Organisation for Economic Co-operation and Development

ENV/CBC/MONO(2025)18

Unclassified

English - Or. English

31 October 2025

ENVIRONMENT DIRECTORATE
CHEMICALS AND BIOTECHNOLOGY COMMITTEE

Cancels & replaces the same document of 17 October 2025

OECD Guidance Document on the Generation, Reporting and Use of Research Data for Regulatory Assessments

Series on Testing and Assessment No 417

This document aims to enhance the consideration and use of research data in regulatory assessments by OECD Member Countries. The aim is to bridge the gap between the increasing amount of non-standard research data and the need for robust scientific evidence to inform regulatory assessments. The guidance targets all stakeholder groups involved in the life cycle of research data, from generation to regulatory use. Research funders, researchers, publishers, reviewers, editors, repository managers, assessors from public and private organisations, and risk managers share responsibilities to improve the regulatory uptake of research data.

JT03575638

Please cite this publication as:

OECD (2025), *OECD Guidance Document on the Generation, Reporting and Use of Research Data for Regulatory Assessments*, OECD Series on Testing and Assessment, No. 417, OECD Environment, Health and Safety, Paris, https://one.oecd.org/document/ENV/CBC/MONO(2025)18/en/pdf

Contact us

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

E-mail: ehscont@oecd.org

About the OECD

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

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

Foreword

At the 6th meeting of the Working Party on Hazard Assessment (WPHA) held on 22-24 June 2022, a proposal from the Joint Research Centre (JRC) of the European Commission to develop a Guidance Document (GD) to improve the use of research data in regulatory assessments was presented. The project was added to the WPHA workplan in Q3 2022. To support the work, delegates of the WPHA were asked for nominations of experts, and the kick-off meeting of the Expert Group took place in December 2022.

Phase I of the project was a scoping exercise, including problem formulation and expected outcomes. To shape the scope of the GD, the JRC hosted a workshop: *Improving the use of academic data in regulatory assessments*, in Ispra (Italy), in October 2022. At the workshop, it was decided that the document should include guidance for integrating non-standard, non-guideline chemical data published in scientific literature and found in various databases (ToxCast, (Q)SAR prediction, etc.) in regulatory risk assessments. It was also agreed that the document should include guidance for setting quality and reporting standards, and guidance for finding and retrieving data. In addition, the guidance is intended to be relevant to the research community generating data, as well as regulators using data in the assessment of chemicals.

The OECD Expert Group met regularly via teleconference in 2023 and 2024 to develop guidance, quality and reporting standards, and case studies illustrating the review and use of academic data in regulatory assessments. In addition, a draft meeting of the Expert Group was held at OECD headquarters in April 2024 to advance a complete draft document.

To promote the project within the academic research community, JRC and Sweden hosted a webinar in January 2024 titled *Good practices and resources to improve the utility of research data in regulatory assessment*. Presenters included members of the Expert Group, along with representatives from European research initiatives (ASPIS and PARC). The webinar attracted over 200 participants and featured an engaging Question and Answer session.

The complete draft GD was circulated for review and written comment to the WPHA in July 2024. Comments were addressed by the OECD Expert Group and JRC and the revised draft *OECD Guidance Document on the Generation, Reporting and Use of Research Data for Regulatory Assessments* was circulated to the WPHA for a second round of comments in January 2025. The revised final draft was approved by the WPHA by written procedure in May 2025.

