Acceptance criteria for new approach methods in toxicology and human health-relevant life science research - part I

Holzer, A.K.; Dreser, N.; Pallocca, G.; Mangerich, A.; Stacey, G.; Dipalo, M.; ... ; Leist, M.

Bench Marks!

Acceptance Criteria for New Approach Methods in Toxicology and Human Health-Relevant Life Science Research – Part I

Anna-Katharina Holzer[1], Nadine Dreser[1], Giorgia Pallocca[2], Aswin Mangerich[3], Glyn Stacey[4], Michele Dipalo[5], Bob van de Water[6], Costanza Rovida[2], Petra H. Wirtz[7,8], Barbara van Vugt[9], Giulia Panzarella[10,11], Thomas Hartung[2,12], Andrea Terron[13], Iris Mangas[13], Matthias Herzler[14], Philip Marx-Stoelting[14], Sandra Coecke[15] and Marcel Leist[1,2]

[1]In vitro Toxicology and Biomedicine, Dept inaugurated by the Doerenkamp-Zbinden Foundation, University of Konstanz, Konstanz, Germany; [2]CAAT-Europe, University of Konstanz, Konstanz, Germany; [3]Nutritional toxicology, University of Postdam, Potsdam, Germany; [4]International Stem Cell Banking Initiative, Barley, Herts, UK; [5]Istituto Italiano di Tecnologia (IIT), Genova, Italy; [6]Division of Drug Discovery & Safety, Leiden University, Leiden, The Netherlands; [7]Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Konstanz, Germany; [8]Biological Work and Health Psychology, Department of Psychology, University of Konstanz, Konstanz, Germany; [9]BioDetection Systems B.V., Amsterdam, The Netherlands; [10]Università "Magna Græcia" of Catanzaro, Catanzaro, Italy; [11]Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany; [12]Center for Alternatives to Animal Testing (CAAT), Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD, USA; [13]EFSA, European Food Safety Authority, Parma, Italy; [14]Bundesinstitut für Risikobewertung (BfR), Berlin, Germany; [15]European Commission, Joint Research Centre, Ispra, Italy

Abstract

Every test procedure, scientific and non-scientific, has inherent uncertainties, even when performed according to a standard operating procedure (SOP). In addition, it is prone to errors, defects, and mistakes introduced by operators, laboratory equipment, or materials used. Adherence to an SOP and comprehensive validation of the test method cannot guarantee that each test run produces data within the acceptable range of variability and with the precision and accuracy determined during the method validation. We illustrate here (part I) why controlling the validity of each test run is an important element of experimental design. The definition and application of acceptance criteria (AC) for the validity of test runs is important for the setup and use of test methods, particularly for the use of new approach methods (NAM) in toxicity testing. AC can be used for decision rules on how to handle data, e.g., to accept the data for further use (AC fulfilled) or to reject the data (AC not fulfilled). The adherence to AC has important requirements and consequences that may seem surprising at first sight: (i) AC depend on a test method's objectives, e.g., on the types/concentrations of chemicals tested, the regulatory context, the desired throughput; (ii) AC are applied and documented at each test run, while validation of a method (including the definition of AC) is only performed once; (iii) if AC are altered, then the set of data produced by a method can change. AC, if missing, are the blind spot of quality assurance: Test results may not be reliable and comparable. The establishment and uses of AC will be further detailed in part II of this series.

Disclaimer: The positions and opinions presented in this article are those of the authors alone and are not intended to represent the any official position of their respective employing institutions. The article is published under the sole responsibility of the authors and may not be considered as an EFSA, BfR, EURL-ECVAM or another institutions' scientific output.

