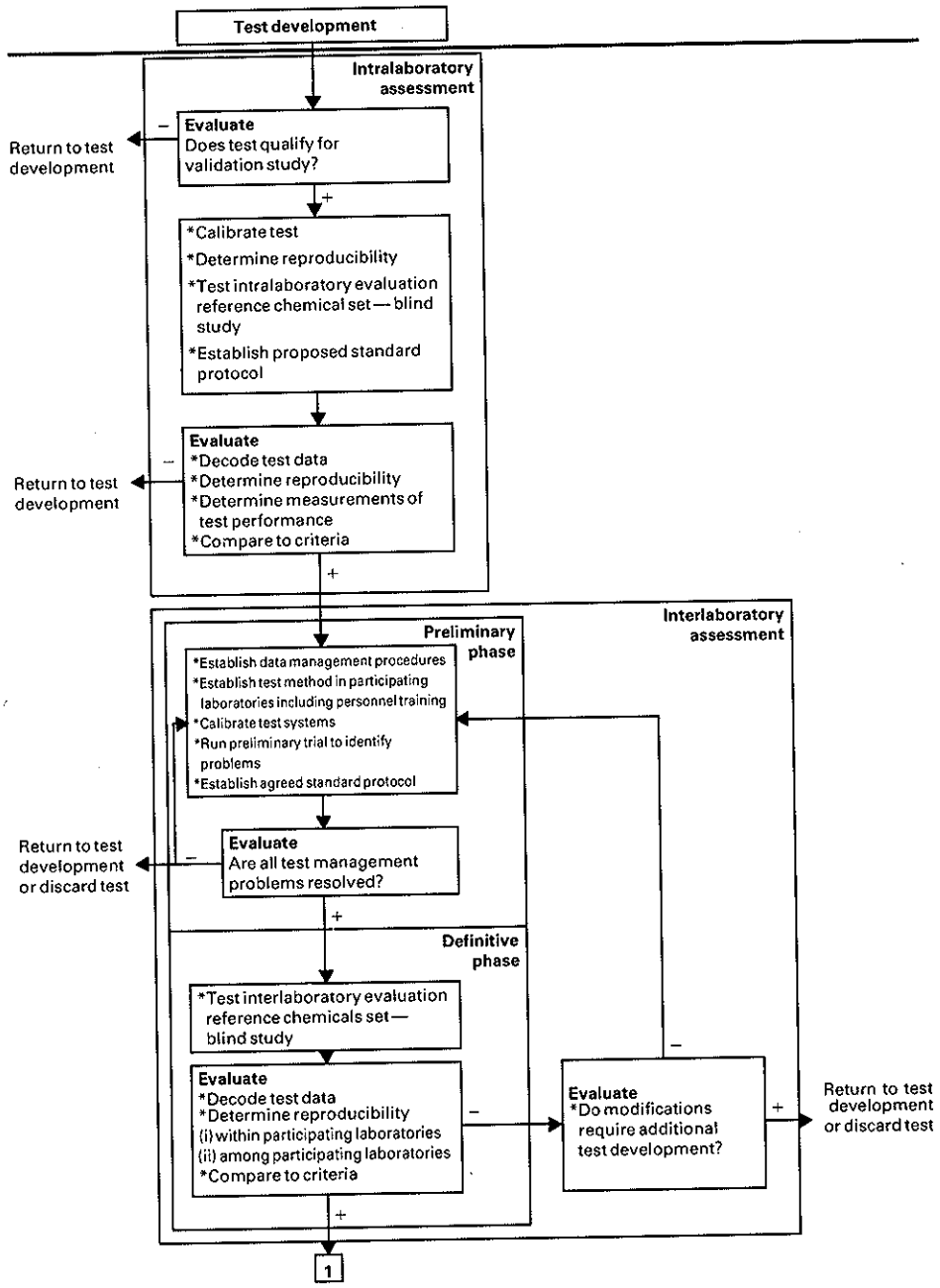
Integration of the Validation Process

Integration can be defined as the process by which information is carried over from one validation step to the next in order to maximise the potential for acceptance of appropriate tests. It may appear that this statement is in conflict with the absolute requirement that a test method must not be altered during validation and the accepted idea that test development precedes validation. However, this is the case only if one looks upon validation as a rigid, linear sequence of events taking place along an axis of time, and not as a number of interacting activities (Figure 4). It is of great importance that the information generated and experience derived from the validation process are not used only to judge test performance. There are ways by which this knowledge can be fed back into the test protocol, thereby optimising the test method for various purposes.

