Bench Marks!

Elements and Development Processes for Test Methods in Toxicology and Human Health-Relevant Life Science Research

Eike Cöllen[1], Yaroslav Tanaskov[1], Anna-Katharina Holzer[1], Michele Dipalo[2], Jasmin Schäfer[3], Udo Kraushaar[3] and Marcel Leist[1,4]

[1]In vitro Toxicology and Biomedicine, Dept inaugurated by the Doerenkamp-Zbinden Foundation, University of Konstanz, Konstanz, Germany; [2]IIT Istituto Italiano di Tecnologia, Genoa, Italy; [3]NMI Natural and Medical Sciences Institute at the University of Tuebingen, Reutlingen, Germany; [4]CAAT-Europe, University of Konstanz, Konstanz, Germany

Received January 4, 2024;

Correspondence:
Marcel Leist, PhD
In vitro Toxicology and Biomedicine
Dept inaugurated by the
Doerenkamp-Zbinden foundation
at the University of Konstanz
Universitaetsstr. 10
78464 Konstanz, Germany
(marcel.leist@uni-konstanz.de)

ALTEX 41(1), 142-148.
doi:10.14573/altex.2401041

Abstract

Many laboratory procedures generate data on properties of chemicals, but they cannot be equated with toxicological "test methods". This apparent discrepancy is not limited to *in vitro* testing, using animal-free new approach methods (NAM), but also applies to animal-based testing approaches. Here, we give a brief overview of the differences between data generation and the setup or use of a complete test method. While there is excellent literature available on this topic for specialists (GIVIMP guidance; ToxTemp overview), a brief overview and easily-accessible entry point may be useful for a broader community. We provide a single figure to summarize all test method elements and processes required in the development (setup and adaptation) of a test method. The exposure scheme, the endpoint, and the test system are briefly outlined as fundamental elements of any test method. A rationale is provided on why they are not sufficient. We then explain the importance and role of purpose definition (including some information on what is modelled) and the prediction model, aka data interpretation procedure, which depends on the purpose definition, as further essential elements. This connection exemplifies that all fundamental elements are interdependent, and none can be omitted. Finally, discussion is provided on validation as a measure to provide confidence in the reliability, performance, and relevance of a test method. In this sense, validation may be considered a sixth fundamental element for practical use of test methods.

Plain language summary

Many laboratory procedures generate data on chemicals, but they cannot be considered complete toxicological "test methods". Here, we give a brief explanation of the fundamental elements of a toxicological test method. We provide an illustration that gives a complete overview of the development of a test method for non-specialists. We introduce the six fundamental elements, i.e., the exposure scheme, the test endpoint, the test system, the purpose definition and the prediction model and describe how they work together. Finally, we discuss the concept of validation. An understanding of these concepts is important for good-quality scientific research and especially for the development and acceptance of alternatives to animal experiments.

1 Setting the scene

It is a broadly accepted fact that test methods are important as research tools in toxicology, pharmacy, pharmacology, clinical chemistry, and many biomedical fields. There is also little dispute on the importance of data from test methods for diagnosis, quality control, efficacy estimates, and setting reference values for human, animal, and environmental safety. Given this background, it is astonishing that the fundamental definition of a test method seems to be clear only to a fraction of researchers and stakeholders in the field. Only a subfraction of these feels confident about the development of a test method or about knowledge on all the steps in the lifecycle, including the use, of a test method.

A common quick-fix for such knowledge gaps is to consult Wikipedia. There, a test method is defined as "*a method for a test in science or engineering, such as a physical, chemical, or statistical test. It is a definitive procedure that produces a test result*". Is this really helpful? Does it not feel like circular reasoning ("*a test*

method is a method for a test...")? Well, it goes a step further: it should produce "*results*". This appears trivial, but it is not. We will return to this issue. One can then also find additional information: "*To ensure accurate and relevant test results, a test method should be 'explicit, unambiguous, and experimentally feasible', as well as effective and reproducible*". Also this contains some empty catchwords (accurate, effective, relevant..) and tautologies (who would develop a test method that is NOT experimentally feasible), but again, the results are important (be they more or less accurate...). Altogether, the "quick-fix" is not a very helpful approach. This somehow reflects the problems mentioned above: The seemingly simple issue of a test method is not that easy to capture.

