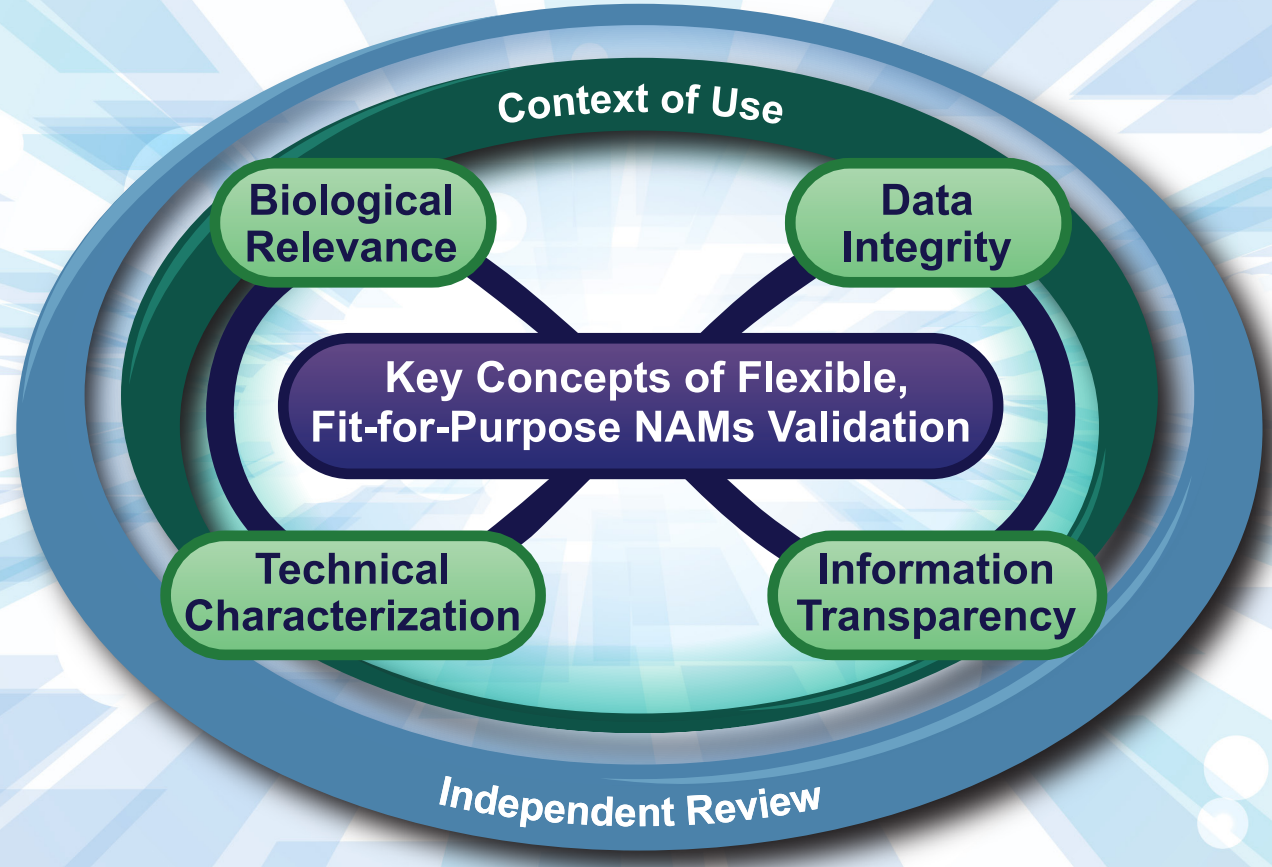
UNITED STATES ICCVAM Advancing Alternatives to Animal Testing

Interagency Coordinating Committee on the Validation of Alternative Methods

# Validation, Qualification, and Regulatory Acceptance of New Approach Methodologies

## March 2024




doi:10.22427/NICEATM-2

# Validation, Qualification, and Regulatory Acceptance of New Approach Methodologies

**A Report of the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) Validation Workgroup**

March 2024

# Table of Contents

# List of Tables

# List of Figures

# Abbreviations and Acronyms

3Rs: replacement, reduction, or refinement of animal use

AOP: adverse outcome pathway

C&E: cause-and-effect

COU: context of use

CPSC: U.S. Consumer Product Safety Commission

DA: defined approach

EPA: U.S. Environmental Protection Agency

EURL ECVAM: European Union Reference Laboratory for the Validation of Alternative Methods

FDA: U.S. Food and Drug Administration

GD: guidance document

GHS: Globally Harmonized System for Classification and Labelling of Chemicals

GIVIMP: Guidance Document on Good In Vitro Method Practices

GLP: Good Laboratory Practices

IATA: Integrated Approaches to Testing and Assessment

ICATM: International Cooperation on Alternative Test Methods

ICCVAM: Interagency Coordinating Committee on the Validation of Alternative Methods

ICH: International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use

IQ/OQ/PQ: Installation Quality/Operation Quality/Performance Quality

JaCVAM: Japanese Center for the Validation of Alternatives Methods

MTS: (3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium)

NAMs: new approach methodologies

NCATS: National Center for Advancing Translational Sciences

NICEATM: National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods

NIEHS: National Institute of Environmental Health Sciences

OD: optical density

OECD: Organisation for Economic Co-operation and Development

SOP: standard operating procedure

TG: test guideline

Tox21: Toxicology in the 21$^{st}$ Century

TSAR: Tracking System for Alternative methods towards Regulatory acceptance

# ICCVAM Validation Workgroup

**Agency for Toxic Substances and Disease Registry**
Moiz Mumtaz, PhD

**Consumer Product Safety Commission**
John Gordon, PhD (Co-Chair)

**Department of Defense**
Donald Cronce, PhD
Natalia Garcia-Reyero Vinas, PhD
Matthew Johnson, DVM, DACLAM (to June 2022)
Emily N. Reinke, PhD (to April 2022)

**Department of Veterans Affairs**
George Lathrop, Jr., DVM, MS, DACLAM

**Environmental Protection Agency**
*Office of Pesticide Programs*
Anna Lowit, PhD
Scott Lynn, PhD
Monique Perron, ScD
*Office of Research and Development*
Kelly Carstens, PhD
Alison Harrill, PhD
Nisha Sipes, PhD

**Food and Drug Administration**
*Center for Devices and Radiological Health*
Jennifer Goode, BS
*Center for Drug Evaluation and Research*
Paul C. Brown, PhD
*Center for Food Safety and Applied Nutrition*
Suzanne Fitzpatrick, PhD (Co-Chair)
Anneliese Striz, PhD
*Center for Tobacco Products*
Jueichuan (Connie) Kang, PhD
*Office of the Chief Scientist*
Tracy Chen, PhD

**National Institute of Environmental Health Sciences**
Warren Casey, PhD
*National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM)*
Helena Hogberg, PhD
Nicole Kleinstreuer, PhD

**National Institutes of Health**
*National Center for Advancing Translational Sciences*
Menghang Xia, PhD

**National Institute of Standards and Technology**
Elijah Petersen, PhD (Co-Chair)

**Occupational Safety and Health Administration**
Janet Carter, MS

**NICEATM Support Contract Staff (Inotiv)**
David Allen, PhD (to December 2023)
Michaela Blaylock
Amber Daniel, MTox
Agnes Karmaus, PhD (to April 2023)

# ICCVAM Agency Representatives

ICCVAM committee membership at the time of publication of this report
*Principal agency representative; +Alternate principal agency representative

**Agency for Toxic Substances and Disease Registry**

* Moiz Mumtaz, PhD
Patricia Ruiz, PhD

**U.S. Consumer Product Safety Commission**

* John Gordon, PhD
\+ Kristina Hatlelid, PhD, MPH
\+ Eric Hooker, MS
\+ Joanna Matheson, PhD

**U.S. Department of Agriculture**

* Jessie Carder
\+ Patrice Klein, MS, VMD, DACPV, DACVPM
Erika Edwards
Ben Green, PhD
Katherine Horak, PhD

**U.S. Department of Defense**

* Shannon Marko, DVM, DACLAM
\+ Natalia Garcia-Reyero Vinas, PhD (Co-Chair)
\+ Saber Hussain, PhD, ATS Fellow, AFRL Fellow
\+ Elaine Merrill, PhD
Matthew Grogg, PhD
Donald Cronce, PhD

**U.S. Department of Energy**

R. Todd Anderson, PhD

**U.S. Department of the Interior**

* Barnett A. Rattner, PhD
\+ Jessica K. Leet, PhD

**U.S. Department of Transportation**

* Steve Hwang, PhD
\+ Rebecca Rothhaas
\+ Ryan Vierling, PhD

**U.S. Department of Veterans Affairs Office of Research and Development**

* Holly Krull, PhD
\+ George Lathrop, Jr., DVM, MS, DACLAM

**U.S. Environmental Protection Agency**

*Office of Chemical Safety and Pollution Prevention*
Charles Kovatch
*Office of Pesticide Programs*
Monique Perron, PhD
William (Bill) Eckel, PhD
Cecilia Tan, PhD
*Office of Pollution Prevention and Toxics*
* Anna Lowit, PhD
Louis (Gino) Scarano, PhD
*Office of Research and Development*
\+ Alison Harrill, PhD
Kelly Carstens, PhD
Grace Patlewicz, PhD

**Food and Drug Administration**

*Center for Biologics Evaluation and Research*
Leslie Wagner
Allen Wensky, PhD
*Center for Devices and Radiological Health*
\+ Jennifer Goode, BS
Simona Bancos, PhD
Rakhi M. Dalal-Panguluri, PhD
*Center for Drug Evaluation and Research*
Paul C. Brown, PhD
Nakissa Sadrieh, PhD

*Center for Food Safety and Applied Nutrition*
* Suzanne Fitzpatrick, PhD (Co-Chair)
Omari J. Bandele, PhD
Patrick Crittenden, PhD
Shruti Kabadi, PhD

*Center for Tobacco Products*
Jueichuan (Connie) Kang, PhD

*Center for Veterinary Medicine*
M. Cecilia Aguila, DVM
Li You, PhD

*National Center for Toxicological Research*
Mugimane (Manju) Manjanatha, PhD

*Office of the Chief Scientist*
Tracy Chen, PhD
Chad P. Nelson, PhD, MSPH

**National Cancer Institute**
* Brian Cholewa, PhD
+ Ron Johnson, PhD
Gary Robinson, PhD

**National Institute for Occupational Safety and Health**
* Stephen Leonard, PhD

**National Institute for Environmental Health Sciences**
* Warren Casey, PhD
+ Stephen Ferguson, PhD
+ David M. Reif, PhD
Nicole Kleinstreuer, PhD

**National Institute of Standards and Technology**
* John Elliott, PhD
+ Elijah Petersen, PhD

**National Institutes of Health**
* Nicolette Petervary, VMD, MS, DACAW
Michael Eichner, DVM, DACLAM

**National Library of Medicine**

* Dina N. Paltoo, PhD, MPH, CPI

**Occupational Safety and Health Administration**

* Deana Holmes, MT
+ Janet Carter, MS

## Disclaimer

This report has been developed as a resource for U.S. federal agencies and stakeholders seeking to establish confidence in new approaches that replace, reduce, or refine the use of animals in testing. The principles described in this report were developed with input from staff from 17 federal agencies, multiple interagency workgroups, the public, and the Scientific Advisory Committee on Alternative Toxicological Methods. As such, this report does not necessarily reflect the opinions or policy of any agency or workgroup. It does not create rights for any person or party and should not be taken as a commitment by any federal agency. This report does not establish any legally enforceable responsibilities. Instead, this report describes the Interagency Coordinating Committee on the Validation of Alternative Methods' (ICCVAM) current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word "should" means that something is suggested or recommended, but not required.

## Executive Summary

New approach methodologies (NAMs) are being developed with increasing frequency and are being utilized to provide regulatory and non-regulatory assessments of the potential toxic effects of chemicals and products on human health and the environment. These NAMs are being used to investigate the biological mechanisms underlying toxicological processes, to assist in the evaluation of new and existing products, and to generate hazard identification and dose-response relationship information for health and environmental hazard classification and risk assessment purposes. This report was developed by ICCVAM to help developers and end users build confidence in NAMs. This confidence can be achieved through the implementation of flexible, fit-for-purpose validation strategies that consider the intended application of the NAM. This report helps to build confidence by describing concepts such as context of use, biological relevance, and technical characterization of NAMs. The report looks at U.S. federal agency regulatory acceptance of NAMs and the potential for international harmonization, and reviews best practices for quality tools and technical assessment of NAMs. The report also emphasizes the need for communication between the developers, end users, and regulatory agencies. The outdated one-size-fits-all strategy for validation does not work in an advanced field like NAMs. ICCVAM is an organization that can support this transition to more modern approaches in an advisory capacity, via coordination of validation efforts, and by establishing opportunities for research collaboration. Therefore, ICCVAM has developed this report, in cooperation with 17 federal regulatory agencies and research laboratories, to help developers, end users, and regulatory agencies build confidence in NAMs so they can be implemented to supplement or replace animal testing for regulatory and non-regulatory purposes.

# 1.0 Introduction

The Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) is composed of representatives from 17 U.S. federal regulatory and research agencies that require, use, generate, or disseminate toxicological and safety testing information. ICCVAM conducts technical evaluations of new, revised, and alternative safety testing methods and integrated testing strategies with regulatory applicability. ICCVAM also promotes the scientific validation and regulatory acceptance or qualification of testing methods that accurately assess the chemical safety and hazards of relevant products in an effort to replace, reduce, or refine (enhance animal well-being and lessen or avoid pain and distress) animal use.

Shortly after its establishment as a standing committee in 1997, ICCVAM published a report, "Validation and Regulatory Acceptance of Toxicological Test Methods" (ICCVAM, 1997), outlining criteria for the validation and regulatory acceptance for new and alternative test methods. Additional guidance was subsequently provided in the 2003 publication, "ICCVAM Guidelines for the Nomination and Submission of New, Revised, and Alternative Test Methods" (ICCVAM, 2003). The principles outlined in these documents were carried forward in developing international criteria described in the Organisation for Economic Co-operation and Development (OECD) Guidance Document (GD) 34, "Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment," (OECD, 2005). These resources described a validation model that is flexible in principle, but in practice has demonstrated various limitations such as being lengthy and resource-intensive. For some contexts of use, methods may not need to undergo every step of this validation process to yield valuable data for a federal agency. Moreover, these documents are not always compatible with many modern approaches to toxicity testing, which place less emphasis on replacement of *in vivo* tests with a single alternative method and more emphasis on integrating results from multiple *in vitro* and *in chemico* assays and *in silico* approaches (e.g., computational models). The 2018 ICCVAM publication, "A Strategic Roadmap for Establishing New Approaches to Evaluate the Safety of Chemicals and Medical Products in the United States" (ICCVAM, 2018), provides a conceptual framework promoting better communication between agencies and test method developers and more flexibility in how confidence is established, to help ensure the adoption of new methods by federal agencies and regulated industries once validated for a specific application or context of use (COU).The text that follows provides more specific insight on establishing confidence in new approach methodologies (NAMs) building upon the principles outlined in the 2018 ICCVAM Roadmap.

In the context of this report, the term NAM refers to any technology, methodology, approach, or combination thereof that can be used to provide information on chemical hazard and risk assessment and supports replacement, reduction, or refinement of animal use (3Rs). This report builds on the principles articulated in existing guidances and documents (Table 1) accepted by the U.S. and internationally to advocate a more flexible approach to building confidence in NAMs. Here we present key concepts of validation, qualification, and regulatory acceptance as they apply to NAMs that include testing in a biological system (e.g., *in vitro*, certain *in chemico*, small model organisms), recognizing that additional considerations may be needed for other types of NAMs, such as computational model predictions (OECD, 2007).

**Table 1.** Examples of Existing U.S. and International Documents Related to Validation, Qualification, and Regulatory Use of NAMs

| **Document Title** | **Reference** |
|---|---|
| Validation and Regulatory Acceptance of Toxicological Test Methods (retired as of publication of this report) | ICCVAM, 1997 |
| ICCVAM Guidelines for the Nomination and Submission of New, Revised, and Alternative Test Methods | ICCVAM, 2003 |
| OECD Series on Testing and Assessment No. 34: Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment | OECD, 2005 |
| Recommended Procedures Regarding the CPSC's Policy on Animal Testing | CPSC, 2012 |
| FDA Predictive Toxicology Roadmap | FDA, 2017a |
| Qualification of Medical Device Development Tools: Guidance for Industry, Tool Developers, and Food and Drug Administration Staff | FDA, 2017b |
| EPA Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the TSCA Program | EPA, 2018 |
| ICCVAM Strategic Roadmap for Establishing New Approaches to Evaluate the Safety of Chemicals and Medical Products in the United States | ICCVAM, 2018 |
| Guidance Document on Good In Vitro Method Practices (GIVIMP) | OECD, 2018 |
| Guidance for Industry and Test Method Developers: CPSC Staff Evaluation of Alternative Test Methods and Integrated Testing Approaches and Data Generated from Such Methods to Support FHSA Labeling Requirements | CPSC, 2020 |
| Qualification Process for Drug Development Tools Guidance for Industry and FDA Staff | FDA, 2020 |
| EPA New Approach Methods Work Plan | EPA, 2021a |
| EPA Strategic Plan to Reduce the Use of Vertebrate Animals in Chemical Testing | EPA, 2021b |
| Advancing New Alternative Methodologies at FDA | FDA, 2021a |

## 2.0 Key Concepts in Flexible, Fit-for-Purpose NAMs Validation

The underlying principles of validation as described in OECD GD 34 remain essential, but the processes used for validation should allow for efficient and timely development of NAMs that are fit-for-purpose, reliable, and provide information relevant to the species of interest. Specifically, OECD GD 34 states that "new test methods undergo validation to assure that they employ sound science and meet regulatory needs," i.e., that the methods are fit-for-purpose, which in addition to informing regulatory decisions, could include screening and prioritization of use cases. The guidance also states that "the validation process should be flexible and adaptable" and that performance must be "demonstrated using a series of reference chemicals" and "evaluated in relation to existing relevant toxicity data" (OECD, 2005). As such, confidence in NAMs should not be considered a universal status bestowed following the completion of a specific process (e.g., considering a NAM to be "validated" following successful completion of a ring trial study); instead, establishing confidence in NAMs should be viewed as an evolving and iterative process that requires communication among method developers, regulatory decision-makers, and validation bodies (ICCVAM, 2018).

