Organisation for Economic Co-operation and Development

ENV/CBC/MONO(2023)14

Unclassified

English - Or. English

4 July 2023

ENVIRONMENT DIRECTORATE
CHEMICALS AND BIOTECHNOLOGY COMMITTEE

Cancels & replaces the same document of 28 June 2023

Report on the WNT Workshop how to prepare the Test Guidelines Programme for emerging technologies

Series on Testing and Assessment
No. 378

JT03523040

OECD Environment, Health and Safety Publications
Series on Testing & Assessment
No. 378

Report on the WNT Workshop how to prepare the Test Guidelines Programme for emerging technologies

IOMC
INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS
A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate
ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT
Paris 2023

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 38 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in eleven different series: **Testing and Assessment; Good Laboratory Practice and Compliance Monitoring; Pesticides; Biocides; Risk Management; Harmonisation of Regulatory Oversight in Biotechnology; Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents;** and **Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

For this and many other Environment,

Health and Safety publications, consult the OECD's

World Wide Web site (www.oecd.org/chemicalsafety/)

or contact:

OECD Environment Directorate,

Environment, Health and Safety Division

2 rue André-Pascal

75775 Paris Cedex 16

France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

Contents

Report of the WNT Workshop on "How to prepare the Test Guidelines Programme for emerging technologies?"

Introduction

1. The Working Party of the National Coordinators of the Test Guidelines Programme (WNT) held a workshop in December 2022 on how to prepare the Programme for emerging science and technologies. A proposal for the workshop and preparatory steps had been presented and agreed in April 2022 at the annual meeting of the WNT. The workshop was attended by 56 participants from 22 delegations (see Annex 1) and chaired by Michael Oelgeschlaeger (Germany).

2. The report summarises the rationale for organizing the event and the preparatory steps taken in the preceding months, the workshop development, the outcome of discussions in breakout group and plenary sessions, followed by the workshop conclusions and next steps.

Setting the scene

3. In April 2022, a concept note for a workshop on emerging science and technologies and opportunities for the OECD Test Guidelines Programme was discussed. Emerging science and technologies cover a vast array of new approach methodologies[1] (NAMs), including but not limited to e.g., innovative ways to combine (data from) *in vitro* methods, biomarker endpoints involving gene expression, etc. Some of these methodologies may be amenable to standardisation. In parallel, regulatory frameworks are evolving to increase use of data coming from NAMs to fulfil regulatory requirements. Emerging science and technologies of common interest will offer opportunities to develop, standardise and harmonise novel methods and approaches into Test Guidelines (TG), especially for complex hazard endpoints.

4. Over the past few years, the WNT has reviewed a number of projects to develop innovative methods, both for stand-alone application and to be used in combination in integrated approaches to testing and assessment (IATA) and/or defined approaches (DA). These project reviews revealed that the WNT faces new issues that may require new ways to conduct the validation, the technical review and the standardization procedures for TG development. In order for these new and emerging technologies to be amenable to TG development, the OECD TG Programme may need to adapt certain processes and guidance, where appropriate, to enable prioritising and adoption of the more promising methods, while

[1] New Approach Methods includes a variety of in chemico, in vitro, and in silico methods that have emerged as a consequence of the expansion in technology used to investigate molecular biology and data science, as well as in vivo approaches that can help bridge data gaps (Chemicals And Biotechnology Committee, November 2022 [ENV/CBC(2023)12]).

maintaining standards to ensure robust and reproducible approaches to address regulatory needs and ensure Mutual Acceptance of Data (MAD). This workshop was an initial opportunity to reflect on current practices and identify areas where further work is needed.

Workshop preparation

5. A Steering Group of National Coordinators was established to help the OECD Secretariat develop preparatory material and consult the WNT as needed. The WNT agreed to organize a series of webinars on priority topics relevant to innovative methods in chemicals testing and assessment. The WNT was consulted to rank topics proposed by the steering group and 6 topics were selected for webinars between June and November 2022.

No	Title	Date	Speakers
1	Extracting the essential principles of validation and good in vitro method practices for NAMs (i.e. NAMs intended to become Test Guidelines)	29 June	Valerie Zuang and Ingrid Langezaal (EC-JRC) Laura Rossi (ECHA) Anna Lowit (US EPA) Deborah Ratzlaff (Health Canada) Elijah Petersen (NIST)
2	Ecotoxicology: new methods and approaches for cross species extrapolation tools	6 July	Henrik Holbech (SDU, Denmark) Dries Knapen (University of Antwerp, Belgium) Carlie Lalone (US EPA) Jessica Ewald (McGill University, Canada) John Colbourne (Birmingham University, UK)
3	Scientific and test method readiness of emerging technologies: criteria, examples and experience	31 August	Miriam Jacobs (UK HSA) Marcel Leist (Uni.Contans) Elise Grignard (PEPPER, FR) Monique Perron (US EPA) Hajime Kojima (NIHS, Japan)
4	4.1 Reproducibility issues (from technical perspective) in toxicological studies (in vitro) and how they affect emergence of new approach methods	23 September	Agnes Karmaus (Inotiv) Katie Paul Friedmann (US EPA) Sofia Batista-Leite (JRC)
	4.2 Probabilistic modelling (making better use of quantitative information from in vitro assays and taking into account uncertainty): example of applications and where they might fit in data interpretation	29 September	Weihsueh Chiu (Texas A&M) Thomas Hartung (John Hopkins University) Joe Reynolds (Unilever)
5	Identifying reference chemicals and building curated datasets: what are the approaches, issues and learnings for the future	24 October	Laura Taylor (US EPA/ORD) Barbara Kubickova(UK HSA)
6	Current practices and challenges encountered in other standard-setting	15 November	Sara Walton (BSI, UK) Claudius Griesinger (EC-JRC)

	organisations		Monica Piergiovanni (EC-JRC) Rusty Thomas (USEPA/ORD)

6. These webinars were intended for informational purposes, and do not represent the collective views of the WNT. The steering group helped identify presenters; WNT members and invited subject matter experts gave presentations. There was no time for discussions during the webinars, but participants could use the chat for questions to the presenters. The Secretariat developed summaries of key points after each webinar, as preparatory material for the workshop. Recordings of the webinars are available on the following URL: https://www.oecd.org/chemicalsafety/testing/webinars-on-emerging-science.htm)

7. A month before the actual workshop, the steering group developed an issues paper to guide discussions at the workshop. The six issues described in Annex 2 were articulated with a short description (problem formulation) and a few questions were developed for consideration by the WNT before the workshop. Preliminary feedback on the issues was collected from the WNT and the responses were shared on e-mail before the workshop with participants.

Workshop development

8. On the morning of the first day of the workshop, the Secretariat gave a few presentations to contextualise the workshop discussions within the OECD Chemicals and Biotechnology Committee on-going conversation on the future of chemicals assessment. A summary of workshop preparations was made, including key points from the webinars.

9. There was appreciation of the utility of webinars. Specific aspects were noted: in relation to the identification of reference chemicals for specific endpoints, the difficulty and labour-intensive task (as automation is hardly possible) that negative chemicals identification represent. Another remark was made in relation to the utility of probabilistic approaches in data interpretation and the need to have more illustrations presented to the WNT to facilitate understanding and uptake. The Secretariat noted that probabilistic approaches of relevance to the WNT would be those that can be integrated in the prediction model in a TG, other than that, they would rather be of relevance to the discussions of the Working Party on Hazard Assessment and the IATA Case Study project.

10. The Secretariat gave a reminder presentation on the Mutual Acceptance of Data system, the reasoning, the scope, the meaning and implications (see Annex 3).

11. Then, members of the steering group presented the issues for discussion during the workshop (see Annex 4). During the first afternoon, four breakout groups were organized to discuss each of the first four issues and propose conclusions, or recommendations for further work. The last two issues in the Table above were directly discussed in plenary on the second day.

12. Issue 1: evaluating test method readiness (presented by Michael Oelgeschlaeger): from the webinars and preliminary feedback received, there appeared to be different understandings of the stage of readiness. A method can be *i)* ready for validation, or *ii)* ready for uptake into the Programme for Test Guideline development, or *iii)* ready for regulatory application. These different contexts determine the level of supporting information expected on a given test method. The first two cases *i)* and *ii)* are at the interface between method providers, and the optimization of the test method (e.g., researchers) and the Test Guidelines Programme, and efforts for clearer communication and minimum readiness criteria were recommended during the preliminary feedback. The dedicated webinar on this topic provided an excellent basis to build clearer communication. To the question of non-transferable/difficult-to-transfer methods, preliminary feedback was that such methods are generally not amenable to Test Guideline development because they will not be broadly available.

13. Issue 2: evolving the concept of performance standards (presented by Valerie Zuang/João Barroso): the term "standard" was defined as a level of quality to achieve, and something used as a measure in comparative evaluations. In the context of the Test Guidelines Programme, Performance Standards are defined by the essential test methods components, list of reference chemicals and target values for reliability and predictive capacity (ref. Guidance Document 34) and have been a means by which to compare similar methods with a previously successfully validated reference method. Experience shows that the concept and its application worked well for alternative methods that have been one-to-one replacement of an animal test result and/or between methods using the same technology and measuring the same endpoint. Recent experience with Key Events-based Test Guidelines and Defined Approaches Test Guidelines showed that while the technique applied (test system, response measured) vary, the biological process addressed may remain a common denominator (i.e., a key event or suite of key events). Regarding the standards to achieve a reliable method, the presentation reminded that guidance provided in the document Good In Vitro Method Practice (ref. GIVIMP) should be followed as it touches upon multiple aspects of in vitro techniques that are generally applicable for any method. Concerning the standard required to achieve recognition of a relevant method, several questions remain unaddressed yet: e.g., how should existing (animal and/or human) data be used to benchmark the standard against? Should there be minimum of reference values and if so, which ones (reliability, sensitivity, dynamic range) to achieve? How should biological/mechanistic relevance be considered in addition to existing reference data?