Perhaps one does not need to deal with a lot of theory about test methods. After all, it is a very practical field, and the main point of a test method is that it delivers results. Thus, why not just go to the lab, use the test method, and get something done. This approach can indeed work well for some. But there are also limitations: It does not apply to methods not yet developed or to methods that need to be adapted or improved. The latter case is more common than one may think: The test items (i.e., the type of "chemicals" that are tested) one has to deal with may change. Methods that work for drugs may not work for industrial polymers; test methods performing well with dissolved chemicals may prove to be difficult to use for gases and aerosols; and methods that work well for pure chemicals may not work satisfactorily for complex mixtures (environmental samples, petrol, pesticide formulations). Another big issue is that materials and equipment that is required for the test method may not be available anymore; for instance, analytical devices may have been replaced. All this has the consequence that the test cannot be done in exactly the same way as before. Another important issue is that data from a test may need to be interpreted, and this may require an understanding of the test method (its functioning, its limitations, etc.).

Back to very practical issues: A typical dispute in modern toxicology is which test methods are best suited to answer a question. Such a discussion may involve a classical toxicologist, who favors animal experiments over all other approaches, and a specialist for *in vitro* systems who advocates the use of modern, stem cell-based scientific methods. The first may claim that rats have provided > 80% of all data in regulatory toxicology and that therefore animals are the best test method. The second may claim that there are obvious species differences, and that human-relevant safety testing requires human cells, which can be generated from stem cells. Ideally, these should be grown into tissues (organs-on-a-chip) and be incorporated into microphysiological systems (MPS), the most human-relevant test method. They both have a point, and still, they are both fundamentally wrong. Their dispute is like an argument about whether carrots or potatoes are the more healthy fruit. Neither animals nor MPS are test methods. Neither can give a "result", which is part of the core definition of a test method. Is there any space for a "yes, but..."? No, there isn't. The mistake of confusing animals or stem cells or MPS with a test method is common, but this does not make the mistake smaller. The conceptual error is fundamental. It is like confusing a motor with a car, or a cook with a restaurant, or a test tube with a SARS-CoV-2 test.

After the failure of theoretical quick-fixes and simplified practical approaches, a third perspective on test methods comes from regulatory texts issued, e.g., by the OECD. The most solid and comprehensive of these is the Guidance on Good *In Vitro* Methods Practice (GIVIMP) (OECD, 2018). This document is definitely worth reading. Especially in industrial laboratories it is also used increasingly as important guidance for the way work is organized and test methods are set up and used. Some stakeholders of the field prefer a more light-weight approach. For an initial orientation, or for entering the field, this can be useful. A short summary on all test method elements and ways to define and describe them is, e.g., given in the ToxTemp document (Krebs et al., 2019). This again refers to all items listed in the OECD guidance document 211, the official, but hard-to-read reference paper on how to describe a test method. Some other publications that give explicit but concise (easy to read) descriptions of all test method documents are available (Leist and Hengstler, 2018; Leist et al., 2012b, 2010; OECD, 2018; Pamies et al., 2022; Schmidt et al., 2017; Krebs et al., 2019, 2020; Pallocca et al., 2022). The main intention here is to provide a graphical overview of this situation. Our objective is to provide the core information in an easily accessible form for a broad audience.

2 Test methods as models to test hypotheses

As mentioned above, neither MPS nor animals are test methods. Both can be a fundamental element of a test method: they can be the test system. Without the other elements, they are a biological item, like a flower, a piece of meat, a bacterium or a yeast extract. Nobody would consider those a test method. Is this some strange theoretical discussion, specific to the field of new approach methodologies (NAMs)? We think it is not. It is rather a universal issue, and it may be exemplified using the classical toxicological approach, i.e., an animal experiment (Pallocca and Leist, 2022; Pallocca et al., 2022): Animals can be used to assess, e.g., developmental neurotoxicity, carcinogenicity or skin sensitization. But... these are quite different test methods. In all cases, animals are the test system. But differences become apparent already here. If one takes a closer look, it is not always the same test system. Animals may be used, e.g., at different ages or from different suppliers or the strain and species can be different. Then, there are different exposure schemes: single dosing, maternal dosing, repeated dosing over a long or short time, priming and challenge dosing. The whole might occur in different caging (enriched environment or not) with different chow, single or group housing, etc. The dosing may be oral, parenteral, by inhalation, etc., and the vehicle/solvent may differ. The endpoints may be various pathological approaches (differences in tissue samples, types of stains used; various microscopic and macroscopic scorings), various functional or clinical endpoints or biochemical measures. The combination of the three fundamen-

Abbreviations: AOP, adverse outcome pathway; MIE, molecular initiation event; MPS, microphysiological systems; NAM, new approach (non-animal) methods (or methodologies)

