Chapter 4
Validation of Alternative *In Vitro* Methods to Animal Testing: Concepts, Challenges, Processes and Tools

Claudius Griesinger, Bertrand Desprez, Sandra Coecke, Warren Casey and Valérie Zuang

Abstract This chapter explores the concepts, processes, tools and challenges relating to the validation of alternative methods for toxicity and safety testing. In general terms, validation is the process of assessing the appropriateness and usefulness of a tool for its intended purpose. Validation is routinely used in various contexts in science, technology, the manufacturing and services sectors. It serves to assess the fitness-for-purpose of devices, systems, software up to entire methodologies. In the area of toxicity testing, validation plays an indispensable role: "alternative approaches" are increasingly replacing animal models as predictive tools and it needs to be demonstrated that these novel methods are fit for purpose. Alternative approaches include *in vitro* test methods, non-testing approaches such as predictive computer models up to entire testing and assessment strategies composed of method suites, data sources and decision-aiding tools. Data generated with alternative approaches are ultimately used for decision-making on public health and the protection of the environment. It is therefore essential that the underlying methods and methodologies are thoroughly characterised, assessed and transparently documented through validation studies involving impartial actors. Importantly, validation serves as a filter to ensure that only test methods able to produce data that help to address legislative requirements (e.g. EU's REACH legislation) are accepted as official testing tools and, owing to the globalisation of markets, recognised on international level (e.g. through inclusion in OECD test guidelines). Since validation creates a credible and transparent evidence base on test methods, it provides a quality stamp, supporting companies developing and marketing alternative methods and creating considerable business opportunities.

C. Griesinger • B. Desprez • S. Coecke • V. Zuang (✉)
European Commission, Joint Research Centre (JRC), Ispra, Italy
e-mail: Valerie.ZUANG@ec.europa.eu

W. Casey
Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM), Washington, DC, USA

C. Eskes, M. Whelan (eds.), *Validation of Alternative Methods for Toxicity Testing*,
Advances in Experimental Medicine and Biology 856,
DOI 10.1007/978-3-319-33826-2_4

Validation of alternative methods is conducted through scientific studies assessing two key hypotheses, reliability and relevance of the test method for a given purpose. Relevance encapsulates the scientific basis of the test method, its capacity to predict adverse effects in the “target system” (i.e. human health or the environment) as well as its applicability for the intended purpose. In this chapter we focus on the validation of non-animal *in vitro* alternative testing methods and review the concepts, challenges, processes and tools fundamental to the validation of *in vitro* methods intended for hazard testing of chemicals. We explore major challenges and peculiarities of validation in this area. Based on the notion that validation per se is a scientific endeavour that needs to adhere to key scientific principles, namely objectivity and appropriate choice of methodology, we examine basic aspects of study design and management, and provide illustrations of statistical approaches to describe predictive performance of validated test methods as well as their reliability.

1 Introduction

What is validation and why do we need it? Validation of alternative methods has been defined as the process by which the reliability and relevance of a particular method is established for a defined purpose (Balls et al. 1990a, b, c, 1995a, b; OECD 2005). This definition has then later been extended to alternative approaches in the wider sense, i.e. not only covering individual methods but also combinations thereof, including strategies for data generation and integration. The reliability relates to the within- and between-laboratory reproducibility as well as to the transferability of the method or approach in different laboratories, whereas relevance relates mainly to its predictive capacity and, importantly, to the biological/mechanistic relevance, traditionally subsumed as “scientific basis”. Judging the overall relevance however also includes aspects of applicability domain and even the level of reliability required in view of the purpose of the method. The defined purpose can be various and range from full replacement of a regulatory test to the generation of mechanistic information relevant to the type and extent of toxic effects which might be caused by a particular chemical (Frazier 1994).

In regulatory toxicity testing, validation is placed between research/development and regulatory acceptance and aims at the characterisation of an *in vitro* test method under controlled conditions which in turn leads to the standardisation of the test method protocol. This aspect of test method development has been summarised in Coecke et al. (2014). Validation generally facilitates and/or accelerates the international (regulatory) acceptance of alternative test methods. In fact, the regulatory acceptance of tests that have not been subjected to prevailing validation processes is discouraged by international bodies (OECD 2005). This is true not only for alternative methods but also for tests conducted in animals. The term “regulatory acceptance” of an *in vitro* test method relates to the formal acceptance of the method by regulatory authorities indicating that the test method may be used as an official tool to provide information to meet a specific regulatory requirement. This includes, but is not limited to, a formal adoption of a test method by EU and/or OECD as an EU test method and included in the EU Test Methods Regulation and/or as an OECD Test Guideline, respectively. Standardisation and international adoption of testing approaches supports worldwide acceptance of data. Under the OECD Test Guideline Programme this is known as Mutual Acceptance of Data (MAD). MAD saves every year an appreciable number of animals and other resources as it avoids duplicate testing.

Three main types of validation processes have been defined: prospective, retrospective and performance standards-based validation—the latter being a form of prospective validation. Prospective validation relates to an approach to validation when some or all information necessary to assess the validity of a test is not available, and therefore new experimental work is required (OECD 2005). Retrospective validation relates to an assessment of the validation status of a test method carried out by considering all available information, either as available in the published literature or from other sources (e.g. data generated during previous validation studies (OECD 2005) or in-house testing data from industry). Validation based on Performance Standards relates to a validation study for a test method that is structurally and functionally similar to a previously validated and accepted reference test method. The candidate test method should incorporate the essential test method components included in Performance Standards developed for the reference test method, and should have comparable performance when evaluated using the reference chemicals provided in the Performance Standards (OECD 2005).

The European Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM) [formerly known as the European Centre for the Validation of Alternative Methods, ECVAM] and its international collaborators published recommendations concerning the practical and logistical aspects of validating alternative test methods in prospective studies (Balls et al. 1995a, b). These criteria were subsequently endorsed by and mirrored in the procedures of the US Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM 1997), and later internationally, summarised in the "Guidance Document 34" of the Organisation for Economic Cooperation and Development (OECD) (OECD 2005).

In 2004, ECVAM proposed a modular approach to the validation of alternative methods (Hartung et al. 2004), according to which the various information requirements for peer-review and as generated during the validation process are broken down into seven independent modules. According to this modular approach, the information requirements can be fulfilled by using data obtained from a prospective study, by a retrospective evaluation of already existing data/information, or by a combination of both.

More recently, the concepts of weight of evidence validation/evaluation (Balls et al. 2005) and evidence-based validation (Hartung 2010) have been introduced; Weight of evidence validation involves the careful analysis and "weighing" of data with regard to their quality, plausibility, etc. in view of concluding whether it supports one or the other side of an argument, in this context whether or not a particular method is useful for a specific purpose. Evidence-based validation essentially refers to the use of tools from evidence-based medicine for purposes of alternative method validation. These may range from systematic reviews (e.g. to determine reference data or analyse a set of existing data) over data grouping and meta-analysis to more probabilistic descriptors of test method performance as are used in medicine, for instance to describe the performance and usefulness of diagnostic tests.

This chapter explores the fundamental concepts behind validation, the hypotheses assessed and information generated, outlines specific challenges of alternative methods validation that relate to the nature of test methods being reductionist proxies for the human situation and provides a detailed discussion of the practical aspects of organising, designing, planning and conducting a validation study and analysing the data generated by appropriate statistical analyses (see also Chap. 5).

2 Validation: Principles, Hypotheses Assessed and Information Generated

This section examines fundamental principles of validation and explores the hypotheses and information generated by validation studies of alternative methods conducted in the context of their envisaged use for the safety assessment of specific test materials such as chemicals (of various chemical and/or use categories) and their integration in integrative approaches (e.g. Integrated Testing Strategies, ITS or Integrated Approaches to Testing and Assessment, IATA). Instead of simply recapitulating commonly accepted concepts of alternative method validation described in OECD guidance document Nr 34 (OECD 2005), we unfold this topic in the following way:

- First we will consider a series of fundamental issues that are necessary for the understanding of some unique features of the validation of alternative approaches.
- Second, we will examine three key concepts and explain their meaning in view of avoiding confusion regarding terminology. These are (a) *validation workflow*, (b) *validation study type (or validation process)* and (c) *the validation information generated through dedicated studies*. These three are often subsumed under the term "validation" but it is important to understand them as separate categories.
- Third, we will discuss the broader concept of 'validation' in view of deducing the central hypotheses assessed by alternative method validation. This will serve to understand the commonalities between validation in general and validation of alternative methods, and sculpt out some specific characteristics of the latter, in particular those constituting major challenges. These challenges include (a) finding appropriate reference data for *in vitro* test method development ("calibration") and validation and (b) the identification of mechanisms that are causative for downstream (i.e. more complex) events and hence should be modelled in reductionist and mechanistically-based alternative methods.
- Finally, we will discuss in more detail the information that needs to be satisfied in order to consider an alternative method valid for a specific purpose. We will put a particular emphasis on the composite nature of judging the overall relevance of alternative methods. This discussion will then lead over to section three and

four that explore the management, planning, design and conduct of validation studies in a manner so as to satisfy these information requirements. Details on EURL ECVAM's specific approach regarding multi-laboratory trials can be found in Chap. 5.

2.1 Fundamental Considerations

2.1.1 Validation in the Current Context Relates to *In Vitro* Methods Used for Toxicity Testing

Validation, i.e. the process of establishing the usefulness and appropriateness of a method for a given purpose, is applicable to a wide range of biological and analytical methods, e.g. in diagnostic medicine, food safety, etc. In the current context, we focus on the validation of biological *in vitro* test methods for toxicity testing of chemicals and safety testing of biologicals (Hendriksen et al. 1998). Therefore, chemicals (or biologicals) are the basic entities used to study and report the performance, utility and applicability of alternative method during validation. Consequently, selecting an appropriate set of test chemicals is of key importance when planning a validation study (see Sect. 4.2). One should however not lose sight of the fact that a mere summary and analysis of testing data would not yet make a complete validation study: a fundamental aspect to consider relates to the biological and physiological processes modelled by the test method and thought to be relevant for the chemicals' adverse effects. This is called the "scientific basis" of a test method and needs to be properly described. This helps judging the plausibility of results obtained with a given test method and supports the assessment of its relevance for a given purpose (see Sect. 2.3.3).

For more than a quarter century validation studies have been conducted in the areas of safety testing of chemicals and biologicals (e.g. vaccines) as well as ecotoxicological toxicity testing. Considerable efforts have been invested in developing internationally agreed validation frameworks, notably by the European Commission's EURL ECVAM, the Centre for Alternative to Animal Testing (CAAT), the US Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) as well as many individual scientists in academia, industry, government and international organisations (Scala 1987; Balls et al. 1990a, b, c, 1995a, b; Frazier 1990a, b; Frazier 1992; Green 1993; Balls 1994; Walum et al. 1994; Fentem et al. 1995; Balls and Karcher 1995; Goldberg et al. 1995; Bruner et al. 1996). This led to the development of reports and guidance documents adopted on international level, such as the OECD report on the harmonisation of validation and acceptance criteria for alternative toxicological test methods (OECD 1996, updated in 2009) which later gave rise to the more complete OECD guidance document on the validation and international acceptance of new or updated test methods for hazard assessment (OECD 2005). These documents reflect the status of international agreement at the beginning of the millennium.

Validation has played and is continuing to play a key role in toxicity testing because of its confidence- and trust-building role. Validation, overseen by impartial actors and subjected to scientific peer review leads to the comprehensive characterisation and transparent documentation of novel test methods. It is an important prerequisite for the international recognition and regulatory acceptance of test methods, e.g. through uptake into relevant legislations outlining official test methods recognised for use in a specific jurisdiction. Examples are the EU's Test Method Regulation EC 440/2008 and, on a global level, internationally accepted guidelines such as OECD's Test Guidelines (TGs). Although formally not relating to legislation per se, data produced in agreement with TGs are binding for OECD member countries due to the OECD agreement on Mutual Acceptance of Data (MAD) dating back to 1981. This stipulates that *"data generated in the testing of chemicals in an OECD Member country in accordance with OECD Test Guidelines and OECD Principles of Good Laboratory Practice shall be accepted in other Member countries for purposes of assessment and other uses relating to the protection of man and the environment*." Notably, combinations of methods as described in OECD guidance documents on "Integrated Approaches on Testing and Assessment" (the OECD term for Integrated Testing Strategies) are not covered by the MAD agreement at present. We will focus in this chapter mainly on validation as a means of characterising and assessing alternative approaches in view of their fitness for regulatory acceptance.

2.1.2 Validation Has Largely Focused on Hazard Testing So Far

Importantly, validation of alternative methods for toxicity testing has mainly focused on predicting potential hazards of chemicals, that is, their intrinsic potential to cause adverse effects in a particular test system (i.e. an animal, cell type, etc.), without providing much information on the potency. Potency relates to the doses required to provoke adverse effects in a whole organism and is key information for a complete risk assessment of chemicals. What are the major bottlenecks concerning methods addressing potency? First, the concentrations that a given cellular population in a human body is exposed to following systemic exposure through the environment are typically not known: hence it is difficult to define appropriate concentrations of test chemicals that should be used in *in vitro* systems—including when validating these. More effort needs to be invested in approaches (including *in vitro* systems) for assessing toxicokinetic processes, i.e. the absorption, distribution, metabolism and excretion (ADME) of chemicals. Reliable data and/or simulations would assist in defining the appropriate range of chemical concentrations to be used in alternative approaches. This has been already highlighted by Balls and colleagues in 1995 (Balls et al. 1995a, b). Second, due to the reductionist nature of alternatives (see below), processes that may influence the human *in vivo* potency (including ADME) are present only to a limited extent in alternative approaches. Hence there is considerable uncertainty regarding the use of concentration-response information from an artificially reduced test system (e.g. a confluent layer of hepatocytes) for predicting potential dose-response relationships (e.g. for hepatotoxicity) via *in vitro–in vivo* extrapolation (IVIVE), even if rooted in mechanistically informed physiologically based kinetic modelling (PBK).

2.1.3 Definition of "Alternative Approaches"

We refer to the definition of "alternative approaches" as suggested by Smythe (1978), i.e. alternatives to *established scientific procedures* which can lead to the *replacement*, the *reduction* or the *refinement* of animal experimentation, thus addressing the 3Rs principle as established by Russell and Burch (1959). Alternative approaches in this sense cover individual test methods, test batteries, strategic combinations of test methods (testing strategies) as well as holistic approaches towards data generation, evaluation and integration. These have been termed "Integrated Testing Strategies (ITS)" or, more recently, "Integrated Approaches to Testing and Assessment (IATA)" and can be composed of testing and non-testing methods. Validation can in principle extend to the assessment of integrated approaches (Kinsner-Ovaskainen et al. 2012). A surprisingly common misunderstanding regarding validation is that it is focusing on one-to-one replacements, i.e. one single alternative that replaces one single traditional animal test. This is however not the case, validation is context-dependent and purpose driven and includes all sorts of assays, also those that address initial mechanisms of action, intermediate effects, pathways of toxicity or modes of action. Further, the term 'alternative method' can relate to *empirical testing methods* (often *in vitro* methods) or methodologies that are not based on empirical testing and therefore referred to as "non-testing methods".

Non-testing methods are essentially approaches employing basic logical and plausibility reasoning or sophisticated mathematical approaches. Examples of non-testing methods include grouping of substances, read-across from one substance to another on the basis of properties such as chemical structure or biological mechanisms, structure-activity relationships (SARs) and quantitative SARs (QSARs). It also includes, in the wider sense, biological modelling approaches including modelling the kinetics of xenobiotics such as physiologically based pharmacokinetic (PBPK) modelling and its applications in toxicokinetics.

With *test methods* we refer to a scientific methodology based on a biological test system (e.g. a cell population, a reconstructed tissue or an excised organ) as well as provisions for handling this system and performing measurements following exposure to chemicals (i.e. the test method's procedure, normally captured in Standard Operating Procedure(s), SOP(s), outlining the related life science or analytical measurement techniques), as well as those relating to data analysis, processing and interpretation.

All alternative methods will need to process the raw data and translate them into toxicologically meaningful information, i.e. the actual results of the test method. This process is often referred to as *data analysis*. The results can then further be converted into predictions of the toxic effects of interest. This is achieved by so-called *prediction models* (Archer et al. 1997; OECD 2005), a description of how to interpret the data or measurements in view of obtaining categorical predictions. This often takes the form of a mathematical function or algorithm The predictions can stretch the entire spectrum of biological organisation, from molecular interactions over mechanisms on organelle or cell level (e.g. signalling pathways) up to mechanisms of cell ensembles, tissues, organs up to the entire organism or (sub)populations. A prediction model can be generically phrased as

$$P = f(x)$$

with P the prediction, f the mathematical transformation of the measured data x. Prediction models can be very simple, for instance in case of *in vitro* skin irritation test methods based on reconstructed human epidermis, a 50 % cell viability of the exposed skin equivalent is taken as a cut-off for ascribing either irritant or non-irritant properties to the test chemical. Notably, not all alternative test methods do feature prediction models, e.g. ecotoxicological assays.

The key components of alternative test methods are schematically summarised in Fig. 4.1.

2.1.4 Alternative Methods are Proxies and Reductionist Models

Typically, life science research is conducted on model systems, which can be further separated into (a) "*proxy*" (*or* "*surrogate*") *systems* and (b) *reductionist systems*, with possible overlap between the two (see below).

First, proxy systems are entities used to study properties of another system: a substantial amount of basic research in biology and biomedicine is conducted on

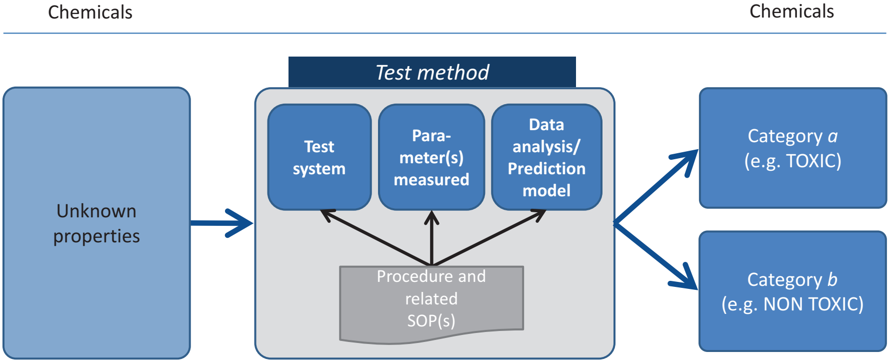
Chemicals
Chemicals
Test method
Unknown properties
Test system
Para-meter(s) measured
Data analysis/ Prediction model
Procedure and related SOP(s)
Category *a* (e.g. TOXIC)
Category *b* (e.g. NON TOXIC)

Fig. 4.1 An alternative test method and its main constituting elements: the biological 'test system', the biological parameters measured in response to exposure of the test system to chemicals and the element of data interpretation and analysis that translate data into toxicologically useful information. This may (but does not need to) include a 'prediction model', i.e. a prescriptive procedure of how to translate the measurements obtained in the test system into categorical predictions. Chemicals with unknown properties can be tested by the method and, using the prediction model or 'classifier', can be assigned to specific categories that can relate to any property ranging from activation of a cellular pathway to a downstream human health effect. Test methods can therefore be seen as "sorting machines" that allow to allocate chemicals with initially unknown properties to distinct categorical classes with defined properties normally relating to the presence or absence of the capacity to trigger a specific biological mechanism related to toxicity

animals (so-called "animal models") with a view to extrapolate the obtained results on animal genetics, biology and physiology to the human situation. Animals are used as proxies for humans with the basic assumption that, with increasing phylogenetic proximity, the results obtained in the proxies are considered more relevant, accurate and less uncertain with respect to reflecting the situation in the "target system", i.e. the human. In many cases, this has proven a successful approach in life science. For example, understanding the dopaminergic system and its involvement in Parkinson's disease has been largely obtained through animal experimentation. Despite these successes, there are limitations with regard to the use of proxies, probably due to phylogenetic differences, including at the level of gene expression, physiological mechanisms, metabolism, etc. that add uncertainty to results from animal studies as models for the human situation. Recent examples from pharmacological preclinical safety trials include the "tegenero incident" (Horvath and Milton 2009; Attarwala 2010) and the unexpected (hepato)toxicity of the antibiotic trovafloxacin (Borlak 2009; Gregory 2014). Secondly, there are reductionist models, such as brain slice cultures, dissociated primary cells and cell lines which are used to study specific physiological processes which recapitulate, at a highly reduced level of complexity, specific mechanisms, structures or other properties of the target system.

Alternative *in vitro* methods represent an interesting blend of these two concepts: they are proxies inasmuch as they are used for human safety testing *in lieu* of humans but they are also highly reductionist methods, since they are modelling only aspects of the target system (e.g. a complete organism or an organ, etc.) and are used to predict the properties of the target system or some of its constituting parts. Consider a barrier model composed of confluent polarised epithelial cells from the gut used to study uptake of substances through this epithelium or the use of monocytic cell lines for studying markers of epidermal inflammation and immune cell activation in the context of skin sensitisation leading to the clinical manifestations of allergic contact dermatitis. Both are highly reduced systems that model key properties thought to underlie higher level ("downstream/apical") effects in the system of interest, the 'target system'. As a major consequence of these facts, both test method development and validation are typically undertaken in relation to proxies or surrogate systems (i.e. animal data) and not the species of interest (Fig. 4.2).

Epistemologically, *in vitro* alternatives used to predict behaviour of chemicals in more complex systems up to the entire organism can be seen as a variety of *explanatory reduction* (Weber 2005). This type of reductionism can be seen as based on the identification of a "difference-making principle" (Waters 1990, 2007) assumed to be a causative (and/or explanatory) factor that is sufficient for studying and explaining features that are emergent at a higher level of organisation. Reductionist systems used to predict such higher level (downstream) properties need to model this "principle" (here: a physiological mechanism; in genetics: the concept of the gene) in order to study potential consequences in the complex target system (e.g. toxicity in humans and human (sub)populations). In the current context, finding the difference making principle is equal to the identification of physiological mechanisms believed to underlie a response (i.e. an adverse effect) in the target system.