Received October 2, 2023;

ALTEX 40(4), 706-712. doi:10.14573/altex.2310021

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)

1 Setting the scene

Imagine the following little scenario: you develop a new scientific method (e.g., an *in vitro* or *in silico* toxicity assay), you describe the method well (according to OECD guidance document 211, e.g., using the ToxTemp form (Krebs et al., 2019)), you develop and establish a definite, logical or systematic plan for executing the test method, detailing all the steps to be taken in a logical order (e.g., a standard operating procedure, SOP), and you deposit it in a curated methods database (see details in the GIVIMP guidance (OECD, 2018) or explanations, e.g., in Krebs et al., 2020). Then, you initiate some form of validation (e.g., a method readiness assessment; Bal-Price et al., 2018; Patterson et al., 2021; Schmeisser et al., 2023). During this process, it becomes evident that something important is missing. Something so important that your intra-laboratory reproducibility may be compromised and that other laboratories may have problems reproducing your results using your method. This scenario sounds like a nightmare. Many nightmares are not real, and when you wake up, the problems are gone.... How is it in this case? Is the above scenario possible?

Unfortunately, the above story is not only realistic, it occurs very frequently. And it will continue to occur if the understanding and use of acceptance criteria (AC) does not become better established. AC are neither novel nor are they a difficult concept. In the field of NAM, they are well-described in the GIVIMP guidance (OECD, 2018), and the setting of AC is a standard module of classical method validation, e.g., performed by EURL-ECVAM in Europe. In practical life, we also unconsciously use a form of AC, e.g., when we buy fruit or vegetables (Fig. 1). In this light, it is astonishing that only few published methods, even those found in curated databases run by high-impact journals or in many other method repositories, have a set of clearly defined AC. As highlighted in the above example, the value of data produced using the method is limited if no AC were applied. The topic of AC thus affects the reliability of a large proportion of all scientific data.

Therefore, it is crucial to understand the concepts behind AC and how to define them. First, it is helpful to rationalize the difference between assay validation and AC, and how the concept of overall quality assurance (QA) differs from the use of AC (Fig. 1). Before we address this, we will clarify some terms used in this context.

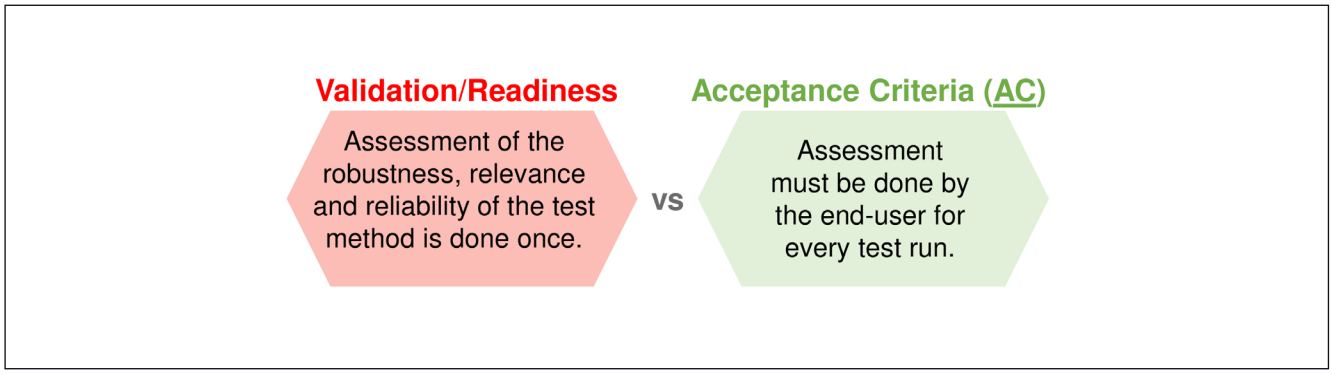

Fig. 1: Acceptance criteria (AC) vs validation
Validation of a method is usually done once, and then it is considered valid from thereon. Validation is meant to provide confidence in the method as such (assuming that everything is done exactly as specified, and that everything works perfectly, and that all material, operators, and equipment perform as during the validation procedure). A daily life analogy may be a study that determines the tastiness and nutritional value of a certain dish (e.g., pizza capricciosa – in general). The concept of AC is fundamentally different. They are meant to make sure that a given run of a test method delivers reliable results (not "in principle", but in real life). A daily life analogy would be whether a particular pizza capricciosa is tasty (or perhaps burnt, or over-salted). Often, models and analogies can be useful, but they have weaknesses. In the case of the pizza, we can obtain direct data on the "test item" and determine their validity (too salty or not). In case of test data on unknown compounds, one does not know whether they are reliable, as crucial information is not available. To obtain confidence that the data are acceptable, one can run known compounds in parallel and determine whether the generated data are as expected (AC fulfilled). If this is the case, one can assume that also data on unknown compounds from the same run are correct. This procedure only works from test run to test run, not in general. For completeness, it needs to be mentioned that the setting of suitable AC can be part of the validation process, while the application of these AC is then part of each test method run.