14. Preliminary feedback received was very constructive and generally indicated that all these questions are worth exploring further when revisiting the concept of Performance Standards. The biological relevance and mechanistic anchoring were generally thought to be pivotal in determine the standard for the method relevance; having a mechanism-specific pool of reference chemicals (including potency ranges and negatives chemicals) was mentioned. Preliminary feedback was consistent on the need for reproducible and transferable methods and methods that generate quantitative information (e.g. concentration/dose-response curves).

15. Issue 3: developing a dedicated section of the Test Guidelines collection for mechanistically relevant and reliable but non-standalone methods (presented by Tim Singer): as noted on multiple occasions, new methods will most likely need to be combined with other reliable sources of information in defined approaches or test strategies, to characterize a biological endpoint of regulatory interest. While such non-standalone methods may have well-established biological relevance, reliability and utility in the context of a battery or testing strategy, without a process to establish these characteristics formally, there is a risk that the confidence to accept such methods into defined approaches for example, may be insufficient, leading to low uptake. The consideration in this issue is whether the development of defined approaches, batteries of assays, IATAs would be impeded if the constituent methods have not been given the status of TG, even if technically validated as biologically relevant, reproducible and transferable? A second consideration was if the TG status would facilitate their integration into defined approaches and testing strategies, is there merit in creating a category of TGs for such non-standalone methods?

16. Preliminary feedback received before the workshop was not unanimous about the merit for a separate category of TGs, nor where would such methods fit if not in a separate category. It was noted that KE-based methods for skin sensitization are already included in existing Test Guidelines as non-standalone methods (TGs 442C, 442D, and 442E). In relation to the possibility of a lighter evaluation/review of non-standalone methods, preliminary feedback was not unanimous concerning the characteristics that are more important to establish than others; the exclusion of predictive capacity was left for discussion during the workshop breakout groups.

17. Issue 4: developing guidance in GD 34 on the validation of batteries of assays (presented by Miriam Jacobs): currently available guidance is minimalistic and mentions that individual elements of a test battery should be validated without specifying how or to what extent. The presentation reminded that for test batteries under development (e.g., thyroid in vitro assays, developmental neurotoxicity, non-genotoxic carcinogenicity), a large number of assays are considered but may not all be necessary to apply all the

time; common understanding around specific scenarios will be needed to develop defined approaches, with the view to keep testing at an acceptable level. As part of key considerations for validating test batteries, the role of quantitative information on variability/uncertainty and probabilistic modelling within the battery will need further consideration and guidance. Similarly, traditional measures of accuracy (% specificity and % sensitivity) would need consideration in the context of methods that generate continuous data in the context of a battery.

18. Preliminary feedback was received on the meaning of "validated" for individual elements of a battery, which for most would be a confirmation of the biological relevance, its reproducibility/robustness by using a set a chemicals representative of the domain of applicability, resulting in data that cannot be contested. From preliminary feedback received, there was support for mathematical/probabilistic modelling to better characterize uncertainties, and to address borderline calls by e.g., confirmatory testing. In relation to accuracy values (defined as accuracy in predicting results of the traditional animal test method), preliminary feedback generated mixed responses that definitely indicated more discussions are needed. Finally, in relation to the biological relevance, the use of key events on adverse outcome pathways and the characterization of normal/adaptive versus maladaptive response should help framing and interpreting the biological relevance of individual assays in the context of a battery.

19. Following these presentations, the workshop participants were invited to split in breakout groups for one hour discussion on each of the four issues. The chair and rapporteur were fixed for the issue they were assigned to, in a way that each group built on what the previous group discussed on the same issue.

Workshop discussions in breakout sessions and plenary, recommendations, and next steps

Evaluating test method readiness

20. Generally, groups agreed that the current procedures under the WNT (SPSF submission, review) to evaluate project proposals and readiness work well. However, efforts towards a better communication with the method developers' community would be welcome. Agreeing and communicating on guiding principles for considering readiness would encourage proposals that better align with the Programme and would anticipate questions from method developers. Such guiding principles should not be too prescriptive and avoid details so as to remain broadly applicable over time and across projects.

21. The information requirements in the SPSF should be made publicly available to let the methods developers familiarize with information to be provided and subsequently reviewed in a project proposal. Groups encouraged promotion of self-assessment by orienting the proponents to key aspects of GDs 34 and 211 and encouraging them to evaluate themselves the gaps; the idea of a web-based tool for self-assessment of method readiness was supported. A brochure directed towards the method developers' community should be developed to explain the expectations, guiding principles, procedures and other important aspects of the TG Programme.

Evolving the concept of Performance Standards

22. Groups discussed the issue of evolving the performance standards. Such evolution could allow innovative methodologies, that are not necessarily similar, to come forward with supporting information to demonstrate compliance. There was general agreement that concrete examples are needed to illustrate the possible evolution of the concept.

23. It was discussed that essential method components could be more generally defined to avoid narrowing down candidate methods; it would be useful to have orthogonal assays covered by the same performance standard (assays that can be used to confirm a result, based on same biology but different

methodology/technology). In some breakout group, it was suggested that the basis for defining the essential aspects could be independent from the technology while focusing on the biology, e.g. using the AOP concept to anchor new key event-based methodologies. Additional essential components could include markers of functional status (maturation, differentiation), stability of the test system.

24. In relation to reference chemicals, some were of the opinion that more important than the number of chemicals would be the coverage of the chemical space (different physical-chemical properties, different applications/sectors) and the identification of negative chemicals. Reference data should be as comprehensive as possible (e.g. including human data when available, it is noted that, not surprisingly, this is most often for pharmaceutical chemicals for human health endpoints), and not solely rely on animal data. Also, at different steps of method development, validation, transfer, use, the importance of components of blind testing should be maintained (or introduced if not yet the case) and kept in the records to maintain trust in the application of the method. Furthermore, as a test method becomes used more, the generation of new chemical data could be used to supplement the original reference chemical set, thus offering some flexibility in the selection of reference chemicals while adhering to agreed selection criteria.

25. It was discussed that target values of specificity and sensitivity would no longer be pertinent target values for methods that are not standalone replacements of an animal method, are mechanism-specific and contribute to a battery. Also, when methods generate continuous rather than categorical data, it will not make sense to determine their individual specificity/sensitivity. However, for all test methods, in vivo or in vitro, reproducibility over time, variability around cut-off values, statistical power of the test to detect the effect size of biological relevance, and the transferability of methodology are and will remain important measures to ensure that the method is robust, that results are reliable, and that the method is accurate and compliant with the target values set in the standard. It will be important to develop further guidance in this regard, and the discussions were very much reflecting that this is work in progress.

Non-standalone methods in a dedicated section of Test Guidelines?

26. For such methods that are not yet Test Guidelines, the biological plausibility and regulatory need envisioned must be clear even if the regulatory use is potentially distant. Many in the groups believed there are limited barriers for incorporating non-Test Guideline-based methods as information sources in a DA; at the same time mechanistic methods could still have their own TG even if there is no immediate regulatory requirement to satisfy.

27. Some groups thought that being a TG is not a pre-requisite for being included in a DA. There was however limited support for a ‘lighter’ process for the validation of non-standalone methods. A separate section in the collection of TGs for non-standalone methods does not seem to be necessary either.

Developing guidance for batteries of assays

28. A question was raised on the level of validation expected for individual elements/information sources of a test battery. There was general agreement that elements of a battery need to be well characterized, while the development of the battery itself can take place in parallel. A proposal was made by some groups to call a technical validation when the biological relevance of a method has been demonstrated (e.g., AOP framework, human relevance), when the reproducibility of the method over time has been shown and the transferability to one or more naïve laboratory(ies) has been successfully implemented, and when proficiency testing and blind testing have successfully been performed. This concept would allow greater efficiency in the validation process and at the same time would require laboratories offering testing services on innovative methods to properly evaluate and train before routine implementation to maintain trust in the laboratory capacity and in the method performance. Some in the groups thought more time is needed for careful reflection on the concept of technical validation.

29. A question was raised on the level of support to revise Guidance Document (GD) 34 for batteries of assays; workshop participants responded positively; a project proposal to revise GD 34 has been submitted (as an SPSF mid-November 2022) and will be the occasion to anchor further discussions on the matter. Beyond the issue of validation, there were discussions in the groups on regulators needs for clarity on the use of data from batteries of assays under various decision-making contexts, and how to deal with variability and uncertainty from individual methods when combined in a battery. The understanding and use of quantitative threshold data to trigger other key events-based assays in the battery, the understanding and use of statistical approaches to determine uncertainty/probability of an adverse outcome are notions that merit further reflection and guided discussions with appropriate subject matter statisticians, and should take into account that uncertainty from in vivo data hasn't been extensively quantified for many hazard endpoints.

30. In the near future, there should also be opportunities to work jointly with the Working Party on Hazard Assessment (WPHA) on IATA case studies where batteries of assays have been proposed or used for regulatory purposes/assessment and revisit the statistical approaches applied, and how uncertainty was addressed in relation to the regulatory decision at stake for that Member Country. This could be done prospectively or retrospectively if there is sufficient material.

Incentivising validation

31. There was general agreement that, regardless of validation processes and changes therein to gain efficiency, awareness-raising on financial support needed for the validation activities, e.g. focus on the conduct of the experimental validation work, not the earlier test method development work, is absolutely necessary to gain trust and regulatory uptake of innovative methods that are based on emerging science. The current funded research initiatives do not include funding the validation activities of emerging methods and approaches. There is a risk that regulatory uptake will be limited (i.e. cost of non-validation) if no resources are spent in demonstrating reproducibility, transferability and determining the coverage of the chemical space those new approaches can be applied for.

32. There was support to develop a public statement from OECD (WNT, with support from the Chemicals and Biotechnology Committee) to call for increased funding for method reproducibility and transferability; the call was issued on 23 January 2023 (LINK). There was also support to organize a workshop around Q4 2023 with the community of testing facilities (CROs, industry), test developers, validation bodies and funding agencies to explore efficient ways to sustain validation of new approaches and methods for regulatory applications, and thereby increase the pace of standardisation.