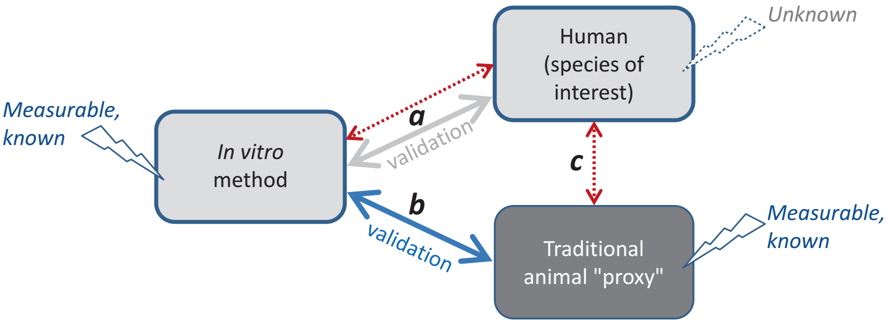

Fig. 4.2 Peculiarities and obstacles specific to validation of alternative methods. The effects of chemicals on the target system of interests (humans) are normally not known (*dotted flash*) or only known to a very limited extent and associated with much uncertainty. (**a**) Alternatives can therefore not be readily validated in relation to the target system (*grey arrow*). (**b**) Since traditionally animals have been used as proxies for humans, a lot of data is available concerning chemical effects on whole animal systems (*blue outlined flash*). In the absence of standardised human data, these data can be used as reference data for validation (*blue arrow*) and related to the measured data—also for developing a prediction model that translates the measured data *in vitro* into a prediction of measured effects in the animal proxy. (**c**) Importantly, the true relationships of the effects measured in the animal proxy and in the *in vitro* method to the target system are often not known (*red dotted arrows*)

In the context of validation, the entire mapping of biological and mechanistic information has been referred to as the "scientific basis" of alternative test methods. This relates to their biological and mechanistic design in terms of recapitulating mechanisms of toxicity (e.g. cellular pathways) or any relevant disturbances of structure or function at the cellular level (mode of action concept) or throughout the different layers of biological organisation. This means that responses from a reduced system need to be extrapolated to a more complex system.

Identifying the key causative factors or events that underlie specific adverse outcomes and which allow predicting these outcomes with sufficient accuracy and reliability is one of the biggest challenges of modern toxicology. The basic assumptions are that it is (a) unnecessary and (b) practically impossible to model all potential mechanisms. Finding those that truly make a difference in view of tilting the homeostatic balance and driving adverse effects is pivotal for developing relevant test methods. While a thorough understanding and description of the scientific basis of alternative test methods has always been part of validation, there are increasing efforts to organise the existing scientific knowledge in a consistent manner so as to improve interactions between various actors within the community (e.g. scientists, test method developers, validators, regulators, legislators, test method users). Identifying and describing physiological key events that can be perturbed by toxicants will allow adjusting chemical design of new substances so as to avoid the interference of substances with known "toxicity pathways" but will allow also help tailoring the scientific development of new test methods and informing their validation. The OECD guidance of describing "Adverse Outcome Pathways" (AOP)

(OECD 2013a, b) is a recent effort to describe the biological events leading to toxicity in humans in a concise but consistent manner. The AOP concept foresees the structuring of key events in relation to the biological level of complexity on which they occur and arranging the different events in causality chains, starting from a molecular initiating event and describing the causal relationships between one key event and another. This approach could improve the identification and description of such key factors and might thus support the development and validation of test methods that map/recapitulate mechanisms and pathways that underlie downstream events or higher-level features of the system. In addition, it has been proposed that evidence-based methods such as systematic reviews could help identifying key causative events triggering the development of biologically relevant test methods (Guzelian et al. 2005; Hoffmann and Hartung 2006b).

2.1.5 Reductionism: Consequences for Test Development and Validation

Above considerations show that the usefulness of alternative test methods needs to be assessed in relation to the target system: *reference points* (=data) need to be derived ideally from the target system that the alternative approach is modelling. It is not sufficient to validate *in vitro* methods in relation to other *in vitro* systems (Goldberg et al. 1995). This is in contrast to many forms of validation where the usefulness of systems is judged in relation to reference points relating to performance of similar systems (e.g. diagnostic tests). Finding accurate reference data (Fig. 4.2) of the actually "true" effect of a given chemical on the species of interest (humans) would be the obviously the ideal approach for assessing the usefulness ("relevance") of alternative methods. However, human data relating to chemical effects (Fig. 4.2) are normally not readily available or need at least to be derived from highly uncertain information (e.g. epidemiological data), involving moreover expert judgement.

This absence of human data makes it very difficult to "calibrate" alternative methods during test method development against the target system whose properties the alternative is intended to predict. This "calibration" typically consists of developing a data analysis procedure for processing the raw measurements into toxicologically meaningful results. This can include a prediction model that translates the measurements obtained in the alternative methods into categorical predictions, either relating to a category system used for hazard labelling relating to adverse health effects (e.g. UN GHS categories) or to a specific mechanism of action or toxicity pathway. To overcome this issue of non-availability of human data, reference data are traditionally taken from proxies or "surrogates", i.e. animal models that have been used in toxicology for many decades (consider for instance the Draize eye and skin tests in rabbits dating to 1944) although these animal models have never been validated themselves (i.e. how well they model or predict the effects in humans or how reliable/repeatable they are).

Moreover, there may be cases where no *reference method* is available, for instance when a method for a new purpose needs to be developed. This would mean that there are no *reference data* at all for the development and validation of an *in vitro* method. The statistical tool of “latent class analysis” (LCA) may be a viable approach to estimate assay performance parameters even when the true state of nature is not known or has not been observed (=is ‘latent’) (Hothorn 2002; Hoffmann et al. 2008). Yet another situation is found in environmental toxicity assessment which uses very few surrogate species/proxies for judging the impact of chemicals on a specific environment, habitat or ecosystem. An example is the use of daphnia or fish as surrogates (i.e. two of thousands of species!) for judging the potential impact on the aquatic ecosystem. For more details on reference points in validation of alternatives see the ECVAM workshop report by Hoffmann et al. (2008).

2.1.6 Modelling the Mechanism Is Necessary But Not Necessarily Sufficient

As outlined above, alternatives that are based on modelling biological events that are assumed to be causative for adverse effects in the species of interest are more credible and useful than methods that show only correlative results with the target system, i.e. without modelling relevant biological mechanisms. This has been pointed out already in early publications on validation (Goldberg et al. 1995; Bruner et al. 1996). It is thus tempting to assume that methods which model such results should quasi automatically produce results that are informative and relevant for downstream health effects. This is however not necessarily the case: biological systems have a great capacity to repair and reset their properties once disturbed (homeostasis). Reduced test methods typically do not model all those homeostatic mechanisms and hence the results can be of limited relevance, especially for health effects that depend on repeated exposure and a variety of stressors (e.g. epigenetic changes involved in cancer). Hence, the modelling of mechanisms in alternative test methods is a necessary precondition of robust and relevant predictions, but it is not necessarily sufficient with respect to the accuracy of such predictions.

2.1.7 Reductionism Requires Integration at Later Stages

A consequence of the fact that alternative *in vitro* methods are *reductionist* models is that in most cases no single method will suffice to describe the properties of the higher-level target system with its complex anatomical and physiological organisation. Consider the health effect of reproductive toxicity: several organs and complex hormonal feedback loops are involved which cannot be modelled by one single reductionist system. Instead, the lack of complexity at the level of individual test methods is sought to be compensated by using a suite of test methods and other information sources that each address different properties of the target system. The complexity of the target system is basically dissected into aspects that can be modelled and experimentally manageable in several reductionist systems. Information from such groups of methods (including also non-testing methods) then needs to be integrated through strategic combinations of test methods in holistic data gathering and evaluation schemes. These, initially, have been termed tier-testing strategies or testing strategies (Balls et al. 1995b) and later referred to as "Integrated Testing Strategies", implemented in the REACH guidance published on ECHA's website from 2007 onwards and including also elements of data collection and evaluation (see also Kinsner-Ovaskainen et al. 2012; Balls et al. 2012). Subsequently, the OECD has introduced the term "Integrated Approaches to Testing and Assessment" (IATA) (OECD 2008; OECD 2014a, b). The communality is that data from various sources, irrespective of whether already available or to be generated, are integrated in order to yield conclusions on whether specific chemicals trigger a particular property of the target system.

This need for data integration has important consequences for validation. Back in the early 90s validation of alternative methods initially intended to establish single replacement methods for addressing an entire health effect (e.g. EC/HO validation study on alternatives to the Draize eye irritation test). This has worked to some extent in topical toxicology where the Draize test for skin corrosion and irritation could be successfully addressed by two sets of *in vitro* test methods, both based on Reconstructed human Epidermis (RhE) (additional methods are available for skin corrosion assessment). However, it is plausible that other health effects of a more systemic nature (often referred to as "complex endpoints" or "systemic toxicity") will require a strategic combination of alternative methods and this has to be taken into account already when validating the individual "building blocks" of such strategies (Bouvier d'Yvoire et al. 2012). This has been discussed in a joint EURL ECVAM/EPAA workshop report (Kinsner-Ovaskainen et al. 2012). We will return to this aspect later in the context of the requirements in terms of chemical number for assessing reliability versus predictive capacity/applicability domain.

The later use of an alternative test method within larger integrative approaches has impacts on validating such a method and the study design needs to take this into account. For example, if a method is used in (strategic) combination with other assays, it is conceivable that the requirements regarding predictive capacity and even reliability are different as opposed to situations where a method would be used as a stand-alone test. The same holds true for a screening assay versus one used to address regulatory requirements as observed by Green (1993).

2.2 Validation: Basic Terminology

Traditionally validation has been seen as a process of assessing the scientific validity of an alternative method. While this is still correct, validation carries additional meanings: for example, when scientists talk about a test method as "being validated" they rather refer to whether or not a method has been shown to be reliable and relevant, i.e. whether the hypotheses that modelling a specific mechanism of action in such a reductionist model indeed picks up, during validation testing, chemicals known to cause specific adverse effects. Therefore, the term validation in the area of toxicity texting incorporates at least three aspects: (a) the formal process of validation or validation workflow; (b) the validation study type (or "validation process") and (c) information generated during validation or the hypotheses assessed during validation.

2.2.1 Validation Workflow

Validation traditionally has been overseen by independent and impartial organisations ('validation bodies') that do not have vested interests. Examples are EURL ECVAM in the EU, NICEATM/ICCVAM in the US, JaCVAM in Japan, KoCVAM in South-Korea, Health Canada and BraCVAM in Brazil. This impartiality is important for the validation of alternative methods that are intended for regulatory acceptance: it ensures that the characterisation and confirmation of validity of test methods is done on the basis of scientific considerations only and independent of specific/vested interests (financial, etc.) of test method submitters. It thus guarantees impartiality, scientific rigour and consistency of approach. All validation organisations follow a practical workflow or process for prioritising test methods for validation, for conducting studies, for subsequent independent peer review and for organising and communicating their main conclusions and recommendations. The above mentioned (supra)national validation bodies in the EU, Japan, Canada, South-Korea and the US work together within the ICATM framework (ICATM = International Cooperation on Alternative Test Methods) and have recently attempted to align and streamline their workflow (see Chap. 14 on international collaboration). A generic validation body workflow comprising four basic steps is shown in Fig. 4.3 and explained below.

Step 1: Evaluation of *in vitro* test methods: Importantly, not all *in vitro* methods that are developed by test method developers will be necessarily validated by validation bodies. Before being able to enter validation, proposed *in vitro* test methods will need to be evaluated against a catalogue of criteria such as: is the test method sufficiently developed to enter validation, in particular is there a "mature" protocol/SOP available and are there some initial data on within-laboratory repeatability and reproducibility (for details see, Sect. 2.3.2)? Does it produce information that could be useful for the intended application, in particular regulatory decision making? Once these criteria have been confirmed, a test method might be considered for validation, a process that involves the use of considerable funding and resources, typically of public funds. Only methods that promise to generate useful information and, in particular, address toxicity effects for which there is no 'alternative coverage' yet, will merit such investment. Essentially, at this step, validation organisations will conduct a cost/benefit analysis in view of prioritisation.

Step 2: Designing and conducting a validation study: Validation involves dedicated scientific studies to determine whether the alternative method appropriately models and, if applicable, predicts the properties of the target system. We will discuss

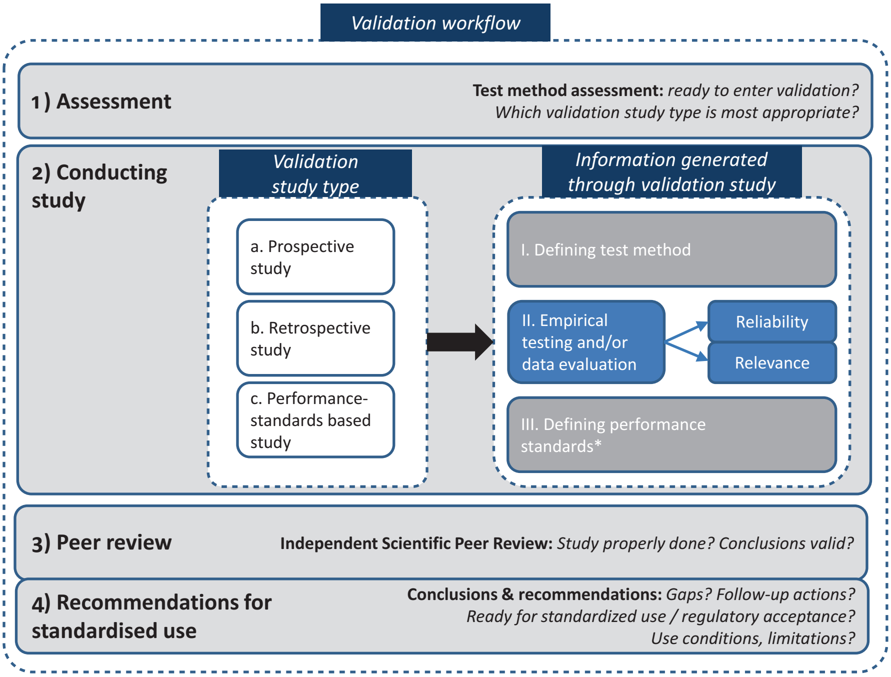
Validation workflow
1) Assessment
Test method assessment: ready to enter validation?
Which validation study type is most appropriate?
2) Conducting study
Validation study type
a. Prospective study
b. Retrospective study
c. Performance-standards based study
Information generated through validation study
I. Defining test method
II. Empirical testing and/or data evaluation
Reliability
Relevance
III. Defining performance standards*
3) Peer review
Independent Scientific Peer Review: Study properly done? Conclusions valid?
4) Recommendations for standardised use
Conclusions & recommendations: Gaps? Follow-up actions?
Ready for standardized use / regulatory acceptance?
Use conditions, limitations?

Fig. 4.3 Schematic outline of the overall validation workflow of independent validation organisations

the various validation study types in more detail in Sect. 2.2.2 below. Irrespective of the validation study type, the information generated relates to three aspects: (i) a better characterisation and definition of the *in vitro* test method (as a result of validation), (ii) the assessment of the key hypotheses of reliability and relevance (which we will explore below in more detail below, Sect. 2.3) as well as (iii) the setting of performance standards and operational criteria (e.g. refinement of test acceptance criteria) that will guide development and validation of future test methods based on similar principles.

Step 3: Independent scientific peer review. Since validation is achieved through scientific experimental studies which, as all scientific endeavours, contain elements of data interpretation and inference to reach conclusions, an essential element of validation is the assessment of the results obtained and the conclusions drawn by an independent group of knowledgeable scientists. This peer review assesses whether key scientific principles such as objectivity and appropriateness of methodology have been observed and will to this end evaluate managerial and study design aspects, ranging from the choice of specific readouts, over composing the data matrix to statistical tools used for analysing the data. In contrast to the peer review of scientific manuscripts undertaken by individual scientists with normally relatively little guidance from the journals/publishers, the peer review of alternative methods, especially when conducted by public validation bodies, needs to be highly consistent in terms of the quality criteria used and the information assessed so as to ensure equality of treatment of submissions from various test method developers which often have commercial interests in the validation of their methods.

Step 4: Final conclusions and recommendations: The peer review will inform on the quality of the study and results and to which extent the conclusions drawn are justified by the data/results obtained. The peer review typically forms the basis for the definition of final conclusions and recommendations on the readiness of the alternative test methods for acceptance into legislation as an officially (and ideally internationally) acknowledged and recognised routine test method for applications that aim at compliance with legislative requirements in view of safety assessment. This process is commonly referred to as 'regulatory acceptance', and, although independent of the validation process, relies on the quality of validation studies. Thus, the final conclusions published at the end of the validation workflow by validation bodies (e.g. "EURL ECVAM Recommendations") inform the relevant stakeholder communities on the characteristics of test methods, identify existing gaps and necessary follow-up activities and therefore prepare and support mainly the scientific aspects of regulatory discussions towards official acceptance. Stakeholders include regulatory end users in governmental agencies, industry end users of the test method as well as civil society organisations (such as animal welfare or environmental activists).

Increasingly, validation is also performed by other actors than validation bodies such as test method developers in academia and industry. These parties may seek independent and impartial evaluation and peer review of their studies by validation bodies that are neutral with respect to the assay (i.e. do not have vested interests). For instance, EURL ECVAM is regularly evaluating 'external' validation studies and having them reviewed by its independent EURL ECVAM Scientific Advisory Committee (ESAC).

2.2.2 Validation Study Types

Validation of alternative methods for toxicity testing is centred on the analysis of testing data relating to a relevant set of chemicals (the so-called "test chemicals"). Testing is normally carried out through formal validation studies that should follow scientific principles and good scientific practice with regard to study design and conduct (see Chap. 5), in particular relating to chemical selection, the statistical planning (e.g. to calculate the power required to derive dependable point estimates such as sensitivity and specificity), but also with regard to the statistical analysis of the study data themselves. This will be explored in detail in Sect. 3.

There are different types of validation studies conceivable that vary in their design. A useful distinction is based on whether the chemicals testing data need to be generated *de novo* (so called prospective studies) or whether they are already existing and are analysed in view of a defined purpose (retrospective studies). Studies can of course also contain both prospective and retrospective elements: make use of newly generated data as well as existing data (see Sect. 2.4 on modular approach).

Prospective Studies

(a) Prevalidation studies

Prevalidation studies are studies conducted in view of assessing whether a test method and associated SOP is ready to merit potential further full validation (Curren et al. 1995) and robust enough to merit the considerable expense of such a study. Prevalidation studies focus on the aspect of transferring the SOP/test method from an experienced laboratory (e.g. test method developer) to naïve laboratories. These studies allow optimising further the SOP based on the experiences during such transfer. Prevalidation studies thus help minimising the risk of transfer problems during full prospective validation studies. Transfer problems due to shortcomings of the SOP or training protocols only create unnecessary cost during full validation studies without contributing to the core goal of a full validation study, i.e. test method characterisation in view of a purpose. Transfer(ability) is assessed through testing a small but conscientiously selected set of chemicals with also challenging properties, such as chemicals that are at the border of the prediction model cut-off (see also Sect. 4.7.2). A major benefit of conducting prevalidation studies is that they produce limited but quality controlled data sets on within-laboratory reproducibility (WLR), between-laboratory reproducibility (BLR) and predictive capacity which may inform about the possible overall *performance* to be expected of a specific test method (see also Sect. 4.6). Like in other validation contexts (e.g. analytical method validation) such *a priori* knowledge and other historical data may support the realistic setting of goals and objectives of subsequent full prospective validation studies, including potential validation acceptance criteria where useful (i.e. if the precise use of the method in a regulatory setting is already known). Although the term prevalidation is not any longer frequently used, there are still studies conducted that adhere to the principles of prevalidation, namely a first check of transferability of SOP from one laboratory to another, identification of pitfalls and improvement of SOP and/or training, if necessary, before embarking on a costly multi-laboratory ring trial.

(b) Full prospective validation studies

These are large-scale studies involving the testing of a sufficient sample of chemicals for characterising a test method in terms of WLR, BLR and predictive capacity and for characterising, with some confidence, its applicability domain and potential limitations. Adaptions of the design of such studies have been suggested and will be discussed below. Such studies create confidence and trust in alternative novel test approaches for regulatory applications when involving a sufficiently large set of test chemicals. When considering the size of the chemical testing set, it is important to separate what would be statistically desirable (in terms of chemical sample size) from what is realistically doable taking also considerations of cost and availability of test materials into account.

(c) Performance standards-based (PS-Based) validation studies

PS-based studies are conducted in relation to a set of predefined "standards", including biological criteria and reference chemicals, as a means of efficiently assessing test methods considered to be sufficiently similar to a previously-validated one. This concept has been initially proposal by Balls (1997). PS are typically defined upon completion of a full validation study. However assessment criteria (factually standards) can also be defined for test method development already and carried over to test method evaluation/validation.

Normally PS-based studies are used to validate, through a smaller scale study involving significantly less chemicals, test methods that are scientifically sufficiently similar to the previously validated "reference methods". The rationale of these studies is that similar biological and operational characteristics will most likely mean that the general performance of a similar method can be assumed to be equivalent to the validated reference method and that it is therefore justifiable to test a smaller set of chemicals instead of repeating a full scale validation exercise. Performance standards typically are composed of three elements: (i) The essential test methods components, defining the test methods and key operational parameters, (ii) a set of "Reference Chemicals" that need to be tested (typically in the range of 20 or so), (iii) target performance values in term of WLR, BLR and predictive capacity. Importantly both the reference chemicals and the target values are defined on the basis of the full validation and the parameters achieved: thus, these values map at a reduced scale the chemical, toxicological and functional spectrum of test chemicals and the values attained during validation. A significant drawback of this study type is that the test chemicals are known beforehand and can be used for test method development.

Retrospective Studies

These studies use existing testing information that can be analysed through data grouping and meta-analysis tools. For a short introduction to meta-analysis see Mayer (2004). Retrospective validation may sometimes be conducted through a 'systematic review'; this term however rather relates to the methodology used. As long as the goal of the systematic review is to characterise a method, and through this, assess its validity for a purpose, it technically constitutes a validation exercise. Retrospective studies require particular attention with respect to the selection of the data through using pre-defined search and selection criteria.