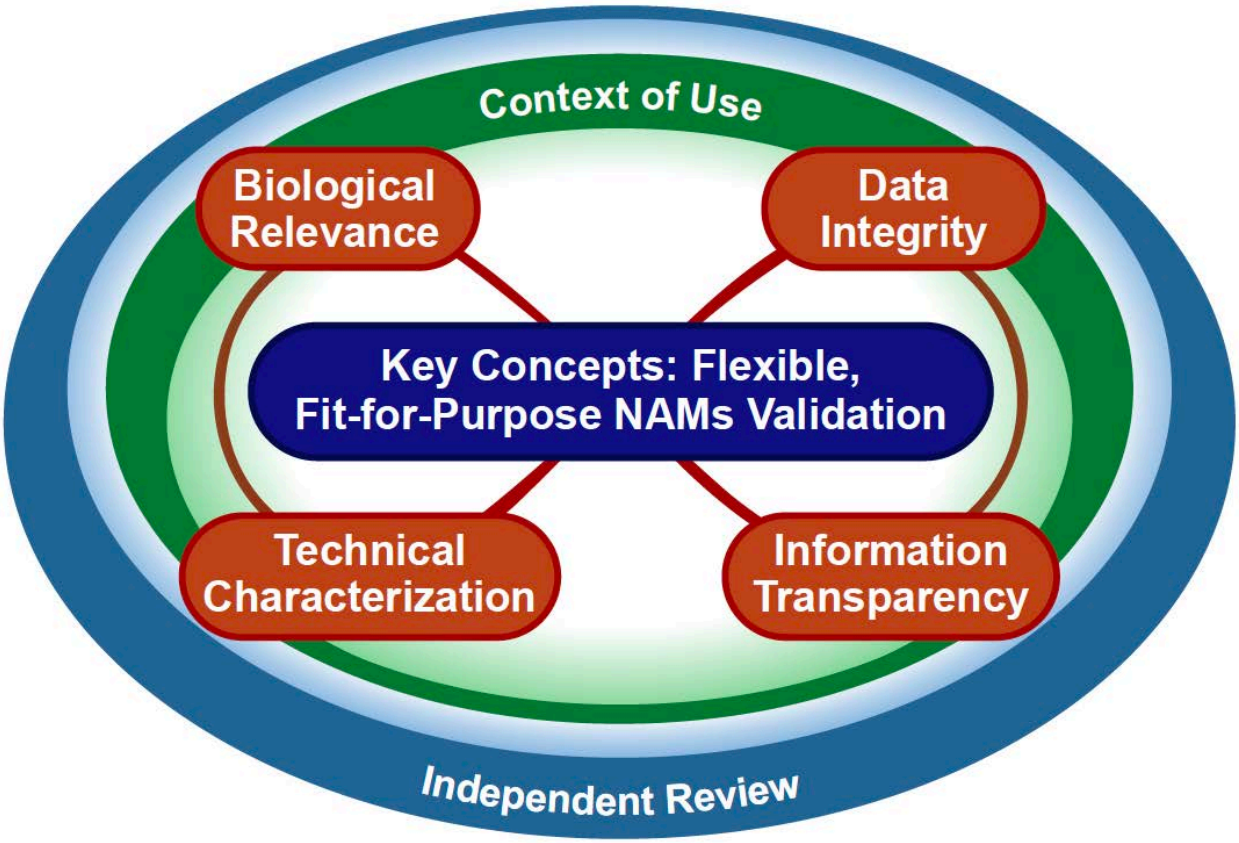



**Figure 1.** Key concepts to consider during development and implementation of flexible, fit-for-purpose NAMs validation strategies. Adapted from van der Zalm et al. (2022).

There are several key concepts that are important to consider when designing and implementing flexible, fit-for-purpose validation strategies, represented in Figure 1. These concepts are broadly applicable to the use of NAMs for various purposes but should be tailored as needed for the specific scenario and application. One overarching concept to consider is COU, or the purpose for which the NAM is intended (e.g., screening, hazard identification, potency evaluation, point of departure for quantitative risk assessment, etc.). The COU will inform the flexibility of the validation process and how the fitness of the NAM for a specific COU is established. Key concepts involved in that flexible, fit-for-purpose validation process are biological relevance, technical characterization, data integrity, and information transparency. As demonstrated in the figure, these aspects are related and interdependent. Finally, all parts of the validation process must be subject to independent review. These key concepts are similar to the essential elements proposed by van der Zalm et al. (2022) in a framework for establishing scientific confidence in NAMs, but we are adapting them here to be useful for a wider range of NAM applications. Each of these key concepts is further detailed in Section 3.0.

# 3.0 Applying the Key Concepts to Build Confidence in NAMs

Validation should be a robust yet flexible process wherein scientific confidence is established by determining the fitness of a NAM for a specific COU. Where possible and appropriate, there should be evidence to support that the use of an alternative method will provide information that is as good as or better than the existing method, and that it will lead a regulatory review to decisions that are as protective for human health. The COU determines the detailed criteria and implementation of the key concepts. In order to determine whether NAMs are fit for a particular purpose, both the COU and the relevant biology the NAM is intended to cover should be clearly defined. The NAM must be well-described and provide technically reliable information that is biologically relevant and/or health-protective for the endpoint or process of concern. The composition of the information provided by the NAM must be sufficiently transparent to allow for independent review to ensure integrity and trustworthiness. Where appropriate and possible, building confidence in NAMs may include demonstrating that the NAM provides information of equivalent or better quality and relevance to the species of interest for regulatory decision-making as compared, either quantitatively or qualitatively, to the information provided by the traditional animal test method. To avoid being constrained by the status quo, the possibility that the NAM may provide better quality and more relevant information for regulatory decision-making than the traditional animal test method must be acknowledged. This requires a validation framework that allows for this possibility and accounts for circumstances in which a comparison to data from traditional animal test methods may not be possible.

## 3.1 Context of Use

Establishing COU, or the intended utilization of a NAM, includes crafting a statement that fully and clearly describes the way the NAM is intended to be used and its regulatory purpose (if applicable). U.S. federal agencies operate under statutes and regulations particular to each agency (see, e.g., Shaffer, 2021), and therefore have different criteria for a NAM to be

acceptable and applicable toward each agency's individual requirements. In some cases, even within an agency (e.g., different centers of the U.S. Food and Drug Administration [FDA], offices of the U.S. Environmental Protection Agency [EPA]) there may be different needs depending on the area of regulation. This is further complicated internationally where regulations in other countries may differ from U.S. regulations such that methods acceptable in one country for a particular COU may not be acceptable elsewhere. Consequently, even though a NAM may be validated to address a specific endpoint for one (or multiple) COU under a particular guidance/regulation, it may not be considered acceptable under a different guidance/regulation among various agencies, regulatory jurisdictions, and countries. To better understand this diverse regulatory landscape, ICCVAM has collected information from a variety of agencies, both domestically and internationally, to characterize agency needs specific to particular toxicity endpoints (NIEHS, 2023a; examples shown in Table 2).

**Table 2.** Manuscripts Produced by ICCVAM Workgroups that Provide Details about Agency Testing Needs for Selected Topics

| **Testing need focus** | **Reference** |
|---|---|
| Acute toxicity testing | Strickland et al., 2018 |
| Ecotoxicity testing | Ceger et al., 2022 |
| *In vitro* to *in vivo* extrapolation | Chang et al., 2022 |
| Nanomaterials testing using NAMs | Petersen et al., 2022a |
| Skin sensitization | Daniel et al., 2018; Strickland et al., 2019 |
| Skin and eye irritation testing | Choksi et al., 2019 |

NAMs with different COUs such as screening/prioritization and hazard identification can be different from those that are designed to answer specific questions regarding toxicological mechanisms and/or support quantitative risk assessments, and as such, different criteria may be developed and applied to evaluating and validating NAMs for each purpose. Many new methods are developed in academic settings to address basic research questions that may be specific or exploratory in nature. Occasionally, it is recognized that one of these methods might have the potential to serve a role other than basic research into more applied areas, such as regulatory decision-making. It is important for developers of new methods to meet with both federal agencies and/or intended users or stakeholders prior to validating a NAM to ensure that validation studies designed to establish confidence will be tailored appropriately to the intended COU.

Other NAMs may be more useful for product discovery and development. A NAM may be adopted for screening new molecules for desirable (e.g., pharmacologic) or potentially undesirable (toxic) activities. These data may be used to make decisions about what molecules

should be taken into further development. Standardization of testing methods can support innovation by focusing effort and resources on new technological or therapeutic discoveries, rather than method development or testing. These uses of NAMs tend to be outside the regulatory authority of most federal agencies. However, it is still recommended that the key concepts to establish scientific confidence in NAMs are applied to ensure that high-quality data are being obtained. It is recognized that these uses of NAMs can both support the 3Rs directly and provide preliminary data and experience that can support further development of a NAM, potentially enabling its use for a regulatory purpose.

### 3.2 Biological Relevance

The relevance of a NAM describes the relationship between the test and the effect in the target species and whether the test method is meaningful and useful for a defined purpose, with the limitations identified. Adequate demonstration of the relevance of a NAM is an important contributor to confidence in a NAM. Biological relevance can be demonstrated in various ways depending on the available information, as detailed below.

Considerations surrounding biological relevance detailed in the following subsections include:

- What type of information does the NAM provide? Is there an understanding of the biology and mechanisms leading to the outcome/endpoint?
- What reference data are available for benchmarking the outcome that the NAM is intended to query?
- What are the considerations on whether, and how, to benchmark outputs of a NAM to an established laboratory method?

Additional considerations surrounding biological relevance may also be important depending on the specific circumstances of the validation application.

#### 3.2.1 Mechanistic Understanding

Consideration of the biology of the species of interest (generally human, but often other species) is important when assessing the relevance of a NAM. Comparisons of the information provided by the NAM to *in vivo* biology should be supported, where possible, by existing mechanistic knowledge (e.g., an adverse outcome pathway [AOP] or toxicologically relevant biological process). Anchoring a NAM to an established AOP via a molecular initiating event or one or more key events can help demonstrate the biological relevance of the NAM. On the other hand, lack of an established AOP for the outcome being predicted by a NAM does not necessarily exclude the NAM from being potentially useful, and such supporting information may come from mechanistic insights provided by *in vivo* data and understanding of the biology of the target species. It is important for a NAM to be comprehensively characterized and clearly describe what biological event is being measured and how it relates to the adverse outcome or hazard of concern.

A description of the NAM should address biological plausibility of the model for predicting *in vivo* outcomes and/or provide a mechanistic linkage to a biological process, mechanism, or AOP. For example, if a NAM is to predict the possibility of a human pharmaceutical to induce fetal malformations or embryo-fetal lethality, there should be an understanding of the mechanisms of embryo-fetal development (e.g., cell migration, differentiation, vasculogenesis, neurulation, gastrulation) and the connection to subsequent developmental adverse effects studied with the model (FDA, 2021b). While the relationship to the *in vivo* effect being predicted may be correlative in nature, it is more challenging to build confidence in NAM predictions from only correlative or empirical relationships; therefore, tests with clear biologic relevance to the process being evaluated are preferred. The absence of an understanding of the biological and mechanistic relevance of a NAM may limit its applicability to boundaries tightly defined by the data used to validate the NAM and make it difficult to extend NAMs to chemical classes outside those used in establishing and validating the NAM.

An AOP is a useful organizing framework to link molecular and cellular perturbations with adverse health outcomes and can be used to develop and anchor NAMs that represent important toxicological processes. Examples exist for endpoints that do have well-established AOPs, such as skin sensitization (Kleinstreuer et al., 2018). For endpoints lacking well-established AOPs, the mechanistic relevance of the NAM can instead be considered based on factors such as biological pathways or processes (Table 3).

**Table 3.** Examples of Endpoints Where Biological and Mechanistic Relevance of NAMs has Been Demonstrated*

| **Endpoint** | **Summary** | **References** |
|---|---|---|
| Skin sensitization | The endpoint has a well-developed human relevant AOP to which defined approaches combining several NAMs are mapped and described in OECD Guideline 497. | Kleinstreuer et al., 2018; OECD, 2021a |
| Endocrine disruption | Established pathway models using complementary NAMs as part of an integrated strategy are available for estrogen and androgen receptor activity. EPA accepts these NAMs for Tier 1 screening in the Endocrine Disruptor Screening Program. | Judson et al., 2015; Kleinstreuer et al., 2017; EPA, 2023 |
| Developmental neurotoxicity | Limited AOPs exist for this complex endpoint. Instead, a battery of NAMs covering critical processes of human neurodevelopment has been developed. An OECD GD on the battery is available that includes Integrated | Crofton and Mundy, 2021; OECD, 2022a; OECD, 2023 |

| **Endpoint** | **Summary** | **References** |
|---|---|---|
| | Approaches to Testing and Assessment (IATA) case studies. | |
| Inhalation toxicity | An alternative approach using an *in vitro* human cell-based assay and computational modeling was used to derive a point of departure for use in EPA human health risk assessment. This approach was also published as an OECD IATA case study. | Corley et al., 2021; EPA, 2021c; OECD, 2022b |
| Eye irritation | A guidance document describing an alternate testing framework for assessing eye irritation potential of pesticides and pesticide products. Available *in vivo*, *in vitro*, and *ex vivo* test methods were reviewed with respect to their relevance to human ocular anatomy and mechanisms of toxicity. | EPA, 2015; Clippinger et al., 2021 |
| Skin irritation | A guidance document proposing an IATA for skin corrosion and irritation using existing information, physicochemical properties, and other non-testing methods. | OECD, 2014 |

* Refer to agency-specific guidance for acceptance of different NAMs.

Ideally the description of the relationship of the NAM to the biologic effect of interest should be based on available information on the relevant biology or mechanism of action for the endpoint of concern in the species of interest. In some cases, this may not be possible, and data from a different species than the one of interest may be the only available basis for comparison. However, the physiology of the species of interest may differ from the existing surrogate test species, further emphasizing the need to incorporate mechanistic understanding into the NAM evaluation. For example, anatomical and physiological aspects of the rabbit eye differ from human (Clippinger et al., 2021) and the windows of susceptibility during brain development differ among species (Smirnova et al., 2014; Tsuji and Crofton, 2012). Consequently, the biological relevance to the species of interest and key exposure considerations should be acknowledged in assessing both the NAM and the existing reference test method. Evaluations of biological and mechanistic relevance might consider, for example, the relevance of the cell type used, the physiological characteristics of the relevant organ/tissue under investigation, or the presence of species-relevant metabolites associated with the test substance (noting that metabolites found in one species might be different from those found in another or produced in a NAM). Additionally, one might examine the ability of a method to assess a particular species-specific mode of action or mechanism (Hartung, 2010; Madia et al., 2021; Parish et al., 2020).

Moving beyond species translation, variation between individuals (e.g., donors of cells or tissues used in *in vitro* assays) may additionally underlie biological variation that can both affect the measured response, as well as potentially add additional insight into inter-individual variability across a population (Harrill, 2020). While not covered extensively in the present document, NAMs users and developers should be aware that observation of an effect (with regards to magnitude and, in some cases, presence of the observed effect) as well as the point of departure or dose at which the effect occurs can be modulated by genetic sequence variation present across a panel of cell lines (Chiu et al., 2017). Use of genetically diverse cell line panels in cell-based tests can offer advantages in understanding the population dynamics of toxicity and phenotypic response (Harrill and McAllister, 2017), uncovering genetic sequence variants that confer susceptibility (Frick et al., 2015), enabling identification of uniquely susceptible subpopulations (Church et al., 2015), and facilitating measurement of toxicokinetic and toxicodynamic variability (Rusyn et al., 2022). While tumor-derived cell lines are sometimes used in *in vitro* models, their genetic composition may not adequately represent genetic variation in the population. Specific recommendations for incorporating diversity in NAMs will depend on context. Accordingly, fit-for-purpose study designs should be developed in cooperation with geneticists to ensure adequate coverage of target species variation, as well as with bioethicists and relevant stakeholder groups if human genetic variation will be queried (see for example ICCVAM, 2022). Specific recommendations for assessing inter-individual variation are beyond the scope of this report and will likely continue to evolve with the emerging science in this arena.

In some cases, complex AOPs or outcomes that are a result of more than one AOP may require that multiple aspects of the mechanism for the adverse outcome be assessed to obtain an adequate prediction of the outcome. NAMs generally have limitations that do not allow them to be predictive for or technically applicable to all chemical classes. The NAM may also be limited with respect to the complexity of the biology it can represent. Combining more than one NAM with differing performances, applicability, and biological coverages into defined approaches (DAs) can enhance their ability to predict outcomes (OECD, 2017). This can be achieved by mapping appropriate NAMs to key events along an AOP to ensure sufficient mechanistic representation to predict the relevant adverse outcome for the species of interest. An example of mapping NAMs to key events along an AOP resulting in a DA to address a regulatory endpoint is found in the OECD "Guideline No. 497: Defined approaches on skin sensitisation" (OECD, 2021a).

### 3.2.2 Reference Compounds

Reference compounds can have a broad range of potential purposes that span from running an assay through evaluating its technical quality and biological relevance as shown in Table 4. A reference compound may work well for one purpose but not be adequate for another. For example, a reference compound may work well for a certain method as a positive control (e.g., a developmental neurotoxicity assay) even if it would not result in the implicated adverse effect in the target species (e.g., due to an inability to pass through the blood-brain barrier) and thus may not be suitable as a biological endpoint reference compound. Depending on the COU, reference compounds may include monoconstituent chemicals, mixtures, or complex extracts (e.g., from medical devices).

**Table 4.** Selected Uses of Reference Compounds

| **Discrete Compound Sets*** | **Purpose(s)** | **Criteria for Compound Selection** |
|---|---|---|
| **Positive Control(s)** | Verify whether the method is performing as expected. | Compounds that have been verified to cause a reliable, measurable, and statistically significant change in a specific assay readout under the method conditions. |
| **Performance Compounds** | Assess the extent to which a new or modified method compares to a similar established method - "me-too". | A set of compounds that reliably elicit a response (or no response) in the already established method. |
| **Proficiency Compounds** | Evaluate the performance of a new lab using an established method. | Compounds that reliably elicit a response (or no response) in the established method; can be a subset of performance compounds. |
| **Biological Endpoint Reference Compounds** | 1) Assess the biological relevance of a method.<br>2) Compare agreement for methods designed to measure an outcome relevant to the same *in vivo* endpoint, but which use a different context (e.g., different species, method, or measurement modality). | 1) Compounds with evidence for *in vivo* effect (positive) or no effect (negative) for the endpoint of interest (ideally from the target organism of interest or a suitable comparator species).<br>2) A common set of compounds that have been tested for the endpoint of interest in at least one reliable method. |

* There may be overlap in eligible compounds across these sets.