Better reporting of validation study results

33. A short introductory presentation provided a recap of existing principles and sources of guidance for reporting *(in vitro*) study results: FAIR[2] principles, RIVER[3] Guidelines (under development), GIVIMP[4]. Some of the typical issues that a validation report should address were mentioned:

[2] FAIR: Findable, Accessible, Inter-operable, Re-usable (source: https://www.go-fair.org/fair-principles/

https://www.nature.com/articles/sdata201618)

[3] RIVER: Reporting in vitro experiments responsibly (source: https://nc3rs.org.uk/our-portfolio/river-recommendations)

[4] GIVIMP: Good In Vitro Methods Practice (source: OECD (2018), No. 286 Series on Testing and Assessment, ENV Publications, OECD, Paris. https://www.oecd.org/chemicalsafety/guidance-document-on-good-in-vitro-method-practices-givimp-9789264304796-en.htm)

34. A project for improving the reporting of academic studies is also underway with the use of SciRap [5] under the Working Party on Hazard assessment, led by the European Commission Joint Research Centre.

35. An important source of information that could be improved is the validation report of test methods that are intended to become OECD Test Guidelines, and open access to those reports. Having sufficient details in the report helps determining the critical steps of the method implementation, understanding how issues arising have been resolved, in particular during the transfer phase.

36. The question of transparency in complex methods (e.g. computerized systems, combination of omics data) was highlighted, with a specific interest in the demonstration of the relevance to human biology. More guidance, as appropriate, could be provided in GD 34 related to elements to address during validation and in the report.

Next steps

37. The workshop proposed that a group should be established to work with the Secretariat on follow-up from the workshop recommendations. While suggestions for additional webinars were proposed by some members and will require scoping, others stressed the importance of depth dialogue. Webinars held in preparation for the workshop will be made publicly available, if individual speakers consent, with a disclaimer that content is for informational purposes and is not necessarily endorsed by the WNT. It was requested that in future, where webinars are intended to be made publicly available, that this be communicated to invited presenters at the outset. A meeting document on technical validation will be prepared for discussion for the next WNT. The WNT call for financial support for validation (in particular to demonstrate transferability) of new methods will be published in early 2023. There was also some interest in investigations on human biological variability for hazard endpoints of regulatory interest.

Annexes

1- List of participants

2- Issues identified for workshop discussions

3- Presentations of issues

[5] SciRAP (Science in Risk Assessment and Policy) http://www.scirap.org/

Annex 1- List of participants

Austria/Autriche

Dr. Martin PAPARELLA	*Toxicologist* *Institute of Medical Biochemistry* *Medical University of Innsbruck*

Canada

Mr. Tim SINGER	*National Coordinator for Canada – OECD Test Guidelines Programme* *Healthy Environments and Consumer Safety Branch* *Health Canada*

Czech Republic/République tchèque

Dr. Petra KUBINCOVA	*Senior Scientist* *Center for Ecology, Toxicology and Analytics* *Research Institute for Organic Syntheses*

Denmark/Danemark

Dr. Sofie CHRISTIANSEN	*Senior Researcher, National Coordinator for Denmark* *Research group for Molecular & Reproductive Toxicology, National Food Institute* *Technical University of Denmark*
Dr. Knud LADEGAARD PEDERSEN	*Advisor, National Coordinator for Denmark* *Chemicals Division* *Danish Environment Protection Agency*

France

Mr. Enrico MOMBELLI	*Ingénieur* *Institut National de l'Environnement Industriel et des Risques (INERIS)*

Germany/Allemagne

Dr. Tanja BURGDORF	
Dr. Michael OELGESCHLAEGER	*Experimental Toxicology and ZEBET, German Centre for The Protection of Laboratory Animals (Bf3R)* *German Federal Institute for Risk Assessment (BfR)*
Ms. Susanne WALTER-ROHDE	*National Coordinator (Environment)* *International Chemicals Management* *German Environment Agency (UBA)*

Hungary/Hongrie

Ms. Tímea TARNÓCZAI	*Department of Chemical Safety and Competent Authorities*

Japan/Japon

Dr. Yoko HIRABAYASHI	*Director of CBSR* *Center for Biological Safety Research* *National Institute of Health Sciences*
Dr. Akihiko HIROSE	*Senior Researcher* *Division of Risk Assessment* *National Institute of Health Sciences*

Korea/Corée

Jin Hee LEE	*Government Scientist* *National Institute of Food and Drug Safety Evaluation*

Ms. Seoyoon CHOI	*Scientific Researcher* *Toxicological Research Division* *National Institute of Food and Drug Safety (NIFDS)*
Ms. Taehee KIM	*Scientific Researcher* *Toxicological Research Division* *National Institute of Food and Drug Safety (NIFDS)*
Ms. Minjeong KIM	*Scientific Researcher* *Toxicological Research Division* *National Institute of Food and Drug Safety (NIFDS)*
So Young YUNE	*Senior Scientific Officer* *Korean Center for the Validation of Alternative Methods (KoCVAM)* *National Institute of Food and Drug Safety Evaluation (NIFDS)*
Nam Hee KANG	*Scientific Officer* *Korean Center for the Validation of Alternative Methods (KoCVAM)* *National Institute of Food and Drug Safety Evaluation (NIFDS)*
Mr. Yongkwon SONG	*First Secretary* *Environment* *Permanent Delegation of Korea to the OECD*

Luxembourg

Dr. Tommaso SERCHI	*Environmental Research and Innovation* *Luxembourg Institute of Science and Technology*

Netherlands/Pays-Bas

Dr. Betty HAKKERT	*Deputy Head Bureau Reach* *Department for Industrial Chemicals, RIVM Centre for Safety of Substances and Products* *National Institute of Public Health and the Environment (RIVM)*
Mr. Andre MULLER	*Senior Scientist, Bureau REACH* *Department for Industrial Chemicals* *National Institute of Public Health and the Environment (RIVM)*

Norway/Norvège

Dr. Sjur ANDERSEN	*Senior Advisor* *Norwegian Environment Agency*
Ms. Christine BJØRGE	*Senior Adviser* *Norwegian Environment Agency*
Ms. Nina TEKPLI	*Senior Advisor* *Norwegian Environment Agency*

Poland/Pologne

Dr. Mariusz GODALA	*Director* *Department for Good Laboratory Practice* *Bureau for Chemical Substances and Preparations*

Sweden/Suède

Dr. Nina AKERBLOM	*National Coordinator OECD Test Guidelines Development, Proposals for Classification and Restriction* *KEMI (Swedish Chemicals Agency)*
Dr. Anne-Lee GUSTAFSON	*National Coordinator OECD Test Guidelines* *Swedish Chemicals Agency*

Switzerland/Suisse

Dr. Lothar AICHER	*Regulatory Toxicology Unit* *Swiss Centre for Applied Human Toxicology*
Mr. Markus HOFMANN	*Deputy Head REACH & Risk Management section* *Chemical Products Division* *Federal Office of Public Health (FOPH)*
Dr. Petra KUNZ	*Scientific Officer, Ecotoxicologist* *Air Pollution Control and Chemicals Division* *Federal Office for the Environment FOEN*

United Kingdom/Royaume-Uni

Dr. Miriam JACOBS	*UK National Coordinator (Human Health) Toxicology* *UK Health Security Agency*
Dr. Hannah LITTLER	*Chemicals, Pesticides & Hazardous Waste | Environmental Quality Directorate* *DEFRA, UK*
Dr. Fatima NASSER	*International Chemicals Team, Environmental Quality* *Department for Environment Food and Rural Affairs*

United States/États-Unis

Dr. Harrill ALISON	*U.S. Environmental Protection Agency (EPA)*
Dr. Nicole KLEINSTREUER	*Deputy Director* *NICEATM* *National Institutes of Health (NIH)*
Mr. Charles KOVATCH	*U.S. National Coordinator* *Office of Chemical Safety and Pollution Prevention* *U.S. Environmental Protection Agency (EPA)*

EU/UE

Dr. João BARROSO	*Scientific Officer - Chemical Safety and Alternative Methods Unit* *Joint Research Centre* *European Commission*
Dr. Edoardo CARNESECCHI	*Data Officer Evidence Management Unit* *Risk Assessment Services (ENABLE) Department* *European Food Safety Authority (EFSA)*
Ms. Sharon MUNN	*Scientific Officer* *EURL ECVAM* *European Commission/DG Joint Research Centre*
Ms. Laura ROSSI	*Scientific Officer* *European Chemicals Agency (ECHA)*
Mr. Andrea TERRON	*Senior Toxicologist EUropean Food Safety Autority*

	Pesticide Unit *European Food Safety Authority (EFSA)*
Dr. Valérie ZUANG	*Scientific Officer/EU NC* *Joint Research Centre* *Commission Européenne - ENTR*

Brazil/Brésil

Dr. Luciene BALOTTIN	*Brazilian National Coordinator (Inmetro)* *Directory of Metrology Applied to Life Sciences* *National Institute of Metrology, Quality and Technology*

International Council on Animal Protection in OECD Programmes

Dr. Vera ENGELBRECHT	*Adviser* *PETA Science Consortium International e.V.*
Dr. Donna MACMILLAN	*Senior Strategist, Regulatory Science* *Research and Toxicology Department* *Humane Society International (HSI)*
Ms. Kristie SULLIVAN	*Vice President of Research Policy* *Secretariat, ICAPO* *Physicians Committee for Responsible Medicine*

Physicians Committee for Responsible Medicine

Dr. Eryn SLANKSTER-SCHMIERER	*Research and Regulatory Affairs*

OECD/OCDE

Ms. Patience BROWNE	*Principal Administrator, Pesticides* *ENV/EHS*
Mme Nathalie DELRUE	*Administrator, Test Guidelines* *ENV/EHS*

Ms. Mar GONZALEZ	*Administrator, Nanosafety and Outreach* *ENV/EHS*
Mme Anne GOURMELON	*Principal Administrator, Test Guidelines* *ENV/EHS*
Ms. Linda RUBENE	*Assistant to Head of Division and Chemicals and Biotechnology Committee* *ENV/EHS*
Ms. Magdalini SACHANA	*Policy Analyst* *ENV/EHS*
Ms. Lesley SMITH	*Assistant* *ENV/EHS*
Mr. Leon VAN DER WAL	*Administrator, Test Guidelines* *ENV/EHS*