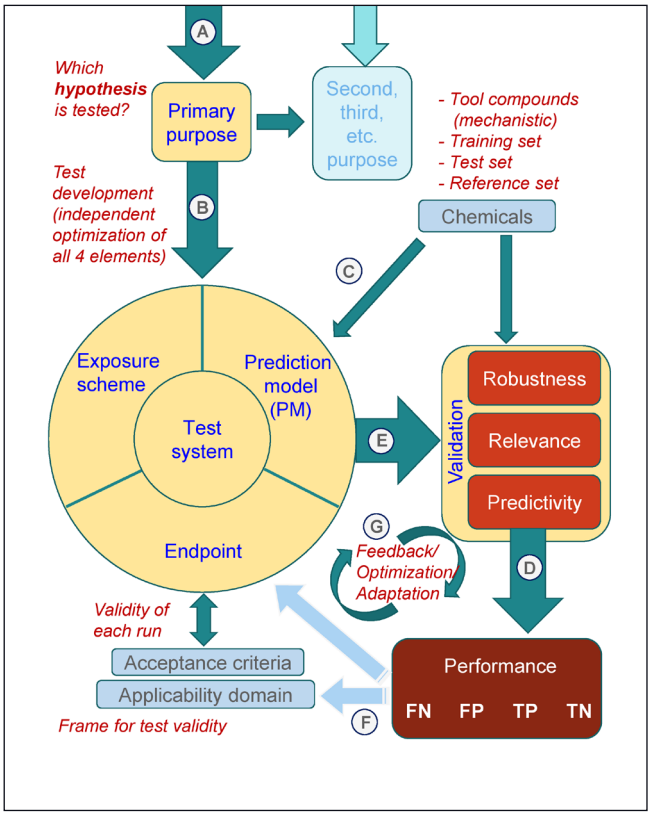

Fig. 1: Overview of test method elements and the test method life cycle

Fundamental elements of test methods are indicated in yellow boxes, and important steps in the life cycle of developing and optimizing test methods are indicated by letters (A-G). Red letters indicate important functions, tools or their significance for the test method. At the start of test method setup, the purpose needs to be defined (A). Then, preliminary versions of the test system, the exposure scheme, and the endpoint can be established and tested against tool compounds (B,C). This should result in some level of predictivity (D). With an increasing set of chemicals and data, a preliminary prediction model may be established and then be used for a validation process (E). The test method may work better for some compound groups and not so well for others. This determines the borders of test method validity, i.e., the applicability domain of the method (F). Even if the method is formally validated and performing well, errors of execution may occur for individual runs. To control for these, acceptance criteria are established and used. Often, new developments, materials or questions require or enable assay optimization and further development (G). This drives the need for a re-adjustment of the validation status, the performance measures, and the prediction model.

tal test method elements (Fig. 1) test system, exposure scheme, and endpoint can thus already lead to hundreds of ways to generate data (all using animals), and still elements of a complete test method are missing. Such an incomplete test method can deliver data, even huge streams of data. It can be used for various displays, such as diagrams, heat maps, etc. Such displays give information on the data structure (e.g., increase of values over time), but they do not yield further information beyond the test performed. For this, complete test methods need to be defined and set up.

Two issues are often forgotten but are necessary for a complete test method: (i) Data are generated for a purpose – otherwise nobody would invest resources and effort; and (ii) data need to be interpreted according to this purpose. After all, what is desired from the whole effort is a gain in information. Data need to be converted into information.

The purpose of a test method can be defined at various levels:

Level 1: What is modelled?

Level 2: Which hypothesis is tested?

Level 3: Are there additional (often super-ordinate) hypotheses that the test information helps to clarify?

In the case of animal-based test methods, it is clear that the main interest is not in whether, e.g., a certain shampoo triggers allergies in rats. Animals are used as models, usually meant to predict some effects in humans. Sometimes, it is forgotten that the animals are used as models (Pallocca and Leist, 2022; Pallocca et al., 2022). The animals are part of a test method, and the purpose of this method is often to provide information on human safety. The whole setup of the test method is in the context of this purpose, and if the data are used in other ways, then the purpose changes. If one of the fundamental elements of a test method changes, then the overall method changes. This means, its performance, relevance, and predictivity change, and one cannot be sure anymore whether the data are useful at all.