A key aspect of demonstrating the scientific validity of a NAM is assessing its performance against existing test methods in use, often via testing biological endpoint reference compounds whose biological activities are well-characterized and understood. Biological endpoint reference compounds generally include compounds that have been demonstrated to produce adverse effects, ideally with a range of potencies, or not to produce such effects (Browne et al., 2019).

Identification of biological endpoint reference compounds with known effects for particular apical outcomes can be challenging. This is particularly true for humans, where data are often lacking, as the field of toxicology has been applied to prevent human exposure to potentially

hazardous chemicals. Success in this effort means that limited human data are available, or such substances have not been adequately characterized or measured in human populations. When such human data are available, they may be from accidental exposures or low-level exposures, both of which often provide limited quantitative data on the exposure levels or outcomes. Case studies can provide indications of hazard but are often unsuited for establishing clear cause-effect correlations.

Despite these challenges, efforts should be made to identify compounds with human effects related to the outcome or mechanism being predicted by the NAM under development. Even if such data are limited in their ability to support quantitative evaluation of the method, they may provide a qualitative check on the relevance of the NAM. High-quality human epidemiological data can also be difficult to obtain and align with toxicological data; however, epidemiological data are useful in identifying environmental chemicals with strong associations to adverse human health effects. Evidence-based frameworks have been developed to allow compilation of vast amounts of information, followed by a transparent and objective selection of narrower, relevant data sets (Wikoff et al., 2020), thereby improving accessibility of data. These frameworks include systematic evidence maps, which provide broad overviews of an evidence base, and subsequent systematic reviews, which provide a more narrow and comprehensive assessment of a particular research question, allowing for more rapid integration of human epidemiological data with other sources of existing information (Wolffe et al., 2019). In combination with mechanistic lines of evidence and legacy animal data, these data might be used to define biological endpoint reference compounds to help establish scientific confidence in NAMs (see Krishna et al., 2021 for an example).

Where possible, compounds that have been shown to cause the relevant effect(s) in the species of interest should be used as positive controls during the assessment of NAM robustness and relevance. Petersen (2021) has described characteristics to consider when choosing positive controls (e.g., chemicals tested using in-process control measurements each time the assay is performed), and many of the same considerations will be relevant for reference compounds to assess a NAM. Both positive and negative biological endpoint reference compounds, performance compounds, and proficiency compounds should be included and should be selected with equal care. Due to historical bias in the literature emphasizing positive results (recognizing not all positives reported in the literature are true positives), reliable reference compounds that are negative for the target endpoint may be more challenging to identify. Data curation efforts focusing on negative results should be strongly considered for publication, and efforts that test compounds up to relatively high concentrations across a broad range of targets may provide valuable information to help resolve this issue.

All COUs may not be covered by one definitive list, and selection of reference compounds should consider the particular regulatory needs for that COU. In some cases, it may also be appropriate to define subsets of biological endpoint reference compounds with respect to the mechanism being evaluated (rather than the apical endpoint) for each NAM within a battery. For example, few compounds qualify as reference compounds for developmental neurotoxicity, and instead each process evaluated in the battery (e.g., proliferation or neurite outgrowth) has its own assay-specific biological endpoint reference compounds. Consequently, a biological endpoint

reference compound may be positive in one assay but negative compound for another assay that is evaluating different mechanisms or processes.

Lists of reference compounds (e.g., for assessing transferability) are required in the formal validation process outlined by the OECD. However, the compilation of reliable reference compounds is time-consuming and resource-intensive. Efforts to facilitate the rapid development of curated lists of reference compounds via rigorous systematic reviews and automation processes where feasible have allowed more robust evaluation of NAM sensitivity and specificity (Judson et al., 2019; Thomas et al., 2019). An example of a biological endpoint reference compound list for NAM qualification can be found in the 2021 FDA revision of the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH), S5(R3) Detection of Reproductive and Developmental Toxicity for Human Pharmaceuticals Guidance for Industry (FDA, 2021b). The list includes compounds with positive and negative outcomes. The guidance notes that these compounds, as well as others, can be used to support qualification of an alternative assay or battery of assays for particular COUs.

### 3.2.3 Comparison to Existing Laboratory Animal Methods

The standard for establishing scientific confidence in a NAM and gaining regulatory acceptance has generally included consideration of whether the NAM can provide information on equivalent or better usefulness, scientific quality, and/or relevance than the existing test method used for regulatory decision-making (as appropriate within each agency's regulatory framework). For example, a criterion in OECD GD 34 for validating any new test method is "the method generates data for risk assessment purposes that are at least as useful as, and preferably better than, those obtained using existing methods. This will give a comparable or better level of protection for human health or the environment" (OECD, 2005). The amended U.S. Toxic Substances Control Act (section 4(h)(1)(B)) includes specific considerations for NAMs and mandates that the EPA encourage and facilitate the "use of scientifically valid test methods and strategies that reduce or replace the use of vertebrate animals while providing information of equivalent or better scientific quality and relevance that will support regulatory decisions" (15 USC §2601, 2016).

Historically, the concept of "equivalent or better" has relied upon a direct comparison with the traditional animal test data. However, to achieve the goal of "better" information, it must be acknowledged that NAMs may not provide the same information generated by the traditional animal test method and the results of the NAM may not directly align with the results of the traditional animal test (for examples, see Clippinger et al., 2021; Hoffmann et al., 2018, 2008; ICCVAM, 2018; Kolle et al., 2017; Petersen et al., 2022b; Piersma et al., 2018; Prior et al., 2019; Sewell et al., 2017). In some instances, the NAM may provide biologically relevant information, mechanistic insights, or sufficiently sensitive endpoints that are adequate for the regulatory decision-making process, and a comparison to data from traditional animal test methods may not be necessary. For example, some pharmacologic or toxicologic targets may not exist in nonhuman species, so animal studies may not be relevant for assessing potential human effects mediated through such targets. Human-based NAMs might provide biologically relevant information in these cases. Furthermore, NAMs often provide mechanistic information rather

than data on apical endpoints measured in animal test methods (e.g., while an observed reduction in body weight in an animal may not elucidate the underlying mechanism of toxicity, NAMs may be able to provide these mechanistic insights). There are also circumstances in which the animal model may be measuring a complex biological endpoint that is relevant to the COU but is not adequately covered by the NAM in question. In these cases, a comparison to traditional animal test methods may be the most expedient option, particularly where the mechanism is not fully understood. The objective of the evaluation of the NAM is to demonstrate that the NAM provides information that leads to a similar regulatory decision as would be made based on existing methods.

Comparisons between NAMs and existing laboratory animal methods, as appropriate, should consider the reliability and reproducibility of the reference animal test methods, including understanding reasons for observed variance in both types of methods, if possible. When available, using reference data from the species of interest allows the assessment of a NAM against the species-relevant response. If *in vivo* time-course data are available, NAMs such as microphysiological systems can be used to simulate toxicokinetics and then computational extrapolation of the microphysiological system can be compared with the animal or human data. Physiologically based pharmacokinetic modeling allows for the estimation of internal concentrations from reference doses and comparison to activity concentrations *in vitro*. These models can also be applied in a reverse dosimetry approach to perform *in vitro* to *in vivo* extrapolation (IVIVE) and predict equivalent administered doses that would result in plasma, or target tissue, concentrations where bioactivity is observed in the NAM. When humans are the species of interest, high-quality epidemiological, clinical, or observational evidence of effects may be useful to provide insight on building scientific confidence for the NAM. However, such human reference data are rarely available for most endpoints and chemicals, and comparisons must often rely upon data from a different species. Data from animal studies can be curated and compared to yield reference standard lists with reproducible, robust, and relevant results. In some contexts, such as ecological applications, sequence homology of molecular targets or conservation of biological mechanisms may be helpful to support a comparison to data derived from different species (Farmahin et al., 2013; LaLone et al., 2016).

Several publications have assessed the results of animal-based reference test methods for a range of endpoints and showed that the results from these tests demonstrate varying level of reproducibility (Browne et al., 2018; Dumont et al., 2016; Karmaus et al., 2022; Kleinstreuer et al., 2018; Luechtefeld et al., 2016; Pham et al., 2020; Rooney et al., 2021). The observed variability could reflect inherent biological factors, reporting errors, protocol differences, or variations in test substance purity, for example. Conflicting results between reference animal test method and NAM data should be explained whenever possible, including reference to the biology of the species of interest. This is necessary to provide a realistic context about the capabilities of the laboratory animal data, and therefore, set appropriate expectations around the maximum performance capacity of NAMs that are compared against the reference test method (Browne et al., 2019).

The relative value of comparing NAMs to legacy animal tests and the biological and mechanistic relevance to the species of interest of the NAMs should be considered based on the quality of the

available data. Realistically, in many cases those seeking to establish confidence in a NAM need to consider the historical use of animal studies and the requirement for comparison to existing methods as one important line of evidence. Under many statutory and regulatory requirements, information derived from NAMs must not be less protective than existing methods. Ideally, the method will be more predictive, and may allow for more rapid and comprehensive generation of relevant data across many chemicals where data may otherwise be scarce or absent, thereby engendering confidence among regulators and stakeholder communities. There may be specific COUs that allow for exceptions to this rule; however, acceptance of such NAMs as qualified methods will be subject to the specific requirements of the relevant agency or regulatory body. The extent of such comparisons may vary depending on the quantity and quality of the available data and the depth of understanding of relevance to the species of interest, of both the NAM and the reference method. When the NAM performance is assessed based on the reference animal test method, the amount of acceptable variance in the NAM must be considered relative to the variance observed in the *in vivo* data and based on the intended COU.

## 3.3 Technical Characterization

Technical characterization is a key component of developing NAMs for widespread use. It includes an assessment of sources of variability in the NAM, designing the assay to include relevant control measurements, evaluating the range of test substances for which the assay can be used (i.e., the applicability domain), and developing a suitable statistical data analysis approach. A key aspect of demonstrating the scientific validity of a NAM is that the assay is sufficiently well-characterized from a technical perspective to ensure that it is robust, reliable, and reproducible. The NAM includes both the test system itself and the method of quantification of the endpoint. There are aspects of technical characterization that overlap with previous key concepts such as biological relevance, which informs the selection of reference compounds for different purposes (Table 4). Examining the performance of a NAM against a well-defined set of biological endpoint reference compounds with relevant bioactivities has already been presented (Section 3.2.2), so this section will focus on technical aspects such as quality tools (Ishikawa, 1985), method development, documentation, and standards.

### 3.3.1 Incorporation of Selected Quality Tools

Overall, the technical characterization of NAMs fits into a framework with overarching steps (Figure 2): 1) initial vetting of the scientific relevance of the NAM, 2) conceptual evaluation, 3) within-laboratory evaluation, 4) statistical data analysis and reporting, and 5) interlaboratory evaluation (if needed[1]). Results from steps 2 through 5 are interrelated, with results from each step potentially impacting all the others. For example, results from interlaboratory testing could reveal that additional intra-laboratory testing should be performed or that the protocol needs to be revised to include a new control measurement.

[1] There could be examples of methods for which interlaboratory evaluation is not needed or not feasible, such as large scale roboticized high-throughput screening assays and machine learning-based *in silico* approaches.

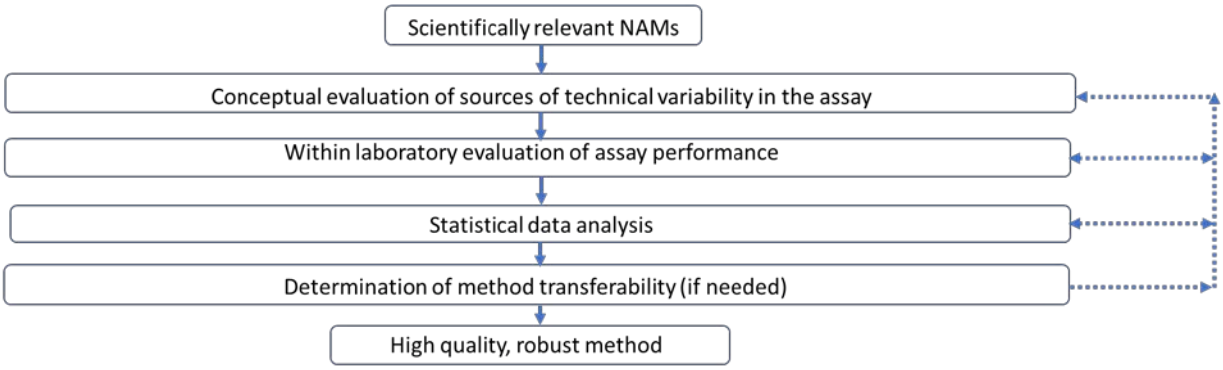



**Figure 2.** Framework for developing robust NAMs. Solid lines indicate steps that should be taken in the suggested order. Dotted lines indicate a direction that can be taken, if necessary, to reevaluate any of the previous steps. Adapted and reprinted with permission from Petersen et al. (2022b).

The initial vetting step includes an evaluation of the biological relevance of an assay, its potential COU, and the potential to reach a sufficient level of technical quality. If the NAM lacks a clear COU or biological relevance or has significant technical quality issues that potentially cannot be resolved, the NAM may not be suitable for regulatory use. In the conceptual evaluation phase, the NAM is reviewed to assess expected sources of variability and design control measurements. In the intra-laboratory evaluation, experiments can be performed to assess the robustness of the assay, establish a typical range for the control measurements (e.g., negative, vehicle, and positive controls), and identify if there are interactions between control measurements and the test result. Information from the conceptual evaluation and the intra-laboratory evaluation enable design of a statistical model to evaluate the mean assay result, characterize variability, and establish criteria for a positive or negative response and the statistical confidence for this determination.

Lastly, it may be necessary for some NAMs to undergo an assessment of transferability through interlaboratory testing. Interlaboratory testing can potentially reveal steps in a protocol that are interpreted differently among laboratories and revisions that can be made to improve the NAM's technical quality. Transferability studies can also identify issues that arise from differences in technical implementation or execution of a protocol (e.g., pipetting in one lab vs. automation in another) that necessitate changes.

Table 5 lists several quality tools that can be helpful in technical characterization of a NAM. A more detailed treatment of this topic is available in Petersen et al. (2022b). Some of the information generated from these tools may remain internal with the test method developer, while other information may be requested to support agency review.

**Table 5.** Quality Tools Often Used in Technical Characterization of NAMs*

| Quality Tool | Description | Benefit/Utility |
|---|---|---|
| Flowcharts | Diagram every step in a protocol. | Optimize identification and coverage of protocol steps that can be monitored by control measurements. |
| Cause-and-Effect Analysis | Review literature and assay background; diagram all expected sources of variability. | Identify key sources of variability and aspects of a method that may be challenging to standardize. Guides robustness testing and selection of control measurements. |
| Control Charts | Control measurements to assess technical performance, including one-time preliminary experiments, periodic measurements at a predetermined frequency, and in-process control measurements made each time the assay is performed. Control charts monitor control measurements across time. | Test for potential biases, evaluate instrument performance and calibration, and measure key sources of variability across time and among experiments. |
| Check Sheets | Record key data, metadata, and control measurements. | Monitor in-process control measurements and support troubleshooting when issues arise, support data analysis and reproducibility. |
| Scatterplots | Plot all data from control measurements and test substance results. | Assess if there is an interaction between different in-process control measurements or between those control measurements and test substance results. |

*Details and examples can be found in Appendix B.

### 3.3.2 Best Practices for Quality Control

This section describes technical factors that federal agencies may consider as best practices for the evaluation of NAMs, including DAs and Integrated Approaches to Testing and Assessment (IATAs). Factors listed here may or may not apply to all NAMs. Furthermore, other factors not

listed in the guidance, such as additional information or testing, may be needed by an agency when evaluating the NAM. Developers and/or sponsors are encouraged to communicate directly with the federal agency to which they intend to submit methods or data to determine additional factors that may be necessary for review.

A submission of a NAM for regulatory evaluation should include a description of any intra-laboratory or interlaboratory studies, if conducted. It should state whether the NAM was compared to a reference method: another NAM, *in vivo* animal data, or human data. Information on the reference method being used for comparison (i.e., reference data set) may also be needed during evaluation. Evaluation of a NAM can be accomplished via several different processes and involves documenting, using specific laboratory investigations, that the performance characteristics of a method are suitable and reliable for the intended application(s). The acceptability of data relates directly to the criteria used to evaluate the method.

In addition to serving to document the performance and characterize the applicability of the NAM, good scientific, technical, and quality practices ensure that the overall evaluation process is efficient and effective, leading to increased confidence in the proposed method. Developers should retain all information necessary to calibrate, operate, and maintain any equipment, such as equipment and software manuals, quality and safety confirmation certificates and warranties, documentation of software versions, modeling algorithms, curated databases, and training sets. Developers should also maintain documentation of suppliers for materials, cells, and reagents if this information is relevant for evaluation of a NAM (discussed in detail in Section 3.3.3). For further details on quality practices, equipment procedures, and documentation to retain, see the OECD Guidance Document on Good In Vitro Method Practices (GIVIMP; OECD, 2018). An Installation Quality/Operation Quality/Performance Quality (IQ/OQ/PQ) report for each instrument used is recommended and is required if the studies are intended to be Good Laboratory Practice (GLP)-compliant.

#### 3.3.2.1 Relevant Information for Cell/Tissue Methods

Documentation of the origins of cells and tissues used in any test method should be maintained. Information retained for review should ideally include (but not limited to):

- Ethical, legal and safety considerations
- Species / strain / sex
- Demographic information (if relevant)
- Source / supplier
- Number of donors
- Organ / tissue of origin
- Cell type(s) isolated
- Sequences of engineered molecular targets, where applicable
- Isolation technique and date
- Biosafety classification
- Cell line identification and authentication

- Serological testing for infectious agents (e.g., mycoplasma testing)
- Number of cell culture passages/population doubling
- Doubling time
- Genetic stability
- Genetic and protein expression information (if relevant)
- Pretreatment

For more detailed information, see the GD on Good Cell and Tissue Culture Practice 2.0 (Pamies et al., 2022) and the OECD GIVIMP document (OECD, 2018).

Documentation about quarantine of new cells and tissues in proper storage conditions in the laboratory prior to use in testing should be provided. Assays utilizing live cells or tissues should include a cell viability assay.

Records and documentation for the performance of all laboratory equipment (plate readers, incubators, refrigerators/freezers, etc.) should be maintained by all participating laboratories.

Registering cells in cell banks can help with traceability and long-term availability. The cell banks should include documentation of cell density and cell passage number. Records of reagent preparation should be maintained (e.g., using check sheets to track any reagent or consumable used). This is also relevant to *in vitro* methods not using cells/tissues. Some examples are (but not limited to):

- Supplier
- Catalog number
- Batch / lot numbers
- Dates of preparation
- Expiration dates
- Analyst name

The sponsor or developer of the study should provide all safety information as it pertains to the method or methodology being developed, as well as all relevant regulations for the use, transport, and disposal of all hazardous materials.