Annex 2- Short description of the six issues identified for the workshop, and questions posed to the WNT

<table>
<tr><td colspan="2">Issue 1: Test method readiness. There are a variety of different test readiness evaluation criteria being employed by various groups examining emerging technologies. As a result, the WNT may be presented with methods deemed ready for test guideline development by potentially widely divergent criteria. Based on experience gained to date, there may be merit in considering whether there are common criteria that could be applied to promote a consistent approach, or alternatively, whether consistency is even needed at all.</td></tr>
<tr><td>Questions for issue 1:</td><td>1. Is the issue of test method readiness clear for you? Do you understand its importance?
2. Would it be useful to have an OECD agreed set of criteria for (self)evaluating method readiness before entering a validation programme?
3. Should the readiness criteria be specific to the proposed context of use? Which criteria already exist that can be used/adapted? (e.g. Bal-Price et al. DNT IVB; ECVAM credibility factors, etc. Jacobs et al 2020)</td></tr>
<tr><td colspan="2">Issue 2: Evolving the concept of Performance Standards. Performance Standards were first described in GD 34 with the following intention: “The purpose of performance standards is to communicate the basis by which new test methods, both proprietary and non-proprietary can be determined to have sufficient accuracy and reliability for specific testing purposes. These performance standards, based on validated and accepted test methods, can be used to evaluate the accuracy and reliability of other analogous test methods (colloquially referred to as “me-too” tests) that are based on similar scientific principles and measure or predict the same biological or toxic effect.” The three elements of Performance Standards described in GD 34 were 1) essential test method components, 2) minimum list of reference chemicals, 3) accuracy and reliability values.
With emerging technologies, methods coming forward for a specific toxicological hazard endpoint may present different unique characteristics (different procedures, different test system characteristics, different readouts), while offering very similar functionalities (e.g. same mechanistic relevance, same physiological processes/responses) and similar accuracy and sensitivity. The challenge is to define a priori a Performance Standard that can be used as a benchmark for various methods/approaches that provide equivalent information based on the intended purpose.</td></tr>
<tr><td>Questions for issue 2:</td><td>1. For new endpoints, is the concept of Performance Standard worth exploring further to enable different test systems/technologies to demonstrate equivalence, and be evaluated for their validity as an OECD Test Guideline method/approach?
2. What is essential for first starting to define a standard? And what could be optional, or developed as the project progresses? : a minimum list of reference chemicals? a first method/technology? A key / molecular initiating event (MIE)/physiological process on an (network of) AOP(s)? A definition of method purpose and how biological relevance could be demonstrated? Anything else?</td></tr>
<tr><td colspan="2">Issue 3: Developing a dedicated section of the Test Guidelines collection for mechanistically relevant and reliable methods that are not stand-alone. We observe that in vitro methods coming forward will not often offer a solution to a regulatory need/question on their own. However, such methods can be</td></tr>
</table>

<table>
<tr><td colspan="2">evaluated and acknowledged for their biological/mechanistic relevance, their reliability and generally their utility in the context of a battery/testing strategy. They are supported by mechanistic/AOP knowledge and/or a Detailed Review/Scoping Paper in a specific area of regulatory toxicology.

If such methods cannot be evaluated and given a status among the user community that qualify them for the characteristics mentioned above, it can be difficult to progress to the stage where they are given a role in a testing strategy or a defined approach but have been nonetheless demonstrated to be biologically relevant, reproducible and transferable and could be used as an information source in a WoE approach, an IATA or a future defined approach

A possibility could be to create an intermediate category of Test Guidelines (e) that can constitute elements of a whole battery of assays, and recognised as such in the dedicated section. In this way, sections 200 and 400 would remain in principle for Test Guidelines that provide a harmonized and accepted methods/approach for a defined regulatory question/need (e.g. as a Defined Approach).</td></tr>
<tr><td>Questions for issue 3:</td><td>1. Do we need such an intermediate repository for methods that are recognised as mechanistically relevant and reliable but not stand-alone, with a view to integrate them in a DA or an IATA?

2. Where do such methods belong? Is it important to give them the status of a standardised method, e.g. in a dedicated section of OECD TGs?

3. If the intention is for such methods to be used in batteries of assays, do they need to be evaluated independently, or would this add an unnecessary step to the process (see Issue 4 below)? Alternatively should we envision a “lighter” independent review process? (e.g. with a narrow focus?)</td></tr>
<tr><td colspan="2">Issue 4: Developing guidance in GD 34 on validation for batteries of assays: there is preliminary guidance in GD 34 that says that all elements of a battery should be individually validated and the document then says comprehensive guidance has not been developed. The future is much about combinations of methods/approaches and further guidance is necessary on how we go about validation of these, and what do we mean exactly with the validation of individual elements.

There will be multiple methods/approaches that can form different combinations for the same toxicological hazard endpoint. For the principle of Mutual Acceptance of Data to remain impactful in reducing duplicative testing (for obvious economic reasons), there is great interest in reaching agreements on testing strategies/batteries with explicit data interpretation procedures.

Taking the examples of in vitro thyroid disruption methods and the DNT in vitro battery, each of them comprises approximately 17-18 assays that each have a substantial cost. Overall, not all assays will need to be implemented all the time for chemicals tested for those toxicological hazard endpoints; there will be a prioritisation screening approaches and specific decision trees developed. Building a common understanding of specific situations/scenarios and subsequently building defined approaches will be important milestones to maintain the costs of testing bearable (case studies might be helpful to contextualise the situations).</td></tr>
<tr><td>Questions for issue 4:</td><td>1. For individual elements composing a strategy or a battery, what does it mean to be “validated”? For these individual elements/information source, what aspects of validation are important for the regulator?

2. In webinars 4.1 and 4.2, we’ve heard about making better use of quantitative information on variability to give an appropriate sense of uncertainty in the outcome. How is uncertainty dealt with at the level of a battery? What can be done to reduce uncertainty within a battery (e.g. computational approaches</td></tr>
</table>

	such as probabilistic modeling? Confirmatory/orthogonal assays?)? Can practical guidance be developed for developers of test batteries? 3. Should accuracy values (% sensitivity/% specificity) apply to individual elements composing a battey/strategy? a. Should reference data be from the relevant species preferably (i.e. human for test systems relevant to human health)? Case specific ? 4. What should accuracy measures mean in the context of continuous data for systemic hazard endpoints? (specificity and sensitivity can be determined when positive/negative outcomes are measured. For systemic toxicity ECx or NOAEC/NOAEL are measured). Some case study examples would be useful here. 5. How can biological relevance be considered when validating batteries of assays, e.g. by linking each assay with a biological process or key event?
Issue 5: Recruiting labs for validation studies (in the absence of financial incentive): OECD could conduct an analysis (via a contractor) of the ecosystem of laboratories, especially contract research organisations, test developers, validation bodies, public-private partnerships, other laboratories doing research and stakeholders, would help better understand who has an interest in taking part in validation studies. Various validation models exist, which are not mutually exclusive. To make validation more attractive, participation in validation studies may need to be rewarded better (academic recognition? facilitation of publication in scientific journals? certification of proficiency in implementing a new method?). Experience from the EU-NETVAL activities could be very valuable. Furthermore, for emerging technologies, the skills and equipment in the labs are becoming more and more specific: not all labs can implement all new methods, some labs are specialised, and it is better to ask a specialised/competent lab to implement a test to be sure that resources will not be wasted.	
Questions for issue 5:	1. Would a dedicated workshop with keyplayers (test developers, validation centres, contract research organisations, academic labs, funding agencies, etc…) be useful to better understand the interests, motivations and how validation studies can be better organised and financially supported in future? a. If such workshop was organised, what preparatory work should be done ahead? Would a scoping study identifying key players be helpful? 2. As an immediate action following the December workshop, would the WNT want to make a public statement calling for better funding of validation studies? Do you have suggestions for text?
Issue 6: Better reporting of study results: we heard on several occasions the need for a more systematic reporting of critical study parameters, quantitative information, different cut-off values where possible (rather than reductionistic response such as POS or NEG), confidence intervals and other quantitative measures of variability within test, publication of SOPs in databases (e.g. TSAR, and other open platforms). There are also several initiatives for the better reporting of academic studies (in vitro or other studies) . Such information could then be useful not only to evaluate the chemical, but also to use the data to build databases, build models, reference chemicals, etc. Moving forward in the field of new	

approach methodologies for systemic toxicities, there is an opportunity to make better use of these methodologies and collect more data.	
Questions for issue 6:	1. Do you support that the collection of additional data as described above, generated during prospective validation studies is important is important for prospective new approaches and methods? 2. Are there common elements to all in vitro methods that should be systematically required in the study report, in line with GIVIMP? 3. Are there other important features that come to mind?

Annex 3- Mutual Acceptance of Data

MUTUAL ACCEPTANCE OF DATA AND THE PRACTICE AS REGARDS OECD TEST GUIDELINES

OECD Secretariat
WNT workshop, 1-2 December 2022

OECD
BETTER POLICIES FOR BETTER LIVES

Why was the MAD system created?

- The MAD framework ensures the generation of high quality and reliable non-clinical test data for regulatory purposes
  - Developed in response to fraudulent studies submitted to regulators
  - GLP provides the quality standards
  - TGs provide the scientific standard
- Regulatory authorities receiving the data under the MAD know that particular quality and scientific standards were followed
  - e.g. they do not have to re-evaluate a test protocol to determine its robustness as it has consensus by countries via the TG programme

What documents are covered by MAD?

- OECD Test Guidelines and OECD Principles of Good Laboratory Practices are covered by MAD
- All other OECD documents are<u>not</u> covered by MAD:
  - Guidance Documents describing certain test methods;
    - The scientific standard for those methods was not met (e.g. validation was considered insufficient as per WNT standard, at that time…)
  - All other documents such as DRPs, Validation reports and peer reviews.

What does the OECD Council Act on MAD (1981) says?

- *"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".*

- Notes:
  - Data requirements are government prerogatives
  - Interpretation of test results is government prerogative
  - No repeat testing for the same data requirement
  - "Acceptance" does not automatically mean "use" of data

Regulatory needs: what is the practice?