Weight of Evidence Validation and Evidence-Based Validation

While the terms prospective and retrospective validation relate to the temporality of data generation, the concept of weight of evidence (WoE) validation relates to the tools for evaluating data sets relevant for a given validation study. Weight of evidence generally relates to the considerations made in a situation where there is uncertainty and which are used to ascertain whether the information/evidence at hand support one or the opposite side of an argument or a conclusion. In the context of validation, WoE considerations can be useful in situations where there is uncertainty regarding the available reference data or in case different and opposing findings from reference methods are available. WoE judgment may of course also be used in case there are several contradictory testing results in retrospective data sets. Possible principles of WoE validation have been summarised by Balls et al. (2005). However, WoE approaches can be tailored to individual needs as long as they are underpinned by the consistent use of a predefined set of criteria relating to quality, relevance, plausibility, etc. of the data. The second element, integrating this information in view of arriving at a final judgement, may depend on the specific case.

Evidence-based validation (Hartung 2010) is a term suggested for validation studies that make full use of data assembly and analysis tools as well as advanced statistical tools as used in (evidence-based) medicine (Mayer 2004). This includes data grouping, the concepts and techniques of meta-analysis as well as the use of likelihood ratios to summarise predictive performance and the consideration of prevalence (Hoffmann and Hartung 2005). This should be seen in the wider context of introducing evidence-based methods from medical research (including systematic reviews) also in toxicology (Hoffmann and Hartung 2006b; Griesinger et al. 2009; Guzelian et al. 2009; Stephens et al. 2013) in order to address in particular issues of variability and uncertainty (Aggett et al. 2007) and make remaining uncertainties transparent (Guzelian et al. 2005). While the evidence-based approaches from medicine can to some extent be used also in toxicology (Neugebauer 2009), there are however important differences between medicine and toxicology which need to be taken into account (Griesinger 2009) (e.g. the focus of prevention in toxicology versus prevention and cure in medicine or the differences of the entities studied through test methods: chemical properties in toxicology and diseased patients in medicine).

2.3 Validation: Hypotheses Assessed and Information Generated

Having outlined fundamental concepts relating to the assessment of alternative methods and validation workflow, we explore the term validation in more detail in the following.

2.3.1 The General Concept of Validation

Validation aims to show whether or not something is valid. "Valid" is rooted in the latin verb *valere*—to be (of) worth. This shows the core goal of any validation: assessing whether something has (some) worth or usefulness. From this, two key characteristics of all validation exercises can be deduced:

- First, the terms "worth" or "valid" are highly context-dependent: something is of "worth" in relation to something or for a specific use, application or performance. Thus, validation always relates to a *specific context* or *purpose*. This purpose may change over time, requiring revisiting or re-conducting validations of systems that have been previously validated in relation to a different purpose. In the context of alternative methods validation, this purpose-oriented aspect is described by the term "**relevance**". Relevance has been described as the *usefulness* and *meaningfulness* of the results of an alternative method (Balls et al. 1990a, b, c; Frazier 1990a, b). We would like to emphasize that it is this rather broad understanding of relevance (Bruner et al. 1996; OECD guidance document Nr. 34, glossary) that we are using here. Unfortunately, relevance has sometimes been reduced to mere aspects of predictive capacity and applicability of an assay. However, judging the overall relevance requires the integration of many types of information and requires also scientific judgement: relevance is a *composite measure* and involves also the biological/mechanistic relevance ("scientific basis") and may also include considerations of reliability of a test method (Bruner et al. 1996). We will discuss this in more detail in Sect. 2.3.3.

- Secondly, a system/method or process is only then fully relevant for an application or purpose, if it is reliable: if it performs in the same manner each time it is applied, irrespective of the operator and in a reasonable independence of the setting within which it is used (e.g. a computer programme should not only work on the developer's computer, but on those of millions of users). This is described by the term "**reliability**". It is immediately intuitive that a test method that is unreliable cannot be relevant for its purpose. Inversely, the purpose of a method will have an influence on the reliability that is requested from of a given test method. For some purposes (e.g. when combining test methods in a battery) a lower reliability may be acceptable than when using an alternative method as a stand-alone replacement test. Thus, reliability may need to be taken into account when judging the overall relevance of a test method for a purpose.

Based on these brief considerations, one can frame the key characteristics of any validation exercise including alternative method validation:

1. Validation is the *process* required to assess/confirm or assess validity for purpose as described under (2)
2. The validation process concerns the *assessment of the value* (*validity*) of a system *within a specific context and/or for a specific purpose*, typically a use scenario or a specific application by examining whether the system reliably **fulfils the requirements** of that specific purpose in a **reliable** manner and is **relevant** for the intended purpose ("fitness for purpose") or application.
3. The validation process is a **scientific endeavour a**nd as such needs to adhere to principles of objectivity and appropriate methodology (study design). Accepting that validation studies are of a scientific nature means that they should be described in terms of assessing clearly described *hypotheses*. These hypotheses include (1) the reliability of an assay when performed on the basis of a prescriptive protocol, (2) the mechanistic or biological relevance of the effects recapitulated. This is measured through testing, during validation, a wide array of chemicals with known properties regarding an adverse health effect: if the modelled mechanism is relevant, this will be reflected in the accuracy of the

predictions or measurements. This will also show whether there are specific chemical classes or other properties for which no accurate predictions can be obtained (applicability/limitations); (3) the predictive relevance, i.e. the appropriateness of the prediction model developed typically on a small set. Obviously, hypotheses 1–3 are related. For practical purposes, they are grouped in *reliability* and *relevance*.

Typically, validation has assessed this "fitness for purpose" outlined in the three hypotheses above by studying (i) whether or to which extent the system fulfils *predefined specifications* relating to performance (for instance sensitivity and specificity of predictions made), (ii) the reliability (and operability) deemed necessary to satisfy the intended purpose as well as (iii) robustness, which is measured *inter alia* through the ease of transferring a method from one to another laboratory which is typically done in prevalidation studies (Curren et al. 1995). Points of reference or predefined standards for predictive capacity and reliability therefore play a key role in validation (Hoffmann et al. 2008). Importantly, the process of validation will inevitably lead to the **characterisation** of the system's performance and, if applicable, its operability, generating useful information even in case the validation goal/objective is not met, the method is not (yet) found fit for purpose or "scientifically valid". Test method validation, should therefore also be seen as a way of characterising a system for future improvement and adaptation. It is this general concept of validation that underlies also the validation of alternative approaches.

2.3.2 Validation of Alternative Methods: Reliability and Relevance

As outlined above, the theoretical basis of alternative method validation can be readily deduced from the general concept of validation: the two key hypotheses assessed by alternative method validation are **reliability** and (overall) **relevance** incorporating biological relevance, relevance (concordance) of predictions for various chemicals (applicability domain) and, at times, reliability.

This definition goes back to discussions at a workshop in Amden, Switzerland in 1990 conducted by the *Centre for Alternatives to Animal Testing* (*CAAT*), USA and the *European Research Group for Alternatives in Toxicity Testing* (*ERGATT*) whose results have been published as CAAT/ERGATT workshop report on the validation of toxicity test procedures (Balls et al. 1990a, b, c). Being sufficiently general, the original definition relating to relevance and reliability provides an appropriate framework for validation of alternatives still today.

In the following we would like to explore how these two hypotheses are addressed in validation studies in more detail:

First, an alternative test method can only be considered useful if it shows ***reliability***, i.e. if it provides the same results or shows the same performance characteristics over time and under identical as well as different conditions (e.g. operators, laboratories, different equipment, cell batches, etc.). In the context of validation studies, reliability has been defined as assessing the (***intra-laboratory***) ***repeatability*** and the reproducibility of results ***within*** **and** ***between laboratories*** over time (Balls et al. 1990a, b, c, 1995a, b; OECD 2005). Repeatability relates to the agreement of results within one laboratory when the procedure is conducted *under identical conditions* (OECD 2005), while reproducibility relates to the agreement of results using the same procedure but not necessarily under identical conditions (e.g. different operators in one laboratory or different laboratories).[1] Reliability assessment is important in view of assessing the performance of methods in their final use scenario, i.e. employed in laboratories across the world. Assessment of within- and between-laboratory reproducibility is often done by means of measuring *concordance* of (i.e. agreement between) predictions obtained with the prediction model. This has the advantage that the reliability is measured on the basis of the **intended results** or **output** generated by the test method, i.e. again under final use conditions. However, it is also important to describe, using appropriate statistical methods, the **intrinsic variability of the parameter(s) measured** (see also Sect. 4.7.2) in the test method (e.g. cell viability, fluorescence as a result of the expression of a reporter gene, etc.). This will allow producing data on reproducibility (or inversely variability) independent of the prediction model and therefore closer to the actual data produced. Such data may be useful in case the prediction model is changed due to post hoc analyses. A *post-hoc* improvement of the prediction model has recently been done on the basis of *in vitro* skin corrosion methods (Desprez et al. 2015). In addition, the transferability of a method is an aspect that needs attention during validation: it relates to how easily a method can be transferred from one experienced laboratory (e.g. test method developer) to naïve laboratories that may have relevant experience with alternative methods but are, at least, inexperienced with the particular SOP associated with the test method (Fig. 4.1). Transferability relates to both the reliability but also the "robustness" of a test method: the more sensitive a method is to slight variations of equipment and operators, the less robust it is. Robustness is important when considering a test method for standardised routine use. A practical way of gauging robustness at early stages is through checking the ease with which a test method can be transferred from one to another laboratory (e.g. in the context of a prevalidation study). Robustness however will also be reflected in the levels of repeatability, and within- and between laboratory reproducibility obtained during validation.

Second, in view of ensuring that an alternative test method is fit for a specific purpose (i.e. the reliable generation of data on the properties of test chemicals) its ***relevance*** *for this purpose* needs to be assessed. This requires that the *purpose is clearly defined* before validation. A surprisingly common shortcoming of validation exercises is that the intended purpose of the test method and, therefore, the goal and

[1] *Repeatability* has been defined as "*the agreement of test results obtained within a single laboratory when the procedure is performed on the same substance and under identical conditions*" (OECD 2005) i.e. the same operator and equipment.

Reproducibility has been defined as "*the agreement of test results obtained from testing the same substance using the same protocol*" (OECD 2005), but not necessarily under identical conditions (i.e. different operators and equipment).

objectives of a validation study are not defined with sufficient precision. This has been already remarked on by Balls and colleagues in 1995 (Balls et al. 1995a, b). Inversely, over-ambitious goals are sometimes set, including the specification of target performance values (e.g. for specificity and sensitivity) which are not sufficiently backed by prior data. Lack of goal setting or defining objectives has a negative impact on the clarity of study design (see Sect. 4): as for all scientific experiments, the objectives of a study will determine the necessary design. Thus, study design is not a 'one-size fits all' issue, but depends on the specifics of the study. With regard to validation of alternatives, relevance for a particular purpose has been defined as assessing the scientific *meaningfulness* and *usefulness* of results from alternative methods (Balls et al. 1990a, b, c, 1995a, b; Frazier 1990a, b). Meaningfulness in this context is crucial and relates to the plausibility of data or predictions and how convincing they are on the basis of a variety of considerations. As observed by Goldberg et al. (1995) and Bruner et al. (1996), hazard predictions from alternative methods that address a specific known mechanism of action or because they closely model a specific tissue are scientifically more credible and are probably more likely to be correct than predictions from a test methods that that does provide correct predictions but does *not* model the biology of the target system or whose relationship with the latter are at least unknown (such assays could be called "correlative methods"). Thus, when judging the overall relevance of a test method, also biological or mechanistic relevance needs to be taken into consideration, i.e. to which extent the alternative model recapitulates key aspects of biology, physiology and toxicity that need to be assessed. This aspect has traditionally been referred to as the "**scientific basis**" of a test method.

2.3.3 Key Information for Relevance: Scientific Basis, Predictive Capacity, Applicability Domain and Also Reliability

As indicated above, relevance is a rather broad term and judgement of relevance is to some extent a subjective process that relies on the evaluation and integration of scientific data. To assess or establish the relevance of a method for a defined purpose requires considering the method's predictive capacity, its applicability domain and limitations, its reliability and, at a more fundamental level, its scientific basis: the biological and/or mechanistic relevance of the test method in view of it being considered a suitable proxy or surrogate for the target system and a model of key causative elements that are involved in emergent properties of the target system (see discussion on explanatory reductionism Sect. 2.1 subpoint 3). Figure 4.4 schematically summarises the information taken into account for judging the overall relevance against the defined purpose.

The four aspects for judging relevance of a method are elaborated in the following:

(a) **Scientific basis** relating to the biological or mechanistic relevance of a test method and its underlying test system. Does it recapitulate a specific tissue

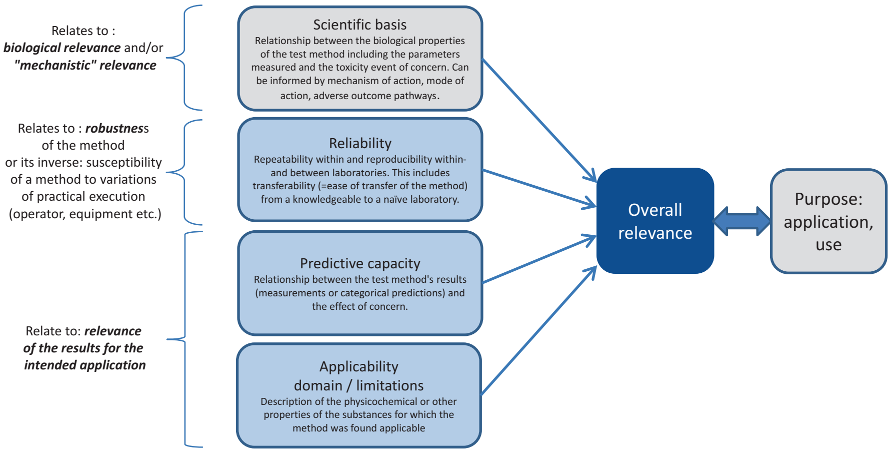
Relates to :
biological relevance and/or
"mechanistic" relevance
Relates to : *robustness*
of the method
or its inverse: susceptibility
of a method to variations
of practical execution
(operator, equipment etc.)
Relate to: *relevance*
of the results for the
intended application
Scientific basis
Relationship between the biological properties of the test method including the parameters measured and the toxicity event of concern. Can be informed by mechanism of action, mode of action, adverse outcome pathways.
Reliability
Repeatability within and reproducibility within- and between laboratories. This includes transferability (=ease of transfer of the method) from a knowledgeable to a naïve laboratory.
Predictive capacity
Relationship between the test method's results (measurements or categorical predictions) and the effect of concern.
Applicability
domain / limitations
Description of the physicochemical or other properties of the substances for which the method was found applicable
Overall
relevance
Purpose:
application,
use

Fig. 4.4 Judging the overall relevance of a method against a specified purpose upon completion of a validation study requires information on the biological and mechanistic relevance (scientific basis) of a test method, its reliability, its predictive capacity and applicability domain. Note that the scientific basis of a method should be defined on the outset of a study (*light grey*) and is not based on empirical testing generated during the study, while information on reliability, predictive capacity and applicability are assessed through the data generated during validation (*boxes in light blue*). Empirical data on the relevance of the results (e.g. an IC_{50} measurement) or categorical predictions (="predictive capacity") in regard of the effect of concern allow falsifying or "verifying" the hypothesis that a particular scientific basis is relevant for predicting an adverse effect. The scientific basis hence is the foundation of a test method. Its description is informed by considerations of mechanisms of action (MOA, relating to the specific biochemical interaction by which a drug/toxin acts on the target system), mode of action (MoA, relating to functional or anatomical changes correlated with the toxicity effect) and adverse outcome pathways (AOP, relating to descriptions of sequences of biological key events that lead from initial molecular interactions of the toxin with the system to downstream adverse health effects of individuals or populations)

architecture, mechanism or action or biological/toxicological pathway? We provide a few examples to illustrate this point:

Reconstructed human epidermis used for skin irritation testing has a high biological relevance for the intended application (prediction of the irritancy potential of chemicals) as it models the upper part of the human skin and is based on human keratinocytes. The predominant readout used for skin irritation testing is cell viability which has some relation to the toxicity mechanisms: it models cell and tissue trauma which is a key event for triggering an inflammatory response in skin leading to the clinical symptoms of irritation (redness, swelling, warmth) (Griesinger et al. 2009, 2014). However, more specific markers that directly probe inflammatory processes would be closer to the toxicity event from a mechanistic point of view (draft AOP in Griesinger et al. 2014).

As another example, transactivation assays for measuring the potential of chemicals to act as (ant)agonists on endocrine receptors (e.g. estrogen, androgen receptors) typically are based on cell lines intrinsically expressing these receptors. Such assays have a high mechanistic relevance as they directly model the mode of action. However, depending on the test system used and the degree of reduction applied (i.e. cell line versus tissue), they have a reduced biological relevance.

(b) **Predictive capacity**: **The relationship between the measurements obtained with the alternative method and the effects in the biological system that the alternative method is supposed to model**. Typically this relationship is captured through assessing the capacity of the alternative method to provide accurate predictions of specific effects in the biological target system. This is called a test method's "predictive capacity". The effects predicted typically relate to distinct categories and constitute "classifiers" (in standard scientific terms one could say that the continuum of effects from non-toxic to highly toxic has undergone a binning procedure; the basis for this binning often relate to decision-rules that relate to regulatory traditions of categorising health effects). These classifiers normally relate to predictions of downstream adverse health effect ("apical endpoint" such as skin or eye irritation and their respective classification and labelling categories), but they may also relate to a specific cellular mechanism involved in toxicogenesis ('toxicity pathway'), to an organ-level effect, etc.

An example of predictive capacity of a health endpoint is *in vitro* skin irritation: skin equivalent models based on human keratinocytes that grow into epidermis-like tissue equivalents in the dish are used to predict the skin irritation effect of chemicals in humans (OECD TG 439 2010; Griesinger et al. 2009). The capacity to predict skin irritation is characterised through an evaluation of test chemicals with known reference properties in the target (or surrogate) system. Here they relate to irritants as defined by classification and labelling schemes such as GHS versus 'non-classified'. The predictive capacity is described by standard statistical measures used for analysing diagnostic or predictive test methods, as long as these methods aim at making categorical predictions of the sort "positive" versus "negative" (=true presence or absence of a property). These are mainly sensitivity (=true positive rate), specificity (=true negative rate) and accuracy (sum of true negatives and true positives over all predictions made); see Fig. 4.5. Importantly, these are all statistical point estimates and they are independent of the balance between positives and negatives in the reference data. Often positive and negative predictive values (PPV, NPV) are also used to characterise the performance of alternatives. However, these values are dependent on the prevalence of positives amongst the test chemicals (see Fig. 4.5) and care needs to be taken when using these descriptors for predictive capacity of test methods after validation studies where normally the balance is 50:50 % (i.e. there is a 50 % prevalence). NPV and PPV only provide meaningful information when either the prevalence of the test chemicals during validation matches the prevalence in the real situation or by taking the prevalence into account when calculating NPV and PPV on the basis of the sensitivity/specificity values obtained during validation using a balanced set (50:50 %). Analogies between the assessment of test methods for chemical safety assessment and those for diagnosing diseases are tempting and hold true

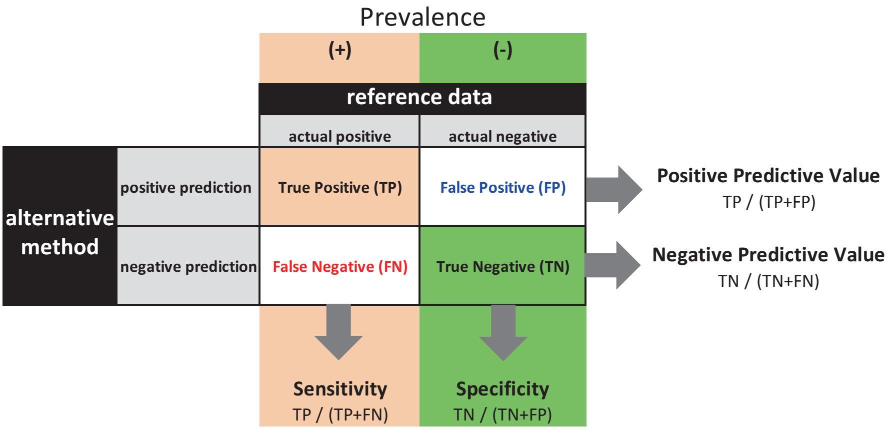

Fig. 4.5 Predictive capacity of a test method is described by assessing its ability to yield correct predictions for classes of properties described by reference data. In the example below, a classical contingency table, there are two categories of the reference data: actual positive and actual negative. The prevalence of chemicals that are ascribed these properties has impact on the statistical analyses and the parameters that are useful. The alternative test method has a prediction model that allows binary classification, either "positive" or "negative". Comparing the results of the alternative method with the reference data allows ascribing to the results of the alternative method the arguments True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). Note that the sensitivity (=true positive rate) and specificity (=true negative rate) are independent of the prevalence of actual negatives/positives. In contrast, both positive and negative predictive values are dependent on this balance

for the most of the statistical issues (Hoffmann and Hartung 2005), but should be used with some care due to obvious differences of the entities examined (diseased people versus chemicals causing adverse effects) (Griesinger 2009) and some issues related to prevalence: while the prevalence of a given disease (e.g. type II diabetes) may be grounded on solid evidence, establishing the 'prevalence' of toxic chemicals with regard to a specific health effect can be challenging. One approach used in the past was to assess the number of entries in chemical registries (e.g. the EU new chemicals database). However it should be noted that the chemicals listed there have already undergone safety assessments and the real prevalence of chemicals when they are being subjected to test methods may be different. Further, other measures in addition to NPV and PPV may be useful when expressing the quality of binary classifications, in particular in cases when actual positives and negatives are highly unbalanced. This includes the "Matthews Correlation Coefficient" (MCC) (Matthews 1975) that indicates the correlation between predictions and observations (actual negatives/positives) on a scale of −1 (no correlation whatsoever) over 0 (random) to 1 (fully correlated).

Assessing the predictive capacity of a test method requires the availability of **reference data** that are used to "calibrate" the prediction model of the method and to assess its predictive capacity during validation. These reference data are often from animal studies and relate to categorical values such as "actual positive" and "actual negative" ascribed to a set of test chemicals. Notably, reference data already carry a considerable degree of simplification due to the reduction of a much more complex reality of a continuum of physiological events into a binary (or other) classification. Reference data therefore need to be used with care, especially when derived from surrogate/proxy animal models, i.e. not the species of interest as is typically the case in toxicology.