Table of contents

Tables

Figures

Boxes

List of abbreviations

AI	Artificial Intelligence
AOP	Adverse Outcome Pathway
AOP-KB	Adverse Outcome Pathway – Knowledge Base
CAS RN	Chemical Abstract Service Registration Number
CRED	Criteria for Reporting and Assessing Ecotoxicity Studies
DA	Defined Approach
DOI	Digital Object Identifier
EASIS	Endocrine Active Substances Information System
ECHA	European Chemicals Agency
EFSA	European Food Safety Authority
FAIR	Findable, Accessible, Interoperable and Reusable (data)
GIVIMP	Good *in Vitro* Method Practice
GLP	Good Laboratory Practice
HAWC	Health Assessment Workspace Collaborative
IATA	Integrated Approaches for Testing and Assessment
IPCHEM	Information Platform for Chemical Monitoring
IRIS	Integrated Risk Information System
IUCLID	Integrated Uniform Chemical Information Database
NAM	New Approach Methodology
NTP-OHAT	National Toxicology Program - Office for Health Assessment and Translation
OECD	Organisation for Economic Cooperation and Development
OECD TG	OECD Guideline for the Testing of Chemicals
OHT	OECD Harmonised Template
PARC	Partnership for Chemicals Risk Assessment
PBK	Physiologically Based Kinetic (models)
PECO	Population, Exposure, Comparator, Outcome
PFAS	Per- and Polyfluoroalkyl Substances
(Q)SAR	(Quantitative) Structure-Activity Relationship
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
RoB	Risk of Bias
SciRAP	Science in Risk Assessment and Policy (tool)
SEM	Systematic Evidence Map
SOP	Standard Operating Procedure
SR	Systematic Review
SSD	Species Sensitivity Distribution
US EPA	United States Environmental Protection Agency
WoE	Weight of Evidence

Executive summary

The *OECD Guidance Document on the Generation, Reporting and Use of Research Data* aims to enhance the consideration and use of research data in regulatory assessments by OECD Member Countries. The Guidance (GD) aims to bridge the gap between the increasing amount of non-standard research data and the need for robust scientific evidence to inform regulatory assessments. Regulatory frameworks for chemicals strive to use all available scientific evidence including data from internationally recognised regulatory standards (e.g., OECD Guidelines for the Testing of Chemicals), and non-standard research data. However, the consideration of research data for regulatory purposes is challenging due to varying reliability and reporting standards.

The GD targets all stakeholder groups involved in the life cycle of research data, from generation to regulatory use. Research funders, researchers, publishers, reviewers, editors, repository managers, assessors from public and private organisations, and risk managers share responsibilities to improve the regulatory uptake of research data.

The document is structured into four main sections, and annexes, which provide detailed resources and case studies. Section 1 introduces the objectives of the Guidance and discusses general principles of data quality, scientific and regulatory relevance, and reliability. Section 2 describes existing resources and good practice to increase the utility of research data in regulatory contexts. This includes targeting regulatory needs, adhering to reporting guidance, and publishing data in accessible formats. Section 3 outlines structured approaches for assessors to identify, screen, evaluate and integrate research data, including systematic review methodologies and tools to evaluate relevance and reliability. Section 4 offers specific recommendations to various stakeholder groups.

In addition, Annex A lists available resources supporting the design, conduct, and report of specific types of research data. Annex B, provides information on available repositories and software for storing, sharing, searching, and screening research or regulatory data. Annex C provides examples of regulatory contexts where research data has been considered in regulatory assessments (non-exhaustive). Finaly, Annex D comprises four case studies as mentioned below (with the leads highlighted in bold):

- Case Study A. *Characterising human health evidence for 500+PFAS: interoperability of workflows*. Developed by **US EPA**, EU EFSA, and Health Canada.
- Case Study B. *Identification of an endocrine disruptor in the EU regulatory context. Identifying best practices on how research data can assist the regulatory assessment of Endocrine Disruptors*. Developed by **EU JRC**, **Sweden**, Germany BfR, and BIAC.
- Case Study C. *The CRED Evaluation Method: A transparent and structured method for evaluation of ecotoxicity data used in risk assessment*. Developed by **Switzerland BAFU**, and Germany UBA.
- Case Study D. *Submission and incorporation of peer-reviewed literature for pesticides approval.* Developed by **EU EFSA**, Australia, Canada, Switzerland BAFU, and BIAC.