- Regulatory needs and data requirements are country prerogatives
  - OECD Test Guidelines intend to address/respond to existing *or foreseeable* regulatory needs in relation to hazard assessment
- What kind of regulatory needs do OECD Test Guidelines address?
  - Determination of specific physical-chemical properties
  - Determination of specific environmental fate and behaviour properties
  - Determination of specific chemical mechanism of action/bioactivity
  - Provision of ADME data for model (QSAR, PBK) building
  - Toxicity data for screening and priority-setting
  - Toxicity data for identification of a hazard, a target organ
  - Toxicity data for characterising a hazard (dose-response curve and NOAEL/LOAEL)
  - Residues chemistry data

MAD and data requirements

- Although data requirements are countries' prerogative and MAD is about avoiding repeat testing
  - If data requirements diverge extensively between countries, MAD will erode as Test Guidelines data will not be accepted across countries having different requirements
  - Therefore, the more compatible/similar data requirements are between countries, the more beneficial MAD will be globally

Data interpretation: what is the practice?

- Although data interpretation of test results is government prerogative
  - OECD TGs often integrate transformation of raw experimental data (e.g. through prediction models, data interpretation procedures,...) to generate a test result that addresses more directly a regulatory need (e.g. identification of a hazard)
  - The data transformation/interpretation procedures implemented in the OECD TG are the outcome of countries' agreement to do so, in order to generate meaningful data, reduce room for (mis)interpretation and maintain a common level playing field;

- It remains countries' prerogative:
  - to use the stand-alone test result to satisfy a data requirement, or
  - to use the test result with other sources of information, in combination with other level of interpretation or criteria, that meet a country specific regulatory need,
  - to not use that test result if their data requirement cannot be satisfied with that test result (alone or in combination)

Annex 4- Presentation of issues relevant to the workshop

TEST METHOD READINESS

GERMAN FEDERAL INSTITUTE FOR RISK ASSESSMENT

20 YEARS BfR German Federal Institute for Risk Assessment

Test method readiness

Test method readiness

Ready for:

- Validation
- Uptake into the TGP WP
- Regulatory Application

Purpose:

- inform researchers / test developers about the expected expectation level of development of a method prior to validation
- Criteria should not be too heavy or too difficult to address, but should allow to clarify the state of development and robustness of the proposed test and to identify what needs to be done before validation

Existing readiness criteria: OECD Projects:

- New Scoping Document on in vitro and ex vivo Assays for the Identification of Modulators of Thyroid Hormone Signalling (ENV/JM/MONO(2014)23)

CATEGORY 1 Initial High Priority Considerations	CATEGORY 2 Assay Performance Considerations
1. Biological Plausibility 2. Extrapolation to humans, or broadly applicable across vertebrates/phyla 3. Availability of Resources 4. Reference Chemicals	5. Within-laboratory reproducibility 6. Between-laboratory reproducibility 7. Assay Variability 8. Accuracy 9. Assay Specificity 10. Assay sensitivity
CATEGORY 3 **Technical Capabilities**	**CATEGORY 4** **Other Practical Considerations**
11. Dynamic Range 12. Concentration test range 13. Detection/Adjustment of confounding factor and/or incorrect/ inconclusive measurements and/or other bias 14. Response Characterization:	15. Technological Transferability/Proprietary elements 16. Transparency of the method 17. Documentation of development and utility of the method.

- Chemical carcinogen safety testing OECD expert group international consensus on the development of an integrated approach for the testing and assessment of chemical non-genotoxic carcinogens (Archives of Toxicology (2020) 94:2899 –2923)

Category 1: Initial high-priority considerations	Category 2: assay performance considerations
• Endpoint addressed and intended purpose of the assay • Biological Plausibility • Extrapolation to humans • Reference chemicals • Availability of a detailed protocol / SOP • Within-laboratory reproducibility Reproducibility	• Assay variability • Accuracy • Assay specificity • Assay sensitivity • Consideration of confounding factors • Data interpretation and prediction model • Limitation of the test method
Category 3: technical capabilities for test methods	**Category 4: other practical considerations**
• Dynamic range • Concentration test range • Response characterization	• Availability of the assay & essential components • Intellectual Property rights • Cost of the assay and essential components • Through-put of the assay • Documentation of development & utility of the method

in silico methods ?

Existing readiness criteria: „Recommendation on Test Readiness Criteria for New Approach Methods in Toxicology: Exemplified for Developmental Neurotoxicity "

ALTEX 35(3), 306 -352. doi:10.14573/altex.1712081

The main categories are scored dependent on the score of the single sub-items.

Phase I	Max. score	UKN2 cMINC
1 Test system	10	9
2 Exposure scheme	3	3
3 Documentation/SOP	5	5
4 Main endpoints	4	4
5 Cytotoxicity	5	5
6 Test method controls	4	4
7 Data evaluation	4	4
Sum	35	34

Phase II	Max. score	UKN2 cMINC
8 Testing strategy	4	3
9 Robustness	4	3
10 Test benchmarks	4	4
11 Prediction model	4	3
12 Applicability domain	3	1
Sum	19	14

Phase III (optional)	Max. score	UKN2 cMINC
13 Screening hits	4	4
Score 0 = D Score 1 = C Score 2 = B Score 3 - 4 = A		
Sum		4

9 Robustness			4
9a Reproducibility	Data available on normal variation; information on factors affecting test variation is given	Historic control data on positive controls show normal range; known artefacts and shortcomings	1
9b Intra-lab	Data available from different operators and different test runs over longer time		1
9c Inter-lab	Data available on transferability / reproducibility in another lab		1
9d Historical controls	Data for PC and NC over time		1
10 Test benchmarks			4
10a Sensitivity (of the test)	Signal noise ratio (S/N) defined; Sensitivity information available	S/N based on adequate data sets. The S/N is used to determine the limit of detection. Additional measures: True positive rate, hit rate, sensitivity to detect a panel of positive controls, etc.	1
10b Specificity (of the test)	Tested with sufficient number and quality of negative controls	Additional measures: true negative rate, etc.	1
10c Acceptance criteria	Clearly defined and documented; Normal range of variation known	E.g., a given positive control has to reduce the main endpoint by at least 25%, otherwise test plate is discarded.	1
10d Response characteristics	Should the response be linear? What are the upper and lower limits?	Additional measures: mono-directional or bi-directional deviation defined, information on accuracy, precision, limit of quantification, etc.	1

The scores of the different phases are evaluated and result in the ranks of readiness.

Phase I Score	Phase I Grading	Phase II Score	Phase II Grading		Explanation of grading
< 7	D	< 4	D	D	Not ready at all
8 - 17	C	5 - 9	C	C	Substantial improvements required to be ready
18 - 28	B	10 - 14	B	B	Improvements required to be ready
29 - 35	A	15 - 19	A	A	Test method is close to ready or ready

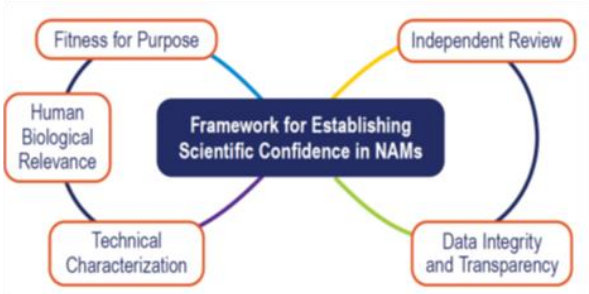
Existing readiness criteria: PEPPER selection of Methods for Validation

Self Assessment Questionnaire (SAQ)

- based on OECD GD , EURL-ECVAM Test pre-submission form
- seven sections that contain important criteria to assess readiness
  - Test method description (test system, the exposure scheme, technical requirements)
  - Operational readiness and data management
  - Reproducibility and transferability
  - Historical data
  - Standard Operating Procedures (SOPs)
  - Regulatory relevance of the method
  - Time and resources availability

Existing readiness criteria: „A framework for establishing scientific confidence in new approach methodologies"

Archives of Toxicology **volume 96**, pages 2865–2879 (2022)



Human biological relevance

- NAM reflects human biology (AOP, KE)
- Concordance with human responses

Technical characterization

- Reproducibility
- Transferability (if applicable)
- Applicability domain
- Limits of detection and quantification
- Accuracy

(not necessarily in comparison with animal data)

Existing readiness criteria: Hajime Kojima (JaCVAM, CBSR, NIHS, Japan)

Checkpoints prior to a validation study

- Objective and purpose of test method
- Need and benefits in comparison to those of the existing test methods
- Comparison of the test method with the essential test method components in the Performance Standards
- Biological and mechanistic relevance
- Test method protocol
- Appropriateness of the validation study management and conduct

Mission during validation study

Training of naive laboratories and transferability
Use of quality assurance system(s) during data generation
Within-and between-laboratory reproducibility
Predictive capacity
Applicability domain and limitations
Completeness of data and documentation

Important role of Experts / VMT to discuss:

- Objective
- Benefit
- Biological / mechanistic relevance
- Study plan, including acceptance criteria
- Chemical selection

Responses to Questions

Is the issue of test method readiness clear for you? Do you understand its importance?

- general acknowledgement of the importance
- In general: Need to distinguish between Test Guidelines and Guidance documents.

Would it be useful to have an OECD **agreed set of criteria for (self)evaluating method readiness before entering a validation programme**? Would fulfilling these criteria be a condition for project proposals to be submitted to the TGP?

- could be part of the SPSF and the evaluation process. Might help to reduced confusion which test methods are acceptable in the work plan and help to communicate the needs of the WNT to developers
- there might be a need for case specific approaches and concepts / difficult to define definitive criteria for all test types. One might develop a set of general criteria that could fit all, and one set of criteria that is specific for the type of test proposed. The criteria should also include open accessibility and transparency as guiding principles
- criteria should not be too strict, allow methods to improve
- development a comprehensive online template, that would be filled in in a stepwise approach

Who should independently evaluate the readiness level? (e.g. The WNT? The submitting country's NC? A dedicated expert group?) The organisation/entity that will coordinate the validation study? And at what stage/time point (e.g. prior SPSF submission or upon SPSF submission prior to WNT decision?)