The same applies to NAMs. For instance, a human neuronal culture may be used. It may take a huge effort to establish such a culture, especially if neurons of a certain type and maturity are required, or if they need to be grown on a certain substrate (like a micro-electrode chip). This is a step in establishing a test system, but more steps are required to set up a test method. The exposure scheme defines how test compounds are added (timing, solvent, medium), whether they are washed out, or whether additive dosing is applied. As endpoints, one may record from single cells or from groups of cells, one may measure changes of the membrane potential, or altered intracellular free ion concentrations (e.g., of Ca^{2+}) or certain types of network behavior. For each of these measures, many variations are possible (e.g., speed of change, amplitude, area-under-the curve). This means that one test system may be the basis of many test methods. The final setup requires a definition of what is being modelled (e.g., the developing or the adult brain, or a certain brain region). One may ask whether the setup is intended to model acute high dose exposure of humans or whether it should model long-term low level exposure, or intermittent exposure.

It is then important to consider which hypothesis is tested: Is a compound acutely neurotoxic? Does a compound sensitize to neurotoxicity? Does a compound disturb network function or network formation? How long is the maximum non-neurotoxic exposure time? Is the metabolite of a compound neurotoxic? Is the toxicity in humans expected to be transient or permanent, functional or structural? These example questions show that adaptations of a test method are required when the purpose of testing changes. Moreover, the testing objective strongly affects the interpretation of the results. Again, only the combination of level *1+2* questions (purpose of the model) provides this information. If a purpose is not defined, then a given preliminary test method cannot tell much

more than how, e.g., a curve of Ca^{2+} concentrations over time looks in a certain cell culture setup under given laboratory conditions. Information beyond the laboratory model can be obtained only by defining what is modelled and which hypothesis is tested.

For the general overview provided by this short paper, an overall understanding of the primary test method purpose is sufficient. For those working at a regulatory level with test method data, or for those preparing such data for regulatory submissions, or in general for the research community dealing with risk assessment, the purpose and interpretation of test methods is more complex: For instance, the primary purpose of a method may be to define whether a chemical can cause neuronal signaling disturbances. A secondary purpose would be to contribute information to the question whether a compound can cause developmental neurotoxicity by leading to transient signaling disturbances. Here, the secondary purpose is more superordinate and more complex. The data from the test method alone may not be fully sufficient to judge on the hypothesis, but may contribute (e.g., together with data on blood-brain barrier permeability, PBK modelling, and data from other models). A tertiary purpose may be to set a safety threshold for the chemical in question or to determine a margin of safety. This is an even more superordinate question. In some cases the test method makes a large contribution to answering such a question, sometimes it is one of many information elements required. For instance, some NAM can answer the question whether a certain hazard exists, but potency information has to be derived from other approaches.

Another short example to re-iterate this point: A skin model may be used with the primary purpose to test whether a chemical is cytotoxic, a secondary purpose may be to test whether the chemical causes skin irritation, and a tertiary purpose would be to ask whether a compound falls into a certain skin irritation class according to the globally-harmonized system (GHS).

The next paragraph will define the last fundamental step required to obtain such information from test methods.

3 The prediction model

Even though test purpose, test system, exposure scheme, and test endpoint may be defined for a procedure, this is STILL NOT A TEST METHOD (Box 1). It may be named a preliminary test method or a bioanalytical procedure. In reality, many putative NAM get stuck at this point. What is the problem here? Some claim that there is no problem: Data are being produced by many screening methods and other approaches that do not have a prediction model. Also, for animal experiments prediction models are not always in place. Is there a conflict of opinions, and are there different ways of weighing the importance of a prediction model?

We suggest that there is less conflict than may appear, and there are two reasons for this:

The first reason refers back to the importance of understanding the purpose of a test method, i.e., that it is usually meant to model something beyond itself and to test a hypothesis. The latter refers to a statement with relevance outside the test method. If these two points are neglected, or not considered important for the ongoing work, then of course data can be generated and used. However, it is important to be aware of the fact that these data mainly refer to the model system. One can answer the question for the test system used whether compound A causes a larger rise in Ca^{2+} than compound B, or whether compound C produces more rapid transcriptional changes in the cell system used than compound D in that model system. If one wants to make statements beyond the model system used, i.e., to refer to certain human populations, or to refer to other model systems, then this purpose needs to be defined, and a prediction model needs to be established. The prediction model, often also termed data interpretation procedure (DIP), converts data from within a model to more generally usable information. This sounds highly theoretical, so we need some examples. Typical testing procedures produce a fluorescence (or absorbance) reading, a relative amount of something (e.g., RNA, a protein or oxygen radicals) or a measure of morphology. In the end, it is a number (ideally with a measure of uncertainty). This number may be compared to other numbers generated in the same system. The prediction model converts such numbers into more general statements. In many cases, it classifies the numbers (e.g., active/inactive; toxic/non-toxic; protective/non-protective). For most toxicological testing in a regulatory context, such a classification is required. The reason is that information is sought on human safety and not on the well-being of cultured neurons or mice in an animal unit. In more basic research or for mechanistic studies, the model itself is the objective of the investigation. In such cases where no broader conclusions are desired, the prediction model (DIP) may be dropped, and there is no conflict. In molecular research the findings from one model system are often claimed to apply to many other (or all other) systems with some similarity to the model used. This is (an unconscious) prediction model, but it is poorly defined (similarity is not measured) and usually not validated. In toxicology, some animal studies also use this approach, with all its conceptional and practical shortcomings.

