INTRODUCTION (John Frazier)

Workshop Overview: Scientific and Regulatory Challenges for the Reduction, Refinement, and Replacement of Animals in Toxicity Testing

Iain F. H. Purchase,* Philip A. Botham,† Leon H. Bruner,‡ Oliver P. Flint,§ John M. Frazier,§ and William S. Stokes†

*Central Toxicology Laboratory, Zeneca Ltd., Alderley Park, Macclesfield, Cheshire SK10 4TJ, United Kingdom; †The Procter and Gamble Company, Health and Beauty Care Europe, Staines, Middlesex TW18 3AZ, United Kingdom; ‡Bristol-Myers Squibb Company, Experimental Pathology, 6000 Thompson Road, Building 32A, East Syracuse, New York 13057-5050; §Armstrong Laboratory, US Air Force, L. A. HSC/OET, 2856 G Street, Building 79, Wright Patterson AFB, Ohio 45433-7400; and NTP Interagency Center for the Evaluation of Alternative Toxicological Methods, Environmental Toxicology Program, NIEHS, MD WC-05, P.O. Box 12233, Research Triangle Park, North Carolina 27709-12233

Received February 19, 1998; accepted February 23, 1998


Public concern for animal welfare has been expressed through legislative control of animal use for experimental purposes since the first legislation was introduced in 1876 in the United Kingdom. Legislative control of animal use has been introduced in virtually every developed country, with major initiatives in Europe (1986) and the United States (1966 and 1985). Advances in scientific thinking resulted in the development of the concept of the three Rs—refinement, reduction, and replacement—by Russell and Burch in 1959. The field has expanded substantially since, with specialist scientific journals dedicated to alternatives, World Conferences organized to discuss the scientific and philosophical issues, and European and U.S. validation organizations being launched. Current scientific attention is focused on validation of alternative methods. The underlying scientific principles of chemical toxicity are complicated and insufficiently understood for alternative methods for all toxicity endpoints of importance in protecting human health to be available. Important lessons have been learned about how to validate methods, including the need to have prediction models available before the validation is undertaken, the need to understand the variability of the animal-based data which is to be used as the validation standard, and the need to have well-managed validation programs. Future progress will depend on the development of novel methods, which can now be validated through international collaborative efforts. © 1998 Society of Toxicology.

Workshop held at the 36th Annual Meeting of the Society of Toxicology in Cincinnati, OH, sponsored by the In Vitro Toxicology Specialty section.

Strategies. The driving forces for these activities involve aspects of technology, economics, and social concern. Legislation in the United States and Europe mandates specific responses by government agencies and commercial enterprises to address these issues. This symposium focused on the nature of these responses in the USA and Europe, emphasising the infrastructure requirement to make progress and the evolution of practical approaches to attain the mandated goals.

The three Rs (refinement, reduction, and replacement). It is generally agreed that the concept of the three Rs originated in the classic work of W. M. S. Russell and R. L. Burch, The Principles of Humane Experimental Technique (1959). A refinement is any alteration in procedures that leads to a decrease in the incidence and/or severity of stress and discomfort to animals used in experimental procedures. Reduction is the development of scientifically justifiable techniques to obtain the same quality of information using fewer experimental animals. Finally, replacement is the introduction of new methods that do not require the direct use of sentient animals in experimentation, either by conducting experiments using tissues derived from sentient animals, or by using nonsentient animals or nonanimal methods (e.g., computational methods, analytical techniques).

Historical background. The issue of animal welfare and biomedical experimentation has a long and involved history. For the purposes of this symposium a brief review of some of the highlights will set the stage (Table 1).

A major legislative milestone that serves as a reference point for many discussions of animal welfare issues is the UK Cruelty to Animals Act of 1876. This legislation provided the first legal basis for the control of animal experimentation. Arbitrarily skipping 80 years of social and political activity brings us to the key event that defines the beginning of the modern era for animal welfare issues—the publication of The Principles of Humane Experimental Technique (Russell and Burch, 1959). This book described the scientific basis for

In vitro mutagenic activity of chemicals that is widely used to identify potential genotoxic carcinogens. This assay is a classic example of an in vitro alternative. About this same time, the U.S. Animal Welfare Act was amended to extend coverage to warm-blooded animals and to require pain relief where appropriate. In 1975, Peter Singer published Animal Liberation. This book presented ethical arguments for humane treatment of animals and introduced the concept of speciesism to support arguments against the use of animals in experimentation.

The 1980s started off with picketing of cosmetic companies by the Coalition of Animal Rights. The pressure brought to bear on the Cosmetic, Toiletries and Fragrance Association resulted in the founding of the Johns Hopkins Center for Alternatives to Animal Testing (CAAT). The initial efforts of CAAT focused on establishing a firm scientific foundation of alternative testing methodologies. In the late 1980s the U.S. federal government regulatory agencies realized that they needed to maintain better communication concerning animal testing issues and the Interagency Regulatory Alternatives Group was established. This group conducted extensive evaluations of alternatives to ocular irritation testing, i.e., replacements for the Draize eye test.

The 1990s can be characterized by the significant legislative actions both in the United States and Europe that have driven developments. In 1993, an amendment to the EC Cosmetic Directive required that cosmetic products must not be tested in animals after January 1, 1998, if validated alternative methods are available. The European Center for the Validation of Alternative Methods (ECVAM) was established and currently operates at the Joint Research Center in Ispra, Italy. ECVAM has become the major focus for alternatives research and validation in Europe. Also in 1993, the First World Congress on the Use of Animals in the Life Sciences was held in Baltimore, Maryland. This Congress brought together all interested parties in the animal welfare issue and was highly successful in advancing the dialogue among the stakeholders. At about the same time, the U.S. Congress passed the National Institutes of Health (NIH) Revitalization Act that required NIH to respond specifically to the alternative testing issue. The National Institute of Environmental Health Sciences (NIEHS) was tasked to fulfill the requirements of Congress. In conjunction with other formal agencies that had a vested interest in animal welfare issues the ad hoc Interagency Coordinating Committee for the Validation of Alternative Methods (ICCVAM) was established. ICCVAM has prepared a report of guidelines for the validation of new testing methods and proposed the establishment of a permanent committee to coordinate validation issues in the United States. In 1996, the Second World Congress was held in Utrecht, The Netherlands, demonstrating the international nature of the animal issues. The high level of interest in the animal welfare issue assures that there will be a Third World Congress in Italy in 1999.

Two very recent events in the United States indicate the importance that the Federal government places on animal issues. The Food and Drug Administration (FDA) has created a Subcommittee on Toxicology under its Science Board to provide advice on the state of the art in toxicity testing and to promote the utilization of the most advanced testing methods in safety evaluation of commercial products under their jurisdiction and the NIEHS is establishing a National Toxicology Program Interagency Center for the Evaluation of Alternative

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timeline of Major Events That Impact on the Response of the Scientific Community to Animal Welfare Issues</td>
</tr>
<tr>
<td>1876</td>
</tr>
<tr>
<td>1959</td>
</tr>
<tr>
<td>1966</td>
</tr>
<tr>
<td>1969</td>
</tr>
<tr>
<td>1970</td>
</tr>
<tr>
<td>1975</td>
</tr>
<tr>
<td>1980</td>
</tr>
<tr>
<td>1981</td>
</tr>
<tr>
<td>1981</td>
</tr>
<tr>
<td>1984</td>
</tr>
<tr>
<td>1985</td>
</tr>
<tr>
<td>1986</td>
</tr>
<tr>
<td>1988</td>
</tr>
<tr>
<td>1991</td>
</tr>
<tr>
<td>1993</td>
</tr>
<tr>
<td>1993</td>
</tr>
<tr>
<td>1993</td>
</tr>
<tr>
<td>1994</td>
</tr>
<tr>
<td>1996</td>
</tr>
<tr>
<td>1997</td>
</tr>
<tr>
<td>1997</td>
</tr>
</tbody>
</table>