(c) **The reliability** of a test method also may influence judgements on its overall relevance. Consider for instance the impact of the practical use scenario of a test method on its relevance judgment: test methods that will be used on their own (stand-alone replacements) will have to show a high degree of reproducibility in order to be judged relevant for the purpose of effectively replacing a traditional animal test. For example, reliability thresholds for single replacement test methods such as skin corrosion and skin irritation are very high. Other test methods on the other hand will be used in conjunction with others, either in parallel, assessing the frequency/mode of predictions obtained from such a "battery" or through strategic step-wise combinations of test methods.[2] In such use cases, test methods with reproducibility performances lower than those of single replacement methods may be nevertheless useful and judged relevant, for instance when used in weight-of-evidence approaches to support plausibility reasoning such as read-across of properties from one chemical substance to another. The relationship between intended use and requirements in terms of accuracy and also reliability was first noted by Green (1993).

Figure 4.6 schematically summarises the three main aspects covered for judging relevance: the scientific basis (triangle or circle) of the alternative test method, that is the mechanism or property recapitulated or modelled by the method and thought to be causally related to an adverse effect in the target system (triangle in target system), the reliability and the accuracy (predictive capacity) of the measurements made in the alternative method with respect to the prediction of properties in the target system. Test methods (a)–(c) have a strong scientific basis since they model mechanism *p* (white triangle) that is either underlying or correlating with property *P* in that system: the predictive capacity shows to which extent the method is able to identify chemicals that activate *p* and which is thought to lead to *P* in the target system. Test methods (d) and (e) have a weaker scientific basis: they do not model mechanism *p* but another one *q*, indicated by a white circle. With regard to the overall relevance of the methods (a)–(e) the following can be said:

Method **(a)** is highly reliably (=always yields the same results) and scientifically relevant, but it is not accurate with respect to the predictions made: for

[2] Such strategic combinations have been proposed in the context of "Integrated Testing Strategies" that were proposed during the implementation of the REACH legislation in the EU (2006–2007) and consisted of steps of data gathering, evaluations and empirical (strategic) testing using several data sources. Later the concept of ITS has been further promoted under the term "Integrated Approaches to Assessment and Testing (IATA) by the OECD (OECD 2008).

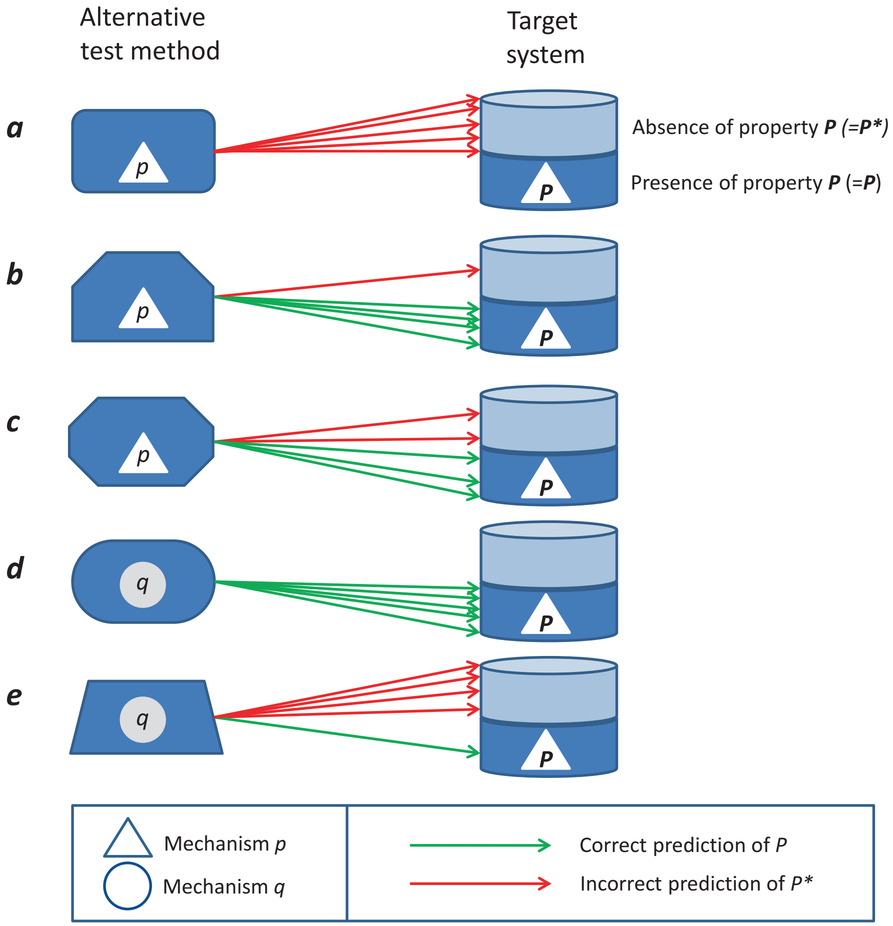
Alternative test method
Target system
a
b
c
d
e
p
q
P
Absence of property *P (=P*)*
Presence of property *P* (=*P*)
Mechanism *p*
Mechanism *q*
Correct prediction of *P*
Incorrect prediction of *P**

Fig. 4.6 Schematic representation of the main aspects impacting on the overall relevance of a test method, i.e. the meaningfulness and usefulness of its data. *Arrows* represent test results from five repeated experiments of the same test chemical. Correct predictions in *green*, incorrect predictions in *red*. The test method's purpose is to predict the presence of property **P** in the target system (e.g. a toxicity pathway). Reference data for the target system are available that have been simplified in two categories: chemicals that trigger P and others that do not trigger P. Thus, the alternative method needs to provide accurate predictions on absence (**P***) or presence (**P**) of property P. Some test methods (**a**–**c**) model the mechanism **p** thought to underlie property **P** (*white triangle*). Other test methods (**d**–**e**) do not model mechanism **p**, but **q**, which is not thought to be causative for **P**. Detailed explanations in the text

chemicals known to activate p, it predicts (*P**) = absence of property (*P*). These wrong predictions are indicated by red arrows. Its overall relevance therefore is very low. **Method (b)** has a strong scientific basis, is reliable and accurate. Its overall relevance is high. **Method (c)** is neither reliable nor accurate, although its scientific basis is relevant. Its overall relevance is low. **Method (d)** is reliable, but its results are more uncertain than those of method (b) since (d) does not model the mechanism of action *p* thought to be related to the occurrence of P in the target system. Thus, although (d) is accurate, its results correlate with rather than predict the adverse effect. **Method** **(e)** is reliable but inaccurate and has a week scientific basis. Its overall relevance is rather low.

(d) Applicability domain and limitations

An additional important aspect for judging the relevance of alternative test methods is applicability. Since test methods are used to assess chemicals, it is the applicability of a test method *to chemicals* that has been traditionally considered under the term "applicability domain". This would cover physicochemical properties, structural groups "chemical categories" or also sectorial use groups (e.g. biocides, pesticides, industrial chemicals, etc.) and such like. The applicability domain cannot be fully defined during validation but only be outlined based on the test chemicals used during validation. The wider the applicability domain, the more useful and hence more relevant is a method.

However, instead of restricting applicability domain only to aspects of chemical structure or physicochemical properties, it is useful to think of the applicability as a multidimensional space that is set up by as many descriptors as needed to describe how a method can be applied (Fig. 4.7). Notably, OECD guidance document 34 goes beyond the mere aspect of *chemical* applicability when defining applicability domain: "*a description of the physicochemical or other properties of the substances for which a test method is applicable for use*" (OECD 2005). Other properties (or descriptors) that may be useful for describing applicability are test outcomes (e.g. only applicable to positives), specific biological mechanisms of action/toxicity pathways.

It is obvious that 'applicability domain' in the above sense always refers to a positive description of what a method is applicable to. Inversely, the term "limitations" can be understood as a negative delineation of applicability, i.e. of "non-applicability". However, in practice, limitations more often relate to simple *technical limitations* and *exclusions* due to technical/procedural incompatibility of test items with a test method. Consider for instance a test methods based on measuring the cell viability using a colorimetric assay: test chemicals that are coloured may interfere with the readout and thus constitute a technical limitation due to incompatibility with the readout. Another example is the use of cells as a test system kept in submerged culture: this will result in a restriction to chemicals that can be dissolved in cell culture medium acting as a vehicle; the limitation would thus relate to insoluble substances such as some waxes or gels.

Thus, while applicability and limitation can be thought of as complementary terms, in reality, it is much easier to describe the limitations of a test method (especially technical limitations relating to compatibility with the test system) than to describe the applicability at the stage of validation. The reason is simply that during a validation exercise, for practical and economic reasons, only a limited number of test chemicals can be assessed: each chemical can be seen as probing with one single entity into the chemical universe composed of a vast space of hundreds of thousands of manufactured and natural chemicals. From each substance one can extrapolate to neighbouring substances within the chemical space (similar structure) or the biological space (similar mechanism of action).

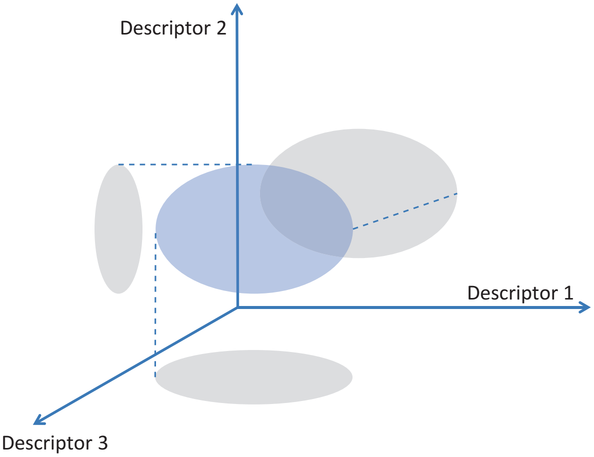

Fig. 4.7 The applicability domain of an alternative method can be seen as the space occupied by the method in a multidimensional coordinate space set up by various descriptors such as chemical structure, biological action, predictive parameters (applicable to negatives or positives only), etc. The space is indicated in *blue* and is a function of the relationships between the various descriptors

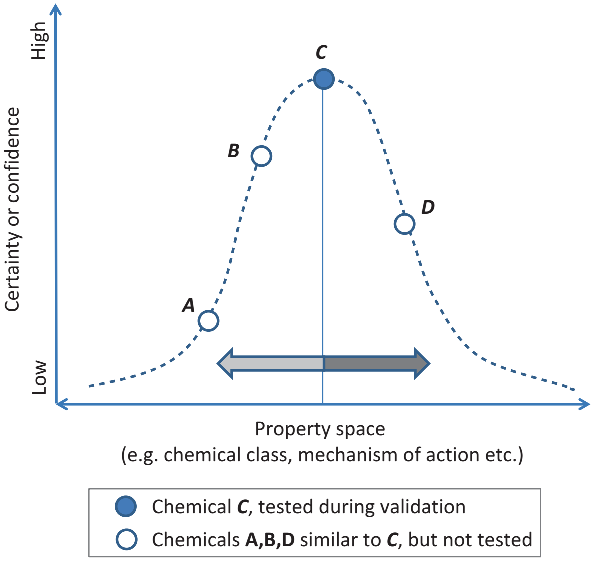
High
Certainty or confidence
Low
C
B
D
A
Property space
(e.g. chemical class, mechanism of action etc.)
● Chemical *C*, tested during validation
○ Chemicals **A,B,D** similar to *C*, but not tested

Fig. 4.8 For practical and economic reasons, validation studies can only empirically test a small sample of the chemical population. From these testing data, inferences can be made on substances with similar properties, e.g. relating to chemical structure or biological activity. Notably the certainty or confidence of these inferences decrease with increasing distance of these chemicals (*A*, *B*, *D*) from the chemical with empirical data (*C*)

The further one moves away from the substance with empirical data, the more uncertain this extrapolation gets (Fig. 4.8). It is clear that it is simply not feasible during a single scientific study to comprehensively delineate the entire space of applicability by testing, so extrapolation and "read across" of results will remain a key aspect of describing the applicability domain. To improve the description of applicability and limitations beyond the scope of validation studies, mechanisms of post-validation surveillance through which end users can report the successful use of test method to new substances as well as report problems, should be used in a more consistent manner and appropriate tools would need to be set up for such reporting.

Finally, since applicability can only be assessed or proven by testing or evaluating existing testing information, the certainty with which the applicability domain is determined is strongly correlated with the number of chemicals that has been assessed. Similarly, the certainty with respect to the predictive capacity is depending on the number of chemicals and minimum requirements in terms of sample size and power calculations for assessing for instance a dichotomous prediction model can be precisely calculated. However, for applicability and predictive capacity one could state that "the more chemicals, the better", i.e. increasing the chemical number will always increase the sharpness and accuracy with which both applicability and predictivity are defined and therefore increase the trust and confidence in the method.

In contrast, this "open-ended" approach regarding chemical number does not hold for reliability assessment: while there is a minimum number of substances statistically required for reliability assessment which can be calculated through statistical methods (sample size/power calculations), this number can be much lower than that required for a more robust description of predictive capacity and applicability domain. In contrast, to the assessment of applicability and predictive capacity, there is no substantial benefit in increasing the number of chemicals for reliability assessment. The different requirements regarding chemical number are schematically depicted in Fig. 4.9. These differences should be kept in mind when

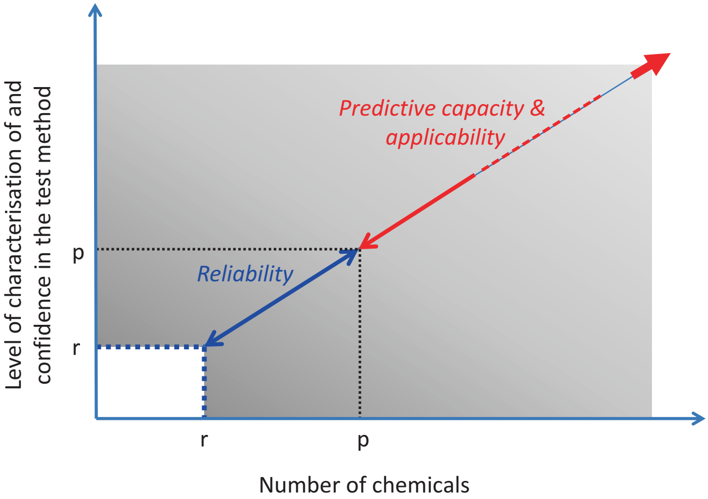
Level of characterisation of and confidence in the test method
p
r
Predictive capacity & applicability
Reliability
r
p
Number of chemicals

Fig. 4.9 The minimum requirements in terms of chemical number for assessing reliability on the one hand and predictive capacity and applicability on the other are different. There is a minimum number of chemicals that is required for reliability assessment in view of satisfying statistical needs (*blue dotted lines*, r). There is, however, no real need to go beyond a certain number of chemicals as defined by statistics to determine reliability since certainty will not increase to a substantial degree. In contrast, while there is also a minimum number of chemicals required for assessing predictive capacity (depending on the number of classifiers used) and applicability (p), the certainty with which these two can be considered characterised will always increase with increasing the numbers of chemicals assessed (*big arrowhead*)

discussing the potential adaptations to validation in terms of "lean processes" (see also Sect. 4.5.1): it is obvious from the above that the different requirements can be exploited in terms of adapting the data matrix generated during validation by dissecting the chemical testing set that has been traditionally assessed for all information requirements (reliability, predictive capacity and applicability domain) into two sets: one for assessing the reliability and a larger one for assessing predictive capacity. We will discuss this in more detail in Sect. 4.5.

2.4 Supporting the Practice of Validation: The Modular Approach

In 2004 EURL ECVAM proposed the "modular approach" to validation (Hartung et al. 2004) that has proven to be a very useful tool for adapting the validation study design not only to the intended purpose but also to the available information. Importantly, this modular approach should not be confused with the one proposed by Goldberg and colleagues in 1995 which relates to validation of *in vitro* methods on the basis of one defined readout against concurrent human data where possible (Goldberg et al. 1995).

Starting from the observation that validation until then had emphasized the process rather than the information requirements, the modular approach suggests to structure the information of scientific basis, within- and between laboratory reproducibility, transferability, predictive capacity and applicability domain into six information modules that need to be addressed during validation so as to allow a test methods to progress to scientific peer review. These modules have been termed (1) test definition (encapsulating aspects of scientific basis and mechanistic/biological relevance), (2) within laboratory reproducibility, (3) transferability, (4) between laboratory reproducibility, (5) predictive capacity and (6) applicability domain. In addition, realising that the definition of performance standards (see also Sect. 2.2) upon completion of validation studies would be helpful for test method development and validation of test methods, based on similar scientific and operational principles (="similar methods" or "me-too" methods), a seventh module of performance standards was added.

Most importantly however, the modular approach introduced a new philosophy towards the practical process of validation, allowing that these information modules be addressed in a flexible temporal order. Thus, test methods do not necessarily have to run through the typical ring-trial type evaluation of classical validation studies but need to address only the information that is judged to be missing. This information can then either be produced prospectively through dedicated new testing or retrospectively through analysis of existing information. For instance, for a specific test method, there may be ample information on predictive capacity, so that validation can focus on defining the test method and assessing mainly the reliability. EURL ECVAM has in recent years conducted several modular studies (see EURL ECVAM webpage, section "EURL ECVAM Recommendations", EURL ECVAM 2012 onwards), notably in the area of skin sensitisation. EURL ECVAM exploited the fact that, for some well-established methods (e.g. Direct Peptide Reactivity Assay, DPRA), there was a wealth of publicly available information on predictive capacity and applicability from user laboratories. This allowed to focus the design of the validation studies on protocol transferability and reliability (within and between laboratories) in order to complete these information modules.

3 Validation Study Management

3.1 Generic Design of a Validation Study

As outlined in Sect. 2.2.2, there are various types of validation studies in terms of the scientific design to assess reliability and relevance. Here we provide a brief outline on the managerial aspects of validation studies (Fig. 4.10).

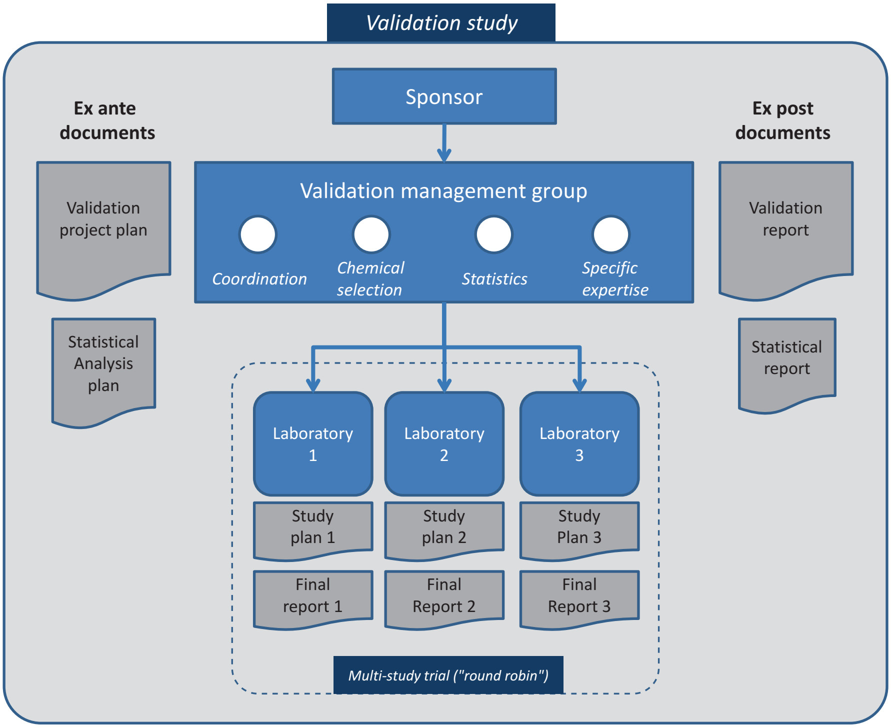

Fig. 4.10 Generic outline of the overall organisation of a prospective validation study: main actors, key documents produced at the outset (ex ante), during testing and upon completion of a study (ex post). Main actors are (1) the sponsor or sponsor consortium, initiating and normally financing the study, (2) the validation management group that is set up by the sponsor in view of managing the science and logistics of the study and composed of experts with different roles and expertise including coordinators, statisticians, chemists and regulators for selecting chemicals and other experts (e.g. in validation, the test method under scrutiny, etc.), (3) the participating laboratories conducting the testing within a round robin or ‘multi study trial’. In case of retrospective studies, the design would be the same, without however the participating labs

3.2 Roles and Responsibilities of Actors

Validation studies are typically initiated by a sponsor or sponsor consortium. The sponsor has an interest in validating the method either because of economic interests and/or in view of legislative requirements necessitating a particular validated alternative method for routine use. The sponsor typically appoints a validation management group to oversee the entire study, i.e. to decide on study design, to manage and coordinate the study execution phase (involving dedicated chemicals testing in case of prospective studies, to analysing the results and concluding on and reporting the main outcomes by writing up the final validation report. The validation management group is composed in view of gathering the expertise needed to conduct the specific study in question. This includes (i) a Chair who is moderating meetings, teleconferences as well as discussions and the decision-making process related to all VMG decisions; (ii) experts with knowledge in the test method under scrutiny and related scientific and regulatory requirements; (iii) statistician(s) that are responsible for suggesting important aspects of the validation study design (e.g. sample size and power calculation) and data analysis; (iv) study coordinator(s) who act as a central secretariat, i.e. ensuring the efficient management and conduct of the overall study (maintaining efficient communication, preparing drafts of key validation study documents, organising meetings, recording key decisions and reports of meetings and teleconferences). Depending on study, the coordinator(s) may or may not participate in the decision making of the group. Finally, among these experts, some can be appointed to define and perform the chemicals selection: identifying and procuring suitable chemicals addressing pre-defined criteria including, importantly, high quality of associated reference data. Importantly, the validation management group, via the coordinator(s), closely interacts with the work of the participating laboratories, each conducting one dedicated laboratory study. The ring trial hence is also referred to as “multi study trial” (see Chap. 5).