The second reason for generation of data from methods without a prediction model is based on a "trick": A "preliminary" or "generic" prediction model is generated by magic (automatically), and often unconsciously. There is nothing wrong with this. It is sound science, as long as data producers and recipients are aware of the procedure. Notably, an automatic prediction model may not always be the best choice, and it is usually not evaluated for performance. Once these shortcomings are realized, data can be handled with the necessary care. One example is the hit definition of screens, which is a generic prediction model. It divides the test outputs into two classes, those values that are translated as hits and those that are defined as non-hits. This is the classical function of a prediction model, i.e., conversion of test outputs into toxicity classes or activity classes. Thus, one may claim to some extent that analytical methods used for screening may be complete toxicological methods. A typical example are mechanistic and molecular initiating event (MIE) assays used, e.g., in the Tox21/ToxCast program. For instance, a kinase inhibition assay is a simple analytical procedure. However, it can be used to classify compounds as kinase inhibitors or not. Moreover, the purpose of the assay can be defined to reach beyond a simple biochemical/pharmacological question. One can claim that in certain settings, the assay results give information on other models, "beyond the original assay". For instance if the argument is used that the assay describes a MIE of an adverse outcome pathway (AOP) or that the assay is part of a systems toxicology

prediction module for a certain toxicity. In such a case, the hit definition would be used to derive more generally applicable toxicological information from the analytical procedure. In such a case, the kinase assay may be classified as a test method with all its fundamental elements. Notably, the same test method may be run in a "screening mode" and in a standard "laboratory mode". With all elements and also the SOP being the same, the prediction model may differ, and thus the outcome may be different. The most typical example of a generic (unconscious) definition of a prediction model is statistical classification. This approach assumes that everything that is statistically different from baseline noise is a signal. In a second step it is then postulated that everything that produces a signal is a hit, and everything that does not produce a signal is a non-hit. This way, most assays can be automatically assigned a prediction model. However, it is not clear whether such hits are biologically or toxicologically meaningful. Only a validation process can show whether such a procedure can lead to relevant results and to a good test performance (predictivity).

As a side note, it may be important to know that in the field of NAM development, enormous resources and efforts have been invested in establishing and validating meaningful prediction models. This may even be considered one of the most important contributions of NAM science to life science research altogether. The application and implementation of NAMs is still challenged by the difficulty of setting up prediction models and obtaining a good and positive validation status because the intention is to replace animal methods that have regulatory acceptance without compromising human safety. On this background it is surprising that such an effort has not been made for animal-based testing. The concept of a prediction model is rarely used in this field, and most interpretations are based on generic prediction models (i.e., everything different from baseline is considered to be an effect). To avoid a too high false-positive rate, some such effects were later defined as non-adverse or as tolerable up to a certain limit. The problems with this system have become particularly evident when defining endocrine disruptors or genotoxic carcinogens, in particular those naturally present in plants used for human consumption and in processed food.

Box 1: Short summary of test method setup

The first fundamental element of a test method is the *purpose*. Note that there can be several layers of purpose. The *exposure scheme*, the *endpoint*, and the *test system* are further fundamental elements. Note that the fundamental elements can be interconnected, but all are essential. For instance, the prediction model does not make sense without test purpose (interconnection), but a test method without a *prediction model* is not a complete test method.

Validation is depicted in Fig. 1 as the sixth fundamental element. Depending on the context of testing and on the interpretation of the word "validation," this categorization is under discussion. Here, validation is seen as any process that is intended to give some information on test method performance and on the usefulness of data from the test method to provide information on something (e.g., human health) beyond the test method itself. Validation (in this sense) does not refer to any regulatory or formalized procedure. An important basic assumption used here is that nobody would apply a test method without believing in some minimum performance/predictivity of the method, i.e., meaningfulness of the data being produced. The basis for such a belief (or better confidence) is validation. In this sense, validation is considered as fundamental to each test method.