Test development is not a component of the validation study, although it may contain elements of validation. Therefore, the first recognised step of validation is the intralaboratory assessment (Figure 2). This will most probably take place in the laboratory where the test was developed and will involve a restricted number of reference chemicals (20–50), of which a subset (5–10) is used for calibration and the rest for evaluation. After testing the reference chemicals, evaluation is carried out according to the criteria outlined above (see section on Criteria for Evaluation of Validation Activities), and a decision is made as to whether the test should be further validated (+) or rejected (−). If a negative decision is made, the test can be returned to the test development stage for further improvement and can then be reassessed. This loop procedure is based on an analysis of why the test failed in the first assessment/evaluation cycle. By correlating *in vivo* toxicity classifications for the reference chemicals with the values generated by the test, outlyers can be identified and the reasons for their poor correlation can be investigated. Thereby, significant knowledge on the shortcomings of the test can be obtained, as well as new knowledge concerning mechanisms of toxicitity. It is thus possible to improve the test performance by further test development, taking this additional knowledge into consideration.

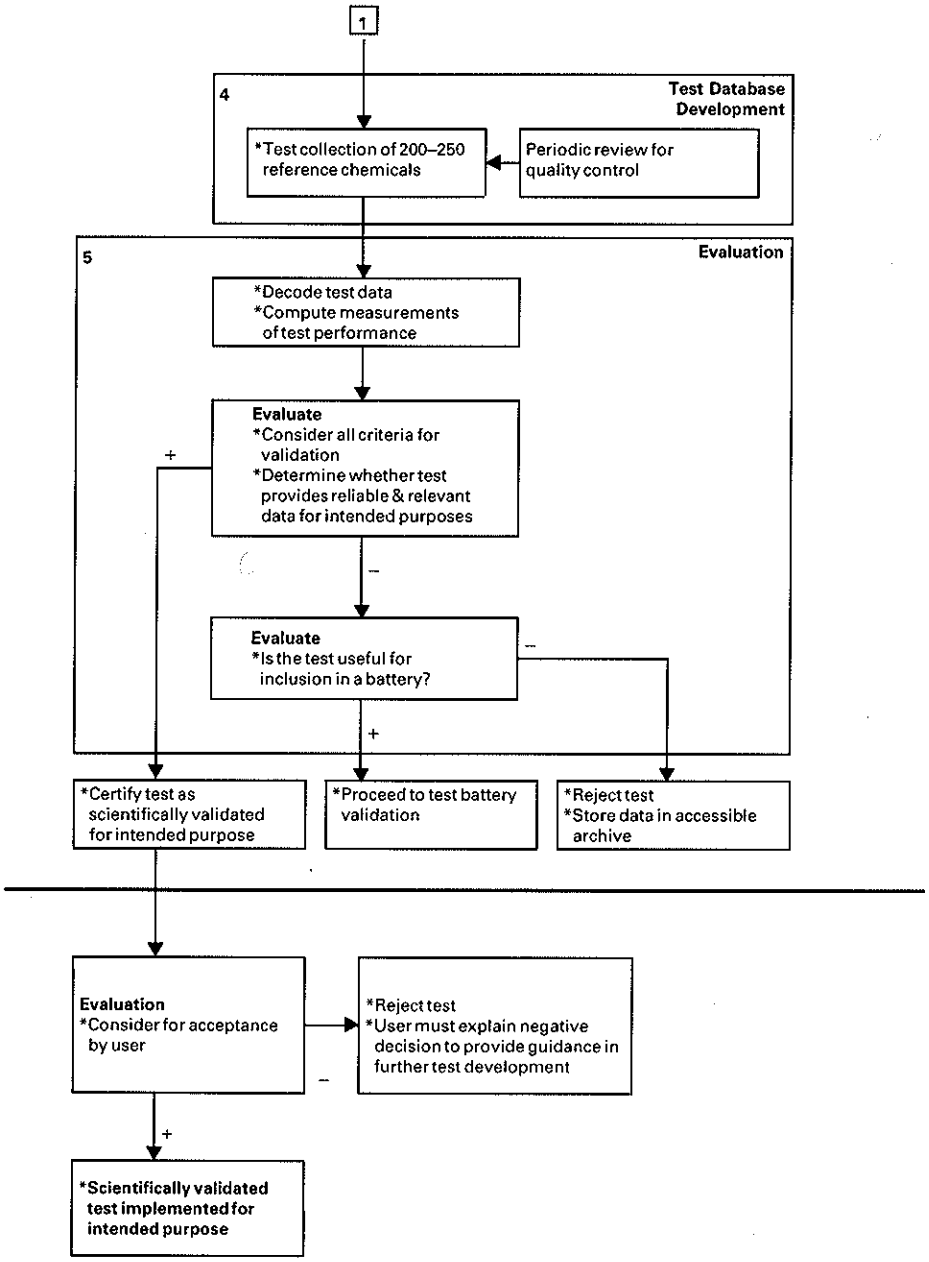
When the intralaboratory assessment results in a positive decision, the test is taken on to the interlaboratory assessment (Figure 4). During a Preliminary Phase, which should conclude with a short test run under the conditions of the Definitive Phase, it is ensured that all technical problems are resolved (see section on Programme Design and Practical Considerations). The Definitive Phase follows successful completion of the Preliminary Phase. This part of the validation assessment is carried out in several laboratories, and involves 10–20 reference chemicals. The number of chemicals to be used depends on the nature of the test that is being validated and on whether or not all the chemicals are to be tested in all the participating laboratories. The reference chemicals used in the interlaboratory