#### 3.3.2.2 Assessing the Analytical Method Used in the NAM

The developer should provide data that clearly demonstrates the substance detected or quantified is the intended chemical or analyte of interest. Table 6 lists and describes best quality control practices relating to analytical methods.

**Table 6.** Analytical Method Assessment*

| **Analytical Method Assessment** | **Description** | **Benefit/Utility** |
|---|---|---|
| Limits of Detection and Quantification | The lowest quantity or concentration of an analyte that can be reliably detected or quantified above the reagent blank or within the standard curve. The highest concentration on the standard curve determines the upper limit of quantification. | Establish the range of an analytical method. |
| Identification of Interference | Identifies when components of the method falsely alter the detected signal. | Determines if there are interactions between method components and prevents reporting of artifactual results. |
| Assessing Analytical Precision | Characterizes precision of the analytical method used and any other tests of precision, such as those assessing performance variability when different personnel use the proposed method or when different instrumentation is used for the method. This can be evaluated by interlaboratory comparison studies. | Builds confidence in the reliability of the analytical method and evaluates sources of laboratory variability. |
| Stability of Materials Used in NAMs | The ability of materials used in the NAM (e.g., test substances, testing apparatus, exposure system, reagents, and analytes) to produce similar and acceptable results over a period of time in a given environment. | Characterizes method materials and ensures reliable data is consistently obtained for a particular method. |

| **Analytical Method Assessment** | **Description** | **Benefit/Utility** |
|---|---|---|
| Robustness Testing | Tests the ability of a method to be reproduced under different conditions or circumstances without the occurrence of unexpected differences in the obtained results. | Determines the range of parameters in which the assay works acceptably. |
| Analysis of Recovery | Tests an extraction method by comparing the results of the extracted samples with those of spiked samples of a similar matrix and/or spiked blanks. | Verifies the efficiency and reproducibility of an extraction method. |
| Technical Analysis of the Applicability Domain | Obtains adequate test method data for chemicals and/or products representative of those relevant to the specific COU for which the test is proposed and clearly describe the physicochemical properties of the applicability domain. | Reduces uncertainty regarding the assay performance for use with different chemicals and/or products and provides methods and criteria for determining when a chemical is within the applicability domain. |
| Positive Control | Identify positive control compound(s) relevant to the endpoint and within the detection window of the assay. | Provides a consistent and trustworthy basis for comparison for test substance results. |
| Reference Standards for Instrument Calibration | Used to calibrate instruments using calibration standards and/or quality control samples. | Ensures reliable measurements and identifies potential sources of uncertainty. |
| Setting Specifications | Set specifications for in-process control measurements based on intra-laboratory and/or interlaboratory test results using a statistical approach. | Ensures sufficiently stringent criteria for control measurements leading to robust test substance results and lack of bias. |

*Details and examples can be found in Appendix C.

#### 3.3.2.3 Assessing Accuracy and/or Concordance of the NAM with Performance Standards

Accuracy and concordance are very similar in definition and are often considered to be interchangeable depending on the context in a given document or passage. Concordance is often defined as the comparison of two methods or tests based on the results obtained, while accuracy is often defined as the comparison of a method or test to a reference method or test result. An assessment of accuracy and/or concordance is often used in the statistical evaluation of test methods and associated data. In this context, accuracy is defined as the proportion of correct predictions among the total number of results. Other statistical parameters used when discussing the accuracy or concordance of methods include sensitivity, specificity, positive and negative predictivity, and false positive and false negative rates.

Whenever possible, well-defined performance standards (e.g., a balanced set of reference substances known to yield positive and negative results) can be used to check response and method validity. Quantitative measures of concordance (i.e., sensitivity, specificity, positive and negative predictivity, false positive, and negative rates) should be reported. When comparing a proposed test method to that of a method with established performance standards (e.g., "me-too" methods for OECD test guidelines [TGs]) that produces functionally and mechanistically similar data, the concordance (including discordant data) of both methods should be evaluated against one another and against a reference method.

#### 3.3.2.4 Standard Operating Procedures and Method Details

It is recommended that the proposed test method have well-documented standard operating procedures (SOPs) to support consistent performance of the method and related laboratory activities such as test system handling and equipment calibration. SOPs should cover all aspects of testing and analysis. The SOPs should include:

- Accountability systems that ensure integrity of test articles (e.g., record keeping, security, and chain-of-sample custody).
- Sample preparation and analytical tools, such as methods, reagents (including, when applicable, the manufacturer, catalog number, lot number, etc.), equipment, and instrumentation.
- Procedures for quality control and verification of results.
- Method details, including complete product description and formulation, exposure system, test substance volume/weight/solubility/dosing protocol, and appropriate exposure/dose range.

The submitting party should also provide a list of operating characteristics and operational criteria for judging test performance and results. Operational information and criteria for the technical systems that comprise the NAM may vary, but the criteria could include quality control charts or other performance standards for all controls, standards, exposures (dose and duration), and experimental groups. A description of the statistical methods used to assess the data should

be included. In addition, developers should describe how experimental uncertainty, statistical uncertainty, interference, and background were assessed.

### 3.3.3 Documentation

This section describes best practices for the documentation of NAMs (including DAs and IATAs). Factors listed here may or may not apply to all NAMs. Furthermore, additional documentation may be needed by an agency to evaluate the NAM. Developers and/or sponsors are encouraged to communicate directly with the federal agency to which they intend to submit methods or data to ascertain additional factors that may be necessary for review. For some agencies, independent peer review of the NAM and associated data may be needed prior to agency review.

For new methods, documentation should include a description of the proposed test method and how it may be relevant for regulatory purposes or would fit into a specific COU. Relevant information would include (but not be limited to) any mechanistic information and biological relevance of the test method and any proposed COU (e.g., applied to risk assessment). Human or appropriate taxa applicability domains should be included in the documentation. Other important aspects for documentation are detailed below.

Additional resources are available that provide guidance on documentation. For *in vitro* methods, developers can consult the OECD GIVIMP document (OECD, 2018).

#### 3.3.3.1 Test Substance Identity and Purity

For all substances tested in the NAM (e.g., controls, reference compounds, other test substances), at a minimum, the substance(s) identity, ideally a unique identifier (e.g., CASRN, SMILES, InChIKey), and information on purity of the substance as provided by the supplier should be reported. If the laboratory has the necessary capabilities, there is added benefit in performing additional quality control measures using analytical methods to assess the purity and identity of the substance. The ability to collect this information may also depend on the COU, as some test substances (e.g., environmental samples that represent complex mixtures with unknown or variable composition) may not be well-characterized.

#### 3.3.3.2 Method Development

A specific, detailed, written description of the method should be developed based on data produced from that method. This can be in the form of a protocol, study plan, report, and/or SOP. Each step in the method should be investigated to determine the extent to which environmental, matrix, or procedural variables could affect the detection and/or quantification of the analytes.

During development close attention should be given to factors such as:

- Reagent selection (with appropriate biological relevance, specificity, and stability).
- Detection method or instrumentation (i.e., performance and calibration procedures readily available).

- Compatibility of disposables (i.e., microtiter plates and other plastics) with the assay measurements and test substances.
- Analysis method / statistical method.
- Steps or processes that could introduce assay variability.

Appropriate steps should be taken to minimize external or matrix effects (or at least characterize those effects) throughout the application of the method, especially if compounds, matrix, or equipment used during development are different from those used during technical characterization of the method.

#### 3.3.3.3 Endpoint and Parameter Measurements

Measurements of each endpoint or analyte should be well tested and documented. Method development for a novel NAM should include demonstration that the method can successfully measure all relevant parameters. Relevant metadata for each experiment should be captured and recorded to link quantitative data to qualitative information, such as experimental conditions or test chemical and concentration, and to track external factors (such as the date or technician conducting the experiment) that may contribute to assay variation or batch effects. These data and metadata should be exported and saved in an accessible format so that they can be referenced during independent reviews of the validation. This topic is also discussed in the sections in Appendix B on control charting and check sheets.

#### 3.3.3.4 Limits of Use

The specific strengths and limitations of the test method should be clearly identified and described. Any potential sources of interference should be listed, and any chemicals or classes of chemicals with the potential to interfere with the test should be identified. The documentation should also identify any known limits on what materials can be tested using the NAM.

#### 3.3.3.5 Well-Defined Endpoint

Data generated by the test method should adequately measure or predict the endpoint of interest, and that endpoint should be clearly defined with an explanation of biological relevance as described in Section 3.2. An example of this would be a NAM that provides information about a specific key event in an AOP. The data should also describe any linkage between the new test method and an existing test method or between the new test method and effects in the target species. Criteria for a positive, negative, or inconclusive result in the NAM should be clearly defined and assessed over time to ensure stability of the system.

#### 3.3.3.6 Building a Statistical Model

Statistical models can be built using data from the intra-laboratory testing. These models can be built using either Bayesian or frequentist statistical approaches. Histograms can be used to assess the distribution of data obtained for in-process control measurements and to evaluate what type of distribution (e.g., normal distribution) fits the data (see Petersen et al., 2022b for details). It is

helpful to develop models to calculate cumulative variability of the NAM from measurements of both test substances and in-process control measurements instead of only using data on variability from the test substances. This information can be used to build a statistical model that can yield a decision (e.g., is the test substance positive or negative) and the statistical confidence for that decision. A simple comparison of the mean value from a test substance assessment to a threshold does not consider the variability of the test results and cannot provide statistical confidence for the decision.

One key concern when developing statistical models for NAMs is how to differentiate between "negative" and "weakly positive" results. The threshold for the statistical model can be informed by *in vivo* data when available (Friedman et al., 2023; Karmaus et al., 2022; Pham et al., 2020). This may require repeated testing of "borderline" compounds to assess the NAM reproducibility; see for example Guideline 497 (OECD, 2021a). For example, a statistical model evaluating a dose-response relationship may be used to assess the point of departure or the concentration that causes a defined effect (e.g., $EC_{50}$ value) and associated data-driven confidence intervals around those values. It may also be relevant to evaluate the quality of a NAM using statistical approaches such as a T-test, Z-factor, or other appropriate statistical criteria (Zhang et al., 1999; Zhang, 2011).

#### 3.3.3.7 Reproducibility of the Assay Results

Documentation of technical reproducibility should be included with the information submitted, where applicable. The reproducibility of the method can be assessed by replicate measurements, including quality controls and samples. This assessment should include discussion of the rationale for the selection of the substances used to evaluate reproducibility (possibly in any intra- and interlaboratory studies conducted), and the extent to which they represent the range of possible test outcomes. Outlier values should be identified and discussed. A quantitative statistical analysis of the extent of any intra- and/or interlaboratory variability or coefficient of variation analysis should be included. Measures of central tendency and variation should be summarized for historical control data (negative, positive, and vehicle where applicable). When testing the same compound(s) multiple times, comparisons can be quantitative (e.g., $EC_{50}$ values obtained) or qualitative (e.g., hazard classification). In cases where the proposed test method is mechanistically and functionally similar to an established test method with existing performance standards (e.g., from an OECD TG), the reliability of the two test methods should be compared and the potential impact of any differences discussed.

#### 3.3.3.8 Data Interpretation Procedure

The data interpretation procedure, including criteria for positive and negative responses, should be clearly described for each NAM. Combining NAMs into DAs requires fixed data interpretation procedures that are objective and do not include expert judgment, ensuring that they will result in the same outcome when applied by different groups (OECD, 2017). Use of computational algorithms, e.g., machine learning models, and software (including version number) should be well-documented to ensure reproducibility of the conclusions.

### 3.4 Data Integrity

Data integrity is a key aspect of ensuring that information derived from NAMs is trustworthy and reliable. Method developers are advised to conduct an internal evaluation of the processes used for the acquisition, transferring, and processing of raw data before those data are submitted to external, independent parties for assessment and peer review to ensure data integrity and credibility of results. Studies should be conducted to the extent possible according to principles of GLP (21 CFR § 58; 40 CFR § 160; 40 CFR § 792; OECD, 1998), where required. Furthermore, evaluating bodies, such as the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) can facilitate assessments of the quality and integrity of the development process of the NAM (NIEHS, 2023b). Other resources are available that provide guidance on maximizing data integrity. For *in vitro* methods, developers can consult the OECD GIVIMP document (OECD, 2018). For digital tools and management of digital data, developers can follow the "FAIR Guiding Principles for scientific data management and stewardship," published in 2016 (Wilkinson et al., 2016).

### 3.5 Information Transparency

Transparency facilitates trust in the use of NAMs and thereby hastens the pace of an agency's regulatory decision-making process and potential regulatory acceptance or qualification. A NAM's relevance to the species, COU, and technical characterization should be transparently communicated to peer reviewers, the scientific community, and to the public. Where appropriate, peer-reviewed articles and information describing the COU, biological relevance, and technical characterization of the NAM should be published in open-access journals and/or summarized in public-facing regulatory documents. Ideally, the principles of the NAM, the protocol, raw data files and scripts used to analyze and graph data, and the reporting standards should be communicated publicly. For NAMs that contain intellectual property, the OECD provides tools to maintain transparency, including reasonable and non-discriminatory terms for licensing commitments (OECD, 2021b). The use of proprietary or patented techniques or equipment in a method can potentially be used to fulfill regulatory testing needs. For some agencies, there may be a need for the NAM developer to convey proprietary or patented information to support regulatory acceptance or qualification; agency-specific guidance can direct test method developers about the agency's information needs.

Partners within the International Cooperation on Alternative Test Methods (ICATM; NIEHS, 2023c) publish information on NAM assessment and peer review via the Tracking System for Alternative methods towards Regulatory acceptance (TSAR) (EURL ECVAM, 2021). TSAR indicates the stages NAMs have reached in terms of acceptance as a recognized standard for use in a regulatory context together with a summary description and accepted protocol(s) or SOP(s). Where available, TSAR also includes relevant records and documents associated with a NAM linked to the different steps of the entire process: submission, validation, peer review, recommendations and regulatory acceptance or qualification. How to interpret the data that a NAM generates, and associated acceptance criteria, should be clearly communicated so that end users understand the process and can apply it in a practical setting.

### 3.6 Independent Review

Information and data supporting the NAM's COU, biological relevance, and technical characterization may be scientifically reviewed by independent third parties (whose members do not have conflicts of interest); however, the necessary level of review will depend on each agency's regulations and policy.

Evaluation of a NAM can be accomplished via several different processes and involves documenting the performance characteristics of a method to determine whether they are suitable and reliable for the intended application(s). A method's reliability often includes (but is not limited to) reproducibility, repeatability, and robustness. However, there may be additional information not listed in this report that may be necessary for the review of some methods. In addition to the performance and applicability of the NAM, good scientific, technical, and quality practices ensure that the independent review process is efficient and effective and leads to increased confidence in the proposed method. Laboratories should retain all information relevant for evaluation of a NAM, such as information necessary to operate and maintain the equipment used in the conduct of a NAM (e.g., equipment and software manuals and quality and safety conformation certificates), as well as documentation of suppliers for materials, cells, and reagents. For studies intended to be GLP-compliant, an IQ/OQ/PQ report for each instrument used may also be relevant. For non-GLP studies, documentation of proper installation and testing to show equipment performs as intended should be retained.

Raw data from and information describing the NAM should be accessible for review by independent third parties and/or regulatory agency decision-makers. The assessment and independent peer review of NAMs may be organized by validation bodies, such as NICEATM, the European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM) and its Scientific Advisory Committee, and the Japanese Center for the Validation of Alternative Methods (JaCVAM). Other bodies or international organizations that can independently review NAMs include the U.S. Federal Insecticide, Fungicide, and Rodenticide Act Scientific Advisory Panel, the European Food Safety Authority Scientific Committee, and OECD. Alternatively, the developer can fund (but not directly lead management of) an independent review of the method. Peer-reviewed publications are useful for sharing assay information with the scientific community and can supplement a more formal review by independent third parties to support acceptance and use of the method in a regulatory context.

The extent of independent review will vary depending on the COU, the regulatory framework, and the specific method being evaluated. Some of the information sent by the developer for independent review would include records from any intra- or interlaboratory studies, including whether the NAM was compared to another NAM, *in vivo* animal, or human data. These evaluations can potentially assist in the interlaboratory transferability of the NAM.

# 4.0 U.S. Federal Agency Acceptance of NAMs

## 4.1 Understanding Regulatory Needs and Decision Contexts

Federal agencies have different authorities to request, obtain, and use toxicology data. These differ according to statute, regulations, and product category. A NAM may be useful and adequate in some statutory or regulatory contexts but not in others. Consequently, developers and users of a NAM must consider the context in which it will be used. Some potential contexts would include but are not limited to DAs, IATAs, or standalone methods.

Whether a NAM will be acceptable for a regulatory or other purpose depends on the nature of the decision to be made, the adequacy of the NAM for its intended use, and the extent to which the regulatory submission depends on the NAM results to support safety, efficacy, and/or risk determination(s). A NAM with high sensitivity and low specificity that identifies a signal of concern may be useful for applications in which a large number of compounds will be screened to prioritize further assessments. Such a NAM may be of limited utility when deciding what dose of a single compound is safe for exposure. In some cases, agencies might look for signals of concern in data-poor situations (hazard identification). In other cases, agencies may need to make more quantitative decisions about compounds (risk assessment) and may be unwilling to accept a high level of uncertainty in outputs from NAMs that are used to inform those decisions. The implementation of performance characteristics should reflect how the NAM will be used and be individualized to each NAM and its COU.

To facilitate this process some agencies may want to evaluate NAMs prior to their use in a regulatory decision-making process. The evaluation will likely focus on the NAM having a well-defined COU. One of the purposes of the evaluation is to enable regulators to apply the results generated using the NAM without needing to re-review all of the underlying supporting data for the NAM (see documents cited in Table 1).

## 4.2 Context of Use Considerations

The purpose of the NAM should be clearly communicated (e.g., hazard identification, potency evaluation, point of departure for quantitative risk assessment etc.), and the NAM should be assessed based on that purpose. It is appropriate to focus application of a newly developed NAM on a single COU. However, additional COUs can be added at later times with appropriate supporting data. A COU generally needs to be focused on a particular regulatory need. Initial efforts at qualifying a NAM may be most successful if the COU is narrow. A COU could be expanded with additional data as appropriate. Regulatory needs differ across agencies so a COU for a particular NAM might also differ. Establishing an appropriate COU for a proposed NAM prior to conducting a full qualification effort is essential.

Determination of an appropriate COU should generally be discussed between the NAM developer and the agency(ies) for which the COU is relevant. Several iterations of a COU may be necessary before an acceptable one is defined. A COU may even change during collection of data as the applicability and limitations of a NAM are further defined.