It should be kept in mind that the development of biological sciences and biotechnology is a major factor in the advancement of alternatives. An early example of the role of basic science is the development of the Ames Assay by Dr. Bruce Ames. Beginning in the late 1960s and into the early 1970s, Dr. Ames developed the bacterial-based assay for determining the mutagenic activity of chemicals that is widely used to identify potential genotoxic carcinogens. This assay is a classic example of an in vitro alternative. About this same time, the U.S. Animal Welfare Act was amended to extend coverage to warm-blooded animals and to require pain relief where appropriate. In 1975, Peter Singer published Animal Liberation.
Toxicological Methods. This center not only will coordinate research and validation activities, but will provide support for the permanent ICCVAM committee. These actions demonstrate the commitment of the U.S. government to support the infrastructure needed to develop, validate, and accept new testing methods that further the goals of the three Rs.

In addition to the specific events described above that have led to the current status of alternative issues in the United States and Europe there are several underlying forces which have had a significant impact on the situation. Advances in the biological sciences and biotechnology have provided tools to develop novel testing methods that would not have been possible even 5 years ago. Second, the commitment of significant resources by the corporate sector to the development and validation of alternative methods has pushed the science forward. These efforts have involved both in-house research and Development activities and support for extracorporate activities. Validation studies, fellowships, and research grants have all contributed to progress. A third factor has been the introduction of scientific journals that focus on in vitro approaches to biomedical problems including toxicology. These journals provide an outlet for the scholarly activities of researchers in the field as well as an archive of achievements. Finally, the establishment of commercial enterprises that either provide products for alternative research or provide services based on alternative methodologies has supported the growth of the alternative movement.

This brief, albeit incomplete, history of the three Rs and the events that impact on that history gives some insight into the complex interactions between social, political, and scientific forces that shape the current state of affairs. The report below will shed more light on some of the more recent progress in the political and scientific communities.

Toxicological decision making. The final point of this introduction relates to the process of making toxicological decisions. Decision makers must consider questions such as the selection of candidate products, considerations of whether a promising new product should be taken to the next level of development, and whether a final product should be marketed. A decision as to the skin irritancy category is certainly an important marketing decision, but is only one component of the toxicological evaluation of a commercial product. The question is, what is the role of alternative testing methods in the safety/hazard evaluation of new products?

In order to answer this question several points must be taken into consideration. These are not new concepts, but they need restatement lest we forget their importance to the argument. First, no single test will answer all questions. Second, tests produce data and data must be converted into information before they become useful. Third, information provided by alternative tests (particularly in vitro tests) is different from that obtained from live animal tests. And fourth, the value of a new test lies in the utility of the information provided by that test in making better (more accurate, more timely, or less costly) decisions. Consider each of these points in more detail:

- No single test will answer all questions. In spite of the fact that most individuals involved in toxicological evaluations will agree with this point, we have not developed validation criteria that evaluate individual tests in any other context than as a stand-alone entity.
- Tests produce data and data must be converted into information before they become useful. As an integral part of any test, the procedure that is used to convert the raw data into useful toxicological information must be defined. This procedure is referred to as the predictive model or algorithm. Inclusion of the predictive model is essential when defining a test, otherwise the test produces useless data.
- Information provided by alternative tests (particularly in vitro tests) is different from that obtained from live animal tests. In vitro toxicity testing data are not directly equivalent to in vivo testing data. All the same, alternative tests provide important information relevant to many toxicological decisions. The extent to which in vitro data can replace in vivo data is still being explored. It must be kept in mind that the in vitro approach is a reductionist approach. It has yet to be proved that the integrated in vivo response can be adequately predicted by a collection of in vitro tests.
- The value of a new test lies in the utility of the information provided by that test in making better (more accurate, more timely, or less costly) decisions. Consider the spectrum of information utilized to make toxicological evaluations. Information provided by in vitro testing constitutes only one source of information. If a new alternative test provides information that allows a decision maker to resolve a toxicological issue faster or more accurately using less information from traditional animal testing sources, then the test can be considered to have merit, even if it does not answer all questions.

The role of alternative testing in toxicological evaluations and the infrastructure to incorporate that information into the regulatory process is continuing to evolve. The following reports discuss the scientific and political activities that are currently shaping the future for the three Rs in animal toxicity testing.

EUROPEAN LEGISLATIVE MANDATE (Iain Purchase)

The development of legislation in Europe and the form of the controls in the United Kingdom provide an instructive view of the issues which are seen to be most important in controlling the use of animals for scientific purposes. Since the initial legislation in the UK in 1876, laws have been passed in many countries which address the control of the way in which animals may be used for experimental purposes. In 1986 new legislation on animal welfare was introduced in the UK and the European Union (EU). To understand the legislative framework in Europe requires an understanding of the way in which the EU legislative machinery operates. The principal instru-
ment for legislation in the EU is the Directive. This provides the framework for legislation in all EU member countries, but only becomes law when the member countries have passed national legislation incorporating the provisions of the Directive. National legislation is mandated to occur within 24 months of adoption of the Directive (by the Council of Ministers of the EU), but often a longer period elapses before the legislation is in place. National legislation must incorporate all the clauses of the Directive (in order to meet the harmonization requirement aimed at providing a level playing field for competition within the EU), but may also be more stringent if it does not interfere with trade.

Animal welfare. The principal EU Directive addressing animal welfare is known as the European Animal Welfare Directive (Council Directive 86/609/EEC). From the point of view of a consideration of alternative methods in research, the most important article within the Directive is Article 7.1 which states: "An experiment shall not be performed if another scientifically satisfactory method of obtaining the result sought, not entailing the use of an animal, is reasonably practically available." Now that the directive is 10 years old, it can be inferred that this requirement covers all countries in the EU (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, The Netherlands, Portugal, Spain, Sweden, United Kingdom).

In the UK, a country with a long experience of high-quality biological research, the 1986 Animals (Scientific Procedures) Act is the legislative instrument (replacing the Cruelty to Animals Act of 1876) which complies with the Animal Welfare Directive. It is one of the most comprehensive pieces of legislation in the world and provides close control of all activities associated with the use of animals for experimental purposes.

There are several individuals who have statutory obligations for animal welfare under the provisions of the 1986 Act. In addition the premises must be licensed through a "certificate of designation." Thus the individual in overall control of the establishment who applies for the "certificate of designation" is known as the Certificate Holder. This individual has overall responsibility for the facility and the work carried out within it. The key control of experimental programs is through a Project Licence, which describes a program of work for a specific purpose for up to 5 years. The Project Licence Holder has a key responsibility to ensure that the work carried out meets the conditions on the Project Licence. The technical and scientific staff directly involved in experimental procedures must have Personal Licences which allows them to carry out certain procedures for which they have been trained and are competent.

The Project Licences provide the main mechanism by which the experimental work is controlled. Applications for a Project Licence require a plan of work and a description of procedures. There is a formal requirement for the Project Licence Holder to consider alternative methods and to consider the cost and benefits of the work. There must also be a named individual who can deputize during absences by the Project Licence Holder. Each year the Project Licence Holders must ensure that the work carried out falls within the licence and must provide the statistical information on the number of animals used within their project licence.