Figure 4: Flow chart for validation process



+ indicates that all applicable criteria involved in a particular evaluation are satisfied and it is possible to continue to the next stage of the process. A − means that the criteria for continuation are not satisfied.

Figure 4: continued



+ indicates that all applicable criteria involved in a particular evaluation are satisfied and it is possible to continue to the next stage of the process. A – means that the criteria for continuation are not satisfied.

assessment consist of the same Calibration Chemical Set that was used in the intralaboratory assessment, a selection of substances from the Intralaboratory Reference Chemical Set used in that assessment, and a subset of substances from the Database Reference Chemical Set. When the assessment and evaluation are completed, it is again decided whether or not validation should proceed to the next stage. If a negative decision is taken, the test can either be rejected or the new information obtained during the inter-laboratory assessment can be incorporated into the test procedure. The nature of the information generated will indicate whether the test performs better in some laboratories than in others. If the reasons for differences can be identified, important technical adjustments can be made. In some instances, it is sufficient to go back to the Preliminary Phase and restart the validation process there, whereas in other cases it will be necessary to go back to test development or discard the test altogether (if the problems are judged to be insurmountable).

When a positive decision is taken after the interlaboratory assessment, the test can proceed to database development and evaluation. Database development consists of blind testing a large collection of reference chemicals (200–250 chemicals, if the full spectrum of chemicals is to be represented; a smaller set can be tested for specialised purposes) and transmission of the coded data to the programme manager. During this stage of the validation study, it is important to maintain a high level of quality control, which is accomplished by periodic reviews of procedures, instrument calibration, and monitoring of results using specific quality control test samples. It is better to identify operational problems at an early stage, rather than wait until the entire project is completed only to find that the data obtained are flawed. All data generated by participating laboratories are transmitted to the programme manager, who decodes the data, and in collaboration with participating investigators, makes appropriate comparisons with the reference classification of the test chemicals.