- Self score in cooperation with NC (in preparation of SPSF)
- Expert group (prior to submission of the SPSF): existing or dedicated/ permanent Expert group
- online interactive tool/scheme

Responses to Questions

Would it be useful to have an OECD agreed set of **readiness criteria for regulatory application** to apply to methods that are not easily transferable (e.g. emerging/costly technologies) but nevertheless amenable to in-house validation/independent review and PBTG development? If yes, please consider addressing sub-questions a-d above for the case of question 3.

- Yes: comparable criteria should apply
  Case-by case decision
  confidence, context of use, and fit for purpose required
- No: TGs should always be transferable to ensure broad applicability and to allow verification of data

Should the readiness criteria be specific to the proposed context of use? Which criteria already exist that can be used/adapted? (e.g. Bal-Price et al. DNT IVB; ECVAM credibility factors, etc. Jacobs et al 2020)

- At least some criteria should be generally applicable to all types of methods
- Published criteria amended to be more generally applicable
- Some flexibility might be possible and case dependent, context of use and how the assay and it's validation approach supports that context of use should be considered

GERMAN FEDERAL INSTITUTE FOR RISK ASSESSMENT

BfR German Federal Institute for Risk Assessment

Thank you for your attention

20 YEARS Identify Risks – Protect Health

German Federal Institute for Risk Assessment
Max-Dohrn-Straße 8-10 ● 10589 Berlin, GERMANY
Phone +49 30 - 184 12 - 0 ● Fax +49 30 - 184 12 – 99 0 99
bfr@bfr.bund.de ● www.bfr.bund.de/en

EVOLVING THE CONCEPT OF PERFORMANCE STANDARDS

Evolving the concept of Performance Standards

OECD WNT Workshop

1-2 December 2022

Valérie Zuang

Joint Research Centre
European Commission

Common definition

standard

noun
plural noun: **standards**

1. a level of quality or attainment.
2. something used as a measure, norm, or model in comparative evaluations.

OECD Guidance Document 34

ENV/JM/MONO(2005)14

Performance Standards for Test Methods

41. The purpose of performance standards is to communicate the basis by which new test methods, both proprietary (*i.e.*, copyrighted, trademarked, registered) and non-proprietary can be determined to have sufficient accuracy and reliability for specific testing purposes.

These performance standards, based on validated and accepted test methods, can be used to evaluate the accuracy and reliability of other analogous test methods (colloquially referred to as "me-too" tests) that are based on similar scientific principles and measure or predict the same biological or toxic effect

European Commission

OECD
BETTER POLICIES FOR BETTER LIVES

A to Z
Search oecd.org

OECD Home | About | Countries | Topics | Français

u Testing of ch

Series on Testing and Assessment:

Element 1. Essential Test Method Components:
These consist of essential structural, functional, and procedural elements of a validated test method that should be included in the protocol of a proposed, mechanistically and functionally similar test method. These components include unique characteristics of the test method, critical procedural details, and quality control measures.

Element 2. List of Reference Chemicals:
These are used to assess the accuracy and reliability of a proposed, mechanistically and functionally similar test method. These chemicals are a representative subset of those used to demonstrate the reliability and the accuracy of the validated test method.

Element 3. Target Values for Reliability and Predictive Capacity (Accuracy):
These are the performance requisites that should be achieved by the proposed test method when evaluated using the minimum list of RC, i.e. reliability and predictive capacity that should be achieved by the proposed test method when testing the RC.

Skin irritation methods using RhE

Within-laboratory reproducibility

An assessment of within-laboratory reproducibility should show in one single laboratory, a concordance of predictions (UN GHS Category 2 and No Category) obtained in different, independent test runs of the 20 Reference Chemicals equal or higher (≥) than 90%.

Between-laboratory reproducibility

An assessment of between-laboratory reproducibility is not essential if the proposed test method is to be used in a single laboratory only. For methods to be transferred between laboratories, the concordance of predictions (UN GHS Category 2 and No Category) obtained in different, independent test runs of the 20 Reference Chemicals between a minimum of three laboratories should be equal or higher (≥) than 80%.

Table 4: Required sensitivity, specificity and accuracy values for similar or modified RhE test method to be considered valid to discriminate skin irritants (UN GHS Category 2) from non-classified (UN GHS No Category)

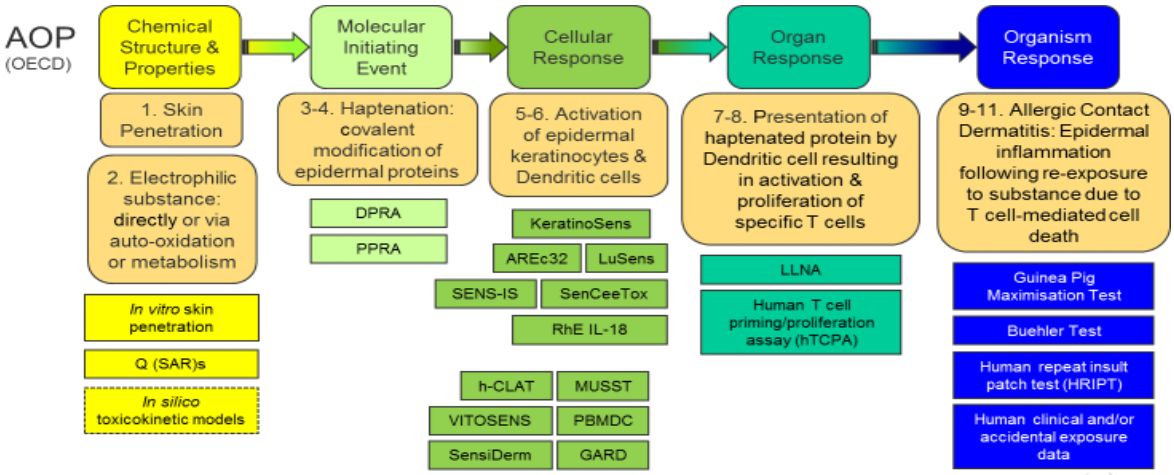
Sensitivity	Specificity	Accuracy
≥ 80%	≥ 70%	≥ 75%

European Commission

Evolution

European Commission

Matching Methods with Key Events/Key Characteristics etc.



12 ways to assess skin sensitisation

	Annex 1 of the second Guidance Document includes the following 12 case studies
1	AOP –based "2 out of 3" wei… hazard identification
2	Sequential Testing Strategy (
3	A non-testing Pipeline appro…
4	Stacking meta-model for skin
5	Integrated decision strategy f
6	Classification consensus mo…
7	Sensitizer potency prediction … KeratinoSens™ data
8	The artificial neural network …
9	Sensitizer potency prediction based on Key event 1+2+3: Bayesian Network ITS/DS for hazard and potency identification of skin sensitizers
10	Sequential testing strategy (STS) for sensitising potency classification based on in chemico and in vitro data
11	Integrated testing strategy (ITS) for sensitising potency classification based on in silico, in chemico, and in vitro data
12	DIP for skin allergy risk assessment (SARA)

Unclassified ENV/JM/MONO(2016)29/ANN1

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development 27-Oct-2016

ANNEX I: CASE STUDIES TO THE GUIDANCE DOCUMENT ON THE REPORTING OF DEFINED APPROACHES AND INDIVIDUAL INFORMATION SOURCES TO BE USED WITHIN INTEGRATED APPROACHES TO TESTING AND ASSESSMENT (IATA) FOR SKIN SENSITISATION

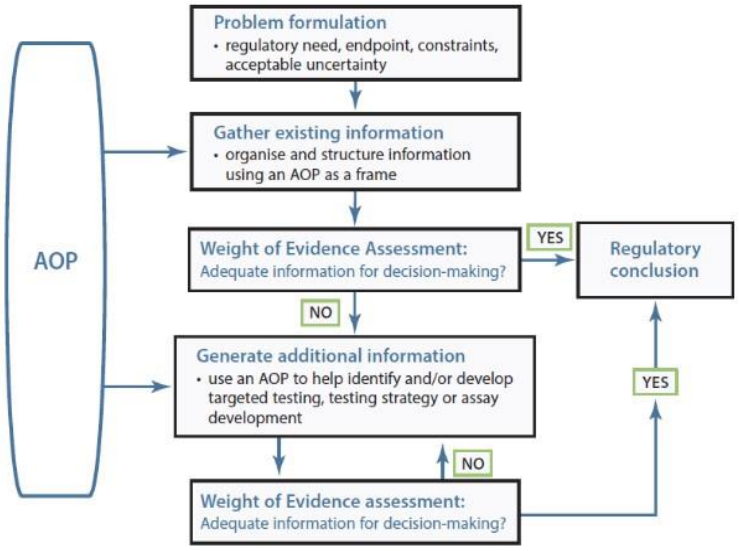
Series on Testing & Assessment
No. 256

OECD
BETTER POLICIES FOR BETTER LIVES

European Commission

Integrated Approaches to Testing and Assessment

An **IATA** integrates and weights all relevant existing evidence and guides targeted generation of new data where required to inform regulatory decisions



Evolution of Performance Standards

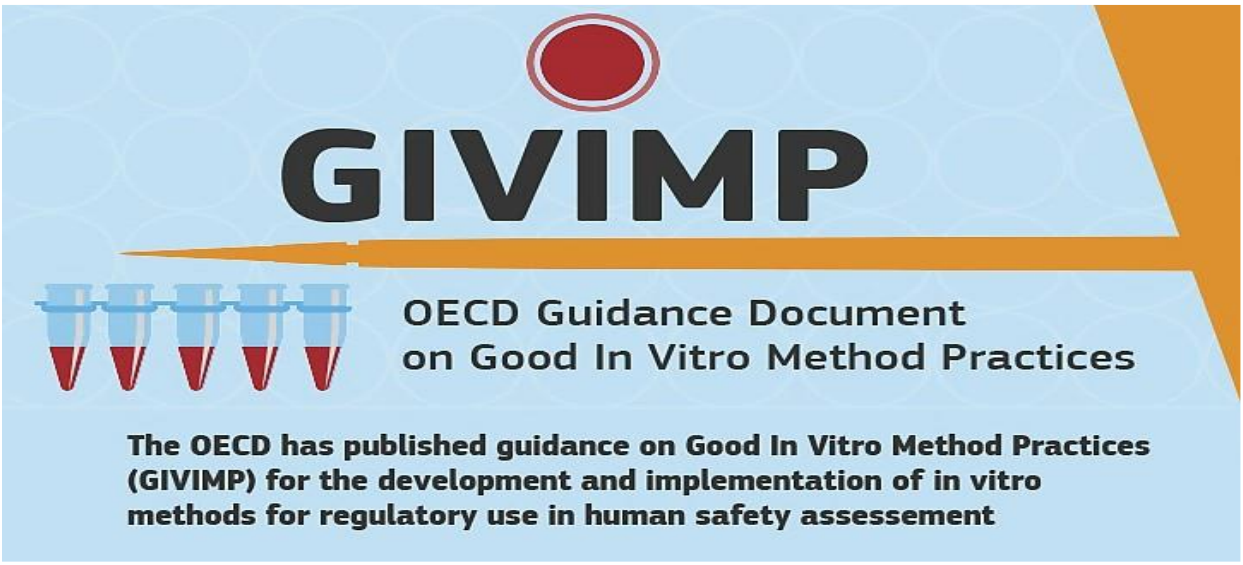
To cover classes of methods that are different from a technical point of view?