The animal welfare issues are addressed through "named" individuals. For each establishment, there must be a veterinarian, who is personally identified in the certificate of designation and is known as the Named Veterinarian. In each area where animals are housed, there must be an individual who is responsible for the day-to-day care and welfare of the animals—the Named Animal Care and Welfare Officer. Each of these individuals must attend registered training courses before they are considered competent to carry out their responsibilities.

The Act is the responsibility of the Home Office. They have an inspectorate of qualified Home Office Inspectors (usually medical or veterinary graduates) who are responsible for advising on the approval of the Project Licences and inspecting the premises and work which is carried out in the establishment. For this purpose, Home Office Inspectors have legal access to all establishments for unannounced visits. There is also an expert committee, The Animal Procedures Committee, which advises the minister responsible for the Home Office on matters related to the Act.

This complicated administrative machinery ensures that experimental work is planned, that certain individuals have particular designated statutory responsibilities including particular welfare responsibilities, that individuals are trained in approved courses, that premises are of an acceptable standard, that alternatives are considered and used where appropriate, that annual reporting of statistics occurs, that independent inspection by qualified inspectors occurs, and that there is an ongoing mechanism for monitoring the overall performance of the Act through the Animal Procedures Committee.

Product safety. In common with most developed countries, the EU has legislation requiring the testing of new and existing products in order to ensure safety to the consumer. The legislation takes the form of Directives which “approximate the laws of Member States” relating to particular products. Thus, there are Directives requiring testing of chemicals used as pesticides, pharmaceuticals, cosmetics, food additives, veterinary drugs, new industrial chemicals, existing chemicals, and many others.

International harmonization of the methods used is an important mechanism for reducing the requirements for duplicating testing merely to meet detailed national regulatory requirements. Methods required in the EU are now harmonized and much reliance is placed on the Organisation for Economic Cooperation and Development (OECD), which, through its Chemicals Directorate, coordinates the agreement on the details of various tests required for establishing the safety of products. The testing philosophy for pharmaceutical products is coordinated by the International Conference on Harmonisation (ICH), a joint activity between Government and Industry in the EU, U.S.A., and Japan.
The scientific and regulatory rationale for the test method, including a clear statement of its proposed use, should be available.

The relationship of the test method's endpoint(s) to the biologic effect of interest must be described. Although the relationship may be mechanistic or correlative, tests with biologic relevance to the toxic process being evaluated are preferred.

A detailed protocol for the test method must be available and should include a description of the materials needed, a description of what is measured and how it is measured, acceptable test performance criteria (e.g., positive and negative control responses), a description of how data will be analyzed, a list of the species for which the test results are applicable, and a description of the known limitations of the test including a description of the classes of materials that the test can and cannot accurately assess.

The extent of within-test variability, and the reproducibility of the test within and among laboratories must have been demonstrated. Data must be provided describing the level of intra- and interlaboratory reproducibility and how it varies over time. The degree to which biological variability affects this test reproducibility should be addressed.

The test method's performance must have been demonstrated using reference chemicals or test agents representative of the types of substances to which the test method will be applied, and should include both known positive and known negative agents. Unless it is hazardous to do so, chemicals or test agents should be tested under code to exclude bias.

Sufficient data should be provided to permit a comparison of the performance of a proposed substitute test method with that of the test it is designed to replace. Performance should be evaluated in relation to existing relevant toxicity testing data, and relevant toxicity information from the species of concern. Reference data from the comparable traditional test method should be available and of acceptable quality.

The limitations of the method must be described; for example, in vitro or other nonanimal test methods may not replicate all of the metabolic processes relevant to chemical toxicity that occur in vivo.

Ideally, all data supporting the validity of a test method should be obtained and reported in accordance with Good Laboratory Practices (GLPs). Aspects of data collection not performed according to GLPs must be fully described, along with their potential impact.

All data supporting the assessment of the validity of the test method must be available for review.

• Detailed protocols should be readily available and confidential.
• The method(s) and results should be published or submitted for publication in an independent, peer-reviewed publication.
• The methodology and results should have been subjected to independent scientific review.

Note. For a new or revised test method to be considered validated for regulatory risk assessment purposes, it should generally meet the above criteria (the extent to which these criteria are met will vary with the method and its proposed use). However, there needs to be flexibility in assessing a method given its purpose and the supporting database. Because tests can be designed and used for different purposes by different organizations and for different categories of substances, the determination of whether a specific test method is considered by an agency to be useful for a specific purpose must be made on a case-by-case basis. Validation of a test method is a prerequisite for it to be considered for regulatory acceptance.

* From National Institute of Environmental Health Sciences (NIEHS, 1997).

**Cosmetics testing.** Public concern about the testing of cosmetics on animals has led to adoption of the Cosmetics Directive in 1993 (Council Directive 93/35/EEC). Included in this Directive is an Article which states "Testing of ingredients or combination of ingredients should be banned as from 1 January 1998." However, it also concedes that alternative methods might not be available by then and states "That date should be postponed where alternative methods of testing have not been scientifically validated." The responsibility for coordinating validation work in the EU falls to the European Centre for the Validation of Alternatives (ECVAM), which works with all parties to develop validation methods and to validate toxicological methods for general use. ECVAM has defined a five-step process for the validation of a new test method, starting with the initial within-laboratory validation and concluding with international acceptance of the test method by the OECD or ICH.

A recent review of the progress in validation (Purchase, 1996) concluded that the number of toxicity tests which could be expected to be fully validated in the next 10 years was fairly limited. This could be predicted from the knowledge of the length of time it takes to validate and then to achieve international acceptance of a test. The overall result would be to reduce the number of animals used to test, for example, a pesticide, by about 5%. The remaining tests would be much more difficult to replace, because the nature and complexity of the tests. In fact, future progress would depend on the general advances in biological and medical sciences which would deliver new understanding of the toxicological mechanisms on which the tests and their alternatives could be based. Progress in this area was likely to be slow and legislators should be realistic in the timetables proposed for advancement.

The Cosmetics Directive has been modified to defer the banning of the use of animals until scientific validation of the test methods has progressed. However, the UK has effectively banned the use of animal tests for finished cosmetics, as the cosmetics Project Licence holders voluntarily surrendered their licences and the government announced that it would not issue any further project licences for finished cosmetics (November, 1997). The testing of cosmetic ingredients may, however, still be undertaken.

**U.S. GOVERNMENT INITIATIVES TO FACILITATE THE VALIDATION AND ACCEPTANCE OF ALTERNATIVE TOXICOLOGICAL TESTING METHODS (William Stokes)**

One of the overarching goals of the U.S. National Toxicology Program (NTP) is the development and validation of improved alternative toxicological test methods. Consistent with this goal, Public Law 103-43 directed NIEHS, a major