[1] https://en.wikipedia.org/wiki/Test_method (accessed 02.10.2023)

Abbreviations
AC, acceptance criteria; CRO, contract research organization; EMA, European Medicines Agency; GIVIMP, Good In Vitro Methods Practice; NAM, new approach methods (or methodologies); PC/NC, positive control/negative control; PM, prediction model; QA, quality assurance; SL, suitability limits; SOP, standard operating procedure

2 Clarification of terms

Some terms are used differently depending on the field of study, the regulatory area, or the professional background. This issue applies in particular to "test methods", i.e., a general description of the methodological basis of new approach methods (NAM; Leist and Hengstler, 2018; Leist et al., 2012b; OECD, 2018; Schmidt et al., 2017; Krebs et al., 2019, 2020; Pamies et al., 2022). The terms "test", "assay", "method", or "test method" are often used interchangeably. Sometimes, especially in colloquial contexts, the term "test system" is used, which, however, more usually describes the test system setup used for a given test method (e.g., the type of cell line used in an *in vitro* test method). For practical reasons, we use here the term test method. This term makes a clear distinction from the term "test system" (which we consider one of the five major elements of a test method, Leist and Hengstler, 2018). According to Wikipedia[1], a test method is "*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*". "*To ensure accurate and relevant test results, a test method should be 'explicit, unambiguous, and experimentally feasible', as well as effective and reproducible*".

All NAM are test methods, but not all test methods are NAM. The term test method can also be used in other fields, like clinical chemistry/medicine, engineering, or psychology. Also, many animal-based methods (e.g., a 90-day study on rats) are test methods (but not NAM, Pallocca et al., 2022; Pallocca and Leist, 2022). Some prefer the term "method" alone. If clear definitions are in place, nothing is wrong with this. We prefer the term "test method" with respect to a broad understanding beyond toxicity testing. The term "method" strongly focuses on the procedure in colloquial language, while "test" is understood to be focused on the result.

Another confusion may arise from the term "non-test methods" in some legislations. This contains also *in silico* methods, like a QSAR-based prediction of genotoxicity. According to the above definitions, these are clearly NAM and test methods. If these definitions apply, it does not matter whether the test system of a test method is a cell culture, an enzyme, or a computer model. However, the focus of this overview is on "classical" NAM, based on cell culture/tissue models. Very complex organoids or *in silico* models may require special considerations and approaches dealt with elsewhere (e.g., Hewitt et al., 2015; Pamies et al., 2022) or in a subsequent part of this series.

Accordingly, some official definitions are that "*a test can be considered an observation or experiment that determines one or more characteristics of a given sample, product, process, or service*"[1]. The purpose of testing involves a prior determination of expected observation and a comparison of that expectation to what one actually observes. Thus, a test is outcome oriented. Also, a method is defined as a procedure or process for attaining an object (goal). For many, the word places a higher weight on the procedure than the outcome. All readers should feel free to use their favorite terminology if the overall understanding is clarified here. The test method we refer to is *a procedure that aims to answer a specific question, is explicitly defined, uses a test system and applies a specific exposure scheme to it, and obtains data by using a specific endpoint. It also includes a data interpretation procedure (prediction model) to define the outcome regarding the question asked (e.g., toxic or nontoxic)*.

The most important term for this article is "acceptance criteria". AC may refer to test methods, method elements, or data. Thus, they may be defined in various ways. A stringent definition is used in GIVIMP or by some validation bodies. Their definition refers to the procedure used to control individual runs of a test method. We will particularly focus on AC following this definition, but also give an overview on how the concept may be applied more broadly. In general, the principle of AC is meant to allow a decision on whether the results from a test method are acceptable or not. Some of the broader sense aspects of AC may also be termed "suitability limits (SL)". The take-home message is that AC/SL can be defined for different elements of a test method or the entire test method or even for the choice of a particular test method to answer a given question. All different types of AC can improve the reliability of data from test methods and confidence in their validity. However, the need and extent of added value to define various types of AC may differ from test method to test method.