There are many validation approaches, but all include a procedure to gain confidence in the robustness (= reliability) of the data generated by the test method. In theoretical terms, a test method is a model that can be used to test a hypothesis and that gives data with a certain relevance and predictivity for the modelled situation (e.g., real world). It would not make much sense to use a test method if one was not convinced that the data somehow predict the situation that is modelled. In many cases, the capacity of prediction is termed "performance", and often it is classified by the number or fraction of false negatives (FN), false positives (FP), true positives (TP), and true negatives (TN) produced by the test method. A frequent misconception is that the performance is mainly determined by the test method (and its fundamental elements). The truth is that these numbers also depend on the reference set of chemicals used to determine the performance of the method. The performance of a method can change if different reference sets are used.

Chemical sets play several roles in the lifecycle of a test method. During the setup, mechanistic tool compounds with known modes of action and with known potencies to interfere with certain biological processes, structures, and signaling pathways can be used to assemble and optimize test elements. They may also be used for mechanistic validation, i.e., for increasing confidence in the mechanistic relevance and predictivity of the test method data with respect to the modelled situation (e.g., the human population). For setting up a prediction model, training and test sets of reference chemicals may be used, and a test set may also be used for many forms of validation.

For comparison of method performance from laboratory to laboratory, or for transfer and modification of a method, reference sets of compounds may be defined, together with the results expected and to be reproduced for each compound. Finally, the quality of test runs may be controlled by using (in each run) a set of chemicals with pre-defined expected outcomes. On this basis, acceptance criteria (AC) for each run may be defined. Often, the setup of a method is not a one-way process but rather an optimization cycle that involves adaptation steps. These adaptations may either improve a performance parameter, or they may ensure that variants of the test method perform in similar ways.

4 Validation as an element of a test method

Validation is used in various fields (from toxicology to informatics; from engineering to psychological testing). It is seen slightly differently from application area to application area, and the processes involved can have major differences. The overall concept is, however, universal: to inspire confidence that something is working (qualitatively) and to provide a measure of how well (quan-

titatively) something is working, e.g., concerning reliability. For predictive methods, also relevance and predictivity can be part of a validation procedure. Ideally, the validation process also results in information on when the method can be used with high accuracy of the results, and for which situations it is more likely to deliver poor or unreliable results. Finally, validation may ensure that a method is transparent, transferable and controllable for each run. For the latter purpose, acceptance criteria may be defined (Holzer et al., 2023). The word validation is used here in its broadest sense as a process to obtain information on how reliable the test method is, how well the prediction model works, and how well the purpose of the test method is fulfilled. Without such a process a test method as such makes no sense, and it cannot be developed further as there is no way to establish whether a change improves it or not.

Highly formalized ways of validation have been developed for regulatory use of toxicological methods (Hoffmann et al., 2016; Hartung et al., 2013; Patterson et al., 2021; Leist et al., 2012a; Hartung, 2007; OECD, 2005; Balls et al., 1995; EMA, 2016). This validation concept is part of the "claim to fame" of the scientific field of NAM, but it has also led to a certain stagnation, due to its high administrative and technical load. For this reason, many lighter or more flexible approaches have also been discussed (Bal-Price et al., 2018; Lanzoni et al., 2019; Schmeisser et al., 2023; Marx-Stoelting et al., 2023; EMA, 2008), and they are being adapted for complex organoids or *in silico* models (e.g., Hewitt et al., 2015; Pamies et al., 2022).

Validation may be seen as independent of the core concept of a test method or as being a fundamental element of the method. Both views are justified, as there are also many ways of doing validation. We suggest that some form of non-formal validation belongs inseparably to a test method (Box 1). After all, only validation can assess whether a method is suitable for its purpose and whether it produces any useful data. Nobody would use a method without believing in these two issues. Thus, for any test method used, at least some form of unconscious validation must have taken place. Perhaps not a good and stringent or highly formalized validation, but the example of animal-based test methods has shown that this is not a strict requirement for the use of data.

In most regulatory settings, it is considered necessary that a NAM test method is formally validated and data from a method cannot be used for regulatory purposes without validation. Therefore, formal validation is mandatory for regulatory use and it is thus an essential part of a regulatory method.

An important trend for the future is not just a higher flexibility of the validation process, but also less reliance on reference chemical sets. For many methods and also for many toxicological fields, only quite limited sets of compounds with reliable known activity in humans are available (Aschner et al., 2017). The number is often not sufficient to define robust prediction models and to define the applicability domain of the tests. For this reason, it has been considered to use biological principles of validation instead, e.g., to judge the relevance of the test system responses, based on detailed evaluation of biological disturbances and the consistency of effects on many levels of biological organization. This approach of a mechanistic or scientific validation is not yet formalized, and many of its principles still need to be better defined (Leist et al., 2012a; Aschner et al., 2017; Bal-Price et al., 2018).