After computation of various measures of test performance (sensitivity, selectivity, predictability, reproducibility), the results are compared to the agreed criteria for satisfactory test performance. Taking into consideration all the factors which will influence the final decision (see Criteria for Evaluation of Validation Activities), a judgement can be made as to whether the test is both reliable and relevant for the intended application. If the judgement is positive, then the test can be certified as "scientifically validated for the specified purpose". If a negative decision is made, i.e. the test is not satisfactory, by itself, for the intended purpose, then consideration as to whether or not the test can be used in a battery can be undertaken. If the answer to this second question is affirmative, then the test should be incorporated into a programme which will select an optimum test battery and validate it for the intended purpose. Often, this may not require any additional testing, since the database already generated may be sufficient for these purposes. If a negative decision is reached at this point, then there is little likelihood that further test development will be of any value and the test should be rejected. In all cases, the data generated in the validation study should be stored in a freely-accessible archive. This should include the *negative* outcome of a validation study, which should be published in order to prevent unnecessary duplication of effort.

This sequence of activities constitutes the validation process. However, successful completion of all phases of the validation exercise does not guarantee that a test will be incorporated into practical testing programmes. Potential users of the tests must make a decision as to whether a scientifically validated test is suitable for their particular purposes. Acceptance at this stage should result in the inclusion of tests in testing guidelines. The possibility exists that the test may be rejected. In this case, it is essential that the user concerned clearly defines the basis for a negative decision, in order to provide guidelines for further test development. Only then can progress toward full and proper utilisation of new tests be attained.

Implementation

International collaboration in the planning, conduct and evaluation of validation schemes is essential and should be encouraged and supported, as should communication of the results obtained and of the state of scientific

validation of new toxicity test procedures.

An essential aspect of the validation process envisaged in this report is the selection of chemicals, for which various criteria have been proposed (see sections on Chemicals Selection and *In Vivo* Toxicological Data for the Classification of Reference Chemicals). This selection process should also be a matter for international concern, as it is desirable that chemicals from the same scientifically-selected sets be widely available. It is therefore suggested that the establishment of an International Chemicals Reference Bank is a matter of urgency. This reference bank should provide:

a) open-access to scientifically-selected reference chemicals of known purity, backed up by comprehensive and authoritative reviews of all the available data on their toxicological activities;
b) safety advice for each chemical.

Many *in vitro* tests exist and are already in widespread use in industry and in research laboratories, and some of them have already been involved in validation procedures. Many more tests are currently in the course of development. *In vitro* toxicology data banks, such as *INVITTOX* (7), should also be developed and supported, so as to guarantee that information concerning test methodology and test validation, including detailed protocols and results obtained in validation studies, can be made freely available to all interested parties.

The concept of validation, and the procedures involved, as discussed in this report, should be fully discussed among the general public and in the scientific and industrial communities in general, but, in particular, among students of toxicology and established applied and fundamental research toxicologists. Special training courses should be organised, and validation should be discussed at scientific meetings, congresses and workshops. The validation procedures proposed in this report should themselves be subjected to evaluation in the near future.

As in other scientific disciplines, the future of applied toxicology will depend on progress made in fundamental research. Adequate funding should therefore be made available for basic toxicology, and any relevant methods developed for basic research purposes, which have potential applicability in *in vitro* toxicity testing, should be further developed, validated and evaluated as non-animal toxicity tests or components of test batteries.

Finally, none of these activities will be worthwhile, if the regulatory and legislative authorities are not willing to recognise scientifically validated new methods and accept and welcome their incorporation into toxicity testing practices. Applications for screening purposes and as adjunct tests currently exist and should be recognised. Eventually, non-animal procedures, which will replace many of the uses of animal procedures currently accepted as means of identifying particular forms of toxic potential and hazard, will undoubtedly be developed, validated and implemented.

Summary and Recommendations

To derive the maximum value from new methods and to ensure their acceptance, toxicity tests must be fully and properly validated. The ultimate aim of the validation process is to make available reliable and relevant methods that can be used for specific purposes in toxicology research and testing. The major steps in the process by which new procedures are developed, validated and accepted, can be formally defined as: *test development; intralaboratory assessment; interlaboratory assessment; test database development; evaluation;* and *acceptance.* Test development (the steps involved in establishing and defining a new procedure), and acceptance (the steps involved in taking the decision to use a particular procedure for a specified purpose), are activities which fall outside the validation process. Thus, validation comprises intralaboratory and interlaboratory assessment, test database development and evaluation.