### 4.3 Evolution of Confidence Based on Experience Gained

Incorporation of a NAM into regulatory use requires sufficient confidence in the method by regulators and the regulated industry. Validation and qualification support this confidence but may not be sufficient to ensure implementation. Education about and experience with a NAM are generally necessary before a NAM will meet widespread acceptance for its purpose. Cost, complexity, availability of reagents and staff trained in the conduct and interpretation of a NAM are some factors that can limit the adoption of a NAM even if validation and qualification data are available.

The performance of a NAM intended to supplement or replace an existing approach will generally need to be compared to the existing approach. There is often high confidence in existing approaches with which there is substantial experience. These existing approaches may not have undergone formal validation but repeated successful use of the existing approach along with assumed inherent validity of testing in animals often builds substantial confidence in the approach. Users need to know that a NAM will be as good or better than existing approaches when using the results to make decisions about safety. Varying levels of uncertainty may be acceptable for different COUs, with increasing confidence needed as one progresses through prioritization and screening to hazard characterization to risk assessment, for example.

One mechanism to build confidence in a NAM is for users to provide data from a NAM in parallel with data from the existing method that the NAM is intended to replace or supplement. Voluntary sharing of information on NAMs within and across industries can help establish a sufficiently large body of data to support confidence in the methods. In time, agencies and industry may see how the NAM can fit into the existing assessment paradigms without compromising safety standards.

## 5.0 U.S. and International Harmonization

Coordination among U.S. federal agencies and more broadly with international regulatory authorities will help ensure harmonization of approaches to validate NAMs and support their application and implementation. ICCVAM and NICEATM play important roles in facilitating communication and collaboration, both domestically and globally.

### 5.1 U.S. Harmonization: Role of ICCVAM and NICEATM

The ICCVAM Authorization Act outlines the following purposes of ICCVAM (42 U.S.C 285l-3, 2000; NIEHS, 2023d):

- Increase the efficiency and effectiveness of U.S. federal agency test method review.
- Eliminate unnecessary duplication of effort and share experience among U.S. federal regulatory agencies.

    - This is accomplished via several means, such as monthly meetings, workgroups, open public meetings (Public Forum) and scientific advisory groups (Scientific Advisory Committee on Alternative Toxicological Methods).
- Optimize utilization of scientific expertise outside the U.S. Federal Government.
    - This is often accomplished by setting up meetings and workshops with ICATM partners as well as with numerous scientists at conferences.
- Ensure that new and revised test methods are validated to meet the needs of U.S. federal agencies.
    - Federal agency scientists and regulators cooperate with developers to ensure that the methods being produced will address a regulatory need.
- Reduce, refine, or replace the use of animals in testing where feasible.

ICCVAM facilitates interagency and international collaborations promoting the development, regulatory acceptance or qualification, and use of alternative tests that encourage the reduction, refinement, or replacement of animal test methods. ICCVAM provides guidance to test method developers, evaluates recommendations from expert peer reviews of alternative toxicological test methods, and makes recommendations on the use of reviewed test methods to appropriate federal agencies. ICCVAM achieves its functions through ad hoc technical workgroups, managed by NICEATM, to perform specific tasks important for the development or validation of alternatives to animal testing. One such ongoing example is ICCVAM's support in coordinating an interlaboratory prevalidation study for a NAM developed by EPA based on an *in vitro* human thyroid microtissue assay for chemical screening (Deisenroth et al., 2020).

NICEATM, an office within the National Institute of Environmental Health Sciences (NIEHS) Division of Translational Toxicology, provides technical, scientific, and operational support for ICCVAM and ICCVAM workgroup activities, peer review panels, expert panels, workshops, and validation efforts. In addition to supporting ICCVAM, NICEATM:

- Conducts test method analyses and evaluations and coordinates independent validation studies on novel and high-priority alternative testing approaches.
- Provides information to test method developers, regulators, and regulated industry through the NICEATM website, the Integrated Chemical Environment, and workshops on topics of interest.
- Supports activities of the NIEHS Division of Translational Toxicology, especially those contributing to the U.S. government's interagency Toxicology in the 21st Century (Tox21) consortium.

The forum for interagency communication provided by ICCVAM and the support provided by NICEATM serve to ensure that limited resources are being leveraged effectively to coordinate U.S. federal agency efforts to validate and qualify NAMs for regulatory application.

### 5.2 U.S. Harmonization: Additional Federal Collaborations to Advance 3Rs

U.S. federal agencies collaborate on NAMs in many ways in addition to their participation in ICCVAM. For example, Tox21 (Tox21, n.d.) is a federal collaboration among EPA, the NIEHS

Division of Translational Toxicology, the National Center for Advancing Translational Sciences (NCATS) within the National Institutes of Health, and FDA. The goal of Tox21 is to develop better toxicity assessment methods to efficiently test whether certain chemical compounds may have the potential to disrupt biological processes in the human body and lead to negative health effects. Tox21 has produced a number of seminal publications and analyses that have been put into regulatory use, for example in the EPA's Endocrine Disruptor Screening Program (EPA, 2023). NICEATM and EPA have also worked closely on multiple studies to reduce or replace the use of animals in regulatory testing. This includes retrospective analyses to eliminate the use of animals for: 1) dermal acute toxicity to pesticides and pesticide formulations (EPA, 2020); 2) determining if using *in vitro* data alone will suffice for dermal absorption factor derivation for human health risk assessment of pesticides (Allen et al., 2021); and 3) determining if the same level of protection of non-target aquatic vertebrates can be achieved with *in vivo* acute toxicity testing on fewer than three fish species (Ceger et al., 2023). FDA also collaborates with NCATS to further develop microphysiological system technologies to promote their advancement and accelerate translational use (FDA, 2023).

### 5.3 International Harmonization

In addition to the international collaborative efforts facilitated through ICCVAM, U.S. agencies independently collaborate internationally to advance the acceptance of NAMs. Examples include ICATM, U.S. engagement with the United Nations subcommittee of experts on the Globally Harmonized System for Classification and Labelling of Chemicals (GHS), engagement with ICH to develop guidances describing the use of NAMs, and participation in the OECD Test Guidelines Programme and Working Party on Hazard Assessment.

ICATM was established as a partnership among validation organizations from the U.S. (ICCVAM), Japan (JaCVAM), European Union (EURL ECVAM), and Canada (Environmental Health Science and Research Bureau within Health Canada). Other participating organizations include the Korean Center for the Validation of Alternative Methods, the Brazilian Center for the Validation of Alternative Methods, and the Chinese Food and Drug Administration and Guangdong Center for Disease Control and Prevention. The overarching goals of this group have been:

- To establish international cooperation in the critical areas of validation studies, independent peer review, and development of harmonized recommendations to ensure that alternative methods/strategies are more readily accepted worldwide.
- To establish international cooperation necessary to ensure that new alternative test methods/strategies adopted for regulatory use will provide equivalent or improved protection for people, animals, and the environment, while replacing, reducing or refining (causing less pain and distress) animal use whenever scientifically feasible.

Within the United Nations, the GHS subcommittee has established a workgroup to update various chapters of the GHS to establish specific criteria for use of NAMs in various hazard classes (e.g., skin irritation/corrosion, eye irritation/serious eye damage, skin sensitization). These efforts have been fruitful in advancing NAMs on an international level and have

established DAs as acceptable methods for hazard determinations. These international efforts are important not only in advancing NAMs but also providing expertise to international regions that may not have sufficient resources in this area.

The value of harmonizing approaches to the use of NAMs is also illustrated in the activities of the ICH. Several guidances developed by ICH describe the use of alternative methods that are considered acceptable approaches by multiple regulatory authorities and industry groups around the world. These methods were assessed by expert working groups within the ICH process and incorporated into the guidances as appropriate. Examples include the use of *in chemico* and *in vitro* methods for the assessment of phototoxicity (ICH, 2013). In addition, the ICH guidance on reproductive and developmental toxicity includes some contexts of use for alternative assays as well as recommendations on the approach to qualify such assays and a list of biological endpoint reference compounds (FDA, 2021b).

The OECD is an international forum for harmonizing regulatory test guidelines and guidance documents and is increasingly focused on validation and use of NAMs. ICCVAM plays an important role in coordinating and contributing to the U.S. position for the OECD Health Effects Test Guidelines Programme. The U.S. National Coordinator represents the United States at the annual meeting of the Working Group of National Coordinators and in other test guideline development activities. In that role, the U.S. National Coordinator solicits input from relevant ICCVAM agencies for OECD TG activities that involve any aspect of the 3Rs. Subject matter experts from ICCVAM agencies serve on multiple OECD expert groups to provide scientific guidance on the development of OECD products such as test guidelines, DAs, and guidance documents. OECD test guidelines are used by stakeholders of the 38 OECD member countries to assess chemical safety. The OECD Mutual Acceptance of Data clause ensures that safety data generated using an OECD test guideline will be accepted by all the member countries, avoiding redundant testing. ICCVAM agencies also contribute to the OECD IATA Case Studies Project, which allows countries to share and collaborate on the use of novel methodologies in IATA for evaluating chemical safety within a regulatory context.

# 6.0 Communication and Training to Encourage Use of NAMs

Communication from agencies about the acceptability of specific NAMs and training on NAMs can facilitate their use. Where appropriate and feasible, agencies could communicate publicly when and how a NAM is acceptable, depending on agency-specific rules and policies. For example, NAMs may be described in GDs or on publicly available web sites. Regulatory agencies can use existing training programs or implement new programs for staff and provide publicly available training, when possible, for new methods. Building confidence in new approaches can begin even before a NAM is validated or qualified through education of the scientific community. These early education efforts can focus on the basic science of the new approaches. As approaches mature and data supporting the validity of an approach accumulates, the educational efforts can shift to familiarizing the community with these data. Training on specific use and interpretation of NAMs can occur when the NAMs have been evaluated for

particular COUs. Entities such as OECD, ICCVAM, and other validation organizations and scientific societies can also provide training and access to information about NAMs that will support confidence in their use.

Interaction between NAM developers, industry users and the regulators can facilitate the development and adoption of methods. Such interactions can occur through participation of all parties in scientific meetings where such methods are discussed and through more formal regulatory interactions according to agency-specific processes. As noted above, development of appropriate COUs and qualification data sets can be an iterative process. COUs and applicability domains can shift during the development and exploration of a NAM as data accumulate. Continued communication between all parties during this process can help assure that an appropriate path to acceptance of a NAM is taken.

Even if a NAM is available for an endpoint and is accepted by a regulatory agency, the sponsor of an application may choose to use other approaches such as a traditional animal test, and thus the NAM may not always be submitted in a regulatory application. There might be several reasons for this that are beyond the control of regulatory authorities. Although a regulatory agency can suggest or recommend substituting a NAM or several NAMs for a traditional test, the sponsor of a compound may not always be obligated to follow the suggestion. Education and familiarity with NAMs among all stakeholders are necessary to build sufficient confidence for NAM adoption.

# 7.0 Conclusion and Implementation

This report is intended to assist method developers, regulated industry stakeholders, and federal agencies in the development, validation, qualification, and acceptance of scientifically relevant NAMs. Here, we have described key concepts that should be considered to allow for efficient and timely development of NAMs that are fit-for-purpose, reliable, and provide information relevant to the species of interest. All the information may or may not apply to any specific method, DA, or IATA. There may also be other concepts that apply to a method, DA, or IATA that are not discussed in this report. It is important for developers to work closely with federal agencies and end users with an eye toward the intended use of the NAM, particularly for risk assessment applications in a regulatory review. Establishing scientific confidence in NAMs and validating or qualifying methods for specific purposes and COUs should be iterative processes that evolve via multi-directional communication among stakeholders.

The field of NAMs is an evolving one, and new considerations on NAM validation and qualification may emerge that were not anticipated at the time of the writing of this report. Consequently, stakeholders engaged in NAM development, validation and qualification may need to remain flexible and open to incorporating considerations not described here. This report will be updated on a regular basis and as needed.

# References

15 USC §2601, 2016. 15 USC §2601: Findings, policy, and intent [WWW Document]. URL https://uscode.house.gov/view.xhtml?req=(title:15%20section:2601%20edition:prelim) (accessed 11.2.22).

21 CFR § 58, n.d. CFR - Code of Federal Regulations Title 21 [WWW Document]. URL https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=58 (accessed 11.3.22).

40 CFR § 160, n.d. CFR - Code of Federal Regulations Title 40 [WWW Document]. URL https://www.ecfr.gov/current/title-40/chapter-I/subchapter-E/part-160 (accessed 5.9.23).

40 CFR § 792, n.d. CFR - Code of Federal Regulations Title 40 [WWW Document]. URL https://www.ecfr.gov/current/title-40/chapter-I/subchapter-R/part-792 (accessed 5.9.23).

42 U.S.C 285l-3, 2000. ICCVAM Authorization Act of 2000 [WWW Document]. 42 U.S.C. 285l-3 Public Law 106-545. URL https://www.congress.gov/bill/106th-congress/house-bill/4281/text (accessed 5.9.23).

Allen, D.G., Rooney, J., Kleinstreuer, N., Lowit, A., Perron, M., 2021. Retrospective analysis of dermal absorption triple pack data. ALTEX - Alternatives to animal experimentation 38, 463–476. https://doi.org/10.14573/altex.2101121

Browne, P., Delrue, N., Gourmelon, A., 2019. Regulatory use and acceptance of alternative methods for chemical hazard identification. Current Opinion in Toxicology 15, 18–25.

Browne, P., Kleinstreuer, N.C., Ceger, P., Deisenroth, C., Baker, N., Markey, K., Thomas, R.S., Judson, R.J., Casey, W., 2018. Development of a curated Hershberger database. Reproductive Toxicology 81, 259–271. https://doi.org/10.1016/j.reprotox.2018.08.016

Ceger, P., Allen, D., Blankinship, A., Choksi, N., Daniel, A., Eckel, W.P., Hamm, J., Harwood, D.E., Johnson, T., Kleinstreuer, N., Sprankle, C.S., Truax, J., Lowit, M., 2023. Evaluation of the fish acute toxicity test for pesticide registration. Regulatory Toxicology and Pharmacology 139, 105340. https://doi.org/10.1016/j.yrtph.2023.105340

Ceger, P., Garcia-Reyero Vinas, N., Allen, D., Arnold, E., Bloom, R., Brennan, J.C., Clarke, C., Eisenreich, K., Fay, K., Hamm, J., Henry, P.F.P., Horak, K., Hunter, W., Judkins, D., Klein, P., Kleinstreuer, N., Koehrn, K., LaLone, C.A., Laurenson, J.P., Leet, J.K., Lowit, A., Lynn, S.G., Norberg-King, T., Perkins, E.J., Petersen, E.J., Rattner, B.A., Sprankle, C.S., Steeger, T., Warren, J.E., Winfield, S., Odenkirchen, E., 2022. Current ecotoxicity testing needs among selected U.S. federal agencies. Regulatory Toxicology and Pharmacology 133, 105195. https://doi.org/10.1016/j.yrtph.2022.105195

Chang, X., Tan, Y.-M., Allen, D.G., Bell, S., Brown, P.C., Browning, L., Ceger, P., Gearhart, J., Hakkinen, P.J., Kabadi, S.V., Kleinstreuer, N.C., Lumen, A., Matheson, J., Paini, A.,

Pangburn, H.A., Petersen, E.J., Reinke, E.N., Ribeiro, A.J.S., Sipes, N., Sweeney, L.M., Wambaugh, J.F., Wange, R., Wetmore, B.A., Mumtaz, M., 2022. IVIVE: Facilitating the use of in vitro toxicity data in risk assessment and decision making. Toxics 10, 232. https://doi.org/10.3390/toxics10050232

Chiu, W.A., Wright, F.A., Rusyn, I., 2017. A tiered, Bayesian approach to estimating of population variability for regulatory decision-making. ALTEX 34, 377–388. https://doi.org/10.14573/altex.1608251

Choksi, N.Y., Truax, J., Layton, A., Matheson, J., Mattie, D., Varney, T., Tao, J., Yozzo, K., McDougal, A.J., Merrill, J., Lowther, D., Barroso, J., Linke, B., Casey, W., Allen, D., 2019. United States regulatory requirements for skin and eye irritation testing. Cutan Ocul Toxicol 38, 141–155. https://doi.org/10.1080/15569527.2018.1540494

Church, R.J., Gatti, D.M., Urban, T.J., Long, N., Yang, X., Shi, Q., Eaddy, J.S., Mosedale, M., Ballard, S., Churchill, G.A., Navarro, V., Watkins, P.B., Threadgill, D.W., Harrill, A.H., 2015. Sensitivity to hepatotoxicity due to epigallocatechin gallate is affected by genetic background in diversity outbred mice. Food Chem Toxicol 76, 19–26. https://doi.org/10.1016/j.fct.2014.11.008

Clippinger, A.J., Raabe, H.A., Allen, D.G., Choksi, N.Y., van der Zalm, A.J., Kleinstreuer, N.C., Barroso, J., Lowit, A.B., 2021. Human-relevant approaches to assess eye corrosion/irritation potential of agrochemical formulations. Cutan Ocul Toxicol 40, 145–167. https://doi.org/10.1080/15569527.2021.1910291

Corley, R.A., Kuprat, A.P., Suffield, S.R., Kabilan, S., Hinderliter, P.M., Yugulis, K., Ramanarayanan, T.S., 2021. New approach methodology for assessing inhalation risks of a contact respiratory cytotoxicant: computational fluid dynamics-based aerosol dosimetry modeling for cross-species and in vitro comparisons. Toxicological Sciences 182, 243–259. https://doi.org/10.1093/toxsci/kfab062

CPSC, 2020. Proposed Guidance for Industry and Test Method Developers: CPSC Staff Evaluation of Alternative Test Methods and Integrated Testing Approaches and Data Generated from Such Methods to Support FHSA Labeling Requirements [WWW Document]. URL https://www.regulations.gov/document/CPSC-2021-0006-0001 (accessed 5.4.22).

CPSC, 2012. Recommended Procedures Regarding the CPSC's Policy on Animal Testing [WWW Document]. URL https://www.cpsc.gov/Business--Manufacturing/Testing-Certification/Recommended-Procedures-Regarding-the-CPSCs-Policy-on-Animal-Testing (accessed 5.4.22).