Key event-based test guidelines grouping test methods which are less similar from a technical point of view but which all predicts the same key event.

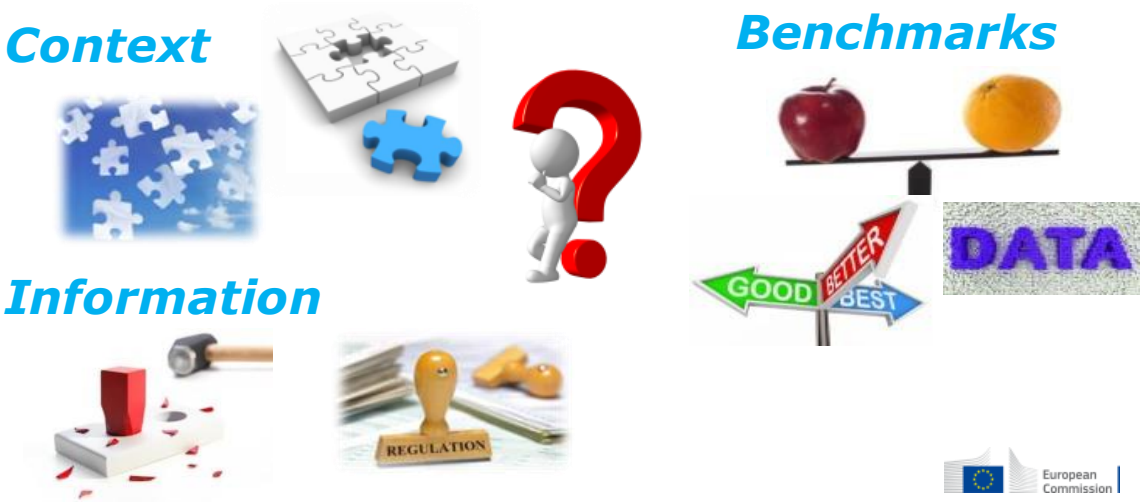
Include the standards in the core TG (and less related to the test method in appendix).

European Commission

Standards to ensure **Reliability**



Standards to ensure **Relevance**



Preliminary WNT replies on issue2

For new endpoints, is the concept of Performance Standard worth exploring further to enable different test systems/technologies to demonstrate equivalence, and be evaluated for their validity as an OECD Test Guideline method/approach?

YES worth exploring

Preliminary WNT replies on issue2 (1/5)

What is essential for first starting to define a standard? And what could be optional, or developed as the project progresses? A minimum list of reference chemicals? A first method/technology? A key / molecular initiating event(MIE)/physiological process on an (network of) AOP(s)? A definition of method purpose and how biological relevance could be demonstrated? Anything else?

Biological relevance/purpose

- Clear mechanism (KE, MIE)
- AOP helpful but not mandatory, other well established concepts of mechanistic relevance exist (e.g. cancer hallmarks/key characteristics, key neurodevelopmental processes etc..)
- The physiological process in combination with a clear definition of method purpose/intended use
- Likely the purpose, biological relevance and some mechanistic anchor would be needed up front, and the rest (from the list in the question) could be developed as the project progresses
- Need to demonstrate functionality with specific, well characterized reference compounds (test component/ purpose dependent) that could be selected from an AOP.
- Actually a definition of method purpose and how biological relevance can be proved is something that should be defined before implementing performance standards
- A key/molecular initiating event (MIE)/physiological process on an (network of) AOP(s) could be developed as the project progresses
- MIE/AOP… the DIP 2 out of 3 for SS is not using this concept, for skin irritation no AOP is used, so apparently this is not necessary.

Preliminary WNT replies on issue2 (2/5)

Biological relevance/Purpose

- This still needs to be explored and discussed but having common mechanisms covered would be helpful. If an AOP exists, common KEs or MIE.
- A definition of method purpose and how biological relevance can be demonstrated is the starting point.
- Starting point to define a standard is to describe the method purpose and biological relevance and then identify a similar method or technology to compare against.
- One should also consider, if the level of standards is use -dependent.
- Purpose of the method is key.
- Definition of a purpose of the method is key
- A mechanistic event, which is of biological relevance (ideally described in an AOP) in combination with a definition of method purpose seems to be essential in order to define a performance standard with a minimum list of reference chemicals.
- Additionally, the PS should acknowledge that different technologies and methods can inform on the same biological endpoint or pathway. The PS should also reflect a more flexible approach for me -too methods while requiring a context of use being considered upfront.

European Commission

Preliminary WNT replies on issue 2 (3/5)

Reference Chemicals

- No long list of reference chemicals
- Reference list can be defined when project progresses
- General reference chemicals not suitable (there might not be a sufficient number of relevant chemicals with reliable in vivo data and test methods might be trained for the reference substances raising questions to which extent the method is applicable for other chemicals)
- Minimum list of reference chemicals: Yes – mandatory. However – such list would be different for different AOs.
- For first starting to define a standard, a minimum list of reference chemicals for each adverse outcome is considered necessary.
- A minimum list of reference chemicals is an essential requirement
- Minimum set of reference chemicals is essential
- One idea that has been discussed: instead of a strict reference list for all approaches, a slightly larger set of ideal chemicals could be provided, and a method could demonstrate accuracy with a selection of types of these reference chemicals, as long as certain specific criteria are fulfilled related to mechanism, potency, physico-chemical properties, or similar as relevant. This flexible reference list could help to make validation easier and less costly, and address practical limitations like chemical availability.
- As the project progresses, when possible, a minimum list of reference chemicals and a key / molecular initiating event (MIE)/physiological process on an (network of) AOP(s) should also be defined, although recognizing that sometimes reference chemicals can not be easily identified for new methods .

Preliminary WNT replies on issue2 (4/5)

Reproducibility/transferability/performance

- Reproducibility and transferability is key. Therefore requirements for reproducibility and transferability are needed.
- Where the technologies are different, the reproducibility as well as the chemical applicability domain of the method may be different. Therefore, requirements for reproducibility/transferability are needed, and performance for the chemical applicability domain.
- The PS should also include the ability to perform a quantitative assessment (dose -response). Certain thresholds are needed to be overcome to advance to the next KE.
- Additionally, the level of uncertainty and lack of reproducibility that is accepted in the current traditional animal test methods should inform the performance standards for newer *in vitro* or *in silico* models.
- Moreover it should be transparent how the method perform within the chemical applicability domain .
- New ways for defining accuracy should be explored, e.g. does the method accurately reflect the current biological understanding of the new endpoint ?

European Commission

Preliminary WNT replies on issue2 (5/5)

Methods first?

- The choice of a first/method technology could result into an inappropriate comparison of performance, especially it should turn out that different methods allow to characterize the same KE but by means of different endpoints.
- First method/technology… as long as it focusses on reproducibility/transferability
- Starting point to define a standard is to describe the method purpose and biological relevance and then identify a similar method or technology to compare against.

European Commission

Thank you

DEVELOPING A DEDICATED SECTION OF THE TEST GUIDELINES COLLECTION FOR MECHANISTICALLY RELEVANT AND RELIABLE METHODS THAT ARE NOT STAND-ALONE

Problem definition

- New methods are **less likely to 'stand-alone'** compared with existing in vivo methods and will need to be combined with other methods in a defined approach, testing strategy or test battery (among other potential approaches) to characterize a biological 'endpoint' of regulatory interest
- Such non stand-alone methods may have **well established biological/mechanistic relevance**, reliability and utility in the context of a battery/testing strategy, but without a process to establish these characteristics formally, there is a risk that the **confidence to accept** such methods into defined approaches, IATAs or batteries **may be insufficient** → low uptake

Key initial considerations

- Do we believe that the development of defined approach -based TGs, IATAs or test batteries would be **impeded** if the constituent methods have **not been given the status of test guidelines** , even if technically validated as biologically relevant, reproducible and transferable?
- If TG status for non-stand-alone but technically validated methods would facilitate their integration into testing strategies, **is there merit in creating a category of TG** for such methods?

Avoiding pitfalls in our discussions

- Ignore the term “intermediate repository” used in the background paper.
  - Was intended to describe “intermediate” (e.g., MIA or KE) events on an AOP, not some transitory or holding position for the TG itself.
- There was no intention to consider developing TGs from methods that lack biological/mechanistic relevance, reliability and utility.
- The issue of where to locate the TGs (e.g., section 4, subsection of section 4) is less important at first than whether these types of TGs should exist at all.