The GD was developed by the *ad hoc* OECD Expert Group on Research Data (see Table 1, below), with the scientific coordination from the European Commission's Joint Research Centre (project lead).

The implementation of the GD is expected to enhance regulatory efficiency and coherence across policy domains and jurisdictions benefiting all OECD Member Countries. The GD will benefit from periodic updates.

Table 1. Members of the *ad hoc* OECD Expert Group on Research Data

Name	Affiliation	Representing
Antonio Franco (project co-lead)	Joint Research Centre (JRC)*	European Union*
Eleonora Chinchio (project co-lead)	JRC	European Union
Effrosyni Katsanou (project co-lead)	JRC	European Union
Andrew Worth (project co-lead)	JRC	European Union
Francisco Sanchez-Bayo	Department of Climate Change, Energy, the Environment and Water	Australia
Michael Beking	Environment and Climate Change Canada	Canada
Nazem El Husseini	Health Canada	Canada
Clotilde Maurice	Health Canada	Canada
Shamika Wickramasuriya	Health Canada	Canada
Tanja Burgdorf	German Federal Institute for Risk Assessment (BfR)	Germany
Enken Hassold	German Environment Agency, (UBA)	Germany
Franziska Kaßner	UBA	Germany
Peter von der Ohe	UBA	Germany
Asako Hotta	National Institute of Technology and Evaluation (NITE)	Japan
Sang Hee Lee	Ministry of Environment - National Institute of Environmental Research	Korea
Betty Hakkert	National Institute for Public Health and the Environment (RIVM)	Netherlands
Anne Kienhuis	RIVM	Netherlands
Caroline Moermond	RIVM	Netherlands
Petra Van Kesteren	RIVM	Netherlands
Marlene Ågerstrand	Stockholm University	Sweden
Anna Beronius	Karolinska Institutet	Sweden
Muris Korkaric	Federal Office for the Environment, (FOEN)	Switzerland
Mireia Marti-Roura	FOEN	Switzerland
Timothy Gant	UK Health Security Agency (UKHSA)	United Kingdom
Morné van der Mescht	Environment Agency	United Kingdom
Anna Lowit	US Environmental Protection Agency, (US EPA)	United States
Jennifer Nichols	US EPA	United States
Kristina Thayer	US EPA*	United States*
Sean Watford	US EPA*	United States*
Fulvio Barizzone	European Food Safety Authority (EFSA)	European Union
Laurent Lagadic	Bayer Crop Science AG	Business at OECD (BIAC)
Steven L. Levine	Bayer Crop Science	BIAC
Ellen Mihaich	Environmental and Regulatory Resources	BIAC
Sandrine Sourisseau	Total Energies	BIAC
William "Jay" West	American Chemistry Council	BIAC
Scott Belcher	North Carolina State University - Center for Health and Human Environment	Endocrine Society
Laura Vandenberg	University of Massachusetts	Endocrine Society
Charlie Stevenson	Cruelty Free International	ICAPO

Note: *- Affiliation listed reflects the author's institution at the time this work was conducted.