---

**TABLE 2**

**Validation Criteria**

- The scientific and regulatory rationale for the test method, including a clear statement of its proposed use, should be available.
- The relationship of the test method's endpoint(s) to the biologic effect of interest must be described. Although the relationship may be mechanistic or correlative, tests with biologic relevance to the toxic process being evaluated are preferred.
- A detailed protocol for the test method must be available and should include a description of the materials needed, a description of what is measured and how it is measured, acceptable test performance criteria (e.g., positive and negative control responses), a description of how data will be analyzed, a list of the species for which the test results are applicable, and a description of the known limitations of the test including a description of the classes of materials that the test can and cannot accurately assess.
- The extent of within-test variability, and the reproducibility of the test within and among laboratories must have been demonstrated. Data must be provided describing the level of intra- and interlaboratory reproducibility and how it varies over time. The degree to which biological variability affects this test reproducibility should be addressed.
- The test method's performance must have been demonstrated using reference chemicals or test agents representative of the types of substances to which the test method will be applied, and should include both known positive and known negative agents. Unless it is hazardous to do so, chemicals or test agents should be tested under code to exclude bias.
- Sufficient data should be provided to permit a comparison of the performance of a proposed substitute test method with that of the test it is designed to replace. Performance should be evaluated in relation to existing relevant toxicity testing data, and relevant toxicity information from the species of concern. Reference data from the comparable traditional test method should be available and of acceptable quality.
- The limitations of the method must be described; for example, in vitro or other nonanimal test methods may not replicate all of the metabolic processes relevant to chemical toxicity that occur in vivo.
- Ideally, all data supporting the validity of a test method should be obtained and reported in accordance with Good Laboratory Practices (GLPs). Aspects of data collection not performed according to GLPs must be fully described, along with their potential impact.
- All data supporting the assessment of the validity of the test method must be available for review.
  - Detailed protocols should be readily available and confidential.
  - The method(s) and results should be published or submitted for publication in an independent, peer-reviewed publication.
  - The methodology and results should have been subjected to independent scientific review.

**Note.** For a new or revised test method to be considered validated for regulatory risk assessment purposes, it should generally meet the above criteria (the extent to which these criteria are met will vary with the method and its proposed use). However, there needs to be flexibility in assessing a method given its purpose and the supporting database. Because tests can be designed and used for different purposes by different organizations and for different categories of substances, the determination of whether a specific test method is considered by an agency to be useful for a specific purpose must be made on a case-by-case basis. Validation of a test method is a prerequisite for it to be considered for regulatory acceptance.

* From National Institute of Environmental Health Sciences (NIEHS, 1997).
component of the NTP, to: (a) establish criteria for the validation and regulatory acceptance of alternative testing methods; and (b) recommend a process through which scientifically validated alternative methods can be accepted for regulatory use. The Director of NIEHS/NTP subsequently established the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM), consisting of representatives from 15 federal regulatory and research agencies, to develop a report recommending criteria and processes for validation and regulatory acceptance. The report, *Validation and Regulatory Acceptance of Toxicological Test Methods* (NIEHS, 1997), was developed using information from participating federal agencies, *Federal Register* (FR) requests for information, a review of the pertinent scientific literature, a 1995 NTP workshop (NIEHS, 1996), and a 1996 OECD workshop (OECD, 1996). Throughout the process of developing the report, broad stakeholder input and comments were sought from interested organizations that included industry, academia, public interest groups, animal welfare organizations, and the international community.

The ICCVAM report describes: (1) the validation and regulatory acceptance criteria that Federal Agencies should employ in considering new and revised test methods (Tables 2 and 3); (2) a series of recommendations to enhance the development, validation, and acceptance of new methods (Table 4); and (3) the establishment and functions of a standing ICCVAM that will replace the ad hoc ICCVAM. The new ICCVAM will function to communicate with test method sponsors throughout the test method development and validation process (Figs. 1 and 2). An NTP Interagency Center for the Evaluation of Alternative Toxicological Methods is also being established to support ICCVAM activities and provide the opportunity for public–private partnerships to facilitate the validation and review of new test methods.

These federal initiatives seek to encourage the development of new methods and improvement of existing test methods, provide effective guidance for scientists and regulatory staff for the validation and evaluation of new test methods, contribute to the increased likelihood of regulatory acceptance of scientifically valid new test methods, and encourage the refinement, reduction, and replacement of animal use in testing when scientifically feasible. The use of validated and accepted improved toxicological testing methods is expected to provide enhanced protection of public health and the environment, and benefit animal welfare by the refinement, reduction, and replacement of animal use.

### NEW TEST DEVELOPMENT I: PRACTICAL SCIENTIFIC ISSUES (Oliver Flint)

**The objective of toxicity testing.** Pharmaceutical and, to a lesser degree, most other consumer products are routinely tested for safety. The objective is to characterize the potentially toxic effects of the active product components, first to protect the consumer, second, to estimate the degree of in-use hazard and, third, to understand the biological mechanism of toxicity induced by the product. Toxicity is a necessary endpoint of any *in vivo* or *in vitro* study that meets this objective. One of the objectives of developing a new test is to reduce or eliminate the distress that is an inevitable consequence of safety testing in animals.

**The nature of preclinical in vivo toxicity tests.** The single greatest advantage of using an animal for product safety testing is that it is an inclusive model of all the factors involved in
### TABLE 4
**Regulatory Acceptance Process Recommendations**

#### Development and validation
- Criteria for validation and regulatory acceptance must be taken into account in the planning and design stages of validation studies.
- Development of novel and innovative test methods that will provide for improved risk assessment should be encouraged and funded. Federal regulatory agencies can and should help to drive innovation.
- Testing batteries and tiered testing strategies should be accommodated in regulatory testing requirements where appropriate, and new methods should be considered for incremental acceptance.
- While both correlative and mechanistic tests can be validated and accepted, mechanistically based methods relevant to the biological or health effects of concern should be encouraged.
- Given the continuing increase in the numbers and types of test methods being developed for varying purposes, the validation process should be flexible and adaptable.
- Test methods should be evaluated by consistent validation criteria and with the same degree of rigor regardless of whether the proposal derives from academia, industry, federal government, or other nations.
- Individuals or organizations developing or proposing new or revised test methods should be in communication with the regulatory agencies that will be asked to review and accept the methods.
- Assessment of the validation status of a new test method should involve relevant federal agencies.
- An efficient and effective process leading to regulatory acceptance of alternative methods should involve regulators at all stages prior to regulatory acceptance: development, validation, and review.
- Current efforts to incorporate validated alternative test methods into regulatory testing strategies should be continued and expanded.
- Federal agencies should continue to hold workshops on validation and acceptance issues of concern.
- Federal agencies should establish internal central clearing systems for evaluation of new or revised methods submitted to the agency, and for the periodic review of methods recommended by the agency.
- Test methods should be periodically reviewed and, where appropriate, revised or replaced, in light of scientific and policy developments. Considerations for such activities include the following:
  - Animal and nonanimal test methods that have the potential to support improved risk assessment and the potential to partially or fully replace existing toxicity tests for some or all of the products regulated should be reviewed and evaluated.
  - Frequency of review should be consistent with scientific activity or progress in that discipline.
  - The process should be efficient and expedient.
  - The process should include outside stakeholders.
  - The reviews and outcomes of the reviews should be made public.
  - Regulations, guidelines, or recommendations should be promulgated for validated and accepted toxicity tests or test batteries.
  - When evaluating the scientific acceptability of new or revised test methods, agencies should establish close links with the relevant scientific community to ensure continuing benefit from shared expertise.
  - Concurrent submission of data from existing and proposed new methods will help facilitate regulatory acceptance of new methods and should be encouraged.
  - Regulatory agency staff should be trained in the evaluation of data from newly accepted test methodologies.
- Federal agencies should establish internal central clearing systems for evaluation of new or revised methods submitted to the agency, and for the periodic review of methods recommended by the agency.
- An efficient and effective process leading to regulatory acceptance of alternative methods should involve regulators at all stages prior to regulatory acceptance: development, validation, and review.
- There should be interagency coordination of the evaluation of proposed test methods that are relevant to the needs of multiple agencies.
- A federal interagency committee on test methods should be established to serve as a forum for the exchange of information, for the coordination of the review and evaluation of test methods, and for related activities. This committee should strive for interagency consistency in review and evaluation processes, and interagency and international acceptance of new and revised methods.
- Federal regulatory agencies should establish consistent processes and criteria for acceptance of new and revised toxicological test methods and should communicate them to interested parties.
- Federal regulatory programs should solicit input from other programs and agencies as they develop and modify test guidelines of general interest.
- Harmonization of hazard classification may be necessary before test guidelines can be harmonized.
- Proposed new or revised test methods relevant to the needs of more than one program or agency should be harmonized as appropriate.
- Interagency differences in test methods that purport to detect the same toxicological endpoints but differ unnecessarily in detail should be identified and harmonized.
- Frequency of review should be consistent with scientific activity or progress in that discipline.
- The process should be efficient and expedient.
- The process should include outside stakeholders.
- The reviews and outcomes of the reviews should be made public.
- Regulations, guidelines, or recommendations should be promulgated for validated and accepted toxicity tests or test batteries.
- When evaluating the scientific acceptability of new or revised test methods, agencies should establish close links with the relevant scientific community to ensure continuing benefit from shared expertise.
- Concurrent submission of data from existing and proposed new methods will help facilitate regulatory acceptance of new methods and should be encouraged.
- Regulatory agency staff should be trained in the evaluation of data from newly accepted test methodologies.