Crofton, K.M., Mundy, W.R., 2021. External Scientific Report on the Interpretation of Data from the Developmental Neurotoxicity In Vitro Testing Assays for Use in Integrated Approaches for Testing and Assessment. EFSA Supporting Publications 18, 6924E. https://doi.org/10.2903/sp.efsa.2021.EN-6924

Daniel, A.B., Strickland, J., Allen, D., Casati, S., Zuang, V., Barroso, J., Whelan, M., Régimbald-Krnel, M.J., Kojima, H., Nishikawa, A., Park, H.-K., Lee, J.K., Kim, T.S., Delgado, I., Rios, L., Yang, Y., Wang, G., Kleinstreuer, N., 2018. International regulatory requirements for skin sensitization testing. Regul Toxicol Pharmacol 95, 52–65. https://doi.org/10.1016/j.yrtph.2018.03.003

Deisenroth, C., Soldatow, V.Y., Ford, J., Stewart, W., Brinkman, C., LeCluyse, E.L., MacMillan, D.K., Thomas, R.S., 2020. Development of an *In Vitro* Human Thyroid Microtissue Model for Chemical Screening. Toxicological Sciences 174, 63–78. https://doi.org/10.1093/toxsci/kfz238

Dumont, J., Euwart, D., Mei, B., Estes, S., Kshirsagar, R., 2016. Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives. Crit Rev Biotechnol 36, 1110–1122. https://doi.org/10.3109/07388551.2015.1084266

Elliott, J.T., Rösslein, M., Song, N.W., Toman, B., Ovaskainen, A.K.-, Maniratanachote, R., Salit, M.L., Petersen, E.J., Sequeira, F., Romsos, E., Kim, S.J., Lee, J., Moos, N.R. von, Rossi, F., Hirsch, C., Krug, H.F., Suchaoin, W., Wick, P., 2017. Toward achieving harmonization in a nanocytotoxicity assay measurement through an interlaboratory comparison study. ALTEX - Alternatives to animal experimentation 34, 201–218. https://doi.org/10.14573/altex.1605021

EPA, 2023. Availability of New Approach Methodologies (NAMs) in the Endocrine Disruptor Screening Program (EDSP) [WWW Document]. URL https://www.regulations.gov/document/EPA-HQ-OPP-2021-0756-0002 (accessed 2.20.23).

EPA, 2021a. EPA New Approach Methods Work Plan [WWW Document]. URL https://www.epa.gov/system/files/documents/2021-11/nams-work-plan_11_15_21_508-tagged.pdf (accessed 5.5.22).

EPA, 2021b. EPA Strategic Plan to Reduce the Use of Vertebrate Animals in Chemical Testing [WWW Document]. URL https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/strategic-plan-reduce-use-vertebrate-animals-chemical (accessed 5.5.22).

EPA, 2021c. Chlorothalonil: Revised Human Health Draft Risk Assessment for Registration Review [WWW Document]. URL https://www.regulations.gov/document/EPA-HQ-OPP-2011-0840-0080 (accessed 3.13.23).

EPA, 2020. Guidance for Waiving Acute Dermal Toxicity Tests for Pesticide Technical Chemicals & Supporting Retrospective Analysis [WWW Document]. URL https://www.regulations.gov/document/EPA-HQ-OPP-2016-0093-0181 (accessed 3.7.23).

EPA, 2018. Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the TSCA Program.

EPA, 2015. Use of an Alternate testing framework for classification of eye irritation [WWW Document]. URL https://www.epa.gov/sites/default/files/2015-05/documents/eye_policy2015update.pdf (accessed 11.13.23).

EURL ECVAM, 2021. EU Reference Laboratory for alternatives to animal testing (EURL ECVAM) [WWW Document]. URL https://joint-research-centre.ec.europa.eu/eu-reference-laboratory-alternatives-animal-testing-eurl-ecvam_en (accessed 11.3.22).

Farmahin, R., Manning, G.E., Crump, D., Wu, D., Mundy, L.J., Jones, S.P., Hahn, M.E., Karchner, S.I., Giesy, J.P., Bursian, S.J., Zwiernik, M.J., Fredricks, T.B., Kennedy, S.W., 2013. Amino Acid Sequence of the Ligand-Binding Domain of the Aryl Hydrocarbon Receptor 1 Predicts Sensitivity of Wild Birds to Effects of Dioxin-Like Compounds. Toxicological Sciences 131, 139–152. https://doi.org/10.1093/toxsci/kfs259

FDA, 2023. MOU 225-23-003 Memorandum of understanding between the National Institutes of Health (NCATS) and the Food and Drug Administration (FDA) for the Microphysiological Systems Program [WWW Document]. FDA. URL https://www.fda.gov/about-fda/domestic-mous/mou-225-23-003 (accessed 5.9.23).

FDA, 2021a. Advancing New Alternative Methodologies at FDA [WWW Document]. URL https://www.fda.gov/media/144891/download (accessed 7.19.23).

FDA, 2021b. S5(R3) Detection of Reproductive and Developmental Toxicity for Human Pharmaceuticals Guidance for Industry [WWW Document]. URL https://www.fda.gov/media/148475/download

FDA, 2020. Qualification Process for Drug Development Tools Guidance for Industry and FDA Staff [WWW Document]. URL https://www.fda.gov/media/133511/download (accessed 7.19.23).

FDA, 2018. Bioanalytical Method Validation Guidance for Industry [WWW Document]. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/bioanalytical-method-validation-guidance-industry (accessed 7.31.23).

FDA, 2017a. FDA's Predictive Toxicology Roadmap [WWW Document]. URL https://www.fda.gov/media/109634/download (accessed 5.5.22).

FDA, 2017b. Qualification of Medical Device Development Tools: Guidance for Industry, Tool Developers, and Food and Drug Administration Staff [WWW Document]. FDA. URL https://www.fda.gov/medical-devices/science-and-research-medical-devices/medical-device-development-tools-mddt (accessed 5.5.22).

Frick, A., Suzuki, O.T., Benton, C., Parks, B., Fedoriw, Y., Richards, K.L., Thomas, R.S., Wiltshire, T., 2015. Identifying genes that mediate anthracyline toxicity in immune cells. Front Pharmacol 6, 62. https://doi.org/10.3389/fphar.2015.00062

Friedman, K.P., Foster, M.J., Pham, L.L., Feshuk, M., Watford, S.M., Wambaugh, J.F., Judson, R.S., Setzer, R.W., Thomas, R.S., 2023. Reproducibility of organ-level effects in repeat dose animal studies. Comput Toxicol 28, 1–17. https://doi.org/10.1016/j.comtox.2023.100287

Harrill, A.H., 2020. ToxPoint: In the era of precision medicine, diversity should not be neglected in chemical safety assessment. Toxicol Sci 173, 3–4. https://doi.org/10.1093/toxsci/kfz232

Harrill, A.H., McAllister, K.A., 2017. New rodent population models may inform human health risk assessment and identification of genetic susceptibility to environmental exposures. Environ Health Perspect 125, 086002. https://doi.org/10.1289/EHP1274

Hartung, T., 2010. Evidence-based toxicology - the toolbox of validation for the 21st century? ALTEX 27, 253–263. https://doi.org/10.14573/altex.2010.4.253

Hoffmann, S., Kleinstreuer, N., Alépée, N., Allen, D., Api, A.M., Ashikaga, T., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Goebel, C., Kern, P.S., Klaric, M., Kühnl, J., Lalko, J.F., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Parakhia, R., van Vliet, E., Zang, Q., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database. Crit Rev Toxicol 48, 344–358. https://doi.org/10.1080/10408444.2018.1429385

Hoffmann, S., Saliner, A.G., Patlewicz, G., Eskes, C., Zuang, V., Worth, A.P., 2008. A feasibility study developing an integrated testing strategy assessing skin irritation potential of chemicals. Toxicol Lett 180, 9–20. https://doi.org/10.1016/j.toxlet.2008.05.004

ICCVAM, 2022. Symposium Webinar: Using New Approach Methodologies to Address Variability and Susceptibility Across Populations [WWW Document]. URL https://ntp.niehs.nih.gov/go/popvar (accessed 3.13.23).

ICCVAM, 2018. A Strategic Roadmap for Establishing New Approaches to Evaluate the Safety of Chemicals and Medical Products in the United States [WWW Document]. URL https://ntp.niehs.nih.gov/iccvam/docs/roadmap/iccvam_strategicroadmap_january2018_document_508.pdf (accessed 5.3.22).

ICCVAM, 2003. ICCVAM Guidelines for the Nomination and Submission of New, Revised, and Alternative Test Methods [WWW Document]. URL https://ntp.niehs.nih.gov/iccvam/suppdocs/subguidelines/sd_subg034508.pdf (accessed 5.3.22).

ICCVAM, 1997. Validation and Regulatory Acceptance of Toxicological Test Methods [WWW Document]. URL https://ntp.niehs.nih.gov/iccvam/docs/about_docs/validate.pdf (accessed 3.7.23).

ICH, 2013. ICH Harmonized Tripartite Guideline Photosafety Evaluation of Pharmaceuticals S10 [WWW Document]. URL https://database.ich.org/sites/default/files/S10_Guideline.pdf

Ishikawa, K., 1985. What is Total Quality Control? The Japanese Way. Translated by Lu, David J., 1st ed. Prentice-Hall.

Judson, R.S., Magpantay, F.M., Chickarmane, V., Haskell, C., Tania, N., Taylor, J., Xia, M., Huang, R., Rotroff, D.M., Filer, D.L., Houck, K.A., Martin, M.T., Sipes, N., Richard, A.M., Mansouri, K., Setzer, R.W., Knudsen, T.B., Crofton, K.M., Thomas, R.S., 2015. Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. Toxicol Sci 148, 137–154. https://doi.org/10.1093/toxsci/kfv168

Judson, R.S., Thomas, R.S., Baker, N., Simha, A., Howey, X.M., Marable, C., Kleinstreuer, N.C., Houck, K.A., 2019. Workflow for defining reference chemicals for assessing performance of in vitro assays. ALTEX 36, 261. https://doi.org/10.14573/altex.1809281

Karmaus, A.L., Mansouri, K., To, K.T., Blake, B., Fitzpatrick, J., Strickland, J., Patlewicz, G., Allen, D., Casey, W., Kleinstreuer, N., 2022. Evaluation of variability across rat acute oral systemic toxicity studies. Toxicological Sciences 188, 34–47. https://doi.org/10.1093/toxsci/kfac042

Kleinstreuer, N.C., Ceger, P., Watt, E.D., Martin, M., Houck, K., Browne, P., Thomas, R.S., Casey, W.M., Dix, D.J., Allen, D., Sakamuru, S., Xia, M., Huang, R., Judson, R., 2017. Development and validation of a computational model for androgen receptor activity. Chem. Res. Toxicol. 30, 946–964. https://doi.org/10.1021/acs.chemrestox.6b00347

Kleinstreuer, N.C., Hoffmann, S., Alépée, N., Allen, D., Ashikaga, T., Casey, W., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Göbel, C., Kern, P.S., Klaric, M., Kühnl, J., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Strickland, J., van Vliet, E., Zang, Q., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (II): an assessment of defined approaches *. Crit Rev Toxicol 48, 359–374. https://doi.org/10.1080/10408444.2018.1429386

Kolle, S.N., Van Cott, A., van Ravenzwaay, B., Landsiedel, R., 2017. Lacking applicability of in vitro eye irritation methods to identify seriously eye irritating agrochemical formulations: Results of bovine cornea opacity and permeability assay, isolated chicken eye test and the EpiOcular; ET-50 method to classify according to UN GHS. Regulatory Toxicology and Pharmacology 85, 33–47. https://doi.org/10.1016/j.yrtph.2017.01.013

Krishna, S., Berridge, B., Kleinstreuer, N., 2021. High-throughput screening to identify chemical cardiotoxic potential. Chem Res Toxicol 34, 566–583. https://doi.org/10.1021/acs.chemrestox.0c00382

LaLone, C.A., Villeneuve, D.L., Lyons, D., Helgen, H.W., Robinson, S.L., Swintek, J.A., Saari, T.W., Ankley, G.T., 2016. Editor's Highlight: Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS): A Web-Based Tool for Addressing the Challenges of Cross-Species Extrapolation of Chemical Toxicity. Toxicol. Sci. 153, 228–245. https://doi.org/10.1093/toxsci/kfw119

Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016. Analysis of public oral toxicity data from REACH registrations 2008-2014. ALTEX 33, 111–122. https://doi.org/10.14573/altex.1510054

Madia, F., Pillo, G., Worth, A., Corvi, R., Prieto, P., 2021. Integration of data across toxicity endpoints for improved safety assessment of chemicals: the example of carcinogenicity assessment. Arch Toxicol 95, 1971–1993. https://doi.org/10.1007/s00204-021-03035-x

NIEHS, 2023a. Testing Regulations and Guidelines [WWW Document]. National Toxicology Program. URL https://ntp.niehs.nih.gov/go/837330 (accessed 7.19.23).

NIEHS, 2023b. Funding Opportunities for Test Method Developers [WWW Document]. Funding Opportunities for Test Method Developers. URL https://ntp.niehs.nih.gov/go/alt-funding (accessed 5.9.23).

NIEHS, 2023c. International Cooperation on Alternative Test Methods [WWW Document]. International Cooperation on Alternative Test Methods. URL https://ntp.niehs.nih.gov/go/icatm (accessed 5.9.23).

NIEHS, 2023d. About ICCVAM [WWW Document]. About ICCVAM. URL https://ntp.niehs.nih.gov/go/iccvam (accessed 5.9.23).

OECD, 2023. No. 377: Initial Recommendations on Evaluation of Data from the Developmental Neurotoxicity (DNT) In-Vitro Testing Battery, OECD Series on Testing and Assessment. OECD Publishing, Paris.

OECD, 2022a. No. 364: Case study on the use of Integrated Approaches for Testing and Assessment for DNT to prioritize a class of Organophosphorus flame retardants, Series on Testing and Assessment. OECD Publishing, Paris.

OECD, 2022b. No. 367: Case Study on the use of an Integrated Approach for Testing and Assessment (IATA) for New Approach Methodology (NAM) for Refining Inhalation Risk Assessment from Point of Contact Toxicity of the Pesticide, Chlorothalonil, OECD Series on Testing and Assessment. OECD Publishing, Paris.

OECD, 2021a. Guideline No. 497: Defined Approaches on Skin Sensitisation, OECD Guidelines for the Testing of Chemicals, Section 4. OECD Publishing, Paris.

OECD, 2021b. No. 298: Guiding Principles on Good Practices for the Availability/Distribution of Protected Elements in OECD Test Guidelines, Series on Testing and Assessment. OECD Publishing.

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

OECD, 2017. No. 255: Guidance Document on the Reporting of Defined Approaches to be Used Within Integrated Approaches to Testing and Assessment, OECD Series on Testing and Assessment. OECD Publishing, Paris. https://doi.org/10.1787/9789264274822-en

OECD, 2014. No. 203: New Guidance Document on an Integrated Approach on Testing and Assessment (IATA) for Skin Corrosion and Irritation, OECD Series on Testing and Assessment. OECD Publishing, Paris.

OECD, 2007. No. 69: Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment. OECD Publishing, Paris. https://doi.org/10.1787/9789264085442-en.

OECD, 2005. No. 34: Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment, OECD Environment, Health and Safety Publications Series on Testing and Assessment. OECD Publishing, Paris.

OECD, 1998. No. 1: OECD Principles of Good Laboratory Practice, Series on Principles of Good Laboratory Practice and Compliance Monitoring. OECD Publishing, Paris.

Pamies, D., Leist, M., Coecke, S., Bowe, G., Allen, D.G., Gstraunthaler, G., Bal-Price, A., Pistollato, F., Vries, R.B.M. de, Hogberg, H.T., Hartung, T., Stacey, G., 2022. Guidance document on Good Cell and Tissue Culture Practice 2.0 (GCCP 2.0). ALTEX - Alternatives to animal experimentation 39, 30–70. https://doi.org/10.14573/altex.2111011

Parish, S.T., Aschner, M., Casey, W., Corvaro, M., Embry, M.R., Fitzpatrick, S., Kidd, D., Kleinstreuer, N.C., Lima, B.S., Settivari, R.S., Wolf, D.C., Yamazaki, D., Boobis, A., 2020. An evaluation framework for new approach methodologies (NAMs) for human health safety assessment. Regulatory Toxicology and Pharmacology 112, 104592. https://doi.org/10.1016/j.yrtph.2020.104592

Petersen, E., 2021. Characteristics to consider when selecting a positive control material for an in vitro assay. ALTEX. https://doi.org/10.14573/altex.2102111

Petersen, E.J., Ceger, P., Allen, D.G., Coyle, J., Derk, R., Garcia-Reyero, N., Gordon, J., Kleinstreuer, N.C., Matheson, J., McShan, D., Nelson, B.C., Patri, A.K., Rice, P., Rojanasakul, L., Sasidharan, A., Scarano, L., Chang, X., 2022a. U.S. federal agency interests and key considerations for new approach methodologies for nanomaterials.

ALTEX - Alternatives to animal experimentation 39, 183–206. https://doi.org/10.14573/altex.2105041

Petersen, E.J., Elliott, J.T., Gordon, J., Kleinstreuer, N.C., Reinke, E., Roesslein, M., Toman, B., 2022b. Technical framework for enabling high-quality measurements in new approach methodologies (NAMs). ALTEX - Alternatives to animal experimentation. https://doi.org/10.14573/altex.2205081

Petersen, E.J., Uhl, R., Toman, B., Elliott, J.T., Strickland, J., Truax, J., Gordon, J., 2022c. Development of a 96-well electrophilic allergen screening assay for skin sensitization using a measurement science approach. Toxics 10, 257. https://doi.org/10.3390/toxics10050257

Pham, L.L., Watford, S., Pradeep, P., Martin, M.T., Thomas, R., Judson, R., Setzer, R.W., Paul Friedman, K., 2020. Variability in in vivo studies: Defining the upper limit of performance for predictions of systemic effect levels. Comput Toxicol 15, 1–100126. https://doi.org/10.1016/j.comtox.2020.100126

Piersma, A.H., van Benthem, J., Ezendam, J., Kienhuis, A.S., 2018. Validation redefined. Toxicol In Vitro 46, 163–165. https://doi.org/10.1016/j.tiv.2017.10.013

Prior, H., Casey, W., Kimber, I., Whelan, M., Sewell, F., 2019. Reflections on the progress towards non-animal methods for acute toxicity testing of chemicals. Regul Toxicol Pharmacol 102, 30–33. https://doi.org/10.1016/j.yrtph.2018.12.008

Rooney, J.P., Choksi, N.Y., Ceger, P., Daniel, A.B., Truax, J., Allen, D., Kleinstreuer, N., 2021. Analysis of variability in the rabbit skin irritation assay. Regulatory Toxicology and Pharmacology 122, 104920. https://doi.org/10.1016/j.yrtph.2021.104920

Rösslein, M., Elliott, J.T., Salit, M., Petersen, E.J., Hirsch, C., Krug, H.F., Wick, P., 2015. Use of Cause-and-Effect Analysis to Design a High-Quality Nanocytotoxicology Assay. Chem. Res. Toxicol. 28, 21–30. https://doi.org/10.1021/tx500327y

Rusyn, I., Chiu, W.A., Wright, F.A., 2022. Model systems and organisms for addressing inter- and intra-species variability in risk assessment. Regul Toxicol Pharmacol 132, 105197. https://doi.org/10.1016/j.yrtph.2022.105197

Sewell, F., Doe, J., Gellatly, N., Ragan, I., Burden, N., 2017. Steps towards the international regulatory acceptance of non-animal methodology in safety assessment. Regul Toxicol Pharmacol 89, 50–56. https://doi.org/10.1016/j.yrtph.2017.07.001

Shaffer, R.M., 2021. Environmental Health Risk Assessment in the Federal Government: A Visual Overview and a Renewed Call for Coordination. Environ. Sci. Technol. 55, 10923–10927. https://doi.org/10.1021/acs.est.1c01955

Slezák, P., Waczulíková, I., 2011. Letter to the Editor: Reproducibility and Repeatability. Physiol Res 60, 203–205.