3 Non-biological examples of AC

In everyday life, we know many acceptance criteria, some explicit, some implicit, some often not recognized as such, but important to rationalize (Fig. 1): Starting a car has AC, like checking seat belt closure and positioning of mirrors. If AC are not fulfilled, the main procedure (i.e., starting of the car) may still be initiated, but driving without a seat belt and adjusted mirrors is considered irresponsible and therefore not acceptable. In the future, car AI systems may prevent the ignition process if AC are not met. Another everyday example is the expiry date of food. It does not give any information on the actual quality of the food; it provides only limited information on the freshness of the item, and it is not related at all to the quality of the food production process. However, it ensures a minimum quality standard for the particular food item (i.e., the milk carton one considers buying) considered (not milk as a food class).

The way AC are used in everyday life is not fundamentally different from many scientific AC. Another example refers to driving license tests. In some countries, one of the AC for enlisting for the driving license test is the requirement of having attended a first aid course. This aspect is not related to the quality of driving. However, the test cannot be taken if the AC is not fulfilled.

Another example that gives indications of typical AC characteristics is the result of a clinical chemistry blood test. The report provides the test method, the test result, the range of values considered normal, and a flag for values that did not pass the AC (i.e., that are outside the statistically defined normal range).

4 Necessity of acceptance criteria

Critics might remark that AC are "yet another trouble/work, when setting up a test method". Anything that requires extra effort needs a justification. So, why and when are AC needed? The answer becomes clearer if the question is asked in a different way: what is lost (or missed), if no AC are set? For the answer, two levels need to be considered: one concerns the data produced (within a lab or project); the second concerns the use of such data by external stakeholders.

First negative consequence of not applying AC: Without AC, a lot of nonsense data (wrong test results) may be produced all the time and everywhere. Of course, also data from test methods without AC may be highly reliable, but the important point is that this cannot be known for sure if potentially critical factors (e.g., viability parameters in a cell culture system) have not been assessed. Therefore, the fact that there are good data from test methods without AC does not prove they are not needed. Many test methods may be highly robust and therefore insensitive to critical factors (i.e., their results will almost always be acceptable because any theoretically existing acceptance criteria would be fulfilled in practice by most test method runs). Or AC are not made explicit but are implicitly present. An example of this situation (typical for academic work) is that data are filtered or discarded if they "look very strange"; or if cells looked strange, contaminated, or dead. Also, if the analytical method does not run well (machine problems, operator problems, etc.), data will be discarded, filtered or flagged. If the assay protocol was violated or if controls showed strange behaviors, data will be flagged. Hundreds of these examples are known in the scientific community, and the adherence to these implicit rules (often unconsciously) can lead to high data quality. It can, but it does not always. And when it does, it does not do so consistently to the same degree. The reason is because such implicit AC are not explicitly defined and are therefore highly prone to subjectivity, varying for one researcher between Monday and Friday (depending, e.g., on the level of optimism), and bound to differ between operators, labs, and institutions. They may also give rise to (confirmation) bias (e.g., unexpected/"unwanted" results may seem "strange" more often than expected/ "desired" ones). Differences are common concerning the quantitative levels considered acceptable, the stringency used for applying implicit AC, or the measures taken when they are not fulfilled. In some settings, such variability is acceptable. In others, it is problematic.

Second negative consequence of not applying AC: If data are used for decisions important for human health or associated with heavy financial burdens, then decision-makers require data sets they can trust, i.e., with more defined levels of variability and uncertainty. This means that explicit AC are necessary to comply with regulatory requirements in such cases. The situation is aggravated by the fact that modern cell-based test methods have become ever more complex. Where simple cell lines were common, now genetically modified cells, stem cells and their differentiated progeny, and also complex organoid cell systems are used (OECD, 2018). For all of these, the factors that can vary have multiplied, and thus, there are new demands for quantitative QA of all elements of a test method and each test run.