1 Introduction

1.1 Objective and scope of this Guidance

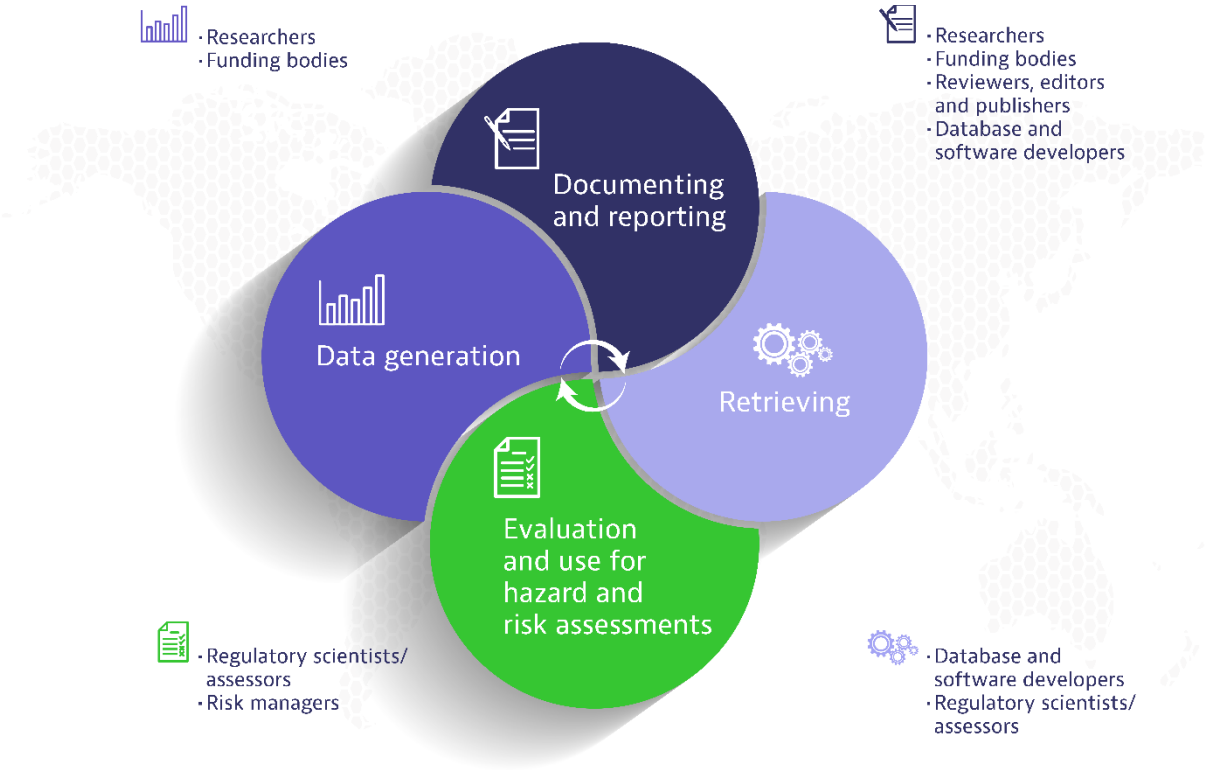
Regulatory systems strive to make best use of all scientific evidence to inform assessment and management of chemicals. Generally, regulatory frameworks promote the use of data generated using internationally recognised standards (e.g., OECD Guidelines for the Testing of Chemicals, OECD TGs), when establishing regulatory information requirements. Typically, companies generate guideline data to comply with regulations or the requirements of certain government programmes. Research data from other sources (often referred to as "non-guideline") informs regulatory assessment processes alongside data generated to comply with regulatory information requirements.

"Research data" is defined here as any scientific data generated in a research context that could potentially inform hazard, exposure, and/or risk assessments of chemicals. Scientists from academia, public and private research institutes, industry, or non-governmental organisations can generate research data. The consideration of studies conducted according to internationally agreed test method guidelines, as well as non-guideline research data is necessary to comply with the legal requirement to take all scientific evidence into account when conducting assessments. Relevant and reliable research data add to the scientific evidence base and may be given just as much weight as guideline studies in regulatory assessment frameworks.

The overall objective of this Guidance is to improve the utility and uptake of research data in regulatory assessments (including hazard classification and risk assessment). The guidance aims to raise awareness of the benefits and available resources to improve the value of research data for regulatory consideration, and to improve the use of research data in regulatory assessment and decision-making. Several groups are involved in the life cycle of research data, from data generation to regulatory use (Figure 1.1)