Recommendation 1. *The purpose of a validation study should be fully defined, particularly in relation to the level of assessment (toxic potential, toxic potency, hazard or risk), and in relation to the type of test required (screening, adjunct or replacement), the type of toxicity to be evaluated, and the chemical spectrum of interest.*

Recommendation 2. *Tests should only be considered for inclusion in validation studies, if the specific purposes for which they have been developed are well defined and are*

consistent with the overall objectives of the validation study.

Recommendation 3. *Tests must have been adequately developed, standardised and documented, and a need for them in relation to the availability of other tests must exist, before they should be considered eligible for validation.*

Recommendation 4. *Relevant national and international agencies, industry associations, and data banks should be kept informed of all validation programmes, from early planning through to completion.*

Recommendation 5. *The results of validation programmes, together with full test protocols and details of the statistical methods employed, should be published in the peer-review literature, and data sets should be made available to all interested parties.*

In an ideal validation program, a test would be validated against a diverse set of chemicals representing all the foreseeable areas of chemistry to which the test might be applied. In statistical terms, the validation process would be most effective if the chemicals to be tested were randomly distributed among the classes of chemicals whose activity was to be predicted by the test. In practical circumstances, tests are generally validated for selected classes of chemicals. Therefore, for a test to be considered to have been validated for chemical groups not hitherto included in the validation process, further validation studies would be required.

Various sets of reference chemicals are required for the validation process, namely:

Reference Set 1: a calibration set for use in test development and intralaboratory assessment;
Reference Set 2: an intralaboratory reference set for use in the blind trial phase of intralaboratory assessment;
Reference Set 3: an interlaboratory reference set for use in the definitive phase of interlaboratory assessment; and
Reference Set 4: a database reference set for use in test database development.

Recommendation 6. *The four sets of chemicals selected for validation studies should form the basis of a Chemical Reference Bank, to facilitate the provision of chemicals and reference data for the validation of tests internationally. The establishment of an International Chemicals Reference Bank is a matter of urgency. The Bank should provide open-access listings of scientifically-selected chemicals, backed by toxicological data reviews, safety advice, and a source of chemicals of known purity and stability.*

Recommendation 7. *Any particular test should be validated against the most appropriate collection of reference chemicals, bearing in mind the specific purpose for which the test is proposed.*

Recommendation 8. *The toxicological classification of reference chemicals in terms of their toxic properties should be carried out by a panel of expert toxicologists, taking into account all the available relevant data. Final classifications should be fully documented with respect to both acceptable data and criteria, and the statistical evaluation procedures used.*

Recommendation 9. *When classifying reference chemicals for use in validation studies, consideration should be given to the numerous factors which affect the generation and quality of reference data and the expression and evaluation of toxicity in the animal investigated.*

Recommendation 10. *It is highly desirable that industry should play an active role in validation, specifically by supplying data not generally available to the scientific community at this time.*

Recommendation 11. *Schemes currently being developed for assessing human toxicology data and making them more readily available to validation studies, should be welcomed and supported.*

Recommendation 12. *Information toxicologists should be encouraged to investigate ways of integrating the information obtained in experimental toxicology, human toxicology and veterinary toxicology, and incorporating this data into the validation process for reference classification of chemicals, whenever this is feasible.*

Recommendation 13. *Wherever it is desirable and practicable, methods for the collection, collation, evaluation and expression of experimental, human and veterinary toxicological data should be standardised in ways which have been agreed upon by, and are acceptable to, the toxicological community as a whole.*

A demonstration that a test gives

reproducible results is an essential aspect of its development. Determination of reproducibility within and among laboratories is the principal aim of interlaboratory validation, and consists of a preliminary phase (including standardisation of procedures) followed by a definitive phase (including an extensive blind trial with coded chemicals).