The answer to why we need AC is thus straightforward for some areas: otherwise, the data from the cell-based test methods will not be acceptable to key decision-makers (e.g., chemical regulators). In parallel with regulatory demands, also academic research is more and more committed to rules that ensure the reproducibility of research data. In this context, it has become clear that it is not sufficient to define increasingly exact protocols or SOPs as handling rules. The necessary additional element to ensure data quality is that there are quantitative criteria for when test methods and data therefrom are acceptable, i.e., in which cases they can be used with sufficient confidence. Notably, this must be distinguished from other important quality issues such as (i) the validation of a test method before it is applied and (ii) following best practices for assay development and application (overview given in GIVMP (OECD, 2018)).

To summarize, AC criteria are necessary for different reasons and serve several purposes. The following three are the critical target purposes:

1. *Avoidance of unnecessary work.* This follows the "fail early principle in QA", meaning that there is no point continuing with a method if, e.g., half the cells are dead (or not performing).
2. *Codification of experience and performance standards*. In this context, AC aid in the transfer of a test method. They ensure that the data generated meet some pre-defined criteria. This aspect is important to ensure that data generation is similar on a study-by-study and lab-by-lab basis, thereby providing the necessary degree of standardization required, e.g., for regulatory purposes.
3. *Performance monitoring*. The rate of fulfillment of AC allows monitoring within test series and over a long time across many test series. This feature is a fundamental element of quality control.

5 Overview of the "AC family"

AC can refer to various test method elements, to the overall test method, or to generated data. There is no limit, and no absolute requirement, for many of these categories, but it is important that potential usefulness is considered during the method choice, establishment, validation, and application. To facilitate an overview, we outline the different types of AC and related concepts that may be considered.

The first principle is that AC need to make sense in the data usage context and the method's overall objective. They need to be set so that a method's most important variables are controlled to increase confidence in the data. In different scientific cultures (e.g., academic vs regulatory or engineering vs biology), there are different views on whether AC are an inherent method feature. These views are influenced by the stringency of requirements for confidence in the data generated and by the connection between method developer and user. In an academic setting, the method user may be identical to the developer. Thus, the opera-