Feedback received in advance

- Question 1: do we need such a category?
  - Many indicated yes, with mixed opinions on whether this should be a full category or just a list of methods.
  - Not unanimous though as there were some views that a ‘special category’ is not necessary and that TG status is not a prerequisite for inclusion in a battery.
- Question 2: Where do such methods belong?
  - A majority of respondents indicated a separate section was warranted, but not consensus on where (new section, subsection of 400 or 500)

Feedback received in advance

- Question 3: is there a need for a full evaluation or is a “lighter” approach possible?
  - Mixed views about whether a full evaluation is needed, with some indicating no if to be used in a battery and others suggesting certain characteristics are more important to establish than others (e.g., some suggesting variability, others biological/mechanistic relevance, transferability and reproducibility)
  - Should the predictive capacity be excluded from the validation package for these types of methods→ for discussion in breakouts

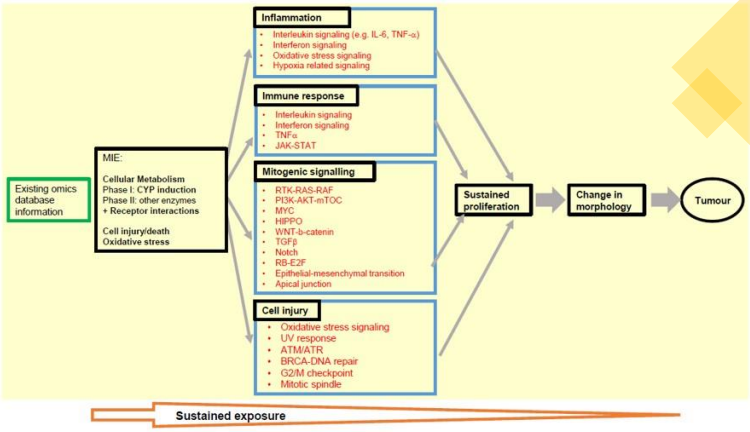
DEVELOPING GUIDANCE IN GD 34 ON VALIDATION FOR BATTERIES OF ASSAYS

Problem definition

- Developing guidance in GD 34 on validation for batteries of assays:
- Currently GD 34 states that
  - all elements of a battery should be individually validated
  - But comprehensive guidance has not been developed.
  - Now and in the near future we will continue to move towards combinations of methods/approaches
  - How do we intend to go about validation of these?
  - What do we mean exactly by the validation of individual elements?
- There will be multiple methods/approaches that can form different combinations for the same toxicological hazard endpoint.
- For the principle of MAD to remain impactful in reducing duplicative testing (for obvious economic reasons), there is great interest in reaching agreements on testing strategies/batteries with explicit DIPs.

Human health Endpoint examples

- in vitro **thyroid disruption** methods, **DNT** battery, **NGTxC m**ethods
  - ➢ 17-18 +/- assays, each may have a substantial cost.
- Overall, not all assays will need to be implemented all the time for chemicals tested for those toxicological hazard endpoints;
- prioritisation screening approaches and specific decision trees will be developed
- Building a common understanding of specific situations/scenarios and subsequently building defined approaches will be important milestones to ensure that the costs of testing are feasible whilst ensuring sufficient health and environmental protection



Case study illustration: Readiness for<u>pre-screening</u>: Identification of pivotal omics markers to be monitored in assay tools that address the key events of inflammation, immune response, mitogenic signalling and cell injury, in the NGTxC IATA

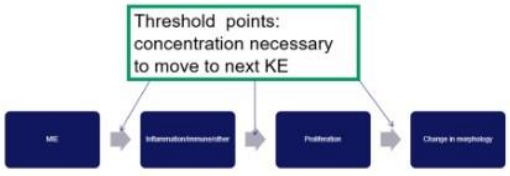
Oku et al IJMS 2022

Key initial considerations

- 1. For individual elements composing a strategy or a battery, what does it mean to be "validated"? For these individual elements/information source, what aspects of validation are important for the regulator?
- 2. Webinars 4.1 and 4.2: , making better use of quantitative information on variability to give an appropriate sense of uncertainty in the outcome.
  - How is uncertainty dealt with at the level of a battery?
  - What can be done to reduce uncertainty within a battery e.g. computational approaches such as probabilistic modelling and Dempster-Shafer Theory applications?
  - Confirmatory/orthogonal assays?
  - Can practical guidance be developed for developers of test batteries?

Addressing accuracy

- 3. Should accuracy values (% sensitivity/% specificity) apply to individual elements composing a battery/strategy?
- Should reference data be from the relevant species preferably (i.e. human for test systems relevant to human health)? Case specific ?
- 4. What should accuracy measures mean in the context of continuous data for systemic hazard endpoints? (specificity and sensitivity can be determined when positive/negative outcomes are measured. For systemic toxicity ECx or NOAEC/NOAEL are measured).
- 5. How can biological relevance be considered when validating batteries of assays, e.g. by linking each assay with a biological process or key event?



Case study e.g. from NGTxC IATA development

Feedback received in advance

- Question 1: For individual elements composing a strategy or a battery, what does it mean to be "validated"? For these individual elements/information source, what aspects of validation are important for the regulator? do we need such a category?
  - For most: Validation confirms the biological relevance of the method, its robustness and reproducibility. Domain of applicability. Strong enough that decision cannot be overturned/contested.
  - In a battery, can look at how to identify which elements do NOT add to the predictive power?
  - Differences depending upon regulatory applications
  - ICAPO: reproducible in -house.
- Question 2: Addressing uncertainty. Can practical guidance be developed for developers of test batteries?
  - Overall support for mathematical/ **probabilistic modelling to reduce uncertainties** e.g.s available-DASS, DST etc ; endpoint specific guidance on cases by case basis; where to compare with animal models, and where this is no longer really appropriate; addressing borderline calls, confirmatory testing with orthogonal assays; context of use dependent

Feedback received in advance

Question 3: accuracy values (% sensitivity/% specificity) apply to individual elements composing a battery/strategy?

a. Should reference data be from the relevant species preferably (i.e. human for test systems relevant to human health)? Case specific

- Mixed views/'yes' and 'no'. sensitivity and specifity are assay and context dependent not to be applied regardless. Case specific.
- Human data =gold standard.

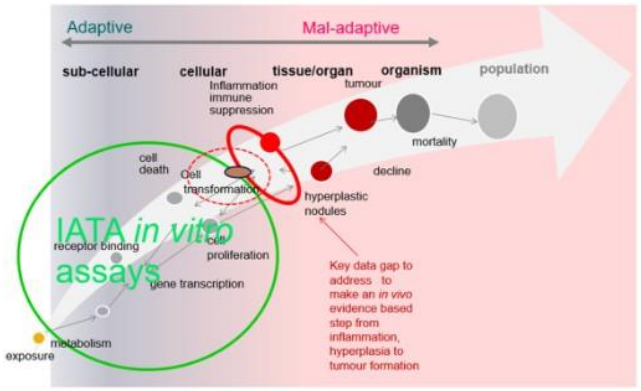
Question 4: What should accuracy measures mean in the context of continuous data for systemic hazard endpoints? (specificity and sensitivity can be determined when positive/negative outcomes are measured. For systemic.

- mixed views: in vitro, **QIVIVE may be insufficient** for NOAELs (adversity in vivo); orthogonal regression analysis; **case studies, lessons learned eg from dermal absorption GN, SARA;** PBPK needed for risk assessment (as opposed to hazard assessment) 'protective rather than predictive'…but the predictions are intended to be protective!

Feedback received in advance

Question 5: How can biological relevance be considered when validating batteries of assays, e.g. by linking each assay with a biological process or key event?

- Pretty harmonised response !
  - **AOP KE's clear biological and mechanistic basis, clear boundaries:**
  - **what is normal/adaptive vs <u>maladaptive</u>;**
  - totality of data combined rather than one to one comparison



E.g. from NGTxC IATA work

HOW TO INCENTIVISE PARTICIPATION IN VALIDATION STUDIES?

Issue paper proposed a dedicated workshop with stakeholders

- Feedback received:
  - Generally supportive of a workshop, preceded by a survey/scoping exercise and survey
    - Stakeholders could include funding agencies, regulatory agencies, CROs, not-for-profit labs, validation centres, test developers,…
    - Topics could include:
      - (scoping) analysis of current practices in validation and subsequent use of new test methods;
      - Resource for training and transfer new techniques/methodologies;
      - Resource for making technology, kits, cells more available for use;
      - Maintaining a network of competent laboratories for specific technologies?
      - Key players interests and incentives for taking part in validation studies;
      - Where/how funding can be aligned or awarded to resolved questions related to methods robustness and reliability (formulated by an OECD group such as the WNT)?
      - Fundraising for validation: how to succeed in mobilizing adequate amounts (e.g. Malta initiative)?

Opportunity for an immediate action to call attention

- Generally supported, key messages could include (for discssion):
  - Biological/mechanistically relevant, robust and transferable (=valid) methods are needed to continue to address regulatory needs as they evolve towards less animal testing, while benefitting from the Mutual Acceptance of Data globally through TG and GLP;
  - Need member countries/authorities/regulatory agencies to financially support the transfer of new methods/technologies that are identified in on-going projects and future TG projects;
  - Successful transferability of robust/mature new methods and emerging technologies ensures their equal access and broader use/implementation.
- Language for a public statement needed (some inspiration from ICAPO response?)

BETTER REPORTING OF VALIDATION STUDY RESULTS

Issue paper asked whether elements from validation studies need more systematic reporting

- Feedback received:
  - General agreement for maximising utility of data/information generated during validation studies, without increasing cost/efforts substantially;
  - Recognition that all necessary guidance for reporting already exist
    - FAIR principles,
    - RIVER Guidelines (under dev. by NC3Rs),
    - GIVIMP Guidance (very comprehensive),
    - on-line tools like SciRAP (www.scirap.org)

* FAIR: Findable, Accessible, Inter-operable, Re-usable
* RIVER: Reporting in vitro experiments responsibly
* GIVIMP: Good Invitro method practices

What needs more systematic reporting?

- From validation studies?
  - Handling/maintenance of test system, cell function state?
  - Applicability/limitations of the method/interference with media components?
  - Selection approach for proficiency substances and expected values/range?
  - Demonstration of validation of computerised system when relevant?
  - Prediction model/data interpretation procedure
  - Statistical approach used and power calculations that underly the N;
  - All measures of variability: variability of the control/normal range of response; cut-off value(s); borderline ranges around cut-off values, SD, CV, confidence interval, z factor?
  - Reporting of negative results?