#### Table 4: Regulatory Acceptance Process Recommendations

<table>
<thead>
<tr>
<th>Intra- and interagency coordination and harmonization</th>
</tr>
</thead>
<tbody>
<tr>
<td>A consistent, coordinated process of involvement and communication among all stakeholders (e.g., researchers, developers, users, regulators, and the public) at all stages (development, prevalidation, validation, review, regulatory acceptance, and implementation) will facilitate the validation and acceptance of new test methods.</td>
</tr>
<tr>
<td>Validation and regulatory acceptance should include the opportunity for input by interested stakeholders inside and outside of government.</td>
</tr>
<tr>
<td>The regulatory acceptance of new and revised test methods by agencies should be communicated to scientists and to various national and international organizations in journals, workshops, and the Federal Register, and by other means.</td>
</tr>
<tr>
<td>Agency regulations and guidelines should be readily available to the public.</td>
</tr>
</tbody>
</table>

#### International harmonization
- U.S. agencies should attempt to harmonize guidelines through international organizations, such as the OECD, where appropriate.
- U.S. agencies should encourage harmonization of test guidelines across international organizations, e.g., between U.N. Transport and OECD, as appropriate.

---

**Note.** To increase the efficiency of reviews of proposed new and revised methods and to increase the likelihood of adequate scientific consideration of new methods, the above considerations should be incorporated into the processes leading to regulatory acceptance of new test methods.

---

\(a\) From National Institute of Environmental Health Sciences (NIEHS, 1997).

human exposure. Dosing can be achieved by the intended route, and the chemical is distributed and modified by physiologically and biochemically appropriate mechanisms that determine the concentration of the chemical or its reactive metabolite(s) in all adversely affected target organs. Orally ingested compounds, unless they are very volatile, are first distributed to the liver, where chemical modification(s) usually render them more water soluble so that they can be more readily eliminated with other bodily waste products. In some cases, liver metabolism converts foreign chemicals to long-lived metabolites that are either very toxic in their own right (cyclophosphamide to acrolein and phosphoramidase mustard; Garattini, 1985a), or can be converted to a toxic reactive metabolite in a distant organ (nephrotoxicity of S-cysteine conjugates of haloalkenes; Chen et al., 1990). These complex interactions are not available in new in vitro systems (see below). The value of animals for establishing safety can be overstated, however, because animals do not always predict human toxicity. An example is the antiviral drug fialuridine (FIAU) which induced fatal hepatic failure in patients, undetected in preclinical animal studies (Colacino, 1996). Follow-up studies identified the woodchuck as a potential model for FIAU-induced hepatotoxicity, but this animal would not ordinarily have been selected for drug development. Another example is 6,8-diethyl-5-hydroxy-4-oxo-4H-1-benzopyran-2-carboxylic acid (DHBC), an antiallergy drug which induced mild reversible hepatotoxicity in clinical trials that was not predicted by laboratory animal tests (Clarke et al., 1985). Chemicals, like DHBC, which induce potentially serious but nonfatal toxicity in humans, not predicted by preclinical safety testing, are infrequently reported in the literature, but are all too familiar to the pharmaceutical toxicologist. Interspecies differences are inevitable because of species-specific differences in pharmacokinetics, organ perfusion rates, and drug metabolism (Garattini, 1985b; Monro, 1990).

The value of new in vitro tests. Since it is impossible to establish the reliability of animal data until humans are exposed, it is essential to evaluate the human risk of a new chemical by any and all scientifically valid means. One approach is to compare the effect of a drug in cultured human and animal tissue, providing information about the potential species-specificity of drug-induced adverse effects, otherwise unavailable in the preclinical phase of safety testing. Precision-cut tissue slicing has been a very useful tool in this respect (Smith et al., 1985), and has been increasingly used for interspecies comparisons of toxicity (Fisher et al., 1991, 1995; Price et al., 1996), metabolism (Steensma et al., 1994; Connors et al., 1996), liver enzyme induction (Glockner et al., 1995; Heinonen et al., 1996; Lake et al., 1996a), and genotoxicity (Baumann et al., 1996; Beamand et al., 1996; Lake et al., 1996b). Using the same preparative technique in all species, slices are relatively simple to prepare from almost any organ, including important targets of toxicity, such as the liver (Smith et al., 1985), lung (Fisher et al., 1994; Price et al., 1995a,b), and kidney (Ruegg, 1994). The slice system has proven to be a robust indicator of human metabolism and toxicity. In one study, species differences of coumarin metabolism were replicated in male Sprague–Dawley rat, female DBA/2 mouse, male Dunkin–Hartley guinea pig, male Cynomolgus monkey, and human liver slices (Steensma et al., 1994). In another study, the highly variable response of humans to valproic-acid-induced hepatotoxicity was correlated with the range of toxicity induced in slices prepared from nine human livers (Fisher et al., 1991).

An in vitro system is derived by isolation of tissues or cells

![Figure 1](https://academic.oup.com/toxsci/article-abstract/43/2/86/1657431/17/March/2019)
from the animal. It is unusual for all the differentiated properties of the intact tissue in vivo to be maintained in culture. Liver cells, for example, rapidly lose their ability to metabolize drugs (Lubinski et al., 1994), unless genetically modified to express specific cytochrome P450's (Pfeifer et al., 1993). The process of isolation of specific cell properties therefore continues after the cultures are established. This can be an advantage if the objective of the study is to establish the exclusive effect of a drug on a particular functional system such as an organelle, a receptor, or a group of enzymes. For example, the role of specific hepatic enzymes in the bioactivation and toxicity of a drug should be easily demonstrable by comparing a drug's effect on the viability of control, nonmetabolizing hepatocytes and cells transfected with active cytochrome P450's, such as those established by Pfeifer et al. (1993). In vitro assays of cell viability are relatively simple and direct compared to the more traditional approach of indirect inference from metabolite analysis, followed by observing the effect of induction or inhibition of specific cytochrome P450's on drug-induced toxicity.