Smirnova, L., Hogberg, H.T., Leist, M., Hartung, T., 2014. Developmental neurotoxicity - challenges in the 21st century and in vitro opportunities. ALTEX 31, 129–156. https://doi.org/10.14573/altex.1403271

Strickland, J., Clippinger, A.J., Brown, J., Allen, D., Jacobs, A., Matheson, J., Lowit, A., Reinke, E.N., Johnson, M.S., Quinn, M.J., Mattie, D., Fitzpatrick, S.C., Ahir, S., Kleinstreuer, N., Casey, W., 2018. Status of acute systemic toxicity testing requirements and data uses by U.S. regulatory agencies. Regul Toxicol Pharmacol 94, 183–196. https://doi.org/10.1016/j.yrtph.2018.01.022

Strickland, J., Daniel, A.B., Allen, D., Aguila, C., Ahir, S., Bancos, S., Craig, E., Germolec, D., Ghosh, C., Hudson, N.L., Jacobs, A., Lehmann, D.M., Matheson, J., Reinke, E.N., Sadrieh, N., Vukmanovic, S., Kleinstreuer, N., 2019. Skin sensitization testing needs and data uses by US regulatory and research agencies. Arch Toxicol 93, 273–291. https://doi.org/10.1007/s00204-018-2341-6

Thomas, R.S., Bahadori, T., Buckley, T.J., Cowden, J., Deisenroth, C., Dionisio, K.L., Frithsen, J.B., Grulke, C.M., Gwinn, M.R., Harrill, J.A., Higuchi, M., Houck, K.A., Hughes, M.F., Hunter, E.S., Isaacs, K.K., Judson, R.S., Knudsen, T.B., Lambert, J.C., Linnenbrink, M., Martin, T.M., Newton, S.R., Padilla, S., Patlewicz, G., Paul-Friedman, K., Phillips, K.A., Richard, A.M., Sams, R., Shafer, T.J., Setzer, R.W., Shah, I., Simmons, J.E., Simmons, S.O., Singh, A., Sobus, J.R., Strynar, M., Swank, A., Tornero-Valez, R., Ulrich, E.M., Villeneuve, D.L., Wambaugh, J.F., Wetmore, B.A., Williams, A.J., 2019. The next generation blueprint of computational toxicology at the U.S. Environmental Protection Agency. Toxicol Sci 169, 317–332. https://doi.org/10.1093/toxsci/kfz058

Tox21, n.d. Tox21 - Toxicology in the 21st Century [WWW Document]. Tox21. URL https://tox21.gov/ (accessed 5.9.23).

Tsuji, R., Crofton, K.M., 2012. Developmental neurotoxicity guideline study: issues with methodology, evaluation and regulation. Congenit Anom (Kyoto) 52, 122–128. https://doi.org/10.1111/j.1741-4520.2012.00374.x

van der Zalm, A.J., Barroso, J., Browne, P., Casey, W., Gordon, J., Henry, T.R., Kleinstreuer, N.C., Lowit, A.B., Perron, M., Clippinger, A.J., 2022. A framework for establishing scientific confidence in new approach methodologies. Arch Toxicol 96, 2865–2879. https://doi.org/10.1007/s00204-022-03365-4

Wikoff, D., Lewis, R.J., Erraguntla, N., Franzen, A., Foreman, J., 2020. Facilitation of risk assessment with evidence-based methods – A framework for use of systematic mapping and systematic reviews in determining hazard, developing toxicity values, and characterizing uncertainty. Regulatory Toxicology and Pharmacology 118, 104790. https://doi.org/10.1016/j.yrtph.2020.104790

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wolffe, T.A.M., Whaley, P., Halsall, C., Rooney, A.A., Walker, V.R., 2019. Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. Environ Int 130, 104871. https://doi.org/10.1016/j.envint.2019.05.065

Zhang, J.H., Chung, T.D., Oldenburg, K.R., 1999. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. J Biomol Screen 4, 67–73. https://doi.org/10.1177/108705719900400206

Zhang, X.D., 2011. Illustration of SSMD, z score, SSMD*, z* score, and t statistic for hit selection in RNAi high-throughput screens. J Biomol Screen 16, 775–785. https://doi.org/10.1177/1087057111405851

# APPENDIX A: GLOSSARY

**DISCLAIMER: This glossary is intended to support the use of terminology within this report. The meaning or use of a particular term may vary across regulatory agencies.**

**3Rs**: Principles of humane experimental technique, specifically regarding the reduction (i.e., minimizing the number of animals used), replacement (i.e., use of a lower-order species or non-animal test), or refinement (i.e., minimizing the pain or distress that research animals might endure due to a particular technique) of the use of animals in research and chemical safety testing.

**Accuracy**: The closeness of agreement between a test method result and an accepted reference value.

**Adverse outcome pathway (AOP)**: A structured representation of sequential events that occur at different levels of biological organization resulting in an adverse effect when an organism is exposed to a substance.

**Applicability domain**: The types of chemicals that can be tested using a method, or the types of chemicals for which the results produced by that method are considered acceptable.

**Assay**: The experimental system used. Often used interchangeably with “test” and “test method”.

**Balanced accuracy**: A statistical metric used to account for an imbalanced data set where one “class” appears much more than the other (e.g., higher number of negatives than positives). Balanced accuracy is calculated as the arithmetic mean of sensitivity and specificity.

**Biological relevance**: A measure of appropriateness for assessing the effects of a chemical within the taxa of interest.

**Concordance**: The closeness of agreement or consistency between two variables. Concordance may be used to qualitatively describe the biological relevance of a NAM compared with the reference animal test method and/or human reference data when available. Concordance may also be used to quantitatively describe the proportion of all chemicals tested that are correctly classified as positive or negative, and the term is often used interchangeably with “accuracy” in this context.

**Context of use (COU)**: A clearly articulated description delineating the manner and purpose of use for a particular method, approach, or application.

**Curated database/list**: A structured set of well-characterized and reliable information that is carefully compiled and maintained.

**Defined approach (DA)**: Consists of input data generated with a defined set of information sources and a fixed data interpretation procedure to derive a result that can either be used on its own or together with other information sources within an IATA to satisfy a specific regulatory need. A defined approach to testing and assessment can be used to support the hazard identification, hazard characterization, and/or safety assessment of chemicals (OECD, 2017).

**Discordance**: The proportion of all chemicals tested incorrectly classified as positive or negative.

**Endpoint**: The biological or chemical process, response, or effect assessed by a test method.

**False negative**: A substance incorrectly identified as negative by a test method relative to the specified reference data.

**False positive**: A substance incorrectly identified as positive by a test method relative to the specified reference data.

**Fit-for-purpose**: Matching the type and certainty of information provided by a NAM (or set of NAMs) with the type and certainty of information needed for a given decision (EPA, 2021a).

**Good Cell Culture Practice (GCCP)**: A set of principles developed for practical use in the laboratory to ensure the reproducibility of *in vitro* (cell-based) work and enhance quality of scientific data (Pamies et al., 2022).

**Good In Vitro Method Practices (GIVIMP)**: A comprehensive framework described in the OECD Guidance Document on Good In Vitro Method Practices (OECD, 2018) that provides recommendations for the development, validation, regulatory acceptance or qualification, and use of *in vitro* methods.

**Good Laboratory Practices (GLP)**: Regulations promulgated by authorities such as the U.S. EPA, U.S. FDA, and OECD that describe record keeping and quality assurance procedures for laboratory records that will be the basis for data submissions to national regulatory agencies.

**Hazard**: The potential for an adverse or harmful health or ecological effect.

**Hazard classification**: Assignment of a chemical or product hazard to a category based on the results of a standard test method for a specific toxicity endpoint; most commonly used for labeling purposes.

**Hazard identification**: That part of risk assessment associated with the determination of whether exposure to a particular substance is or might be associated with adverse health or ecological effects.

**Integrated Approach to Testing and Assessment (IATA)**: An approach based on multiple information sources used for the hazard identification, hazard characterization and/or safety assessment of chemicals (OECD, 2017).

**Interlaboratory reproducibility**: A measure of whether different qualified laboratories using the same protocol and test chemicals can produce qualitatively and quantitatively similar results. Interlaboratory reproducibility is determined during the prevalidation and validation processes and indicates the extent to which a test method can be transferred successfully among laboratories.

**Mechanistic relevance**: A measure of appropriateness for assessing the biochemical process or pathway by which a chemical may exert an effect.

**Negative predictivity**: The proportion of correct negative responses relative to the defined reference data among substances testing negative by a test method. It is one indicator of test method accuracy. Negative predictivity is a function of the specificity of the test method and the prevalence of negatives among the substances tested.

**New approach methodology (NAM)**: A broadly descriptive reference to any technology, methodology, approach, or combination thereof that can be used to provide information on chemical hazard and risk assessment and that supports replacement, reduction, or refinement of animal use (3Rs).

**Performance**: The accuracy and reliability characteristics of a test method (see “accuracy”, “reliability”).

**Performance standards**: Standards, based on a validated test method, that provide a basis for evaluating the comparability of a proposed test method that is mechanistically and functionally similar. Included are (1) essential test method components; (2) a list of reference chemicals selected from among the chemicals used to demonstrate the acceptable performance of the validated test method; and (3) the comparable levels of accuracy and reliability, based on what was obtained for the validated test method, that the proposed test method should demonstrate when evaluated using the minimum list of reference chemicals.

**Positive predictivity:** The proportion of correct positive responses among substances testing positive relative to the defined reference data by a test method. It is one indicator of test method accuracy. Positive predictivity is a function of the sensitivity of the test method and the prevalence of positives among the substances tested.

**Potency**: A measure of the relative biological or chemical activity of a substance. The potency of a single substance can differ for different biological or biochemical effects.

**Precision**: The closeness of individual measurements of an analyte after multiple analyses, often of a single sample. Precision is often expressed as the coefficient of variation.

**Protocol**: The precise step-by-step description of a test method, including the listing of all necessary reagents and all criteria and procedures for generating and evaluating test data.

**Qualification**: A conclusion that the results of an assessment using a validated model or assay can be relied upon to have a specific interpretation and application in product development and regulatory decision-making.

**Quality control**: A set of activities or samples used to verify that the quality of the product or method is maintained as intended.

**Reference compounds**: Chemicals selected for use during the research, development, or evaluation of a proposed test method because their response in the reference test method or the species of interest is known (see "reference test method").

**Reference test method**: The accepted test method used for regulatory purposes to evaluate the potential of a test substance to be hazardous to the species of interest.

**Reliability**: A measure of the degree to which a test method can be performed reproducibly within and among laboratories over time. It is assessed by calculating intra- and interlaboratory reproducibility and intra-laboratory repeatability.

**Repeatability**: The consistency of test results obtained when the procedure is performed on the same substance under identical conditions within a given time period; "…the closeness of the agreement between independent results obtained with the same method on the identical subject(s) (or object or test material), *under the same conditions*" (Slezák and Waczulíková, 2011).

**Reproducibility**: The consistency of individual test results obtained using the same test protocol and test samples; "…the closeness of the agreement between independent results obtained with the same method on the identical subject(s) (or object, or test material), but *under different conditions* (different observers, laboratories etc.)" (Slezák and Waczulíková, 2011).

**Recovery**: A quantitative method of verifying the efficiency and reproducibility of an extraction method by comparing the results of the extracted samples with those of spiked samples of a similar matrix and/or spiked blanks.

**Risk assessment**: Evaluation of the potential adverse health and environmental effects to a target species from exposures to exogenous agents.

**Robustness**: The ability of a method to be reproduced under different conditions or circumstances, without the occurrence of unexpected differences in the obtained results.

**Sensitivity**: The proportion of all positive chemicals that are classified correctly as positive in a test method. Sensitivity may also be defined in the context of detection limits as the lowest analyte concentration that can be measured with acceptable accuracy and/or precision.

**Specificity**: The proportion of all negative chemicals that are classified correctly as negative in a test method. It is a measure of test method accuracy. This word is also used to describe the ability of an analytical method to detect a specific analyte.

**Stability**: The ability of a test material (e.g., test substance, testing apparatus, reagent, or analyte) to produce similar and acceptable results over a period of time in a given environment.

**Standard curve**: A quantitative method of plotting assay data to determine the concentration of a substance in an unknown sample by comparing the unknown to a standard sample of known concentration (often using the positive control material).

**Standard operating procedures (SOPs)**: Formal, written procedures that describe how specific laboratory operations are to be performed. These are required by GLP guidelines.

**Target organ**: The organ for which information on the potential toxicity of a chemical is sought.

**Target species**: The species for which information on the potential toxicity of a chemical is sought.

**Test**: The experimental system used; used interchangeably with “test method” and “assay”.

**Test method**: A process or procedure used to obtain information on the characteristics of a substance or agent. Toxicological test methods generate information regarding the ability of a substance or agent to produce a specified biological effect under specified conditions. Used interchangeably with “test” and “assay”. See also “validated test method” and “reference test method”.

**Test method developer**: The organization or individual that initially devises a test method and ensures its reproducibility and suitability for the intended use.

**Test method sponsor**: The organization or individual that puts forward a test method submission for consideration; may also be the same organization or individual as the test method developer.

**Test method submission**: Compendium of supporting documentation for a test method that is proposed for a regulatory or other defined application. A test method submission generally includes records of validation studies that have been completed to characterize the usefulness and limitations of the test method for a specific proposed regulatory testing requirement or application as well as other adequate documentation of the scientific validity prepared in accordance with test method submission guidelines.

**Transferability**: The ability of a test method or procedure to be accurately and reliably performed in different, competent laboratories. See also “interlaboratory reproducibility”.

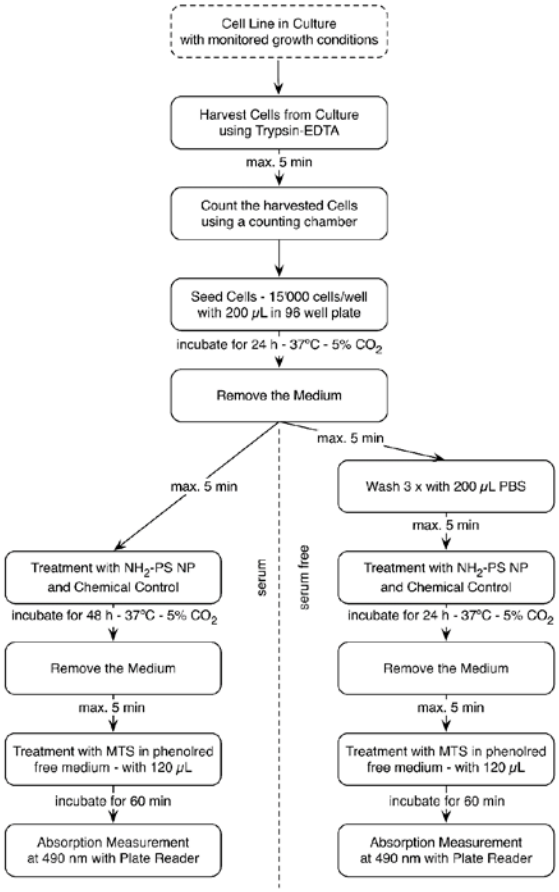
**Validated test method**: An accepted test method for which validation studies have been completed to determine the accuracy, reliability, and relevance of this method for a specific proposed use.

**Validation**: The process by which the accuracy, reliability, and relevance of a procedure are established for a specific purpose. Validation for one specific purpose does not imply validation for other specific purposes. Further qualification may be needed for a particular context of use.

# APPENDIX B: QUALITY TOOLS

## 1.0 Flow Charts

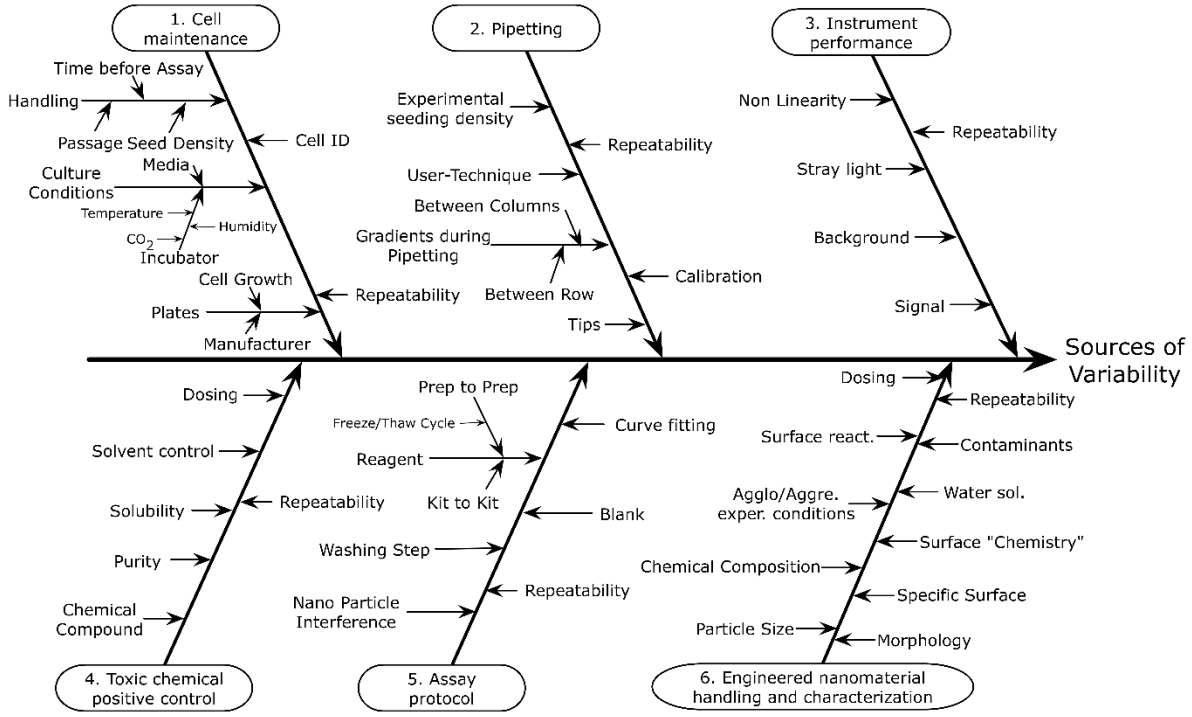
In a flow chart, every step in a protocol is diagrammed (see Figure S1 for an example). This can be helpful in designing control measurements for the experiment by ensuring that each step is covered and tracked when feasible. It is possible for more than one control measurement to cover a single step, and conversely a control measurement may cover multiple steps. In addition, comparing control measurements among NAMs may reveal similar steps among different NAMs. For these steps, the sources of variability will also likely be similar.