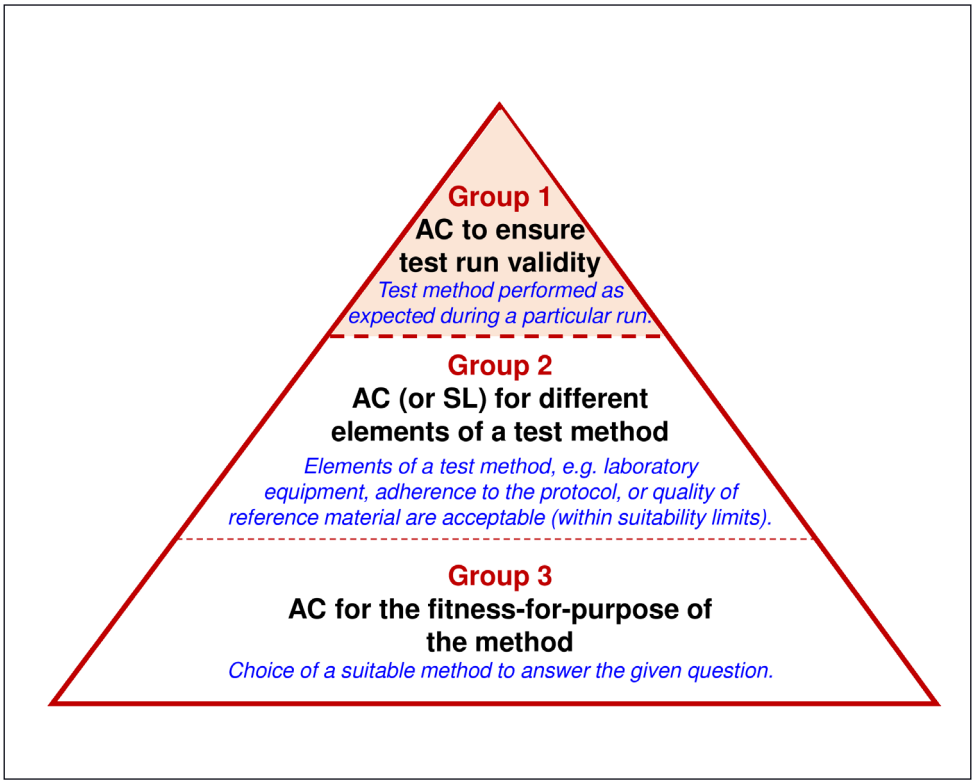

Fig. 2: Different categories of acceptance criteria (AC)
The overall concept of AC may be applied broadly. This includes the definition of what is acceptable as an outcome and setting action measures, depending on whether this criterion (AC) is fulfilled or not. Group 1 refers to the core definition of AC, e.g., for NAM. These AC have the main purpose to assure the validity of individual test method runs. Depending on the field, the test method, or the application, other AC may be defined. One group refers to some elements of a test method, another to the suitability of the method to answer the question at hand. Including these groups in a general thinking is useful for many purposes. To distinguish them better, they may be labelled suitability limits (SL) or fitness-for-purpose criteria.

tor (user) would have intimate knowledge of the method's performance and variability and he/she would apply many intrinsic AC. This would lead to a reduced requirement for formalized and explicit AC. On the other side of the spectrum is a NAM used in a regulatory setting and performed by a contract research organization (CRO). The data recipients (e.g., regulatory authorities) neither know the method well nor can they judge the performance, the operators, etc. In such cases, formal and explicit AC are highly valued to ensure quality and inspire trust in the data.

Roughly, the broad panel of potential AC may be assigned to three groups (Fig. 2).

Group 1: AC to assure the validity of test runs

If a narrow definition is applied, this is the core application of AC. *Vice versa*, these AC can take such an important quality assurance role that they need to be defined within the method description (SOP). They would also be checked during the validation process and be an essential part of the validated method. Such criteria are essential for allowing, e.g., lab-to-lab transfer with sufficient confidence that the data from different labs are comparable. These AC mainly refer to the data generated by known control substances (Aschner et al., 2017). In addition, they may include basic confirmations that the test system is functioning, and that the generated data and the overall data structure are not entirely outside the range known from historical controls.

Notably, separate research fields have established their own terminology and procedures. In many cases, the term AC is not applied, although the principle and way of thinking is similar. For instance, data may be generated by the questionnaire method in psychology or nutrition research. It is common practice to have such test methods validated (and to publish the validation results). The analogy of an AC is the following: For the assessment of test method runs (sample of filled questionnaires in a particular study), reliability and consistency criteria are set (e.g., Cronbach's alpha), and questionnaires can be discarded based on the outcome of this assessment (e.g., Cronbach's alpha < 0.6).

Group 2: AC to ensure that different test elements are within a suitable range

This set of AC is often defined less stringently and may also be termed "suitability limits" (SL). For instance, it helps to define whether chemical purity or apparatuses used are still within an SL. In a broader sense, this group may also deal with the acceptability (or better: SL) of protocol deviations. Even with a perfect protocol, some form of deviation will always happen. The question is to what extent a deviation can be accepted.

One example is the exposure time. An SOP may define a measurement after 24 h, but it usually does not specify which deviation from this time is acceptable (e.g., ± 10 min or ± 1 h?). By themselves, deviations do not mean that AC are automatically missed. However, the degree of deviation should remain within acceptable limits, or some action needs to be taken. In this sense, the concept of setting suitability limits for protocol deviations has clear similarities with the concept of AC.