Proceeding to the next stage, namely, test database development, depends on a successful outcome of the interlaboratory assessment. This stage involves conducting tests on a large spectrum (200–250) of carefully selected and coded chemicals. It is not necessary that every chemical be tested in every participating laboratory or even in a single laboratory, (e.g. overlapping data sets from several laboratories is sufficient). Smaller sets of chemicals may be tested for more-specific applications of the proposed test. Computer-based and theoretical methods, such as SAR procedures, should not be validated with reference chemicals that form all or part of the learning set upon which the procedure in question is based.

It is difficult to establish a common set of criteria for use in determining the outcome of validation programmes for all types of procedures. Nevertheless, assessment of test performance and evaluation of success in relation to purpose, are essential features of any such determination. Assessment of test performance includes consideration of reproducibility, correlation with reference classification, and logistical aspects. Performance criteria will be affected by whether the test in question is designed to be used as a screening test, an adjunct test or a replacement test. Acceptance may ultimately be based on the degree of phenomenological or mechanistic similarity to an existing animal test or relevant *in vivo* toxicological endpoint. Validation is an ongoing process, and evaluation of the success or failure of a particular procedure will be based on the conditions existing at a given time. Validation must not be looked upon as a rigid, linear sequence of events, but as a number of interacting activities. A detailed scheme for the integration of these activities is proposed (see section on Integration of the Validation Process).

In all likelihood, single tests will not provide sufficient data to serve as replacements. Therefore, batteries of validated tests should be constructed, for example, to combine tests which reflect different mechanisms of toxicity or respond to different classes of chemicals. However, when constructing batteries, the individual tests should be independent of one another and should not be redundant.

New methods for toxicity testing will continue to be developed at a rapid rate in the foreseeable future. In order to make the validation process more efficient, an administrative framework must be established at an international level to support these activities. Within this framework, chemical banks and data banks are high priorities.

Recommendation 14. *The concept of validation should be fully discussed among the general public, in the scientific and industrial communities, and, in particular, among toxicologists, for whom special training courses in the application of validation should be organised.*

Recommendation 15. *The regulatory and legislative authorities should be encouraged to welcome scientifically-validated methods and to accept their incorporation into toxicity testing practices.*

The introduction of new methods, as well as the application of existing methods to new purposes, requires scientific validation. The establishment of acceptable procedures to attain this goal are still under discussion among scientists. However, the issues are more focused now than in the past. Agreement among academic, industrial and government scientists on the basic principles of validation will facilitate the process of technology transfer, i.e. getting new and existing methods from the research laboratory into the practical testing environment.

Acknowledgements

Although the authors accept full responsibility for the opinions expressed in this report, they are necessarily grateful to the following for their constructive criticism of the manuscript: Professor Angelo Carere, Dr Richard Clothier, Dr Michael Dickens, Dr Bjorn Ekwall, Professor Victor Feron, Dr Paul Garvin, Dr Steven Gettings, Dr Gerald Guest, Dr Alan Goldberg, Dr Roy Goulding, Dr Richard Hill, Dr Herman Koëter, Professor Robert Kroës, Dr Barry Margolin, Dr Iain Purchase, Dr Robert Roth, Dr Andrew

Rowan, Dr Glyn Volans, Professor Friedrich Wurgler, Dr John Yam and Dr Errol Zeiger.