The scope of the Guidance reflects the broad definition of research data given above. The emphasis is on primary data as opposed to secondary data (reviews, meta-analysis). That includes data from human and environmental observational studies, data obtained using experimental methods (e.g., *in vivo*, *in vitro*, omics, monitoring), and computational (*in silico*) methods (e.g., (Quantitative) Structure-Activity Relationships ((Q)SARs), Physiologically Based Kinetic (PBK) models). More specifically, the focus is mostly on toxicity, ecotoxicity and human observational studies. Studies generating research data do not usually follow national or international regulatory standards, such as those adopted by the OECD (OECD TGs). In fact, rigid study design may not serve hypothesis-driven research well due to different aims, resources, and ethical constraints. Nonetheless, research data may add valuable evidence by addressing endpoints, (eco)toxicological pathways and species that are not necessarily covered by regulatory standards. This includes studies for which regulatory standards do not exist (e.g., most non-animal methods, epidemiological studies) and studies focused on the development or evaluation of new methods. The development of new approach methodologies (NAMs) in toxicology and ecotoxicology, in particular, is generating an increasing amount of published, peer reviewed non-standard research data.

Figure 1.1. Steps and main groups involved from production to regulatory use of research data



Note: Groups are defined by their function. Different stakeholder groups may share similar functions. For example, "regulatory scientists/assessors" includes registrants, consultants, and public authorities.

Accessibility of research data is obviously essential for assessors. Research data are typically published in peer reviewed scientific literature but can also be found in grey literature (e.g., dissertation theses, scientific reports). Public funding policies nowadays generally adhere to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) to ensure that research data is openly accessible and reusable to the largest extent possible (e.g., OECD recommendation concerning access to research data from public funding (OECD, 2018b; OECD, 2020a), EU research programme policy (European Commission, 2017)). While referring to generic open science policies, some funding programmes promote or mandate specific solutions for publishing research. For example, EU-funded programmes currently require participants to publish in full open access journals and make data available in repositories to receive funding. It is important that such repositories are accessible and sustainable. Publication of research data in the grey literature may complicate findability and accessibility but is one way to publicly share data that would otherwise be unavailable to assessors (e.g., from industry research programmes).

Research studies published in scientific journals are generally subject to an independent peer review of the study methods, results, and potential impact (or relevance) of the research. Assessing study reliability and relevance in a regulatory context serves different objectives compared to journal peer review processes. Regulatory relevance, in particular, is context-specific and inherently different from scientific relevance (Rudén et al., 2017). Thus, assessors need to identify, understand, and evaluate the reliability and relevance of an increasing amount of research data generated by a wide variety of different methods and models, with variable reliability and reporting standards. Compared to the appraisal of studies following internationally adopted standards (e.g., OECD TGs), assessing reliability and relevance of research data is generally more technically challenging and time-consuming.

Numerous guidance documents on good practices, reporting standards, and tools for different types of research data and underpinning methods (e.g., *in silico*, *in vitro*, *in vivo*, omics approaches) are available. These were developed by international organisations (e.g., OECD), national authorities, scientific societies, independent research groups, and communities of practice[1]. These resources facilitate the regulatory uptake of research data. Guidance is also in place across countries and policy areas to aid assessors with the identification, screening, evaluation, and integration of scientific evidence, including research data. Annex A and Annex B provide a non-exhaustive list of such resources.

Although most of the existing resources are publicly available, the generation and publication of research data do not always follow guidance and reporting standards. Scientists may not be aware of existing good practice and reporting standards, may not appreciate the utility of research data in a regulatory context, or may lack incentives to follow these practices when publishing research. The implementation of practices that increase regulatory reliability and relevance requires time and is often an unrewarded task for scientists, even though it would make the review process easier and would support consistency in the outcomes of assessments.

Assessors regularly use guidance and tools to facilitate structured screening and evaluation of research data in regulatory contexts. Even if many core considerations for use of data in regulatory applications are similar across countries or within the same jurisdiction across policy domains (general chemicals, pesticides, biocides, etc.), approaches for identifying information and criteria for regulatory consideration of research data have mostly developed independently and are not harmonised.