The key documents to be defined at the outset of the study are:

- The validation project plan which can be seen as the major blue-print or roadmap of a study. The validation project plan outlines the goal and objectives of the study and defines the test method in sufficient detail. The document determines the SOP versions that must be used during testing and lays out in sufficient detail the relevant scientific, managerial and logistical steps in view of conducting the study (see Sect. 4.4 for more details). This includes aspects relating to data analysis, handling problems and deviations. It includes contributions from specific experts of the management group, e.g. from the chemical selection committee which will outline the test chemicals to be studied and their associated reference data or from the statistician, describing the sample size calculations conducted in view of addressing the study goal and objectives).
- **The statistical analysis plan**, outlining the data handling, analysis, interpretation and reporting. This plan can be part of the project plan or a stand-alone document.

Key documents during the validation study are:

- The laboratory **study plans** and **final reports** (requirements under GLP) that outline all the relevant SOPs required (not only that of the test method, but also those relating to equipment and other issues of the local laboratory) and that define how the testing data will be reported in agreement with the quality assurance measures in operation at the laboratory.

Key documents upon completion of a validation study are:

- The **statistical report** summarising the analysis of the data and the statistical findings. This report can be a stand-alone document or be part of the validation report. Important is that the statistical analysis and its conclusions are not influenced by the VMG (who may be biased with respect to the decisions it took during the study) and is conducted solely on the basis of the data available.
- The **validation report** that summarising the entire validation study (referring where necessary to other documents, e.g. the statistical report), the problems encountered and which has to clearly outline results obtained, the conclusions drawn and take clear position with respect to whether or not the study goal has been achieved.

4 Validation Study Design

Having discussed the key actors, the key documents and the generic organisation of a validation study in Sect. 3, we now explore the most important elements to be addressed during validation study design. These typically would be captured in a validation project plan (see Fig. 4.9).

4.1 Number of Chemicals, Sample Size and Power Considerations

Conclusions drawn on the basis of empirical testing can be considered solid scientific insight only if they can be generalised beyond the *single experimental result*. The assessment of the capacity of an alternative test method in view of obtaining predictions on the effects of chemicals cannot be done on an infinite number of chemicals, but, for practical and economic reasons, on the basis of a restricted number. This should however be sufficient to allow such generalisations, taking also into account the restricted reproducibility of scientific experimentation. Thus, empirical testing will be restricted to a *sample* of the population (chemical substances). In the following we discuss this ‘sample size’ problem, that is, the problem of concluding from the relative frequency of events *in a sample* to the relative frequency *in the* *entire population*. We equate here sample size with number of chemicals since the goal of validation is to make inferences on the ability of a test method to predict the properties of chemicals. It is however noted that the term may also reflect the sample size of two or more distinct populations or simply to the number of observations or replicates.

The number of chemicals used for the validation study needs to be determined by statistical means so as to allow adequate quantitative metrics in view of the validation study goal and objectives. The quantitative metrics relate to mainly the within-laboratory and between-laboratory reproducibility (WLR and BLR) and predictive capacity; for the latter the number of categories predicted (dichotomous/binary or more; see Fig. 4.1) will be an important factor influencing the sample size/power calculations.

The sample size, here the number of chemicals, should be large enough to represent sufficient statistical power for comparing two (or more) populations by a statistical test on the basis of a measured parameter; the latter can be a *mean* or a *proportion*. Two types of errors can be encountered, type 1 and type 2. Both types are taken into consideration for the sample size calculation:

- The type 1-error is the error that consists in rejecting the null hypothesis H_0 of equality of the parameter when H_0 is true. It represents therefore the false positive cases. The probability that this type of error occurs is usually denoted by α.
- The type 2-error is the error that consists in not rejecting the null hypothesis H_0, i.e. accepting H_0, when H_0 is false. This type 2-error represents therefore the false negative prediction. The probability that this type of error occurs is usually denoted by β. The power of the statistical comparison is defined by $1-\beta$.

In the case of *in vitro* test methods, predictions typically consist of categorical outcomes relating to specific mechanisms (e.g. activating estrogen receptors) or entire health outcomes (e.g. in Skin Corrosion Tests, Category 1A, Category 1B/1C, and Non-Corrosive). The value of WLR is typically obtained by calculating the proportion (i.e., fraction in percentages) of chemicals that have concordant predictions throughout the runs used in one laboratory. The test chemicals represent the population for which the calculation of the sample size is required. This WLR is the measured parameter over the population of chemicals. For defining the sample required, the expected values (target value, here relating to WLR) is an important aspect to be defined prior to testing. The target value should be based on prior testing of a small set of chemicals (e.g. in the context of a so-called "prevalidation" study) or can be derived from other historical information. The formula to be used, for calculating the sample size, is the one based on proportions and will include this target values as well as α and 1- β values.

The following equation shows the advantage of simultaneously taking into account the targeted WLR value and the lower limit of this value (i.e. WLR should not go below this value). The target value is represented by π, the error by δ, the lower limit by π-δ.

Z_α and Z_β respectively represent the Z distribution values for the probabilities α and β. This formula was proposed by Flahault et al. (2005) and can be derived by the one presented by Lachin (1981).

$$Number\ of\ chemicals = \frac{\left(Z_\alpha\sqrt{(\pi-\delta)(1-\pi+\delta)}+Z_\beta\sqrt{\pi(1-\pi)}\right)^2}{\delta^2}$$

Such calculation of the number of chemicals needed, prior to testing, plays an important role in the validation study as this sets up the level of confidence—and conversely deals with the uncertainty towards the obtained values of WLR.

Statistical considerations also apply to the calculation of the number of chemicals needed to reach target values of BLR. Similarly to WLR, BLR is a proportion—the fraction of chemicals for which concordant predictions have been made over the participating laboratories. The difference δ accepted for the target value π plays a critical role in the formula: when δ decreases, n increases according to the inverse of δ square root. For instance, a target value of WLR of 90 %, i.e. $\pi=0.9$ with a power of 80 %, i.e. $1-\beta=0.8$ ($Z_\beta=0.842$), and a risk $\alpha=0.05$ ($Z_\alpha=1.645$) will result in a different sample size whether the value of δ is 0.1 or 0.2. If $\delta=0.2$ the total number of chemicals needed is 25; if $\delta=0.1$ the total number of chemicals needed is 83 and therefore much higher.

Therefore, the assumptions (or the certainty of preliminary target values) play a critical role for calculating the number of chemicals to be assessed in a validation study. These assumptions cover not only the target values of WLR or BLR, but also the underlying statistical formulae used for the calculation (normal approximation to the binomial law).

4.2 Selection of Test Chemicals and Associated Reference Data

For above said reasons, the selection of chemicals used in validation studies is critical and the success or failure of a validation study may largely depend on it. This includes issues of both *number* and *nature* of chemicals selected. Ideally, a high number of chemicals should be selected to represent different chemical classes and, depending on the purpose of the validation study, also different chemical use categories, such as e.g. industrial chemicals, food additives, pharmaceuticals, cosmetic ingredients, pesticides, etc. Ideally, the following information on the selected chemicals should be known and compiled: use applications, *in vivo* data sources, substance supply, chemical classes, physical chemical properties and GHS classifications (if applicable).

Chemical selection has traditionally focused on mono-constituent substances of high purity, ensuring correspondence of documented *in vivo* data to sample material acquired for *in vitro* testing. Nevertheless, acknowledging the REACH definitions, 'pure mixtures' (multi-constituent substances with negligible impurities) have also been admitted, provided composition was reported quantitatively and consistent with the material used for the *in vivo* study.

In general, a principal requirement for chemical selection is the availability of complete and quality assured supporting *reference* data sets, for comparative evaluation of *in vitro* mechanistic relevance and/or method predictive capacity. These reference data are typically from surrogate animal studies ("*in vivo* data"), but can also be derived from other sources. In areas where the mechanisms of action is not preserved across species, (e.g. metabolism, CYP induction), the availability of human reference data for the mechanism studied is essential. Human toxicity data however are often problematic with respect to their availability and their quality (see Sect. 2.1).

The availability of human reference data for many areas in toxicokinetics and toxicodynamics is often limited to pharmaceuticals since this is the only sector where testing is performed in humans after pre-clinical toxicological testing. Human data from the pharmaceutical and other sectors can also be obtained from selected scientific references and poison control centres. In such registries, human data derived from clinical case studies, hospital admissions, and emergency department visits can be found. Although this information is not acquired systematically, it represents a potential source of human toxicity and toxicokinetic data available for commonly encountered chemicals. Thus, the clinical information is used as a basis for comparison with *in vitro* values.

Another source of more reliable human toxicological data may be obtained through the testing on human volunteers for some areas of local toxicity, such as skin and eye irritation. Human volunteers for skin irritation testing produce concentration-effect curves for fixed endpoints, while in the case of eye irritation, testing is, for ethical reasons, limited to minimal mild effects (redness, itchiness). A more recent technology to obtain human data is the human microdosing. This technology seems promising for obtaining human toxicity data as only extremely low amounts of chemical need to be given to the human volunteers. These external amounts could well remain below current threshold of toxicological concern (TTC) values. However this area needs to be further explored and it is stressed that all experiments with human volunteers need to be carefully considered for their ethical implications before being conducted.

In general, the selected chemicals should be (1) commercially available, (2) stable after fresh preparation of a stock solution, (3) soluble in saline, or in solvents that are used in concentrations not affecting the mechanism of interest and (4) not precipitate for defined time frames when used under standard operating procedures.

Experience has shown that all laboratories should use the same solvent and the same non-cytotoxic highest concentration of the test item over a defined period as defined in the standard operating procedure during a validation study.

Another prerequisite is to use defined chemicals (that is by their Chemical Abstracts Service (CAS) formulas or their generic names) rather than proprietary mixtures or coded industrial products. Studies performed with defined chemicals allow for between-laboratory testing and clear definition of critical components of the validation study.

4.3 Defining the Data Matrix Required

Once the sample size has been determined, it is advisable to determine the precise data matrix that would be required for a statistically appropriate analysis of the performance characteristics of the test method during validation. By data matrix we simply mean the number of data points required for each test chemical in view of characterising the performance of the method. Aspects of lean design (see Sect. 4.5.1) can be taken into account when defining the data matrix.

A typical example of a data matrix can be defined by

- X number of laboratories testing…
- Y number of chemicals in …
- Z number of experiments

The terms "experiment" and "run" or "complete run" are sometimes used interchangeably. Importantly, these terms usually relate to all the measurements and data processing/analysis required to generate a *final result* for a given test item (either a toxicological measure or a categorical prediction). Thus, runs or experiments relate to the intended application of the test method in practice when routinely assessing test items.

In the following section we present some illustrative examples from test methods in the area of topical toxicity testing (mainly skin irritation) as summarised in a background document on *in vitro* skin irritation testing (Griesinger et al. 2009), see Fig. 4.11:

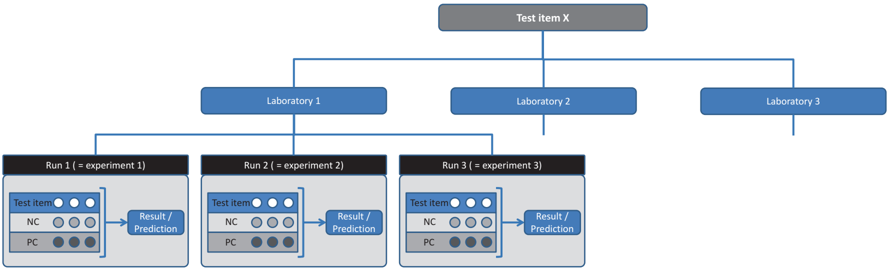

Fig. 4.11 Schematic depiction of a possible data matrix for one given test item (X) in the context of a validation study. The example is based on *in vitro* skin irritation testing. Each test item is tested in three laboratories. In each laboratory, three experiments (=runs) are being conducted. A run is the experiment that will yield the final result of the test method as intended in practice, i.e. either a final toxicological measure or a categorical prediction. Thus, a run incorporates all steps necessary to produce this information and thus includes the testing of the test item, the controls, as well as all data analysis as required. This can include conversion of the result into categorical predictions by means of a prediction model. The three runs conducted in each laboratory can be used to assess the within-laboratory reproducibility (e.g. by assessing concordance of run predictions). Runs are based on several replicate measurements (circles) whose results normally are being averaged and analysed for variability as a measure for the quality of the data underlying the experiment or run. Variability measures such as Standard Deviation (SD) or Coefficient of Variation (CV) can be used to define "Test Acceptance Criteria", i.e. quality criteria for accepting or rejecting an experiment based on replicate measures

4.3.1 Number of Replicates

The replicates are the repeated individual measurements of the parameter of interest, for a given test chemical, and together with other relevant measurements (e.g. controls) constitute the data underlying a run, i.e. the actual result of the test method when used in practice. Each replicate measures the parameter of interest (Griesinger et al. 2009). Replicate measurements can be used to calculate mean and standard deviation (SD) values. The SD value can be used to further calculate the coefficient of variation (CV) defined, in percentages, by $CV = (SD/Mean) \times 100$. SD and CV are quantitative indicators of variability. Measures from all replicates are usually averaged to derive the final result or prediction for the test item tested. Importantly, the use of replicate measurements allows assessing the quality of the experiment: the variability of these replicate measurements should be below a pre-defined threshold (e.g. a SD value), otherwise the run result is considered invalid (or "non-qualified"). The SD thus serves as a tool for applying a "Test Acceptance Criterion) (TAC)". In the example of skin irritation testing, the SD derived from three tissue replicates must be equal or below 18 %. Importantly, the TAC must be set on the basis of a sufficiently large set of historical testing data and the number of replicates required to assess within-experimental variability should also be based on sufficient previous data. Typically, during a validation study, the number of "replicates" will follow the provisions of the test method protocol intended for later application. However, when defining the validation data matrix, it should be carefully assessed whether the number of replicates can be reduced ("lean design"), e.g. by analysing historical data sets and assessing the impact of such reduction. Importantly, the number of replicates is specific to each test method and, unlike for the number of runs or laboratories, no general recommendations can be made.

4.3.2 Number of Runs (Experiments)

A run is the actual experiment that provides a final result on a given test item. A run (or experiment) thus consist of (1) testing the test item itself and, concurrently, all necessary controls (e.g. positive control, negative control) (Griesinger et al. 2009) and (2) performing all necessary data processing and analysis steps to generate a final results for the test item. This may, where applicable, include the conversion of the toxicological result into categorical predictions by means of a prediction model.

In a validation study, typically three runs (or experiments) are performed in each laboratory. Since each run provides a final prediction, the between run-concordance (=agreement between) such predictions can be used to assess the within a laboratory repeatability and within-laboratory reproducibility of the test method.

Predictions at run level may also be used for deriving a final prediction per chemical in one laboratory. This has typically been done by simply determining the "mode" of predictions and settling unequivocally on a final prediction per chemical. If this approach is used, the number of runs needs to be an odd number (e.g. three runs).

4.3.3 Number of Laboratories

For the same considerations as described above for the number of runs, three laboratories are usually participating in a validation study. The involvement of several laboratories allows evaluating the reproducibility of the test method between laboratories. The between-laboratory reproducibility can be calculated as described in Sect. 4.7.2.

4.4 Validation Project Plan

The validation project plan serves as a driver and a reference for the conduct of the validation study. It covers an extensive range of topics relevant for conscientiously planning the scientific and managerial aspects of the validation study. It takes into account logistical and practical considerations and sets up timelines. The project plan defines the test methods under validation, the goal and objectives of the study, it describes the actors involved and their respective roles and responsibilities, and defines specific stages/timelines of the study.

A typical project plan can include the following main sections:

1. Definitions: this section provides definitions of the test methods studied during validation, outlining (1) the test systems (e.g. reconstructed human tissue of multi-layered epithelium) used as well as (2) determining the associated protocols/SOPs and the precise version numbers to be used during the study.
2. Validation study goal and objectives: goal and objectives of the study should be clearly outlined. Typically the goal of a study corresponds to a regulatory requirement and often to the prediction of specific hazard classes or categories of chemicals (e.g. Category 2 of eye irritant in the United Nations Global Harmonized Systems for classification and labelling, UN GHS). Therefore, this section should explicitly mention the name of the regulation addressed. If several regulations are concerned it should be specified how the study will relate to these. The objectives would be more detailed aims, such as validation for identification of negatives or for a specific class of chemicals in view of filling an existing methodological gap, etc.
3. *In vitro* test methods: this section provides a detailed scientific characterisation of the *in vitro* test methods undergoing validation. This relates to the scientific basis, the test method's mechanistic and biological relevance, as well as historical aspects relating to test method development (test method development, optimisation, previous assessments including prevalidation studies, etc.).
4. Validation management group (VMG): the VMG is the body that oversees and manages the validation study (see Sect. 3.2). The validation project plan should outline the expertise required in view of ensuring an efficient conduct of the study. Typically a VMG consists of (i) a Chair responsible for chairing meetings, facilitating decision making and representing the VMG; (ii) relevant

experts with specific expertise required for the study; (iii) statistician(s); (iv) study coordinator(s) acting as focal contact point and running the study secretariat. Moreover, depending on study, a VMG subgroup dedicated to selection of test items and associated reference data, Moreover, observers or liaisons may participate (e.g. representing other validation bodies). Also, representatives of the laboratories can be involved for specific agenda items of VMG meetings related to technical and/or experimental issues. The specific role of each of the above mentioned categories of participants and the way they interact together should be clearly explained and may be supported by a schematic figure. In order to maintain an impartial and unbiased study, the VMG must not include members directly involved in the development of the methods undergoing the validation process. However, the VMG may consult the test method developer if necessary.

5. Validation study coordination and sponsorship: this part of the validation project plan defines sponsors of the study as well as the activities that should be covered by the study coordinators, including logistical aspects (e.g. coding and distribution of chemicals), communication (e.g. frequency, means), organisation of VMG meetings, teleconferences, minutes, etc. This section should also describe the allocation of financial resources, e.g. purchasing of test chemicals and other relevant service contracts (e.g. statistical support).
6. Chemicals selection: The process and criteria for selecting test chemicals should be detailed in this section. Chemical selection can be done by *ad-hoc* experts or by a dedicated VMG chemical selection group (CSG). Experts can include members of the validation study coordination, independent scientists, liaisons and representatives of the competent authorities. Moreover, since *in vitro* methods will be evaluated against reference data, this section should also stipulate criteria for the selection of such data associate with the test chemicals. To this end, the type of reference data and the sources of these data (e.g. databanks, literature, etc.) are specified. Eligible chemicals are usually compiled in table format (e.g. classification of selected chemicals according to the UN GHS for skin corrosion). Number of chemicals needed for the validation study, obtained from sample size calculation (see paragraph 4.1), will be mentioned as well as the proportions of distinct classes/categories (e.g. negative vs positive, solids vs liquids, etc.). In terms of procedure, the CSG proposes the list of eligible chemicals to the VMG. This latter may also take into account the availability of the chemicals to be tested, especially those commercially available versus the proprietary ones as well as other practical factors such as potential health effects of test chemicals: since validation studies are conducted under blind conditions, substances with specifically high risks can be excluded (e.g. "CMR substances" with carcinogenic, mutagenic and reproductive toxicity effects) as long as these are not related to the health effect of concern to the study.
7. Chemical acquisition, coding and distribution: This section should outline the provisions regarding acquisition, coding and distribution of the test chemicals. This should be accomplished by a person affiliated to a certified ISO 9001/GLP

structure. Individuals involved in this process must be independent from those conducting the testing. The process should foresee a purity analysis of the chemicals and the provision expiry dates. In laboratories testing different versions of one protocol (e.g. separate protocols for testing solid and liquid chemicals), codes of chemicals will be different for each version.

8. Receipt and handling of chemicals: this part of the validation project plan tackles the shipping of the coded chemicals, the storage time and conditions as well as health and safety measures related to their handling.
9. Participating laboratories: This section should outline the requirements of the participating laboratory, e.g. study director, quality assurance officer/unit, study personnel and a safety officer. This section also includes a description of how laboratories, within a group, may communicate together and when the VMG should be involved in these discussions. For instance, during the testing phase, the participating laboratories must not contact each other without approval of the VMG.
10. Laboratory staff: the validation project plan specifies the roles of the study directors, the quality assurance officers/unit, the study personnel and the safety officers. The study director should be an experienced scientist in the field and acts as main contact point of the VMG. He/she is responsible for preparing each necessary report. The quality assurance officers will assure that compliance with any quality requirements (e.g. GLP) is respected. The quality officer needs to be independent from the study director direction and from the study personnel conducting the experiments. The experimental team will perform the testing. It should be trained, experienced and competent for the specific techniques. The safety officer is in charge of receiving the coded chemicals and transmitting them to the responsible person of the laboratory. He/she is in charge of the sealed material data sheets (MSDs) corresponding to the test chemicals and their codes. These will be disclosed only in case of accident.
11. Validation study design: this section of the project plan includes details on each type of assay taking part in the validation study. For instance, number of chemicals, runs and replicates should be clearly defined. Specific technical aspects of the test methods are tackled. For instance, if there are two different protocols for a given test method with different exposure times, those will be mentioned.
12. Data collection, handling and analysis: this part of the validation project plan describes how final reports and the reported data are forwarded to the biostatistician. He/she will decode the chemicals and proceed to the analysis (see paragraph 4.6, Statistical analysis plan) and produce a biostatistical report to the VMG. This report should present the results (predictive capacity, within- and between laboratory reproducibility, quality criteria) as well as how data were analysed and the statistical tools used. Data analysis strategy should be developed, before the end of the experimental phase, by the biostatistician in a statistical analyses and reporting plan. This latter will be submitted to the VMG for approval.

13. Quality assurance good laboratory practices: it is usually desirable that the validation study complies with OECD good laboratory practices (GLP) in order to facilitate international acceptance of the validation study and its outcomes (OECD GD 34 2005). This allows full traceability of the study at all levels of its experimental phases.
14. Health and safety: the laboratories should comply with applicable (and required) health and safety statutes. The safety officer of each laboratory is designated as the contact point for these questions.
15. Records and archives: provisions should be made for the appropriate archiving of raw data, interim and final reports of the validation study (where, how many copies, by which means) as well as for the management of the archiving.
16. Timelines: defines the critical timelines that should be respected. Timelines are established for each critical phase of the validation study (e.g. chemical eligibility, approval of the validation project plan, approval of the validation study design, dates of testing, etc.).
17. Documents and data fate: proprietary questions in relation with the documents and data generated are described. This also covers the confidentiality of these elements and whether and to which extent information can be disclosed.