References

1. Goldberg, A.M. & Frazier, J.M. (1989). Alternatives to animals in toxicity testing. *Scientific American* **261**, 24–30.
2. Balls, M. & Clothier, R. (1989). Validation of alternative toxicity test systems: lessons learned and to be learned. *Molecular Toxicology* **1**, 547–559.
3. Frazier, J.M. (1990). *Scientific criteria for validation of in vitro toxicity tests*. Environment Monographs, no. 36, 62 pp. Paris: OECD
4. Ashby, J. & Purchase, J.F.H. (1977). The selection of appropriate chemical class controls for use with short-term tests for potential carcinogenicity. *Annals of Occupational Hygiene* **20**, 297–301.
5. Volans, G.N. & Wiseman, H.M. (1988). Surveillance of poisoning — the role of Poison Control Centres. In *Surveillance in Health and Disease* (eds W.J. Eylenbosch & N.D. Noah), pp. 225–271. Oxford: Oxford Medical Publications.
6. Ekwall, B., Bondesson, I., Castell, J.V., Gomez-Lechon, M.J., Hellberg, S., Hogberg, 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 tests in rodents. *ATLA* **17**, 83–100.
7. Warren, M., Atkinson, K. & Steer, S. (1989). Introducing *INVITTOX*: the ERGATT/FRAME *in vitro* toxicology data bank. *ATLA* **16**, 332–343.
8. Frazier, J.M. & Bradlaw, J. (1989). Technical Report No. 1: *Technical problems associated with in vitro toxicity testing systems*, 19 pp. Baltimore: Johns Hopkins Center for Alternatives to Animal Testing.
9. Pet-Edwards, J., Rosenkranz, H.S., Chankong, V. & Haimes, Y.Y. (1985a). Cluster analysis in predicting the carcinogenicity of chemicals using short-term assays. *Mutation Research* **153**, 167–185.
10. Pet-Edwards, J., Chankong, V., Rosenkranz, H.S. & Haimes, Y.Y. (1985b). Application of the carcinogenicity prediction and battery selection (CPBS) methodology to the Gene-Tox database. *Mutation Research* **153**, 187–200.
11. Pet-Edwards, J., Haimes, Y.Y., Chankong, V., Rosenkranz, H.S. & Ennever, F.K. (1989). *Risk Assessment and Decision Making Using Test Results: The Carcinogenicity Prediction and Battery Selection Approach*, 221 pp. New York: Plenum Press.
12. Benigni, R., Pellizzone, G. & Guiliani, A. (1989). Comparison of different computerized classification methods for predicting carcinogenicity for short-term test results. *Journal of Toxicology and Environmental Health* **28**, 427–444.
13. Chankong, V., Haimes, Y.Y., Rosenkranz, H.S. & Pet-Edwards, J. (1985). The Carcinogenicity Prediction and Battery Selection (CPBS) method: a Bayesian approach. *Mutation Research* **153**, 135–166.
14. Hellberg, S., Bondesson, I., Ekwall, B., Gomez-Lechon, M.J., Jover, R., Hogberg, J., Ponsoda, X., Romert, L., Stenberg, K. & Walum, E. (1990). Multivariate validation of cell toxicity data: the first 10 MEIC chemicals. *ATLA* **17**, 237–239.

Appendix A: Terminology

There is much confusion among scientists as to the precise definition of certain terms used in the discussion of test validation. An attempt has been made to define these terms and their associated meanings and to use them in a consistent manner throughout this report. If these terms prove acceptable to the scientific community, it is hoped that they will become universally adopted.

Adjunct refers to a test which has sufficient value to be incorporated into advanced decision making, but alone is not sufficient to justify a final decision.

Biological endpoint refers to the biological processes, responses or effects assessed.

Calibration involves the use of a specified set of chemicals (called the *calibration set)* to standardise the response of a test, in order to verify reproducibility of test performance in the same laboratory or in different laboratories.

Endpoint measurement refers to the techniques used to assess biological endpoints.

Hazard is a quantitative expression of the adverse effects elicited by a chemical under defined conditions of exposure.

Integration is the process by which information is carried over from one validation step to the next, in order to maximise the potential for acceptance of useful tests and to identify and reject inadequate tests.

Interlaboratory assessment is the stage of validation which establishes whether or not a test can be successfully transferred from one laboratory to another. It may include two phases: a *preliminary phase,* which identifies operational problems in the procedures for the assessment, and a *definitive phase,* in which a reference chemical set is evaluated by the accepted standard protocol.

Intralaboratory assessment is the first stage of validation, aimed at establishing the feasibility of using the proposed test to accomplish the intended purpose.

Method refers to the general manner in which exposure, endpoint measurement or data analysis are performed.

Potency is a measure of the relative toxicity of a chemical and can be used for the ranking of chemicals and/or their classification.

Potential refers to the inherent toxicological properties of a chemical, i.e. the possibility that toxicity can occur, with no concern for its likelihood or severity.