**Supplemental Figure 1**. Flow chart describing the modified MTS (3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium) protocol for an interlaboratory study. This figure is reprinted with permission from Elliott et al. (2017).

## 2.0 Cause-and-Effect Analysis

Cause-and-effect (C&E) analysis is a conceptual tool that can be used to identify possible key sources of variability and display them using C&E diagrams, sometimes known as fish-bone diagrams (see Figure S2 for an example). The process of developing these diagrams can include brainstorming and reviewing the literature on an assay. Each branch of the C&E diagram indicates a key source of expected variability. C&E diagrams can also support development of new NAMs, because shared branches of C&E diagrams (e.g., performing measurements using the same type of cytotoxicity assay) can be made more quickly, will likely require similar control measurements, and will likely have similar variability mitigation strategies. Analysis of C&E diagrams can help identify aspects of a method that may be challenging to standardize (e.g., an instrument challenging to calibrate or unstable assay reagents). C&E diagrams can help guide robustness testing and the selection of control measurements so that, ideally, the sources of variability in each branch and subbranch of the C&E diagram are analyzed.



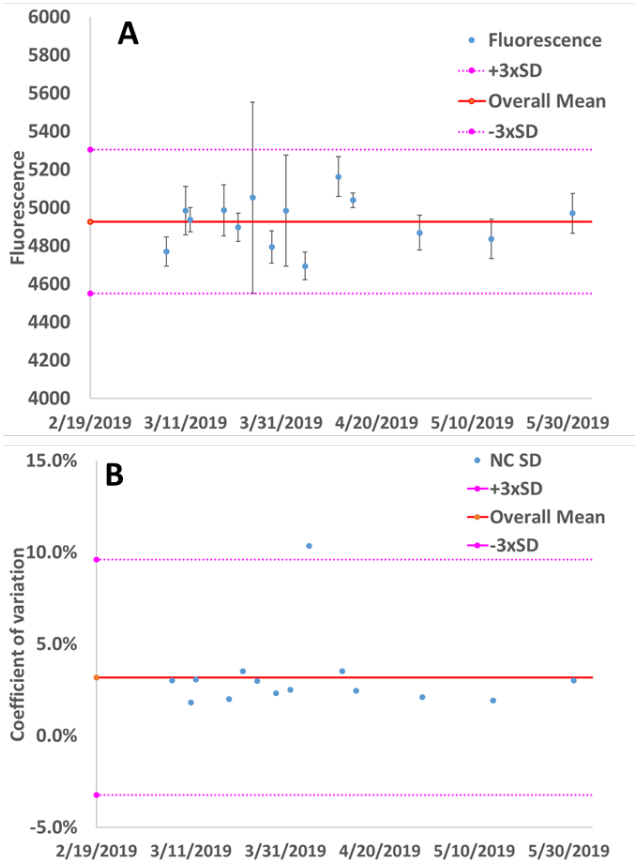

**Supplemental Figure 2.** Cause-and-effect diagram of an MTS cytotoxicity assay designed for use with engineered nanomaterials. Reprinted with permission from Rösslein et al. (2015).

## 3.0 Control Charts

Control measurements can be one-time preliminary experiments (e.g., to test for potential biases), periodic measurements performed at a predetermined frequency (e.g., evaluation of instrument performance and instrument calibration), and in-process control measurements made each time the assay is performed. For some assays that allow for a limited number of samples to

be analyzed concurrently (e.g., an inhalation assay with exposure to aerosolized chemicals using a flowthrough exposure system), periodic measurements can be helpful. One-time preliminary experiments can be helpful to evaluate if a test substance may cause a bias, for example, if a nanoparticle may adsorb a key assay reagent. In-process control measurements can be used to measure key sources of variability each time the assay is performed. For example, one common in-process control measurement is the positive control. This control measurement can reveal if a maximal response (100 % effect) is reached and can also be used to demonstrate the sensitivity of the assay response to chemical concentrations that yield lower, more moderate responses. Key considerations for selecting a positive control have been described by Petersen et al. (2021). Additional common in-process control measurements are a control with no cells and no additional assay reagents and a control with cells and additional assay reagents added but without test substances. It may not be possible to include all potential in-process control measurements, and there may be tradeoffs in terms of what in-process control measurements to include.

Key sources of variability can be monitored across time and among experiments using control charts that display the mean values and variability of in-process control measurements (see Figure S3 for an example). This can be helpful to assess if there are systematic changes in the mean or variability values across time which suggest that something in the assay may have changed (e.g., instability of a reagent). Reviewing the check sheets can be helpful to identify why changes occurred.



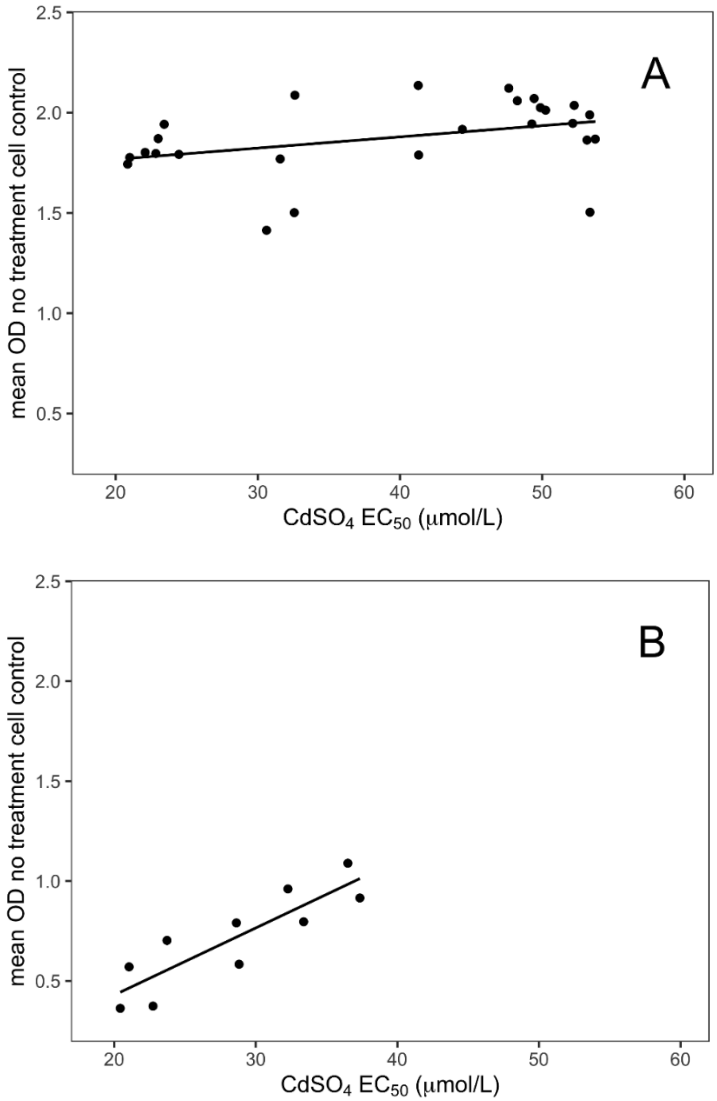

**Supplemental Figure 3**. Control charting data for the electrophilic allergen screening assay fluorescence method for the negative incubator control: (A) mean and (B) coefficient of variation for all experiments depending on the date they were performed. In graph B, one value is an outlier for the coefficient of variation and outside of the specifications for this study (overall mean ± 3 times the average standard deviation value). Also, there is not a systematic trend with either the mean or coefficient of variation values across time. This figure has been modified and reprinted with permission from Petersen et al. (2022c) while the figure caption is modified and reprinted from Petersen et al. (2022b).

## 4.0 Check Sheets

To record key metadata and fulfill requirements for GLP, check sheets can be used. If GLP is not used when a method is being developed, principles laid out in GLP such as recording key data using check sheets will be important if the NAMs are intended for regulatory submissions, when GLP-compliance may be required. Alternative relevant resources during assay development are the guidance document on GCCP 2.0 (Pamies et al., 2022) and GIVIMP (OECD, 2018) that are specific to cell-based work with additional quality management lists to ensure reproducibility and high-quality scientific data. Check sheets can also be useful for monitoring the in-process control measurements and supporting troubleshooting when issues arise. For example, they can be used to document changes in lot numbers for reagents and consumable supplies (e.g., pipettes and microwell plates). This can be valuable for troubleshooting if one specification is not met more frequently than other specifications. It is also possible for check sheets that include data calculators to evaluate if data from an assay run meets all the specifications and to perform statistical evaluations of the assay results. Electronic notebooks and laboratory information management systems may be used instead of written check sheets (with advantages such as ease of storage, transferability among laboratories, and searchability).

## 5.0 Scatterplots

Scatterplots can be used to assess if there is an interaction between different in-process control measurements or between those control measurements and test substance results (see Figure S4 for an example). The test substance results should be independent of the results for in-process control measurements within the range set by specifications.




**Supplemental Figure 4.** Correlation of cadmium sulfate ($CdSO_4$) $EC_{50}$ values determined using the MTS assay with mean optical density (OD) values of A549 cells not exposed to a test substance or the positive control.These data either show a lack of an interaction between the $EC_{50}$ values (part A) or an interaction (part B) depending upon the range of mean OD values. The solid lines are linear regression fits. The slope in part B is statistically different from 0, indicating that the $EC_{50}$ value is correlated with the OD values. Modified and reprinted with permission from Petersen et al. (2022b) and the figure caption is reprinted with permission from Petersen et al. (2022b).

# APPENDIX C: ANALYTICAL METHOD ASSESSMENT

The information in Appendix C was derived from several references, including "Validation and Regulatory Acceptance of Toxicological Test Methods" (ICCVAM, 1997), "Bioanalytical Method Validation Guidance for Industry" (FDA, 2018), OECD GD 34 (OECD, 2005), and the OECD GIVIMP document (OECD, 2018). The information in Appendix C is intended to provide a general framework for method assessment. Some aspects may or may not be applicable to all methods.

## 1.0 Limits of Detection and Quantification

The limits of detection and quantification define the range of an analytical method. The analytical method is often tested either at the lower limit of quantification or at the limit of detection, whichever is more appropriate for the system. This process determines the lower end of the range of an analytical method by establishing the lowest quantity or concentration of an analyte that can be reliably detected (limit of detection) or quantified (lower limit of quantification) above the reagent blank or within the standard curve. The highest concentration on the standard curve determines the upper limit of quantification.

## 2.0 Identification of Interference

Method developers should document interfering substances, which can come from critical and non-critical components of the method, including any interference with the detected signal (e.g., fluorescence/absorbance, luciferase, enzymatic) of the method. Interference can also come from consumables, such as certain plastics in endocrine disruptor test methods as well as other reagents in the method. Potential interfering substances include, but are not limited to, endogenous matrix components, metabolites, decomposition products, and other xenobiotics. If the method is intended to quantify more than one analyte, all intended analytes should be tested to ensure that there is no interference. Blank samples are often used to test for interference.

## 3.0 Assessing Analytical Precision

It is helpful to include a description of the precision of the analytical method used and any other tests of precision, such as those assessing performance variability when different personnel use the proposed method, or when different instrumentation is used for the method, as well as participating in interlaboratory comparison studies when possible.

## 4.0 Stability of Materials

Stability as it relates to materials used in the NAM (e.g., test substances, testing apparatus, reagents, and analytes) refers to a material's ability to produce similar and acceptable results over a period of time in a given environment. The stability of the NAM itself is discussed earlier (Appendix B, Section 3.0) with evaluation of control charts across time. The effects of sample collection, handling, and storage conditions should be evaluated. Materials that can be affected by stability issues should be given expiration dates appropriately. The stability of a test

substance, reagents, and testing apparatus (e.g., plastic microplate) should be ensured to avoid interferences from degradation products and changes to the applied actual dose. Stability studies performed on any of the components should be included in reporting. The stability of any chemical mixtures or prepared samples being prepared before the day of the study should be evaluated for stability before use in development, validation, or qualification studies.

The chemical stability of a given mixture or matrix under specific conditions for given time intervals is assessed in several ways. Pre-study stability evaluations should cover the expected sample handling and storage conditions during the conduct of the study, including conditions at the test site, during shipment, and at all other secondary sites. The stability of an analyte in a particular mixture, matrix, and container system is relevant only to that mixture, matrix, and container system and should not be extrapolated to other systems. Stability testing should evaluate the stability of the analytes for long-term (frozen at the intended storage temperature) and short-term (bench top, room temperature) storage, and after freeze and thaw cycles and the analytical process. Conditions used in stability experiments should reflect situations likely to be encountered during actual sample handling and analysis. If, during sample analysis for a study, storage conditions changed and/or exceeded the sample storage conditions evaluated during technical characterization of the method, stability should be established under these new conditions. Independent stability studies should be conducted and cover any condition a critical reagent may encounter. Stability testing of a reagent or material should be considered for a variety of contexts, including during freeze-thaw cycles and use on the benchtop, in stock solutions, in processed samples, and over long-term studies.

## 5.0 Robustness Testing

Robustness is the ability of a method to be reproduced under different conditions or circumstances without the occurrence of unexpected differences in the obtained results. One aspect of a NAM's robustness is the consistency of the control charting results across time as described in Appendix B, Section 3.0. Robustness testing is often used to detect changes in results from unintended variations in experimental reagents or protocols. Robustness testing is recommended for all aspects of test methods, and ranges for all parameters and measurements should be established whenever and wherever possible.

For example, an incubation time of 5 minutes was established as optimal for a study, but after robustness testing, data passed all quality control requirements at 5 minutes plus or minus 30 seconds. Therefore, the robustness-tested acceptable incubation time would be 5 minutes ± 30 seconds.

Some study parameters that should have an established acceptance range include:

- Incubation times
- Incubation temperatures
- pH
- Sources of reagents

- Cell densities (if applicable)
- Experimental conditions
- Analysis software

The design of the model will determine what experimental conditions will require acceptance ranges. For example, models with flowing media will need acceptance ranges for parameters related to flow. More complex models are more likely to have study parameters requiring control.

When applicable, results of robustness testing for critical and non-critical reagents should also be reported. Different suppliers (whenever practical) should be tested to determine if a reagent should be purchased from one supplier or if multiple suppliers can be used. This can also be helpful for avoiding supply chain disruptions or if a manufacturer stops making a specific product.

Instrumentational robustness testing should also be conducted whenever applicable. Cross-laboratory validation often involves different brands of instruments with different performance parameters and capabilities, which can add to the variability of the data or change parameters of the method performed.

## 6.0 Analysis of Recovery

Studies performing tests on extractions from a given matrix should perform recovery studies to verify the efficiency and reproducibility of the extraction method. Recovery studies are often performed by comparing the results of the extracted samples with those of spiked samples of a similar matrix and/or spiked blanks.

## 7.0 Technical Analysis of Applicability Domain

There should be adequate test method data for chemicals and/or products representative of those relevant to the specific COU for which the test is proposed. Documentation for the NAM, DA, or IATA should clearly describe the physicochemical properties of the applicability domain. This would include any limitation of the method, such as for chemicals of a specific molecular weight range, volatility, solubility, or stability.

While the applicability domain should be analytically evaluated during the initial development of an assay, additional data may become available through broader usage of the method. These data could reveal interferences that may need to be further addressed technically (Petersen et al., 2022a). If an assay is not evaluated for chemicals with certain physicochemical properties, there will be greater uncertainty regarding the assay performance for these chemicals both from analytical (e.g., whether there are biases that impact the assay performance) as well as potential concordance (e.g., whether the assay yields similar results as an *in vivo* assay) considerations. Therefore, it may also be valuable to assess the assay concordance across chemicals with

different physico-chemical properties to increase confidence in the applicability of the assay with a broader range of chemicals.

## 8.0 Positive Control

At least one positive control should be part of the assay development. These controls can be used in conjunction with standard curves or as a single concentration for the positive control. The positive controls should be relevant to the endpoint of detection for the assay. The concentration of the positive control should lie within the detection window for the assay. Often, a positive control concentration is chosen to yield a response between 50% and 80% of the maximum detection range if only a single concentration is tested and data on the resolution of the maximum dose is available. A standard curve is often produced using the positive control for the system, with a standard curve defined as the relationship between assay response and known concentrations of the analyte. The standard curve should be reproducible over time. Standard curves should be prepared in the same vehicle or solvent as the intended test samples of the study. When the same mixture or matrix cannot be obtained, surrogate mixtures or matrixes can be used with proper documentation and justification.

## 9.0 Reference Standards for Instrument Calibration

When applicable, instruments should be calibrated using calibration standards and/or quality control samples. The source of the standard or quality control sample should be documented, including purity, stability, source, lot number, certificate of analysis, and expiration date. There are several types of standards such as U.S. Pharmacopeia compendial standards and commercially supplied materials obtained from a reputable commercial source with documented purity.

Ideally, the control standard for some instruments (e.g., mass spectrometers) should be identical to the analyte of interest in the assay. Often this is not possible, and an established chemical form of known purity can be used.

## 10.0 Setting Specifications

Specifications can be set for in-process control measurements based on intra-laboratory or interlaboratory test results. There are multiple approaches that can be used. For example, commonly used methods apply a 95% confidence interval to the in-process control measurements or use the mean ± two or three times the standard deviation value, but other statistical approaches are often used as well. Setting specifications involves balancing different objectives. It is important to have the specifications sufficiently stringent to exclude data indicating that the assay is not working as expected or that the data for an in-process control measurement is within a range where it could bias the test substance result. It is also important to not set the specifications so tightly that tests with test data in what would be considered a normal range have an overly high number of assay runs fail to meet the specifications.