Finally the validation project plan should also make provisions for retesting in case of non-qualified (invalid) runs so that this can be implemented in the study plans for the laboratories under supervision of the individual study directors. In particular this should address how often experiments relating to one chemical can be repeated, i.e. how many retesting runs are permissible. Typically, the validation coordinator prepares an example of a study plan that can be adapted by the laboratories in compliance with their own specific laboratory procedures (see Chap. 5).

4.5 Adaptations of Validation Processes

The modular approach (Sect. 2.4) can be regarded as an important adaptation of the classical validation approach. Traditionally information on reliability and the judgement of relevance followed a rather rigid sequence towards producing a comprehensive data matrix. The modular approach introduced a significant degree of flexibility with regard to the generation of the information. Two further adaptations have been under discussion recently namely approaches to reduce the data matrix without compromising the adequacy of the validation study ("lean design") and, secondly, the use of automated equipment (e.g. automated platforms, medium- and high-throughput platforms) for generating empirical testing data. Third, some methods used for prioritisation have been developed on custom-made automated platforms and some aspects of validation cannot be always applied to such assays (e.g. transferability assessment). These three adaptations are briefly discussed below.

4.5.1 Lean Design of Validation Studies

As discussed in Sect. 2.3.3(d), the requirements in terms of sample size for assessing reliability and for assessing predictive capacity and applicability domain are different. This can potentially be used in view of adapting the data matrix in order to reduce both cost for test chemicals, test systems and the labour involved. As a general consideration, it is conceivable to assess the reliability of a test method using a small set but statistically sufficient set of chemicals in three laboratories, while assessing the predictive capacity (e.g. in terms of a dichotomous prediction model requiring a higher sample size) with more chemicals but only in one laboratory or by testing subsets of this larger set in various laboratories. A feasibility study of this approach has been conducted by Hoffmann and Hartung (2006a, b) using the data set of the EURL ECVAM skin corrosion validation study (Barratt et al. 1998; Fentem et al. 1998). Using resampling techniques it was shown that the number of test runs could be reduced by up to 60 % without compromising significantly the level of confidence. While this result is promising it should be noted that the reproducibility of these methods was very high and this has probably led to the remarkable reduction rates of the data matrix that were possible. It still needs to be evaluated to which extent the lean design can also be useful for other test methods and other use scenarios in particular.

4.5.2 Automated Testing as a Data Generation Tool for Validation

Validation studies normally assess test methods on the basis of manually executed SOPs. This ensures that validated test methods and their associated protocols are universally usable, also by laboratories that do not have automated platforms at their disposal. This however does not mean that automated methodology (e.g. relating to liquid handling steps in a manual method) could not be used during validation. Automated or robotic platforms can greatly accelerate the generation of testing data and allow the economical testing of a larger numbers of test items in shorter a time. This supports a more complete characterisation of the predictive capacity and applicability (see Sect. 2.3.3) of a test method (Bouhifd et al. 2012). An important prerequisite to use automated approaches for validation is to ensure that the automated protocol is equivalent to the manual one in terms of the results and/or predictions it generates. There may be variations that need to be assessed with attention (e.g. smaller exposure volumes, slightly different application regimes with regard to the test chemicals etc). When used for additional data generation during validation, automated testing represents rather a technical than a conceptual adaptation of the validation process.

4.5.3 High-Throughput Assays for Chemicals Prioritisation

In the context of alternative *in vitro* testing methods, high-throughput assays (HTAs) are those using automated protocols to test large chemical libraries over a range of concentrations. Chemical prioritization is often the objective when using HTAs which aims to identify those chemicals in large libraries that may exert a specific mechanism of action with the potential to lead to particular adverse effects. While these HTAs are not intended for global use by end users (e.g., via OECD test guidelines), data generated via HTAs may be used by regional agencies and international organizations to inform regulatory decision-making, especially as part of a weight-of-evidence approach. Consequently, it is important to consider whether adaptations of standard validation approaches may be appropriate for use with HTAs.

The principals of validation outlined in Sect. 2 are applicable to all alternative methods, including HTAs. However, the unique nature of the automated assays and the resulting volume of data generated using HTAs differ significantly from traditional "manual" methods, and these aspects need to be taken into account during the validation process.

Most HTAs are performed using highly automated processes developed on custom-built robotic platforms and are therefore not amenable to traditional "ring-trial" studies used to demonstrate transferability of the method. Transferability, one of the assessments of reliability along with inter-laboratory repeatability, is important because (i) it provides independent verification of results obtained using the same method in another laboratory and (ii) it allows a statistical assessment of between laboratory reproducibility (BLR, see Sect. 4.7) that can be used in an overall assessment of how robust the protocol is when used in different laboratories. The statistical characterization of method transfer is generally not germane to HTAs due to the highly customized and unique nature of these assays, Judson et al. (2013). However, the ability to confirm independently the results of the HTAs remains an extremely important aspect of method validation and deserves careful consideration. Since many HTAs are adapted from previously existing low-throughput methods (i.e. manual protocols), the most straightforward approach to confirm results from HTAs is *via* use of performance standards developed for mechanistically and procedurally similar assays (see Sect. 2), the latter without regard of the equipment used to execute specific procedures (i.e. protocol steps), i.e. manual or automated.

In the event that the HTA is measuring a unique event or utilizing a proprietary technology, data generated in other assays measuring activity in the same biological pathway may be useful in confirming or at least supporting results of the HTA assay undergoing validation. If a number of chemicals produce consistent results across several different key events in a given biological pathway, then the activity of those chemicals may be able to serve as a reference for other (new) assays that target key events in the same pathway. For example, if the HTA undergoing validation measures one key event in a signaling pathway (estrogen receptor dimerization, for example), then data generated in other assays measuring different key events in the same pathway (e.g., ligand binding, DNA binding, mRNA production, protein production, cellular proliferation) may potentially be used to establish confidence in the HTA data.

Another critical aspect to consider when validating HTAs is the volume of data generated by these methods, which necessitates increased reliance on laboratory information management systems (LIMS) and automated algorithms for data analysis. Although data management and statistical analysis (see Sect. 4.7) are important components of all validation studies, the large amount of data associated with HTAs often results in analysts being "disconnected" from the data, which has the potential to lead to wide-scale misinterpretation of the results. With this in mind, the validation of data management tools and the statistical approaches employed become paramount.

4.6 Ex Ante *Criteria for Test Method Performance*

Clear criteria relating to desired or expected performance defined at the outset of validation (before data generation) can support an objective evaluation of the results and conclusions of a validation study and in particular to which extent its goals have been met. These criteria can be fixed values or ranges relating to specificity, sensitivity and within- and between-laboratory reproducibility. They should be based on reliable empirical data from prevalidation or derived from other relevant data sets such as in-house (non-blinded) testing in the test developer's laboratory. Importantly, the performance criteria should relate to the intended purpose of the test method, i.e. its practical application, e.g. whether the test will be used in pre-regulatory screening or for the generation of data for regulatory dossiers in response to legislative requirements (Green 1993). Moreover, the use scenario is a key factor to be considered: for instance, will the method be a stand-alone or be merely part of an integrative approach? Ex ante performance criteria have been used by EURL ECVAM when validating *in vitro* skin corrosion methods (Fentem et al. 1998), using ranges of sensitivity and specificity that were subdivided in bands of acceptability. This approach was recently used again by EURL ECVAM when validating *in vitro* methods for eye irritation testing (EURL ECVAM 2014).

4.7 Statistical Analysis Plan

The statistical analysis plan includes a series of calculations that aim to demonstrate two main features of the test method to be validated. The first one deals with the reliability of the method and covers two main parameters: the within-laboratory reproducibility and the between-laboratory reproducibility. This second feature is the predictive capacity of the method. Below we outline the basic statistical approaches that can be used to describe these. Most of the relevant literature to describe predictive capacity deals with evaluations of diagnostic tests during clinical trials (i.e. versus a gold standard test). Most of the concepts and tools can be applied also to predictive toxicity tests, although there are important differences with regard to the entities tested and the nature of predictions obtained (see Sect. 2.1.4). An overview of statistical evaluations of test methods can be found in Pepe (2003).

4.7.1 Statistical Evaluation of the Information Provided by Alternative Test Methods

Fundamental Considerations

Two basic groups of test methods can be distinguished with regard to the results they provide: Test methods that provide meaningful toxicological information without transforming these into categorical predictions and those that convert measurements into distinct categorical predictions by means of a prediction model.

(1) Results are measures of some sort but no categorical predictions: Examples include assays that provide *in vitro* concentration-response curves and thus information about *in-vitro* potency. Generally, ecotoxicological test methods provide results that are not in form of categorical predictions. An example is the Fish Embryo Toxicity Test (FET) which yields an LC_{50} value (concentration that leads in 50 % of the animals in the observation group to lethality).

(2) Results are categorical predictions: The final measurements are converted into categorical predictions. These, in most cases, are dichotomous (or binary) predictions of the general form "toxic" versus "non-toxic". Test methods used for hazard identification in relation to categorical systems such as the United Nations Globally Harmonised System (UN GHS) for classification and labelling (C&L) of chemicals will need to produce categorical predictions to be useful in practice. The categories in this case relate to downstream ("apical") health effects such as skin corrosion, acute oral toxicity, etc. However, categorical predictions do not necessarily need to be tied to C&L classes or apical health effects. Categories can in principle relate to events at any level of biological organisation (e.g. activation of a given pathway). When considering and using categorical information from any toxicological test method (irrespective of whether it is a traditional animal test or an alternative method) one should keep in mind that the distinct categories (as defined for purposes of C&L) have been set as an arbitrary convention to simplify risk management and transport of chemicals. Unlike other testable properties that may come in two classes (e.g. absence or presence of a disease marker), toxicity and hazard are continuous events and categorical differences do not exist in reality. This is especially important when considering data close to the cut-off of a prediction model (see Fig. 4.13, Sect. 4.7.2). Chemicals close to the cut-off can lead to an apparent high variability (or low reproducibility) of the test system and impact on the predictive capacity. It can be useful to consider such data close to the cut-off as "inconclusive" results which need to be further processed by expert judgement (i.e. ascribing one of the two categories). This judgement can be aided by additional statistical measures (e.g. Confidence Intervals) and/or other sources of toxicological information (read-across, QSAR, etc.).

In this chapter we will focus on statistical measures of predictive capacity of categorical predictions. Statistical analyses of the results from non-categorical methods need to be defined on a case-by-case basis. To return to the example of the Fish Embryo Toxicity test: in this case the predictive relationship between

LC_{50} values of embryonic fish and LC_{50} values from adult and juvenile fish was assessed by means of orthogonal regression (Belanger et al. 2013) providing information on slope, intercept and range of concentrations over which the correlation held.

Predictive Capacity (PC)

The predictive capacity of tests that provide categorical predictions informs about test method performance in terms of correct and incorrect predictions in comparison to pre-selected reference data that are considered "true" and referred to a "actual positives" and "actual negatives". These data can be derived from the species of interest (Goldberg et al. 1995) or from other reference methods, typically surrogate animal methods. The latter has been, for reasons outlined in Sect. 2.1.4, typical practice during validation. The predictive capacity gives quantitative information on test method performance in terms of translating the actual measurements obtained (e.g. cell viability, quantified gene expression) into predictions of a defined effect (e.g. a pathway or a downstream health effect). The predictive capacity therefore reflects the final outcomes of the test method when applied as intended.

The predictive capacity serves as a tool for policy makers and regulators to evaluate to which extent the test method considered is likely to accurately predict the biological effect(s) of interest. Based on the predictive capacity and duly considering its intended use scenario, regulators can decide whether or not a given method is ready to be implemented in regulation as a routine tool for contributing to risk assessment. Due to the fact that alternative methods have primarily focused so far on hazard identification (Sect. 2.1), the predictions often relate to categories of classification and labelling as defined in international classification systems such the United Nations (UN) Global Harmonized System for Classification and Labelling (GHS).

For calculating the predictive capacity, the final outcomes/predictions of the test method are compared to those from a reference method or to other reference data. The reference method is usually an *in vivo* test method (see Sect. 2.1.4), but comparison can also be performed against human data if available.

Sensitivity and Specificity

Typically, test methods provide binary outcomes (see Fig. 4.1). This is true for most diagnostic tests in medicine but also for most alternative methods. Binary (=dichotomous) predictions here relate to diagnostic results of yes/no (absence or presence of a property) or predictions on causative properties of test items in the system of interest, i.e. "positive" = causing a toxicity effect or negative = not causing this effect (or at least at a threshold below concern = "cut-off").

To characterise the diagnostic or predictive capacity of methods with binary outcomes, the sensitivity (Se) and specificity (Sp) of the test method is calculated. To this end, the binary predictions of the alternative test method are compared to binary predictions obtained from the reference data, typically the *in vivo* test method, for the same set of test chemicals. Predictions from the reference method are considered as actual positive or actual negatives.

As defined by OECD Guidance Document No. 34, the *sensitivity* is the proportion of positive chemicals for the endpoint considered that are correctly identified by the test method (true positive predictions) as compared to the actual positives of the reference method; conversely, the proportion of positive chemicals wrongly predicted as negative corresponds to the false negatives. The *specificity* is the proportion of negative chemicals that are correctly identified by the test method (or true negative predictions); conversely the proportion of negative chemicals wrongly predicted as positive is the false negative rate. Additionally, the accuracy of the test method is the proportion of correct predictions made—in comparison to the reference data—over all predictions obtained.

Two-by-two contingency tables are useful tools for summarising the outcomes of test methods in relation to the actual positives and actual negatives of the reference data. These tables show the frequency of each type of prediction: a positive prediction of the test method for a test item considered an actual positive is termed “true positive” (a). Accordingly the outcomes “false negative” (c), “true negative” (d) and “false positive” (b) are determined. Additionally the number of chemicals assessed is shown (see Table 4.1 and Fig. 4.5).

The fraction (P) of chemicals that produce a positive result in the reference method, over the total number (N) of chemicals is often named ‘prevalence’. Conversely the fraction of chemicals that produce a negative result in the reference method is (1 − P). Therefore, the number of chemicals producing a positive result in the reference method is P × N and the number of chemicals producing a negative result in the reference method is (1 − P) × N (Table 4.1). Denoting the reference method by R, for which the outcome can be positive or negative (respectively R+ and R−) and the test method by T, for which the outcome can be positive or negative as well (respectively T+ and T−), the prevalence P can be expressed as the probability in the reference data set that the outcome is positive and formulated as $P = P(R^+)$ and $1 - P = P(R^-)$.

The calculation of sensitivity and specificity can be formulated with the use of Bayes’ formulas as follows:

$$Se = P\left(T^+ \mid R^+\right) = \frac{P\left(T^+ \cap R^+\right)}{P\left(R^+\right)} \Leftrightarrow Se = \frac{a}{a+c} \tag{4.1}$$

$$Sp = P\left(T^- \mid R^-\right) = \frac{P\left(T^- \cap R^-\right)}{P\left(R^-\right)} \Leftrightarrow Sp = \frac{d}{b+d} \tag{4.2}$$

Those equations show that the proportion of actual negatives and actual positives in the sample do not influence the calculations of Se and Sp. One can also say that both are independent on the prevalence (number of actual positives) in the sample. That means that Se and Sp are indicators directly related to the intrinsic features of the test method.

Table 4.1 Two-by-two contingency table for binary outcomes, providing types of predictions and their respective proportions

	Reference + (actual positive)	Reference − (actual negative)	
Test+ (positive prediction)	a = P × N × Se	b = (1 − P) × N × (1 − Sp)	a + b
	True Positive prediction	False Positive prediction	
Test− (negative prediction)	c = P × N × (1 − Se)	d = (1 − P) × N × Sp	c + d
	False Negative prediction	True Negative prediction	
	a + c = P × N	b + d = (1 − P) × N	a + c + b + d = N

× = multiplication sign

Table 4.2 Three-by-three contingency table for three possible outcomes, providing types of predictions and their respective proportions

	Reference Category 1	Reference Category 2	Reference Category 3
Test	**a**	**b**	**c**
Category 1	Correct prediction as Category 1	Underprediction as Category 2	Underprediction as Category 3
	Rate = (a/n_1) × 100	Rate = (b/n_2) × 100	Rate = (c/n3) × 100
Test	**d**	**e**	**f**
Category 2	Overprediction as Category 1	Correct prediction as Category 2	Underprediction as Category 3
	Rate = (d/n_1) × 100	Rate = (e/n_2) × 100	Rate = (f/n3) × 100
Test	**g**	**h**	**i**
Category 3	Overprediction as Category 1	Overprediction as Category 2	Correct prediction as Category 3
	Rate = (e/n_1) × 100	Rate = (h/n_2) × 100	Rate = (i/n3) × 100
	a + d + g = n_1	b + e + h = n_2	c + f + i = n_3

× = multiplication sign

Positive and Negative Predictive Values

Apart of sensitivity and specificity, two other quantitative indicators can be calculated: Positive Predictive Value (PPV) and Negative Predictive Value (NPV). They correspond to the *probability* that a chemical produces a positive result in the reference method when the outcome of the test method is positive (PPV), and the probability that a chemical produces a negative result in the reference method when the outcome of the test method is negative (NPV). Using Bayes' formulas, they are respectively calculated as:

$$PPV = P\left(R^{+} \,|\, T^{+}\right) = \frac{P\left(T^{+} \cap R^{+}\right)}{P\left(T^{+}\right)} \tag{4.3}$$

$$NPV = P\left(R^{-}|T^{-}\right) = \frac{P\left(T^{-}\cap R^{-}\right)}{P\left(T^{-}\right)} \tag{4.4}$$

They respectively result in:

$$\begin{aligned}PPV = P\left(R^{+}|T^{+}\right) &= \frac{P\left(T^{+}\cap R^{+}\right)}{P\left(T^{+}\right)} = \frac{P\left(T^{+}\cap R^{+}\right)}{P\left(T^{+}\cap R^{+}\right)+P\left(T^{+}\cap R^{-}\right)}\\ &= \frac{\left(P\cdot N\cdot Se\right)/N}{\left(\left(P\cdot N\cdot Se\right)/N\right)+\left(\left(1-P\right)\cdot N\cdot\left(1-Sp\right)/\mathrm{N}\right)}\\ &= \frac{P\cdot Se}{P\cdot Se+\left(1-P\right)\cdot\left(1-Sp\right)}\end{aligned} \tag{4.5}$$

$$\begin{aligned}NPV = P\left(R^{-}|T^{-}\right) &= \frac{P\left(T^{-}\cap R^{-}\right)}{P\left(T^{-}\right)} = \frac{P\left(T^{-}\cap R^{-}\right)}{P\left(T^{-}\cap R^{-}\right)+P\left(T^{-}\cap R^{+}\right)}\\ &= \frac{\left(\left(1-P\right)\cdot N\cdot Sp\right)/N}{\left(\left(\left(1-P\right)\cdot N\cdot Sp\right)/N\right)+\left(\left(P\cdot N\cdot\left(1-Se\right)\right)/\mathrm{N}\right)} = \frac{\left(1-P\right)\cdot Sp}{\left(1-P\right)\cdot Sp+P\cdot\left(1-Se\right)}\end{aligned} \tag{4.6}$$

It is obvious that both positive predictive value (Eq. (4.5)) and negative predicted value (Eq. (4.6)) depend on the prevalence (P), unlike sensitivity and specificity.

Therefore PPV and NPV calculations do represent the performance of the test method per se but for a specific set of chemicals in terms of the relative proportion of actual negatives and actual positives. They give only post-testing indications on how predictions were made for the set of chemicals that has been used; those indications would be different with another set of chemicals (e.g. where the prevalence of positive chemicals would be different—see Sect. 2.3.3). In contrast, the calculations of Sensitivity and Specificity are representative of the intrinsic test method performance, independent of the prevalence, i.e. the fraction of chemicals producing positive results. The examination of both sensitivity and specificity allows capturing the test method performance. This simultaneous evaluation of sensitivity and specificity can also be done when performing Receiver Operating Characteristic (ROC) analysis, as described below.

Considerations for More Than Binary Outcomes

When the possible outcomes of a test method are not binary and thus provide more than two types of prediction, sensitivity and specificity *sensu stricto* are not used but similar calculations are performed. For instance, when the prediction model yields three (sub-)categories, the resulting contingency table is therefore a three-by-three table, covering nine possible predictions. Still, predictions performed by the *in vitro* method are compared to those from the reference data (e.g. the *in vivo* reference method or other relevant data relating to the toxicity event). Consider a situation with three categories, category 1 relating to the most severe effect, category 2 to intermediate effects and category 3 to the least severe effects. For predictions regarding category 1, the three possible outcomes are: correct predictions into category 1, under-prediction into category 2, and under-prediction into category 3. For category 3, the three possible outcomes are: correct predictions into category 3, over-prediction into category 2, and over-prediction into category 1. For the middle category 2, the three possible outcomes are: correct predictions into category 2, under-predictions into category 3, and over-prediction into category 1. For each of these nine predictions it is possible to calculate their respective rates in percentages within the category predicted by the reference method.

Accuracy

Whether the outcome is binary or not, the accuracy of the test method—also referred to as 'overall accuracy'—can additionally being calculated. It is defined by the total number of correct predictions divided by the total number of predictions performed.

When examining the most common case of binary outcome (see Table 4.1), the overall accuracy (OA) is also related to the Prevalence (P) by the following formula:

$$\begin{aligned} OA &= P\left(\left(T^{+} \cap R^{+}\right) | \left(T^{-} \cap R^{-}\right)\right) = P\left(T^{+} \cap R^{+}\right) + P\left(T^{-} \cap R^{-}\right) \\ &= P\left(T^{+} | R^{+}\right) \cdot P\left(R^{+}\right) + P\left(T^{-} | R^{-}\right) \cdot P\left(R^{-}\right) = Se \cdot P + Sp \cdot \left(1 - P\right) \end{aligned} \tag{4.7}$$

The same result is obtained when calculating the OA using the expressions in Table 4.1 cells.

$$OA = \frac{a + \mathrm{d}}{N} = P.Se + \left(1 - P\right).Sp = P\left(Se - Sp\right) + Sp \tag{4.8}$$

If Se > Sp, then from the above formula (Eq. 4.8) it follows necessarily that Se > OA > Sp

If Se < Sp, then it is derived from the same formula that necessarily that Sp > OA > Se

Additionally, when Se > Sp and if P increases, the OA increases as well; if P decreases, the OA decreases. When Se < Sp, if P increases, the OA decreases; if P decreases, the OA increases.