Selection of tissue or tissue cell population is a significant problem for in vitro prediction of drug safety, since all organs are potential targets for tissue-specific, drug-induced toxicity. Drugs do not conveniently target one particular organ, such as the liver or the kidney. For example, the immune-suppressant FK506 induces both pancreatic (Doi et al., 1992; Hirano et al., 1992) and kidney toxicity (Ryffel et al., 1994). Interestingly, FK506 targets both the exocrine and the endocrine pancreas, while newer analogues may not be intrinsically toxic to pancreatic endocrine cells (Mollison et al., 1996). Cultures of pancreatic cells are used in retrospective studies of drug-induced toxicity, but are not normally considered in predictive toxicity studies. Nevertheless, pancreatic toxicity can be fatal and is induced by a number of drugs, including furosemide, methylprednisolone, metronidazole, estrogen, tetracycline, and valproic acid (Braganza, 1993). Similarly, the adrenal gland is almost never considered for in vitro toxicity studies, yet bilateral necrosis of the adrenal cortex, a common target for drug-induced lesions, is inevitably fatal, and many important drugs can induce lesions in the adrenal cortex; examples are clotrimazole, phenytoin, ketoconazole, suramin, tamoxifen, and zimelidine (Thomas, 1993).

In a typical preclinical toxicity study, this problem is solved by taking 50 or more different tissues for histopathological examination from each animal. This approach is not viable for in vitro toxicity studies. First, there would be an impossible increase in the labor involved. Second, satisfactory techniques for the culture of all adult tissues are not available; finally, comparison with human tissue would be difficult since routine availability only applies to those tissues destined for transplantation, such as the kidney, liver, lung, or skin.

One solution to this problem is to restrict the number of tissues selected for in vitro study to the liver and the kidney. Successful drug development is most frequently threatened by preclinical and clinical toxicity in these tissues, for two important reasons: drugs usually achieve their highest concentrations in these organs and the liver has the highest concentration of drug metabolizing enzymes. Although human liver and kidney tissue are available for these studies, availability has been reduced because of an increasing demand for organ transplantation and research use of tissue rejected for transplantation. It has become clear that immortalized human cells, by introduction of SV40 viral genes for example (Steinberg, 1996), may be the only solution to this problem of human tissue scarcity.

NEW TEST DEVELOPMENT II: THE DISCIPLINE OF VALIDATION (Leon Bruner)

Validation is the process by which the reliability and relevance of an alternative method are established for a particular purpose (Balls et al., 1990a). Reliability has been defined as establishing the reproducibility of results and reproducibility of toxicity hazard predictions within and between laboratories and over time (Balls
Initiating the validation process. The first step in a validation program (Fig. 3) is to confirm that the alternative method has been developed sufficiently to be entered into the validation process. An alternative method is ready for a validation when the following three elements have been established: (1) the method is relevant for its intended purpose, (2) the protocol and standard operating procedures have been completed, and (3) the reliability measures to be assessed have been defined. An initial review of the relevance of the alternative method should be completed prior to the start of any validation program. If there is little scientific basis supporting the method, the resources used to conduct the validation study may be lost since it is unlikely that a new method would ever be accepted for routine use. Study protocols and SOPs provide an important source of documentation about the conduct of the study. These documents must be written clearly so that all factors affecting the results, the collection of data, and interpretation of the alternative method results are available before the study begins. There are two reliability measures that must be defined prior to the start of a validation study. The first is the reproducibility of individual results from the alternative method. This measure provides the basis for determining the number of laboratories, the number of test substances, and the range of toxicity that should be included in a validation study (Bruner et al., 1996a). The second is the reproducibility of toxicity predictions from an alternative method in the form of a predefined Prediction Model (Bruner et al., 1996a). It is not possible to assess the validity of an alternative method unless the Prediction Model has been defined (Bruner et al., 1996b).

Measuring alternative method reliability. The tool used to confirm the reliability of an alternative method in the validation process is the validation study. The steps that need to be completed in a practical validation study are shown within the shaded box in Fig. 3.

The design of a validation study is crucial to its success, not only in terms of testing reliability, but also in retaining credibility and gaining acceptance by regulatory agencies. The factors that need to be considered include: (1) study management, (2) the participating laboratories, (3) test substances, (4) data collection and analysis, and (5) compliance with Good Laboratory Practice (GLP) [see the following references for details: (Balls et al., 1995a,b; Barratt, 1995a,b; Brantom et al., 1995; Bruner et al., 1996a; Curren et al., 1995; and Chamberlain and Barratt, 1995)].

Once the design phase of the program is complete, each of the materials must be tested in the alternative method as defined by the SOPs and Protocols (Balls et al., 1995a). When the data from testing in the alternative method have been provided, the reproducibility of assay results and toxicity predictions obtained from the alternative method must be assessed. If the results are reproducible, and if the predictions of toxicity lie within the limits defined by the Prediction Model, then it would provide strong evidence supporting the reliability of the alternative method. If the method is not reliable, then two courses of action may be considered (Fig. 3). Additional research should be pursued if it appears there is merit in further developmental work. Alternatively, the method may be abandoned if further effort is unlikely to be fruitful.

Assessing alternative method relevance. Once the reliability has been confirmed the relevance of the alternative method must be assessed. The establishment of scientific meaningfulness (i.e., relevance) is important because conclusions derived from fundamentally sound alternative methods have a higher probability of being correct. The specific factors that must be considered during the evaluation of relevance are as follows:

- Assess the theoretical best performance from the alternative method. Ideally, alternative method results should provide nearly perfect predictions of the toxic endpoints measured in vivo. However, there are important technical factors that prevent this ideal from being reached. Computer simulations based on the known performance characteristics of the in vivo test and the alternative method can be used to establish benchmarks for objectively judging the performance of an alternative method. A practical example of this process is found in Bruner et al. (1996a).
- Assess performance relative to the in vivo test that will be replaced. Another benchmark that can be used to assess the relevance of an alternative method is to compare its performance against that of the in vivo toxicity test it will replace. If the capacity of the alternative method to predict an in vivo toxicity endpoint is at least equivalent to the capacity of the in vivo test to predict its own result, then it would provide strong evidence supporting the relevance of the alternative method. See Bruner et al. (1996a) for a practical example.
- Assess the mechanistic basis for an alternative method. The mechanistic basis supporting use of an alternative method must be assessed (Frazier, 1994; Flint, 1992). A strong common mechanistic basis for both the alternative methods and in vivo tests is important because fundamentally sound assays are more likely to provide correct predictions of in vivo toxicity.
- Assess the technical limitations imposed on the alternative method. The restrictions placed on the use of an alternative method from a technical point of view must be considered. It is important to assess how well these limitations are defined (e.g., assay limited to water soluble test substances), and whether available data support the recommended limitations. Additionally, it is important to consider whether the use of an alternative method is so restricted that its use is ultimately impractical.
- Assessing other relevance factors. It is important to consider experience gained in use of an alternative method outside formal validation programs. Although the quality of data from other sources may be variable or not collected under blind conditions, it may provide additional useful insights into an
FIG. 3. The validation process. The flow chart depicts series of steps that may be used as a guide to design and conduct a validation program. The steps proceeding down the left side of the chart represent the validation process. The pathway within the shaded box represents the validation study process. The steps proceeding up the right side of the chart depict the steps associated with improving the performance of the alternative method and defining another Prediction Model prior to inclusion of the method in a subsequent validation study. Any new method, whether it is based on a fundamental understanding of toxic mechanisms or based on empirical correlations, may be assessed for validity using this approach.
alternative method's overall performance. If such data are consistent with the results obtained from a validation study, it would provide further evidence supporting the relevance of the alternative method for a given purpose.