5 Conclusion

This article refers to a classical NAM, i.e., an *in vitro* method that uses some cell culture construct (2D or 3D) or tissue explant to predict, more or less directly, some form of toxicity, e.g., phototoxicity, eye corrosion, genotoxicity. This concept may be extended to many applications in toxicology and beyond, e.g., prediction of infection or organ damage from blood tests, prediction of pyrogenicity in quality control processes, prediction of bioactivity of hormones or toxins, prediction of toxicokinetic parameters (blood-brain barrier permeability, hepatic clearance, etc.) or prediction of elements of a complex toxicity (inhibited neurite growth for developmental neurotoxicity or loss of contact inhibition for carcinogenicity, etc.).

However, it is not clear whether the concept will remain valid for all future applications and all methods that may be considered as NAM. For instance, it may be a difficult task to define predictivity for assays that only address some partial aspect of a complex biological event. This is linked to problems in defining the purpose, the prediction model, and the validation strategy, but solutions have been proposed (Bernasconi et al., 2023).

Other NAMs that will require adaptations (or perhaps a different concept altogether) are *in silico* methods. Here, the purpose and the validation process may be clear, and the other elements may be mainly fused with the prediction model. Perhaps other fundamental elements need to be considered? Perhaps the parametrization and underlying assumptions of a physiologically-based kinetic model may be considered a proxy of the test system? How would one deal with future artificial intelligence (AI)-based models, where the internal working may be based on machine learning, but not be understood by the operator?

Even faced with such problems, it appears important to consider also in the future all concepts and fundamental elements of a test method as outlined here. There must be very good reasons to drop or substitute elements. Rather than dropping any of them, they may find new definitions and versions adapted to new tasks and challenges. At present, it is hard to imagine any test method without a purpose, a prediction model, and some form of validation. It is also hard to imagine reasons why concepts like predictivity and relevance should be dropped. However, they may be re-interpreted, or the methods to assess them may change. A method may be relevant for an AOP key event or MIE, rather than an adverse outcome. Predictivity may be established independent of chemicals, e.g., by using principles of biological validation and demonstrating that relevant pathways work in similar ways in the NAM, as in the situation to be modelled (e.g., human pathology). At present, efforts are underway to redefine the concepts of validation to make them more flexible. Similarly, reporting on NAM data is becoming more formalized, but care needs to be taken not to become too rigid about the setup of NAMs and their classical elements in a situation where this concept may change. On the other hand, the classical concept still applies to a majority of NAMs, and awareness and appreciation of their fundamental elements is a foundation for good science and quality of data produced with NAM (Krebs et al., 2019, 2020; Leist et al., 2010).

References

Aschner, M., Ceccatelli, S., Daneshian, M. et al. (2017). Reference compounds for alternative test methods to indicate developmental neurotoxicity (DNT) potential of chemicals: Example lists and criteria for their selection and use. *ALTEX 34*, 49-74. doi:10.14573/altex.1604201

Balls, M., Blaauboer, B. J., Fentem, J. H. et al. (1995). Practical aspects of the validation of toxicity test procedures. *Altern Lab Anim 23*, 129-146. doi:10.1177/026119299502300116

Bal-Price, A., Hogberg, H. T., Crofton, K. M. et al. (2018). Recommendation on test readiness criteria for new approach methods in toxicology: Exemplified for developmental neurotoxicity. *ALTEX 35*, 306-352. doi:10.14573/altex.1712081

Bernasconi, C., Bartnicka, J., Asturiol, D. et al. (2023). Validation of a battery of mechanistic methods relevant for the detection of chemicals that can disrupt the thyroid hormone system. *Publications Office of the European Union*. doi:10.2760/862948

EMA (2008). Qualification of novel methodologies for drug development: Quidance to applicants. EMA/CHMP/SAWP/72894/2008. https://bit.ly/3RPtNLE

EMA (2016). Guideline on the principles of regulatory acceptance of 3Rs (replacement, reduction, refinement) testing approaches. EMA/CHMP/CVMP/JEG-3Rs/450091/2012. https://bit.ly/3TP7bxv

Hartung et al. (2007). Food for thought … on validation. *ALTEX 24*, 67-80. doi:10.14573/altex.2007.2.67

Hartung, T., Hoffmann, S. and Stephens, M. (2013). Mechanistic validation. *ALTEX 30*, 119-130. doi:10.14573/altex.2013.2.119