In other words, the OA is influenced by the prevalence P and always takes values that are necessarily between Se and Sp, whatever the value of P is—except for the particular case of Se = Sp, then OA = Sp = Se. During the validation process, the OA is sometimes used and reported. However using the OA does not reflect the intrinsic performance of the test method—in contrast to Se and Sp—as it depends on the prevalence P. or instance, an overall accuracy of OA = 0.78 could correspond to three different cases such as: {Se = 0.9; Sp = 0.7; P = 0.4} or {Se = 0.9; Sp = 0.5;

P = 0.7} or {Se = 0.9; Sp = 0.3; P = 0.8}. Therefore, the single use of OA is not a very useful tool to describe the concordance of a test method against a reference method (or reference data).

Likelihood Ratios

As demonstrated above, typical measures characterising test method performance relate to the prevalence-independent measures of sensitivity, specificity and overall accuracy taking into consideration the number of chemicals tested. However, likelihood ratios can be useful for assessing and reporting test method performance. For binary tests, one distinguishes likelihood ratio positive (LR^+) from likelihood ratio negative (LR^-). Likelihood ratios are routinely used in medicine in the context of describing how informative diagnostic tests are. However they have not been used much in toxicity for describing how informative a particular test result is from a specific test method.

Positive and negative likelihood ratio are defined as follows:

$$LR^+ = \frac{P\left(T^+ \mid R^+\right)}{P\left(T^+ \mid R^-\right)} = \frac{P\left(T^+ \mid R^+\right)}{1 - P\left(T^- \mid R^-\right)} = \frac{Se}{1 - Sp} \tag{4.9}$$

$$LR^- = \frac{P\left(T^- \mid R^+\right)}{P\left(T^- \mid R^-\right)} = \frac{1 - P\left(T^+ \mid R^+\right)}{P\left(T^- \mid R^-\right)} = \frac{1 - Se}{Sp} \tag{4.10}$$

In the expressions of LR^+ and LR^- (Eqs. (4.9) and (4.10)) the prevalence P is absent. That means that both likelihood ratios are not influenced by the prevalence P. In that sense, they are not mere ratios (re-)using sensitivity and specificity values. They represent probabilistic indicators reflecting how likely it is that a type of prediction is true. The LR^+ indicates the probability of a positive result being a true positive. In terms of performance, it is desirable that LR^+ is as high as possible which corresponds to high rate of true positive and/or low rate of false positive. Conversely, the LR^- should be as low as possible which corresponds to high rate of true negative and/or low rate of false negative. In medicine, likelihood ratios are often translated into qualitative stratified ratings ("qualitative strength") using cut-offs of LR's. These ratings aid the communication of test method strength. Mayer (2004) for instance lists four categories corresponding to "excellent", "very good", "fair" and "useless".

ROC as Means of Evaluating Optimal Cut-Off

Variations of the cut-off value are usually examined at the stage of test method development, but can be taken into account at any point in time. Desprez et al. (2015), have recently provided an example of a post hoc analysis of prediction models used for skin corrosion sub-categorisation and, on the basis of the analysis, proposed improved prediction models. For prediction models using cut-off values for assigning the predictions "negative" or "positive", any variation of the cut-off value will result in changes of the Se and Sp, in opposite directions. Thus, depending on the intended application it is possible to set a cut-off (i.e. within the prediction model) so that it optimises either sensitivity or specificity. To systematically assess the impact of shifting the cut-off, a useful approach consists in obtaining a Receiver Operating Characteristic (ROC) curve which provides quantitative indications of the predictive capacity.

A ROC curve is a graphical representation of test method performance: the x-axis represents values of (1-Specificity) and the y-axis represents values of the Sensitivity when monotonic variation of the cut-off value is applied for binary predictions (Fig. 4.12). The best theoretical performance of the method is obtained when both Se and Sp are close to 1 i.e., when Se is close to 1 and 1 – Sp close to 0. The area under the ROC curve is necessarily between 0 and 1, and the best performance of the method is obtained when this area is close to 1. In contrast to the simple use of single values of Se and Sp, the ROC curve represents all possible values of Se and 1-Sp for all possible cut-offs. The ROC analysis will thus consist of finding the cut-off that will maximize the value of Se and minimize the value of 1 – Sp (i.e. maximize the value of Sp). Usually the diagonal line—defined by the points (0; 0) to (1; 1)—is also represented. The shape of the ROC curve gives also an indication of the test method performance; it should have a hyperbolic shape, that is it should be as far away as possible from the midline and follow as closely as possible a curve that would link the points (0; 0) to (0; 1) and (0;1) to (1; 1). Such a curve would lead to an area under ROC close to 1, i.e. the best possible result.

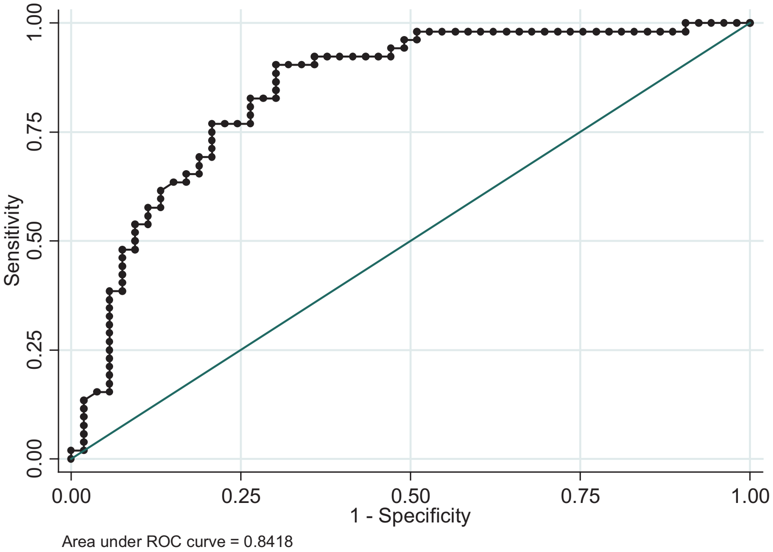

Fig. 4.12 Theoretical example of receiver operating characteristic (ROC) curve

4.7.2 Statistical Evaluation of the Within- and Between Laboratory Reproducibility

Within laboratory reproducibility (WLR) (or intra-laboratory reproducibility) gives information on the extent to which a test provides the same results over time when conducted in the same laboratory (OECD 2005), while the between-laboratory reproducibility (BLR) addresses this question for different laboratories (OECD 2005). In a more general manner, WLR and BLR may not only focus on the obtained prediction but may also examine the variability (e.g. standard deviation) of the measured endpoint of the test method.

Reproducibility and Variability Within One Laboratory

Within-Laboratory Reproducibility

The OECD Guidance Document No. 34, on the validation of new test methods (OECD 2005), provides a definition of the "within laboratory reproducibility" (WLR) or "intra-laboratory reproducibility". The WLR aims to determine the "extent that qualified people within the same laboratory can successfully replicate results using a specific protocol at different times". Typically, the experiment is performed over several runs that are independent and the WLR is assessed considering the agreement between the predictive results of these runs. The WLR is part of the indicators that measure the test method reliability (together with the between laboratory reproducibility, see below).

Several ways can be used to assess the WLR. Classically, the WLR is calculated on the basis of the fraction of chemicals (in %) for which concordant predictions in all runs were made (Eq. (4.11)) either over all chemicals with valid test results in the laboratory (see Eq. (4.11)) or over all chemicals included in the study. Whether to relate the number of concordant predictions to one or the other ideally should be defined at the outset of the study.

$$WLR = \frac{Number\ of\ chemicals\ for\ which\ concordant\ predictions\ are\ made\ in\ all\ runs}{Total\ number\ of\ chemicals\ used\ for\ these\ runs} \quad (4.11)$$

The advantage of this type of calculation is that it uses the final outcome or result of the method as used in practice and is easy to perform.

However, it should be kept in mind that an analysis of the reproducibility (or inversely variability) of the underlying measurement (e.g. normalised cell viability) allows assessing reproducibility without potential misleading results of substances close to the cut-off of the prediction model: obviously, measures that are close to the cut-off value defined for deriving predictions may show low variability between each other and yet result in different predictions which would be interpreted as "non-reproducibility" (Fig. 4.13). Notably, the closer

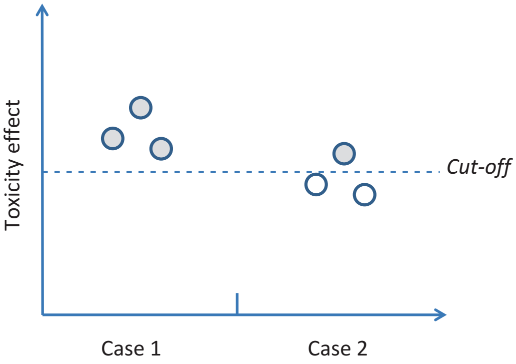

Fig. 4.13 Vicinity of measurements to an arbitrary categorisation cut-off may lead to non-concordant predictions that are interpreted as non-reproducibility although the dispersion between the individual runs (*circles*) is very similar in case 1 versus case 2

measurements are to the cut-off, the greater the influence of random variations that tilt results in one or the other direction (i.e. positive or negative prediction). In these borderline cases, the assessment of WLR based on concordant predictions may not capture accurately the reproducibility of the test method and when interpreting reproducibility via concordance of predictions the vicinity of values to arbitrarily fixed cut-offs needs to be taken into account. It is therefore also useful to assess and quantify the variability (dispersion) of the actual measurements before application of the prediction model.

Variability

In addition to assessing the agreement of predictions it is useful to study the variability of the measurements obtained, e.g. over runs. Variability can be studied by examining medians, means, as well as standard deviation (SD) values, and coefficient of variations (CV) of the measured parameters. Observation of the SD value helps establishing a threshold: data points for which the SD values are below this threshold have a low variability and are considered concordant. Analysis of variance (ANOVA) can further be performed and would compare the variability of the parameter over the runs. However before performing an ANOVA, some conditions regarding the data should be verified first. These conditions are that (i) the groups of comparison (i.e. the runs) are independent, (ii) the distribution of the data is normal, and (iii) the equality of variance in the compared groups is verified. This ANOVA can be combined with pairwise comparisons that help determining which pairs of runs are eventually significantly different (e.g. if four runs were performed, six pair comparisons should be done between runs 1 and 2; 1 and 3; 1 and 4; 2 and 3; 2 and 4; 3 and 4).

When the conditions of the ANOVA are not verified, the analysis can be performed on the basis of non-parametric statistical tests, such as the Kruskal-Wallis or Mann–Whitney tests as those statistical tests are based on the ranks of the parameter (Van Hecke 2012). For instance, when three runs are performed, the Kruskal-Wallis test helps to find out whether significant differences are globally observed on more than two groups of data (However, in some cases the performance of non-parametric tests might result in a loss of statistical power compared to ANOVA (Ferreira et al. 2012)). Although this is still a matter of debate, the data transformation—when applicable—may be worthwhile to obtain normally distributed data, and therefore allow ANOVA to be performed.

The assessments of the reproducibility of a test method based on concordant predictions (i.e. after application of the prediction model) and variability of the measured parameter (without using the prediction model) are complementary. They both give valuable quantitative information. The assessment of concordant predictions provides information on the WLR and BLR of the test method for its intended use and is therefore necessary for regulatory purposes. The assessment of the measured parameter is also necessary to capture variations of this parameter over runs, especially to identify borderline cases (when the measured parameter approaches the cut-off value) and therefore helps in understanding how predictions were performed and may help identifying chemicals for which predictions have been problematic. Moreover, defining variability independent of the prediction models may support later adaptation of the prediction model if necessary (Desprez et al. 2015).

Between Laboratory Reproducibility

The between laboratory reproducibility (BLR) is also called inter-laboratory reproducibility and has been also defined in the OECD Guidance Document No. 34. The BLR provides information on the reproducibility of a test method in different laboratories, i.e. under slightly different conditions. Together with within-laboratory reproducibility and the ease of transferring a method from one experienced to less experience laboratories (“transferability”), BLR informs on the robustness of a test method, i.e. its “resilience” towards minor variations in terms of equipment, operator and aspects such as shipment of cells, etc.

The way to assess BLR is similar to the one for assessing WLR and it can therefore be based on the fraction of chemicals that led to concordant predictions in all different labs (see Eq. (4.12)) either over all chemicals with valid test results in the laboratories (see Eq. (4.12)) or over all chemicals included in the study. Whether to relate the number of concordant prediction to the one or the other ideally should be defined at the outset of the study.

$$BLR = \frac{Number\, of\ chemicals\ for\, which\, concordant\ predictions\, are\, made\, in\, all\, laboratories}{Total\, number\, of\ chemicals\ used\ for\, these\, laboratories} \tag{4.12}$$

Similarly to what has been said before, it can be useful to assess also the variability between laboratories using medians, means, standard deviations and coefficient of variations of the measured parameter.

4.7.3 Providing Confidence Intervals Instead of a Single Point Estimates

The use of a single value (i.e. point estimate) does not entirely capture the uncertainty related to the use of a test method and its predictions. The key values of Sensitivity, Specificity, WLR and BLR may be given within a confidence interval (CI), for example at 95 % (CI_{95}). Calculating and reporting CIs takes this into account and communicates the uncertainty associated with a point estimate, thus improving the description of test method performance.

Consider the use of CI in the following example of WLR: In theory, if the whole population of chemicals would be tested, the obtained WLR would be the exact WLR (WLR_{ex}). However, for validation studies only a very limited number of chemicals(= a representative sample of chemicals) is used. In terms of statistics, this set of test chemicals is a sample of the entire population of existing chemicals. Therefore the WLR value (WLR_{est}), obtained with this set of test chemicals, is an estimated value of the exact one (WLR_{ex}). The CI_{95} represents the range of WLR values for which the probability to find the exact one is 95 %, that also means that the probability of not including the exact value in this interval is 5 %. Any value included in the CI has the same probability than any other to occur, including also the mean (see Fig. 4.14). Obviously, the sample size plays a critical role. The greater the sample, the narrower the confidence interval.

For instance, a test that classifies 20 chemicals and for which 17 out of 20 chemicals are concordantly predicted has, according to previous definition, a WLR rate of (17/20) × 100 = 85 %. The CI_{95} for this value is [62.1–96.8 %], following binomial distribution. If we now consider a set of 60 test chemicals, for which the WLR rate is also 85 % i.e., 51 out of 60 chemicals have concordant prediction, then the CI_{95} is [73.4–92.9 %]. This CI is therefore much narrower than the previous one that has the same mean value of WLR.

Any value of this CI has the same probability to occur and the mean value of 85 % is included in this interval. If the whole population of chemicals was tested

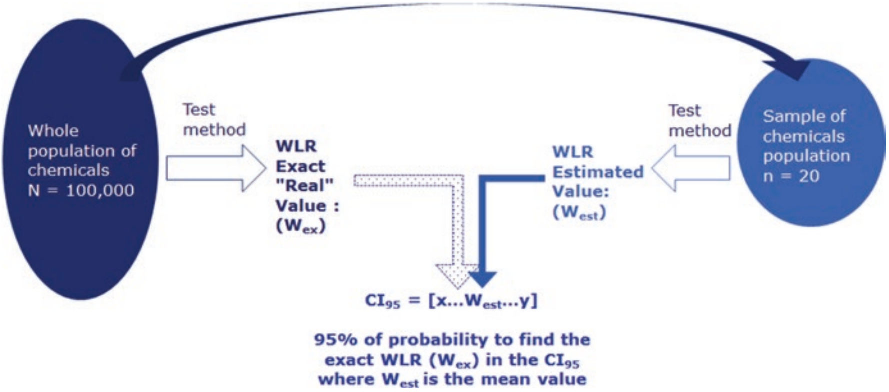
Whole population of chemicals N = 100,000
Test method
WLR Exact "Real" Value : (W_{ex})
Test method
Sample of chemicals population n = 20
WLR Estimated Value: (W_{est})
CI_{95} = [x...W_{est}...y]
95% of probability to find the exact WLR (W_{ex}) in the CI_{95} where W_{est} is the mean value

Fig. 4.14 WLR estimation with confidence intervals

then we would get the exact value of WLR. The CI_{95} represents the range of value of WLR for which there is 95 % of chances to find the exact WLR value.

In the examples above, where WLR = 85 %, the use of the CI helps to quantify the uncertainty of this value with an error risk of 5 %. When comparing the CI for 20 test chemicals, which is [62.1–96.8 %], and the one for 60 test chemicals, which is [73.4–92.9 %] it becomes clear that the extent of the first one is much greater than the second: the larger the CI, the higher the uncertainty. In other words, the central value of 85 % in the first CI is framed by a much more extended range of values compared to the second one.

4.7.4 Using All Experimental Observations for PC, WLR and BLR

Up to now we have presented analyses that are based on the assumption that, for BLR analysis, there is *one final prediction per laboratory* that can be used to determine concordance of predictions between laboratories (Eq. (4.12)) and that there is one final prediction per chemical so as to calculate the predictive capacity of the assay for the sample tested (Table 4.1, Fig. 4.5). However, this is normally not the case since during validation studies, experiments are typically conducted in triplicate in each of the laboratories which creates nine available experimental predictions for each chemical. The reason for this data-rich matrix is the need to assess within-laboratory reproducibility of the predictions of experiments, i.e. the final outcome of the test method as used in practical application. For BLR and PC however, the data matrix results in the following problems: which of the three from each laboratory should be used for determining BLR and PC within the laboratory, and which of the nine predictions available should be used to determine PC for the assay? To address these issues, it has been common practice in many validation studies to derive “final calls” (i.e. a final prediction). For BLR per one final call per laboratory was derived either by (a) calculating the mode of predictions of the three experiments (hence per laboratory normally an odd number of experiments is performed, e.g. three) or (b) by calculating an average value of the final measurement (before application of the prediction model) of the three experiments which was then converted into a final prediction by applying the prediction model as usual. These final laboratory predictions were then analysed for their concordance, i.e. in exactly the same way as WLR had been established. The percentage value of concordant predictions *between* laboratories is then communicated as the between-laboratory reproducibility of the assay. Similarly, final singular predictions (“final calls”) were produced per chemical in view of calculating the predictive capacity (PC), i.e. the sensitivity, specificity and accuracy of a test method. To this end, the mode of the final laboratory predictions (determined for BLR analysis, see above) was determined yielding one *final call per chemical*. This created the basis for calculating one point estimate for each of the predictive indicators sensitivity, specificity, accuracy based on exactly the number of chemicals analysed during the study. Although this analysis may appear a straightforward way of simplifying the artificially inflated validation data matrix, mentioned approach has a fundamental disadvantage: instead of using the data from the test methods as it would be used in practice, results from experiments are artificially "aggregated" or "condensed" by means of averaging or, basically, majority voting (mode of predictions). Moreover, this approach leads to a loss of information on the experimental level.

This problem is of course not specific to toxicological data sets but is encountered in many disciplines including biology, medicine (e.g. evaluation of diagnostic test methods, clinical trials), epidemiology, etc. where large sets of (non-independent or not fully independent) observations are available. Standard statistical literature has been cautions with regard to fully using all observations (Colton 1974), mainly because this may be misleading with regard to the actual sample size that should be reflected in the analysis (Colton gives an example of 800 blood pressure measurement in a drug study which were however based on 10 measurements weekly over an 8-week treatment course in only ten patients, which would be the actual sample). In our example, using all observations would mean calculating the sensitivity and specificity on the basis of *all* predications generated during the study, i.e. nine observations per chemical times the sample of chemicals tested. So if 100 chemicals have been tested, the sample would appear to consist of 900 and not the 100 that have been tested in reality. Thus the actual sample is overstated and it also misleadingly narrows the confidence intervals. More recently more publications have addressed the issue of using how all observations can be used or, in particular, to which extent each observation contributes statistically to the overall analysis in such cases. The statistical technique of *Generalized Estimating Equations* (GEE) (Hanley et al. 2003) can be used in such situations and its applicability to validation data sets should be considered. Another way to estimate the WLR, BLR and predictive capacity is the *bootstrapping* technique (Holzhütter et al 1996; Hoffmann and Hartung 2006a). The data from all experiments performed during the validation study are resampled over a large number of times e.g., 1000 times, and on the basis of the resampling the studied parameter is estimated. The idea is that the entire population of chemicals cannot be studied and the sample size is deemed to be limited (e.g. 20 chemicals) for the estimation of the parameter. Therefore the performance of resampling on the sample itself and repeated many of times may better capture the variability of the parameter than the approach based on a single value. For instance, for the WLR, the resampling can be performed at the level of the different runs of a given laboratory. Then WLR is calculated on the basis of concordant results obtained in this resampling. The resampling procedure can also be done at the level of the chemicals. For the BLR, the principle would be the same, e.g. resampling over the results obtained in all laboratories.

5 Conclusions

In this chapter we have explored the fundamental concepts underlying the validation of *in vitro* test methods for hazard/safety assessment of chemicals or biologicals and have summarised the major challenges as well as established processes and tools for validation. Validation sits between the development of novel test methods and their routine use for safety assessment by industry and regulators. The aim of validation is to provide a robust, transparent and trustworthy scientific basis regarding the characterisation of a test method in view of its application for a particular purpose ("fitness for purpose"). From this it follows that there can never be a final or ultimate validation of a test method: validation is context-dependent. Validation studies and subsequent recommendations support regulators, policy makers and stakeholders when considering whether or not to formally adopt (i.e. into legislation) a given test method for a specific use in relation to legislation that aims to protect human health (e.g. workers, consumers) and the environment.

We stress that the term validation incorporates various meanings: it relates to the formal process of assessing and establishing fitness-for-purpose of a test method (often conducted by impartial governmental or supranational organisations), but also to the scientific study type for achieving this and, last, to the testing of a set of hypotheses. These are: (1) that the scientific basis of a test method is relevant for an adverse outcome or toxicity pathway in a target system, (2) that a given protocol associated with the test method allows its reproducible use and (3) that a given prediction model allows making sufficiently accurate predictions on adverse outcomes. All of these hypotheses are assessed through empirical testing of test chemicals (prospective studies) or the evaluation of existing information (retrospective studies). Consequently validation studies are scientific endeavours that need to be conducted in agreement with key scientific principles such as objectivity and appropriateness of methodology. These relate to statistically informed sample size calculations, conscientious selection of test chemicals, ex ante criteria for test method performance and the independence and/or impartiality of some of the actors (at least the scientific peer review).