Predictive value is a measure of test performance. Positive predictive value is defined as the proportion of all chemicals tested which are predicted as positives by a test under evaluation and are, in fact, true positives based on the reference classification. Negative predictive value is defined as the proportion of all chemicals tested which are predicted as negatives by a test under evaluation and are, in fact, true negatives based on the reference classification.

Prevalence describes the proportion of chemicals in a population of chemicals that are capable of eliciting a particular biological response or effect.

Procedure refers to a test or battery of tests.

Protocol refers to the precise step-by-step description of methods.

Reference chemicals are chemicals selected for use in the validation process. Several sets of these chemicals are required for the different stages of the validation process.

Reference classification is the toxicity classification of chemicals selected for use in a validation process. This classification is usually based on human or animal *in vivo* data and can be either qualitative (involving categories, such as moderately irritating) or quantitative (involving measured values such as an LD50).

Relevance of a procedure describes whether a test is meaningful and useful for a particular purpose.

Reliability of a procedure describes reproducibility within and among laboratories and over time.

Replacement refers to any non-whole-animal procedure which provides a sufficient basis for making a definitive toxicological assessment or decision which previously was based on an *in vivo* procedure.

Risk is the probability that an event will occur given a particular condition of exposure. It is expressed as the product of the hazard and the likelihood of exposure, where exposure is estimated for a specific population.

Screening is a preliminary testing activity, which implies that there will subsequently be more definitive testing. It is normally used for preliminary decision making and priority setting.

Sensitivity is a measure of test performance, and describes the proportion of all chemicals tested which are classified as positive for a particular toxicological endpoint that are predicted as positive by the test being evaluated.

Specificity is a measure of test performance, and describes the proportion of all chemicals tested which are classified as negative for a particular toxicological endpoint that are predicted as negative by a test being evaluated.

Test and **assay** both refer to the combination of biological system, exposure protocol, endpoint measurement and data analysis. Test should be used in preference to assay.

Test database development is the stage of validation in which data are accumulated when proposed tests are applied to a large set of reference chemicals.

Test development is the process by which the components of a test — biological system, exposure protocol, endpoint measurement and data analysis — are defined, integrated and optimised for a specific purpose, and is normally carried out in the laboratory of origin.

Validation is the process by which the reliability and relevance of a procedure are established for a specific purpose.

Appendix B: Calculations of Performance Measures

The following formulae are provided as examples of methods used to determine specificity, sensitivity and predictive value in the case of a binary classification for test results and reference chemicals (i.e. outcomes can only take on two values, + or −). For calculations of predictive value, an estimate of the prevalence of chemicals which elicit the biological effect of interest must be made. For tests which produce continuous responses (e.g. calculations of the ID50, EC50, etc. for cytotoxicity tests) calculations of sensitivity and specificity, and therefore predicative value, are not well-defined. Computer-based SAR analyses may provide a form of sensitivity and specificity, but it is important not to use the same data set or an overlapping subset of the reference data base to determine these values. SAR activities require both a training data set and a second independent validation data set.

Sensitivity

$$Sn = \frac{a}{a + b}$$

Specificity

$$Sp = \frac{c}{c + d}$$

Positive Predictive Value

$$PV(+) = \frac{a}{a + d} = \frac{Sn^* Pr}{(Sn^* Pr) + (1 - Sp) * (1 - Pr)}$$

Negative Predictive Value

$$PV(-) = \frac{c}{b + c} = \frac{Sp^* (1 - Pr)}{Sp^* (1 - Pr) + (1 - Sn) * Pr}$$

where:

a = number of positive chemicals correctly predicted by a test *(true positives)*

b = number of positive chemicals incorrectly predicted by a test *(false negatives)*

c = number of negative chemicals correctly predicted by a test *(true negatives)*

d = number of negative chemicals incorrectly predicted by a test *(false positives);*

Pr = prevalence of chemicals capable of eliciting a given biological effect among the universe of chemicals to be tested;

PV(+) = positive predictive value, which is an estimate of the likelihood that a positive outcome in the test correctly identifies a toxic chemical under the proposed conditions of use;

PV(−) = negative predictive value, which is the likelihood that a negative outcome in the test will correctly identify a non-toxic chemical under the proposed conditions of use.