- Concluding the relevance assessment (second diamond, Fig. 1). Once all of this information is assembled and assessed, the reviewers of a validation program must render a final judgment about whether a method is relevant or not. If the data from the validation study, performance relative to simulated benchmarks, compensation for lack of mechanistic understanding, and other supporting information are judged adequate, then it may be concluded that the method is relevant. If the alternative method is judged not relevant, then the reasons for the rejection should be clearly stated so that the deficiencies can be identified and resolved in follow-up research. Alternatively, the alternative method may be abandoned if additional work is unlikely to solve the problems.

Concluding validation program. If an alternative method is judged both reliable and relevant at the end of the process, then the new assay should be considered validated. Once validated, the alternative method may be used routinely in the safety assessment process and may be considered for acceptance by regulatory authorities (Fig. 3).

NEW TEST DEVELOPMENT III: PITFALLS AND PROBLEMS OF TEST VALIDATION (Phil Botham)

Validation is the procedure by which the reliability and relevance of a procedure are established for a specific purpose. This definition, which now has general support throughout the alternatives community, was first published in the report of the CAAT/ERGATT workshop on the validation of toxicity test procedures, held in Amden (Amden I—Balls et al., 1990a). Amden I proposed a set of principles for the validation process, and also made several practical recommendations on how validation studies should be conducted, including guidance on the selection of tests, the number of chemicals to be used at each stage of the process, and the criteria for evaluation of test performance. Amden I was followed by a second workshop (Balls et al., 1990b) which proposed a set of principals for the independent evaluation of validated alternatives, and for their incorporation into the regulatory framework, where appropriate.

Since 1990, these reports, particularly Amden I, have been criticized for being too rigid and impractical. While they were never intended to be a set of rules, experience with the design and conduct of validation studies in the early 1990s highlighted that Amden I's proposals were, in some cases, too theoretical and that a more time- and cost-effective approach was needed. The most significant of these studies was the European Commission/British Home Office (EC/HO) validation study on alternatives to the Draize eye irritation test.

The EC/HO study. In November 1991, a meeting was held in London, under the auspices of the EC and the HO, which included representatives of industrial companies and academia, where progress in the development of alternative methods for assessing eye irritation was discussed and steps toward their validation were considered. It was concluded that several new methods of sufficient promise were available and that a formal validation study could be undertaken. This conclusion was based to a large extent on the successful use within a number of European chemical companies of in vitro methods as prescreens, to detect the most severely irritant materials. The results were used to guide the way in which any subsequent animal testing was conducted. In vitro methods were also being used successfully as part of the safety assessment process in the personal care industry, in some cases as replacements for animal tests.

A Steering Group was formed which formulated proposals for a validation study which were then discussed by meetings of international experts in London in June, and then October, 1992. These meetings endorsed the goal of the validation study, which was "To provide scientifically credible data for proposing to regulatory authorities that one or more non-whole-animal methods should be adopted as a replacement for the Draize rabbit eye irritation test."

Four target steps relating to this goal were agreed. These were to determine whether the data obtained in the study indicated that it would be possible.

1. To replace the Draize eye test for identifying all severely irritating materials.
2. To replace the Draize eye test for identifying severely irritating materials belonging to specific chemical classes.
3. To replace the Draize eye test completely (i.e., to identify all levels of irritancy of materials without regard to chemical class).
4. To replace the Draize eye test for identifying all levels of irritancy belonging to specific chemical classes.

Another important step taken at these meetings was the appointment of a Management Team for the validation study, which comprised Prof. Michael Balls (University of Nottingham; now ECVAM), Dr. Phil Botham (Zeneca), Dr. Leon Bruner (Procter and Gamble), and Prof. Horst Spielmann (ZEBET). Agreement was also reached on the tests which should be included in the study. This was achieved by analyzing the proposed test methods against four selection criteria.

a. The specific purposes for which the test had been developed were well defined and likely to complement, but not duplicate, those of other tests in the set.
b. The test had been adequately developed, standardized, and documented, and a need for it in relation to other candidate tests had been demonstrated.
c. The test had shown promise in one or more previous interlaboratory studies.
d. The test was judged potentially to be capable of routine use in laboratories worldwide.

The Management Team eventually selected nine tests which
met some or all of the four criteria; the red blood cell hemolysis (RBC) test, the EYTEX method, the bovine corneal opacity/permeability (BCOP) test, the hen’s egg-chorioallantoic membrane (HET-CAM) test, the fluorescein leakage (FL) test, the isolated chicken eye (ICE) test, the isolated rabbit eye (IRE) test, the silicon microphysiometer (SM) test, and the neutral red uptake (NRI) test. In line with the Amden proposals, four or five laboratories were invited to conduct each test. The laboratories were selected by taking into account their experience, their adherence to the principles of GLP, and their ability to take part in an international study. A lead laboratory was appointed by the Management Team for each test to take responsibility for issues such as consistent handling of test materials, protocols, and reports (without having access to the actual data from the other collaborating laboratories).

One of the major problems identified by the Steering Group, and an area where it was clearly not possible to follow some of the Amden recommendations, was test chemical selection. An ECETOC Task Force had identified a set of chemicals of known and different degrees of ocular irritation, specifically for use in validation studies (Bagley et al., 1992). This Task Force had been asked to act as a Chemicals Selection Committee for the EC/HO study, and they presented 60 test materials to the meeting held in London in June 1992 (Amden I suggested that for this type of validation study as many as 300 chemicals may be required). One of the reasons for the smaller number of materials was that, quite correctly, rigorous selection criteria had been applied by the Task Force. These were:

1. Single chemical entities available at known high consistency and purity and expected to be stable in storage.
2. In vivo data generated since 1981 in studies carried out to OECD guideline 405 (i.e., at least three rabbits, instillation of 0.1 ml or 0.1 g of material into the conjunctival sac, observations at 24, 48, and 72 h—and beyond where needed to assess reversibility, individual tissue scores, and no anesthesia or rinsing) and following the principles of GLP.

However, their selection was criticized in that the number of chemicals tested in solid form was inadequate, there were insufficient chemicals of moderate to severe irritancy, and no pesticides were included. A search for further materials was made which proved to be extremely difficult. In addition, 14 test substances from the ECETOC database were found to have been tested in vivo with a rinsing of the eyes 1 h after instillation; these materials were therefore retested in vivo. Eventually, a revised set of 60 materials was accepted by the Management Team and this comprised 52 chemicals assessed in 60 different tests (four chemicals were tested at two concentrations, and two at three concentrations). These subdivided into 26 liquids and 34 solids; 20 of the 34 solids were tested in vivo as such, the other 14 as diluted solutions; 12 of the materials were surfactants, 30 were soluble in water, and 18 were insoluble.

The Management Team decided that, for this study, the in vivo data should be expressed as a modified maximum average score (MMAS), a total Draize score modified to include only those observations made 24 h or more after instillation. The main reason for using the MMAS, rather than for example regulatory classification schemes, was to present the in vivo data in the form of an “international currency” rather than in accordance with any particular national or multinational scheme such as that of the EC (i.e., R36, R41, or unclassified).

An independent laboratory (BIBRA International) was appointed to deal with the supply, coding, and distribution of the test materials. The testing laboratories conducted all the in vitro tests blind, and the results were submitted for analysis using the code numbers supplied by BIBRA. This organization was also asked to perform the statistical analyses but by a group that was independent from those who handled the test samples. The results were presented to the Management Team in the form of scatterplots showing MMAS versus each alternative method endpoint (of which there were 27, as some of the in vitro methods had up to five different endpoints). The linear regression line was calculated and plotted together with 95% confidence intervals; correlation coefficients were given both for the MMAS versus each in vitro endpoint and for interlaboratory correlations for the alternative test data. Details of all the results and analyses can be found in Balls et al. (1995a).