We have discussed some major challenges of validated alternative test methods, mainly relating to the fact that these are proxy systems and highly reductionist models. We have also discussed the basic design of test methods. These are based on specific test systems (e.g. a specific cell line or tissue), the measurement of specific parameters as well as a prediction model. These elements of a test method are normally described in the procedure associated with the test method. Prediction models are of key importance for the validation of test methods as they are used to derive a performance characterisation in terms of predictive capacity and applicability. Prediction models are functions that convert the measured parameters into categorical predictions relating to any classification that is relevant of the purpose. Classifications can relate to chemicals being sorted according to their intrinsic potential to activate a toxicity pathway or to downstream (apical) health effects/adverse outcomes. Using the terminology of adverse outcome pathways (AOP), classifiers of alternative test methods can relate to everything from molecular initiating events to adverse effects on population level. Typically however, these relate to categories as defined by classification and labelling systems, for instance the United Nations Global Harmonized Systems for Classification and Labelling (UN GHS). Therefore the conversion of the measured parameter into classes/categorical variables, by means of the prediction model, is a simplification process that renders the outcome of the test more comprehensive but loses resolution with regard to the reality of a continuum of toxicity effects from non-toxic to highly toxic.

The validation process also encompasses the careful examination of the regulatory context. Due to the reductionist nature of alternative methods (i.e. modelling only small aspects of a more complex system), validated methods will be increasingly used in integrated testing strategies (ITS) or integrated approaches on testing and assessment (IATA), bringing together data from a variety of sources. The concept of adverse outcome pathways (AOPs) supports consensus-building on what may be the most important toxicological events leading to a final adverse effect: AOPs provide a description of these so-called "key events" and, to the extent possible, their causal links. In that sense the AOP concept promises to contribute to the identification of knowledge gaps and is expected to expedite the development of new test methods that model upstream mechanisms relevant for the downstream (apical) adverse effect of concern. The AOP concept thus also supports the validation of alternative methods of greater mechanistic and biological relevance and, it is hoped, greater predictive power and overall relevance.

The validation workflow typically includes four steps: assessment of test methods, conduct of validation studies, independent scientific peer review and final conclusions and recommendations. Regarding the practice of validation, the so-called "modular approach" has proven extremely useful: the information generated during the validation studies is systematically assessed through several information modules that all need to be sufficiently satisfied in view of scientific peer review of the validity status of a test method—notably, what constitutes "sufficient" depends on the purpose. The modules include the test definition (i.e. a description of the scientific basis of the method, within- and between-laboratory reproducibility, transferability, predictive capacity, applicability domain and performance standards, defined upon completion of a validation study. All of these modules are informed by testing data on chemicals. Thus, the number of chemicals tested influences the certainty of the data obtained. Therefore calculation of sample size, prior to the conduct of the study, is a prerequisite for enabling the generation of a sufficient amount of data. This relates to the statistical power and the target values defined for the study (e.g. target values of within-laboratory reproducibility or sensitivity). The reliability relies to the reproducibility of the method within a given laboratory, so called within laboratory reproducibility (WLR), the reproducibility over several laboratories, or between laboratory reproducibility (BLR) as well as the ease with which methods are amenable to transfer from one to another laboratory ("transferability"). WLR and BLR are assessed by the proportion of concordant predictions obtained. However this may not capture all the variability observed when using the method and other quantitative tools for assessing data variability before application of the prediction model are useful. The predictive relevance relies to how useful the obtained predictions are for the intended regulatory use. This is quantitatively assessed by the predictive capacity of the method. The predictive capacity uses accuracy values, such as sensitivity and specificity. Reporting confidence intervals helps capturing the uncertainty on the values obtained. ROC curves are another useful tool for assessing in a systematic manner the performance of the test method as a function of variations of the cut-off value of the prediction model.

In summary, validation is a multidisciplinary scientific exercise requiring expertise in a wide range of disciplines and areas, including biology, physiology, chemistry, statistics and regulatory frameworks. All these aspects are necessary for as complete a characterisation of a test method as possible through validation: This will help to understand and describe the extent of certainty and confidence in a test method and the remaining level of uncertainty. Validation will therefore play an ever greater role as new tools and more probabilistic approaches emerge in risk assessment wherein alternative methods are likely to play a central role.

References

Aggett P et al (2007) Variability and uncertainty in toxicology of chemicals in food, consumer products and the environment. Report by the Committee on toxicity of chemicals in food, consumer products and the environment. http://cot.food.gov.uk/sites/default/files/cot/vutreport-march2007.pdf

Archer G, Balls M, Bruner LH, Curren RD, Fentem JH, Holzhütter H-G, Liebsch M, Lovell DP, Southee JA (1997) ATLA 25:505–5016

Attarwala H (2010) TGN1412: from discovery to disaster. J Young Pharm 2(3):332–336

Balls M (1994) Replacement of animal procedures: alternatives in research, education and testing. Lab Anim 28:193–211

Balls M (1997) Defined structural and performance criteria would facilitate the validation and acceptance of alternative test procedures. ATLA 25:483–484

Balls M, Karcher W (1995) Comment: the validation of alternative test methods. ATLA 23:884–886

Balls M, Blaauboer B, Brusick D, Frazier J, Lamb D, Pemberton M, Reinhardt C, Roberfroid M, Rosenkranz H, Schmid B, Spielmann H, Stammati AL, Walum E (1990a) Report and recommendations of the CAAT/ERGAAT workshop on the validation of toxicity test procedures. ATLA 18:313–337 (“Amden I report”)

Balls M, Botham P, Cordier A, Fumero S, Kayser D, Koeter H, Koundakjian P, Lindquist NG, Meyer O, Pioda L, Reinhardt C, Rozemond H, Smyrniotis T, Spielmann H, Van Looy H, van der Venne MT, Walum E (1990b) Report and recommendations of an international workshop on promotion of the regulatory acceptance of validated non-animal toxicity test procedures. ATLA 18:339–344 (“Vouliagmeni report”)

Balls M, Bridges J, Southee J (1990c) Animals and alternatives in toxicology: present status and future prospects. VCH Publishers, New York

Balls M, Blaauboer BJ, Fentem JH, Bruner L, Combes RD, Ekwall B, Fielder RJ, Guillouzo A, Lewis RW, Lovell DP, Reinhardt CA, Repetto G, Sladowski D, Spielmann H, Zucco F (1995a) Practical aspects of the validation of toxicity test procedures. ECVAM workshop report 5. ATLA 23:129-147 (“Amden II report”)

Balls M, De Klerck W, Baker F, van Beek M, Bouillon C, Bruner L, Carstensen J, Chamberlain M, Cottin M, Curren R, Dupuis J, Fairweather F, Faure U, Fentem J, Fisher C, Calli C, Kemper F, Knaap A, Langley G, Loprieno G, Loprieno N, Pape W, Pechovitch G, Spielmann H., Ungar K, White I, Zuang V (1995b) Development and validation of non-animal tests and testing strategies: the identification of a coordinated response to the challenge and the opportunity presented by the sixth amendment to the Cosmetics Directive (76/768/EEC). ATLA 23:398–409

Balls M, Amcoff P, Bremer S, Casati S, Coecke S, Clothier R, Combes R, Corvi R, Curren R, Eskes C, Fentem J, Gribaldo L, Halder M, Hartung T, Hoffmann S, Schechtman L, Laurie Scott L, Spielmann H, Stokes W, Tice R, Wagner D, Zuang V (2005) The principles of weight of evidence validation of test methods and testing strategies. the report and recommendations of ECVAM workshop 58. ATLA 34:603–620

Balls M, Combes RD, Bhogal N (2012) The use of integrated and intelligent testing strategies in the prediction of toxic hazard and in risk assessment. Adv Exp Med Biol 745:221–253

Barratt MD, Brantom PG, Fentem JH, Gerner I, Walker AP, Worth AP (1998) The ECVAM international validation study for *in vitro* tests for skin corrosivity. 1. Selection and distribution of the test chemicals. Toxicol *In Vitro* 12:471–482

Belanger SE, Rawlings JM, Carr GJ (2013) Use of fish embryo toxicity tests for the prediction of acute fish toxicity to chemicals. Environ Toxicol Chem 32(8):1768–1783

Borlak J (2009) Trovafloxacin: a case study of idiosyncratic or iatrogenic liver toxicity—molecular mechanisms and lessons for pharmacotoxicity. In: Hayes W, Griesinger C, Guzelian (eds) Proceedings of the 1st international forum towards evidence-based toxicology. Hum Exp Toxicol 28:119–212

Bouhifd M, Bories G, Casado J, Coecke S, Norlén H, Parissis N, Rodrigues RM, Whelan MP (2012) Automation of an *in vitro* cytotoxicity assay used to estimate starting doses in acute oral systemic toxicity tests. Food Chem Toxicol 50(6):2084–2096

Bouvier d'Yvoire M, Bremer S, Casati S, Ceridono M, Coecke S, Corvi R, Eskes, C, Gribaldo L, Griesinger C, Knaut H, Linge JP, Roi A, Zuang V (2012) ECVAM and new technologies for toxicity testing. In: Balls M, Combes RD, Bhogal N (eds) New technologies for toxicity testing. Springer series "Advances in experimental medicine and biology." Springer/Landes Bioscience 745:154–180

Bruner LH, Carr GJ, Chamberlain M, Curren RD (1996) Validation of alternative methods for toxicity testing. Toxicol *In Vitro* 10:479–501

Coecke S, Bowe G, Millcamps A, Bernasconi C, Bostroen AC, Bories G, Fortaner S, Gineste Jm, Gouliarmou V, Langezaal I, Liska R, Mendoza E, Morath S, Reina V, Wilk-Zasadna I, Whelan M (2014) In: Jennings P, Price A (eds) Considerations in the development of *in vitro* toxicity testing methods intended for regulatory use, Coecke S. et al. 2014. *In vitro* Toxicology Systems series: methods in pharmacology and toxicology. Springer Science+Business Media, LLC, pp 551–569

Colton T (1974) Statistics in medicine. Little, Brown and Company, Boston

Curren RD, Southee JA, Spielmann H, Liebsch M, Fentem JH, Balls M (1995) The role of prevalidation in the development, validation and acceptance of alternative methods. ATLA 23:211–217

Desprez B, Barroso J, Griesinger C, Kandárová H, Alépée N, Fuchs H (2015) Two novel prediction models improve predictions of skin corrosive sub-categories by test methods of OECD Test Guideline No. 431. Toxicol *In Vitro* 29(2015):2055–2080

Draize JH, Woodard G, Calvery HO (1944) Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membranes. J Pharmacol Exp Ther 82:377–390

EURL ECVAM (2014) The EURL ECVAM—Cosmetics Europe prospective validation study of reconstructed human tissue-based test methods for identifying chemicals not requiring classification for serious eye damage/eye irritation testing

EURL ECVAM, European Commission, Joint Research Centre (2012 onwards) Website on EURL ECVAM Recommendations on validated test methods; the site is continuously updated. https://eurl-ecvam.jrc.ec.europa.eu/eurl-ecvam-recommendations

Fentem JH, Prinsen MK, Spielmann H, Walum E, Botham PA (1995) Validation—lessons learned from practical experience. Toxicol *In Vitro* 9:857–862

Fentem JH, Archer GE, Balls M, Botham PA, Curren RD, Earl LK, Esdaile DJ, Holzhütter HG, Liebsch M (1998) The ECVAM International validation study on *in vitro* tests for skin corrosivity. 2. Results and evaluation by the management team. Toxicol *In Vitro* 12(4):483–524

Ferreira E, Rocha M, Mequelino D (2012) Sigmae Alfenas 1(1):126–139. http://www.google.it/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&ved=0CCUQFjAA&url=http%3A%2F%2Fpublicacoes.unifal-mg.edu.br%2Frevistas%2Findex.php%2Fsigmae%2Farticle%2Fdownload%2F99%2Fpdf&ei=4hK5VJS3H8i7ygPf0wI&usg=AFQjCNHE4phiKyiXay4fNnpctO489uddBQ&sig2=tNR25ijrcheBiS87zWLcrA&bvm=bv.83829542,d.bGQ

Flahault A, Cadilhac M, Thomas G (2005) Sample size calculation should be performed for design accuracy in diagnostic studies. J Clin Epidemiol 58:859–862

Frazier JM (1990a) Scientific criteria for validation of *in vitro* toxicity tests. Environment Monographs no. 36. Paris: Organization for Economic Co-Operation and Development; 62 pp. Not available any longer. The document gave rise to the OECD Solna report (OECD, 1996)

Frazier JM (1990b) Validation of *in vitro* models. J Am Coll Toxicol 9:355–359

Frazier JM (1992) Validation of *in vitro* toxicity tests. In: Frazier JM (ed) *In vitro* toxicity testing: applications to safety evaluations. Marcel Dekker, New York, pp 245–252

Frazier JF (1994) The role of mechanistic toxicology in test method validation. Toxicology *In Vitro* 8:787–791

Goldberg, AM, Epstein, LD, Zurlo J (1995) A modular approach to validation—a work in progress. In: Salem H, Katz SA (eds) Advances in animal alternatives for safety and efficacy testing. Taylor & Francis, WA, pp 303–308. Expanded edition of: Animal test alternatives. Marcel Dekker, New York

Green S (1993) Regulatory agency considerations and requirements for validation of toxicity test alternatives. Toxicol Lett 68:119–123

Gregory CD (2014) Cell biology: the disassembly of death. Nature 507:312–313

Griesinger C (2009) Comparing medicine with toxicology— a mapping of knowledge creation, concepts and basic epistemology. In: Hayes W, Griesinger C, Guzelian P (eds) Proceedings of the 1st international forum towards evidence-based toxicology. Hum Exp Toxicol 28(2–3):101–104

Griesinger C, Hoffmann S, Kinsner A, Coecke S, Hartung T (2009) Special issue: evidence-based toxicology (EBT). Preface. Hum Exp Toxicol. In: Hayes W, Griesinger C, Guzelian P (eds) Proceedings of the 1st international forum towards evidence-based toxicology. Hum Exp Toxicol 28(2–3):83–86

Griesinger C, Schäffer M, Worth A, Zuang V (2014) Skin corrosion and irritation. In: Alternative methods for regulatory toxicology—a state-of-the-art review. JRC science & policy report. Publications Office of the European Union. http://bookshop.europa.eu/is-bin/INTERSHOP.enfinity/WFS/EU-Bookshop-Site/en_GB/-/EUR/ViewParametricSearch-Dispatch

Guzelian PS, Victoroff MS, Halmes NC, James RC, Guzelian CP (2005) Evidence-based toxicology: a comprehensive framework for causation. Hum Exp Toxicol 24(4):161–201

Guzelian PS, Victoroff MS, Halmes C, James RC (2009) Clear path: towards an evidence-based toxicology (EBT). In: Hayes W, Griesinger C, Guzelian P (eds) Proceedings of the 1st international forum towards evidence-based toxicology. Hum Exp Toxicol 28(2–3):71–79

Hanley JA, Negassa A, Edwardes MD, Forrester JE (2003) Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol 157(4):364–375

Hartung T (2010) Evidence-based toxicology—the toolbox of validation for the 21st century? ALTEX 27(4):253–263

Hartung T, Bremer S, Casati S, Coecke S, Corvi R, Fortaner S, Gribaldo L, Halder M, Hoffmann S, Roi AJ, Prieto P, Sabbioni E, Scott L, Worth A, Zuang V (2004) A modular approach to the ECVAM principles on test validity. Altern Lab Anim 32(5):467–472

Hendriksen C, Spieser J-M, Akkermans A, Balls M, Bruckner L, Cussler K, Daas A, Descamps J, Dobbelaer R, Fentem J, Halder M, van der Kamp M, Lucken R, Milstien J, Sesardic D, Straughan D, Valadares A (1998) Validation of alternative methods for the potency testing of vaccines. ATLA 26:747–761

Hoffmann S, Hartung T (2005) Diagnosis: toxic!—trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations. Toxicol Sci 85(1):422–428

Hoffmann S, Hartung T (2006a) Designing validation studies more efficiently according to the modular approach: retrospective analysis of the EPISKIN test for skin corrosion. Altern Lab Anim 34(2):177–191

Hoffmann S, Hartung T (2006b) Toward an evidence-based toxicology. Hum Exp Toxicol 25(9):497–513

Hoffmann S, Edler L, Gardner I, Gribaldo L, Hartung T, Klein C, Liebsch M, Sauerland S, Schechtman L, Stammati A, Nikolaidis E (2008) Points of reference in the validation process: the report and recommendations of ECVAM Workshop 66. Altern Lab Anim 36(3):343–352

Holzhütter HG, Archer G, Dami N, Lovell DP, Saltelli A, Sjöström M (1996) Recommendation for the application of biostatistical methods during the development and validation of alternative methods. ATLA 24:511–530

Horvath CJ, Milton MN (2009) The TeGenero incident and the Duff Report conclusions: a series of unfortunate events or an avoidable event? Toxicol Pathol 37(3):372–383

Hothorn LA (2002) Selected biostatistical aspects of the validation of *in vitro* toxicological assays. ATLA 30(2):93–98

ICCVAM (1997) Validation and regulatory acceptance of toxicological test methods: a report of the *ad hoc* interagency coordinating committee on the validation of alternative methods. National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, NC, USA, p 105. NIH Publication No: 97-3981. http://iccvam.niehs.nih.gov/docs/guidelines/validate.pdf

Judson R, Kavlock R, Martin M, Reif D, Houck K, Knudsen T, Richard A, Tice RR, Whelan M, Xia M, Huang R, Austin C, Daston G, Hartung T, Fowle JR 3rd, Wooge W, Tong W, Dix D (2013) Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. ALTEX 30(1):51–56

Kinsner-Ovaskainen A, Maxwell G, Kreysa J, Barroso J, Adriaens E, Alépée N, Berg N, Bremer S, Coecke S, Comenges JZ, Corvi R, Casati S, Dal Negro G, Marrec-Fairley M, Griesinger C, Halder M, Heisler E, Hirmann D, Kleensang A, Kopp-Schneider A, Lapenna S, Munn S, Prieto P, Schechtman L, Schultz T, Vidal JM, Worth A, Zuang V (2012) Report of the EPAA-ECVAM workshop on the validation of Integrated Testing Strategies (ITS). Altern Lab Anim 40(3):175–181

Lachin JM (1981) Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 2:93–113

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405(2):442–451

Mayer D (2004) Essential evidence-based medicine. Cambridge University Press, Cambridge

Neugebauer EA (2009) Evidence-based medicine—a possible model for evidence-based toxicology? In: Hayes W, Griesinger C, Guzelian P (eds) Proceedings of the 1st international forum towards evidence-based toxicology. Hum Exp Toxicol 28(2–3):105–107.

OECD (1996, updated in 2009) Final report of the OECD workshop on the harmonisation of validation and acceptance criteria for alternative toxicological test methods. “Solna Report”. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/mc/chem/tg(96)9&doclanguage=en

OECD (2005) Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment. OECD series on testing and assessment, document No. 34. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono%282005%2914&doclanguage=en

OECD (2008) Workshop on integrated approaches to testing and assessment. Series on testing and assessment No. 88. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2008)10&doclanguage=en

OECD (2013 (rev.), adopted 2010) Test guideline No. 439. *In vitro* skin irritation: Reconstructed human Epidermis test methods. http://www.oecdilibrary.org/docserver/download/9713241e.pdf?expires=1417441398&id=id&accname=guest&checksum=DB5C47DAF73F635E918A565FCCABCA01

OECD (2013) Guidance document on developing and assessing adverse outcome pathways. Series on Testing and Assessment No. 184. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2013)6&doclanguage=en

OECD (2014 (2nd rev.), adopted 2004) Test guideline No. 431. *In vitro* skin corrosion: Reconstructed human Epidermis test methods. http://www.oecdilibrary.org/docserver/download/9714521e.pdf?expires=1417442281&id=id&accname=guest&checksum=BA3044BD757F25BF4559D9D83E8E83A6

OECD (2014) New guidance on an integrated approach on testing and assessment (IATA) for skin corrosion and irritation. Series on testing and assessment No. 203. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2014)19&doclanguage=en

Pepe M (2003) The statistical evaluation of medical tests for classification and prediction. Oxford statistical science series. Oxford University Press, Oxford

Russell WMS, Burch RL (1959) The principles of humane experimental technique. Methuen, London

Scala RA (1987) Theoretical approaches to validation. In: Goldberg AM (ed) *In Vitro* toxicology: approaches to validation. vol 5, Alternative methods in toxicology. Mary Ann Liebert, New York, pp 1–9

Smythe DH (1978) Alternatives to animal experiments. Scolar Press for the Research Defence Society, London

Stephens ML, Andersen M, Becker RA, Betts K, Boekelheide K, Carney E, Chapin R, Devlin D, Fitzpatrick S, Fowle JR 3rd, Harlow P, Hartung T, Hoffmann S, Holsapple M, Jacobs A, Judson R, Naidenko O, Pastoor T, Patlewicz G, Rowan A, Scherer R, Shaikh R, Simon T, Wolf D, Zurlo J (2013) Evidence-based toxicology for the 21st century: opportunities and challenges. ALTEX 30(1):74–103

Van Hecke T (2012) J Stat Manage Syst 15(2–3). http://www.tandfonline.com/doi/abs/10.1080/09720510.2012.10701623?journalCode=tsms20#.VH7SxRAhB8E

Walum E, Clemedson C, Ekwall B (1994) Principles for the validation of *in vitro* toxicology test methods. Toxicol *In Vitro* 8:807–812

Waters CK (1990) Why the antireductionist consensus won't survive the case of classical Mendelian genetics. In: Fine A, Forbes M, Wessels L (eds) Proceedings of the biennial meeting of the Philosophy of Science Association, vol 1. Philosophy of Science Association, East Lansing, pp 125–139

Waters CK (2007) Causes that make a difference. J Philos 104:551–579

Weber M (2005) Philosophy of experimental biology. Cambridge University Press, New York