Group 3: AC to ensure the method is fit-for purpose
This set of AC differs from groups 1 and 2 fundamentally because it is not applied daily or per test run. It refers to the choice of the method as such and whether the method is acceptable for the given purpose. For instance, the AC defined by the European Medicines Agency (EMA) refer to the method's suitability given certain product characteristics. A method suitable for one product may not be suitable for another (e.g., because the requirements for sensitivity, specificity, and linearity of the analysis depend on the product). The applicability domain of a given test method may be considered such an AC. A test method may be principally unsuitable for use with substances of high lipophilicity, for nanoparticles, or for certain chemical structures, etc. If this is defined, it also should be defined which consequences must be taken if a compound is outside the applicability domain (e.g., not running the test). Notably, often applicability domains are not sharply defined or are not comprehensive. There are also many examples of uses of assays outside their applicability domain or outside the area they have been validated for, but the way data from such test runs should be handled is often not generally defined.

6 A recent practical example for AC

Recently, a report was assembled on the validation of a battery of mechanistic methods relevant for the detection of chemicals that can disrupt the thyroid hormone system (Bernasconi et al., 2023). The AC are key components in this validation study, and many details of setting and application of AC are given. The AC considered were specified as falling into the three groups. Group 1 AC are inherent to the method and are established to validate that the method performs as expected (validity of test runs). They are based on data generated from known control substances, and occasionally on the test system's data structure. Group 2 AC confirm that different elements of the test are in a range suitable for accurate results. These suitability limits help determine whether chemical purity or apparatus used are within acceptable limits. Importantly, deviations must remain within allowable limits. Group 3 AC refer to the overall fitness-for-purpose of the method. They assess whether the method is acceptable for its intended use.

7 Conclusion

There is a lot of discussion on reproducibility issues in academic research, as well as on the confidence in data generated from NAM in the field of toxicology. One aspect of this, quite readily improvable, is to provide more complete and comprehensive information on methods used to generate the data (Leist et al., 2010). Another measure is to provide clear evidence for the robustness of the methods used as well as their relevance (or fitness-for-purpose) and predictivity. This may be achieved by standardized descriptions of test methods, e.g., using the ToxTemp (Krebs et al., 2019), by defining an SOP (Krebs et al., 2020; OECD, 2018), and by some form of validation procedure, such as the definition of method readiness (Bal-Price et al., 2018; Lanzoni et al., 2019, Schmeisser et al., 2023; Marx-Stoelting et al., 2023; EMA, 2008) or a formal validation, as specified in various legislations (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). So far, so good. Up to this point, there is also a broad understanding in many subgroups of stakeholders. What is less clear to many is that this is not enough to ensure data reliability. For specialists in the validation procedure and for those needing to deal with test data in a regulatory context, it is necessary to see evidence that the method not only works (is validated) in principle, but also that the acceptability of each test run is verified and documented. This is done by defining and applying AC. AC are so important for the generation of reliable data that they have become part of a method and of the validation procedure in the field of NAMs applied in a regulatory context.

With part I of this article series, we want to communicate the concept of AC to a broader community and promote its application in wider fields. The main message is that adherence to an SOP and comprehensive validation of the test method cannot guarantee that each test run produces data within the acceptable range of variability and with the precision and accuracy determined during the method validation. An extreme consequence of this principle is that the same method, run in different labs without harmonized/standardized AC being applied, may yield different results. Or the same method run in one lab, but with different AC, may also yield different results. The reason for this is that AC are used to discard certain sets of data (those where AC are not met), and thus the overall data structure changes, depending on the setting and application of AC.

We also indicated that the concept of setting and using AC may be applied not only to test runs of whole test methods, but also to individual elements of a test method or even the method choice. This will be further detailed in part II of the series. Moreover, the procedure on how to define and use AC will be detailed.

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., Langezaal, I., Bartnicka, J. 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. https://data.europa.eu/doi/10.2760/862948

EMA (2008). Qualification of novel methodologies for drug development: Guidance to applicants. EMA/CHMP/SAWP/72894/2008. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-novel-methodologies-drug-development-guidance-applicants_en.pdf

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/46pvkOC

Hartung, T. (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. 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

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, Suppl 1*, 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*, OECD Publishing, Paris. 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

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

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

Patterson, E. A., Whelan, M. P., 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

Conflict of interest

The authors have no conflicts of interest.

Data availability

No datasets were generated for this manuscript.

Acknowledgements

This work was supported by the BMBF, INVITE2, the Land BW (BW-3R) and the University of Konstanz. 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).