The scatterplots revealed a broad scattering of the data points around the regression line, and the correlation coefficients ranged from 0.012 to 0.616. Additionally, the 95% predicted confidence intervals were wide, frequently ±40 (i.e., an in vitro score was predictive of an MMAS with a range of 80, on a scale of only 110). This situation was true for the entire test set, and for solids, liquids, solutions, solubles, and insolubles, and for all nine alternative test methods. However, surfactants showed somewhat higher correlation coefficients and lower confidence intervals. In contrast, for many of the methods, the in vitro results obtained in the groups of four or five laboratories were highly correlated with each other.

Thus, the results of this study showed that while the interlaboratory consistency of the alternative methods was generally very good, the precision of the predictions of the in vivo MMAS was very low, and the practical utility of such predictions is questionable.

The Management Team concluded that while the outcome of the study was clearly disappointing, with none of the nine methods being able to fulfill the goal and targets of the study, a number of important lessons have been learned, not only for eye irritation but also for the whole process of validating alternative methods.

Five key areas were identified as possible contributors to the outcome of poor prediction of eye irritation in the EC/HO study:

1. the choice of test materials;
2. the variability of the in vivo data;
3. the in vivo endpoint used in the comparisons;
4. the management of the in vitro test protocols;
5. the statistical analyses used.
**Test materials.** Whilst the goal of the study was set with the needs of regulators in mind, namely to have alternative methods that can be used to test all types of tests substances, experience of the methods within industrial and contract laboratories suggests that they work best when their use is limited to specific classes of test substances that act through similar mechanisms of toxicity. In house, the physical and chemical properties of test substances are also usually known so that, for example, the solubility of a chemical in aqueous tissue culture media will often dictate whether or not an in vitro test is conducted. In addition, in many cases, the in vitro data can be interpreted only by reference to results obtained using benchmark test materials.

**The variability of the in vivo data.** While the in vivo data used in this study were judged against the rigorous criteria developed by the ECETOC Task Force (Bagley et al., 1992) and were thus of the highest attainable quality, there were no data on the reproducibility of the in vivo scores for the majority of the 60 test materials. However, Weil and Scala (1971) showed that the Draize rabbit eye test can give highly variable results and it is therefore possible that the MMAS scores used in the statistical comparisons in this study were subject to high coefficients of variation (CV’s). Bruner et al. (1996b) have performed computer simulations to determine the effect of different CV’s on the highest achievable correlation coefficient for the prediction of the MMAS by an in vitro method. This showed that, if the in vivo data for a reference set of chemicals are as variable as suggested by Weil and Scala, then it is not possible to demonstrate that an alternative method can provide predictions that have high levels of certainty. Noisy in vivo data will therefore obscure adequate performance of any in vitro test.

**The in vivo endpoint used in the comparisons.** The use of the MMAS has been criticized as being inappropriate because it represents a composite score of responses from several tissues that does not accurately reflect the response of the individual ocular tissues, and also because some in vitro alternatives have been developed to predict only broad eye irritation classifications. However, a number of previous studies (referenced in Balls et al., 1995a) have shown very high correlation coefficients between the MMAS and Kay-Calandra, or individual tissue scores. The outcome of the EC/HO study may not, therefore, have differed significantly had these other in vivo endpoints been used.

**The management of the in vitro test protocols.** For some of the nine methods, the protocols evaluated in the EC/HO study had not been used before, as significant changes had been made from the original method in order to establish a common procedure that could be used by all four or five participating laboratories. In one or two cases, even when a common procedure had been adopted, laboratories deviated from the agreed protocol. In addition, the lack of an agreed Prediction Model for each alternative method prior to the initiation of the EC/HO study led to post hoc data fitting, which may have resulted in a less objective assessment of the performance of some of the tests.

**The statistical analyses used.** The analyses used in the EC/HO study were restricted to the relatively simple linear regression techniques and calculation of 95% predicted confidence intervals. No attempt was made to use more complex models, first because the scatterplots suggested that this was unlikely to demonstrate an improvement of the performance of any of the test methods, and second because of the lack of Prediction Models. However, it is possible that other statistical methods may have been capable of detecting relationships which were hidden within the very noisy sets of data.

**The lessons learned from the EC/HO study.** While the outcome of the EC/HO study was clearly disappointing, it has nevertheless proven to be a seminal study, with many lessons learned about the concepts of validation as well as the practical considerations of running validation studies. One of the first beneficiaries of this learning was the COLIPA (European Cosmetic, Toiletry and Perfumery Association) validation study on alternatives to the Draize eye test (Branton et al., 1997). For the first time, Prediction Models were used in a validation study. Ten alternative methods were assessed (five of which were also used in the EC/HO study) using a test substance set comprising 23 cosmetic ingredients (20 of which were common to the EC/HO study) and 32 formulations. Coding and supply of the test substances and statistical analysis of the in vitro data were once more conducted independently. Three methods (the fluorescein leakage test, the red blood cell hemolysis test, and the tissue equivalent assay) each satisfied one criterion of reliability or relevance, but again none of the alternative methods could be confirmed as valid replacements for the Draize eye irritation test across the full irritation scale.

The COLIPA study has therefore shown that even when a greater level of management control and scientific discipline is used in a validation study, the currently available in vitro methods are not able to give predictions of eye irritation which have widespread practical utility. While this may still be a reflection of factors such as the variability of the in vivo data used in the statistical comparisons, it is likely that the area of eye irritation may now need to go back to the stage of test development, so that methods of greater mechanistic relevance can be brought forward for validation.

Another beneficiary of the EC/HO study was a workshop on practical aspects of the validation of toxicity test procedures, held in Ameden (Ameden II—Balls et al., 1995b). This workshop examined the experience gained from the EC/HO and other studies conducted since Ameden I and concluded that there should be five stages in the validation process: test development, prevalidation, validation, independent assessment, and progression toward regulatory acceptance. The pre-validation stage was highlighted as a particularly important component in the evolution of a new test, as it enables greater emphasis to be placed on protocol refinement, transfer, and performance than has been envisaged in Ameden I. The primary
objective of prevalidation is therefore to define and demonstrate the robustness and reproducibility of in vitro test protocols. The findings of Amden II have now been accepted and implemented as a major part of the strategy of ECVAM.

Under the auspices of ECVAM, the Amden II principles have already been used in a prevalidation and validation study of four alternative methods for predicting skin corrosion. The results of the prevalidation study have been published (Botham et al., 1995) and the full validation study will be completed and reported by the end of 1997. ECVAM is also supporting the validation of methods for predicting hematotoxicity and embryotoxicity/developmental toxicity, which hopefully will reach the validation stage in 1998. Another area which has benefited from the new criteria for validation is phototoxicity; ECVAM is now supporting a follow-up validation study of the 3T3 neutral red uptake phototoirritation test which it is hoped will enable the preparation of a draft OECD test guideline.

Recognition of the pitfalls and problems associated with the validation of alternative test methods, and the particular lessons learned from the EC/HO study of alternatives to the Draize eye test, have resulted in the development of a more practical and scientifically based approach to validation. This is now ensuring that the significant levels of human and financial resources required to assess the relevance and reliability of alternative methods are deployed more effectively. It has also enabled the validation of alternatives for assessing skin corrosivity and phototoirritation, and there is considerable optimism that, for these two areas, methods will be available for regulatory acceptance within the next year or two. For eye irritation, however, it is likely that further work on the development of mechanistically based tests will be required and validated alternatives are unlikely to be available for at least another 5 years.

REFERENCES