Hewitt, M., Ellison, C. M., Cronin, M. T. D. et al. (2015). Ensuring confidence in predictions: A scheme to assess the scientific validity of in silico models. *Adv Drug Deliv Rev 86*, 101-111. doi:10.1016/j.addr.2015.03.005

Hoffmann, S., Hartung, T. and Stephens, M. (2016). Evidence-based toxicology. *Adv Exp Med Biol 856*, 231-241. doi:10.1007/978-3-319-33826-2_9

Holzer, A. K., Dreser, N., Pallocca, G. et al. (2023). Acceptance criteria for new approach methods in toxicology and human health-relevant life science research – Part I. *ALTEX 40*, 706-712. doi:10.14573/altex.2310021

Krebs, A., Waldmann, T., Wilks, M. F. et al. (2019). Template for the description of cell-based toxicological test methods to allow evaluation and regulatory use of the data. *ALTEX 36*, 682-699. doi:10.14573/altex.1909271

Krebs, A., van Vugt-Lussenburg, B. M. A., Waldmann, T. et al. (2020). The EU-ToxRisk method documentation, data processing and chemical testing pipeline for the regulatory use of new approach methods. *Arch Toxicol 94*, 2435-2461. doi:10.1007/s00204-020-02802-6

Lanzoni, A., Castoldi, A. F., Kass, G. E. et al. (2019). Advancing human health risk assessment. *EFSA J 17*, e170712. doi:10.2903/j.efsa.2019.e170712

Leist, M., Efremova, L. and Karreman, C. (2010). Food for thought … considerations and guidelines for basic test method descriptions in toxicology. *ALTEX 27*, 309-317. doi:10.14573/altex.2010.4.309

Leist, M., Hasiwa, N., Daneshian, M. et al. (2012a). Validation and quality control of replacement alternatives – Current status and future challenges. *Toxicol Res 1*, 8-22. doi:10.1039/c2tx20011b

Leist, M., Lidbury, B. A., Yang, C. et al. (2012b). Novel technologies and an overall strategy to allow hazard assessment and risk prediction of chemicals, cosmetics, and drugs with animal-free methods. *ALTEX 29*, 373-388. doi:10.14573/altex.2012.4.373

Leist, M. and Hengstler, J. G. (2018). Essential components of methods papers. *ALTEX 35*, 429-432. doi:10.14573/altex.1807031

Marx-Stoelting, P., Rivière, G., Luijten, M. et al. (2023). A walk in the PARC: Developing and implementing 21st century chemical risk assessment in Europe. *Arch Toxicol 97*, 893-908. doi:10.1007/s00204-022-03435-7

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, No. 34*. https://one.oecd.org/document/env/jm/mono(2005)14/en/pdf

OECD (2018). Guidance Document on Good In Vitro Method Practices (GIVIMP). *OECD Series on Testing and Assessment, No. 286*. OECD Publishing, Paris. doi:10.1787/9789264304796-en

Pallocca, G. and Leist, M. (2022). On the usefulness of animals as a model system (part II): Considering benefits within distinct use domains. *ALTEX 39*, 531-539. doi:10.14573/altex.2207111

Pallocca, G., Rovida, C. and Leist, M. (2022). On the usefulness of animals as a model system (part I): Overview of criteria and focus on robustness. *ALTEX 39*, 347-353. doi:10.14573/altex.2203291

Pamies, D., Leist, M., Coecke, S. et al. (2022). Guidance document on good cell and tissue culture practice 2.0 (GCCP 2.0). *ALTEX 39*, 30-70. doi:10.14573/altex.2111011

Patterson, E. A., Whelan, M. P. and Worth, A. P. (2021). The role of validation in establishing the scientific credibility of predictive toxicology approaches intended for regulatory application. *Comput Toxicol 17*, 100144. doi:10.1016/j.comtox.2020.100144

Schmeisser, S., Miccoli, A., von Bergen, M. et al. (2023). New approach methodologies in human regulatory toxicology – Not if, but how and when! *Environ Int 178*, 108082. doi:10.1016/j.envint.2023.108082

Schmidt, B. Z., Lehmann, M., Gutbier, S. et al. (2017). In vitro acute and developmental neurotoxicity screening: An overview of cellular platforms and high-throughput technical possibilities. *Arch Toxicol 91*, 1-33. doi:10.1007/s00204-016-1805-9

Acknowledgements

This work was supported by the BMBF (NeuroTool), INVITE2, the Land BW (BW-3R), and EFSA (TGX-Mapr). It has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements RISK-HUNT3R (No 964537), ToxFree (No 964518), and PARC (No 101057014).