1.2 Target audience and benefits of this Guidance Document

The Guidance aims to serve as a reference point to benefit all stakeholders involved in the life cycle of research data from production to regulatory uptake. Harmonising approaches to consider research data in regulatory assessments brings multiple benefits to the groups shown in Figure 1.1.

- Research funders can use it to design, monitor, and evaluate research programmes.
- Researchers can design, perform and report research to maximise scientific and regulatory value (Figure 1.2).
- Editors, publishers, and reviewers can refine publication policies to improve data accessibility, method reporting, and overall quality.
- Repository managers can enforce data policies and update software for storing and sharing research data.
- Assessors can harmonise assessment workflows for screening and evaluating research data to enhance efficiency and reuse elements like search tools and systematic review outcomes).
- Risk managers can develop policies with a stronger evidence base and reuse assessments across regulatory frameworks, supporting legal requirements to consider all scientific evidence.

In line with the 3Rs principle (replacement, reduction, and refinement of animal studies), many research funding schemes in (eco)toxicology and several regulatory programmes (e.g., US EPA, 2018; European Commission, 2023), prioritise NAMs. Consequently, generating and using research data is important for minimising animal use in regulatory testing.

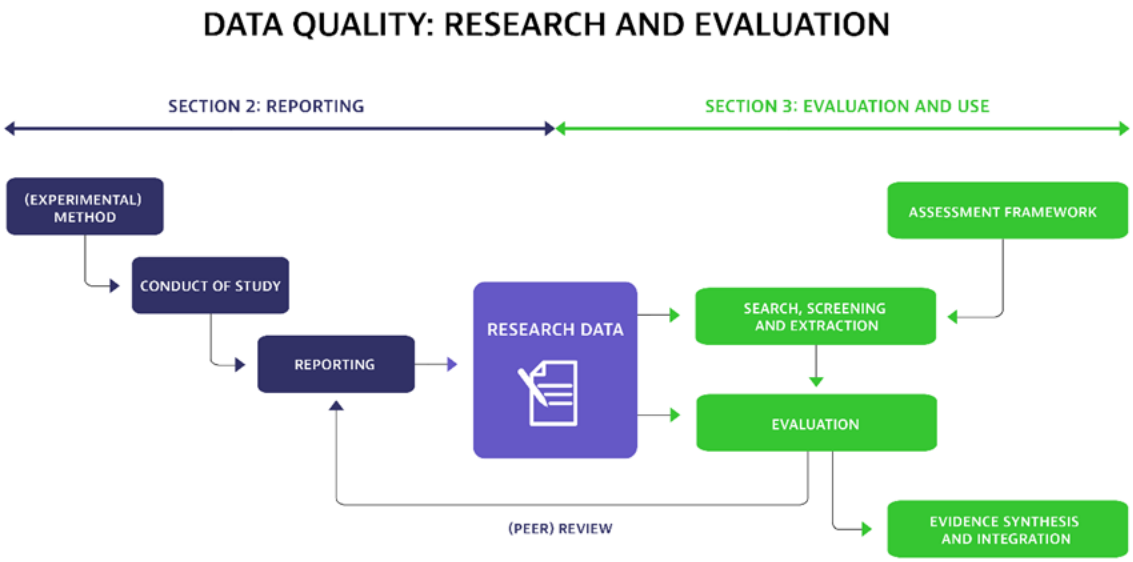
[1] e.g., Equator network, Elixir toxicology. See also Box 2.1. Community resources and professional interest groups.

1.3 Outline

This Guidance Document describes the lifecycle of research data from production to regulatory uptake, highlighting core considerations for researchers, assessors, and other stakeholder groups. Definitions of key concepts and terms are provided in a Glossary of selected terms to ensure accurate interpretation of this Guidance Document. However, it is important to note that specific contexts may define some of these terms differently. Overarching key concepts and definitions are described in Section 1. Section 2 presents considerations for researchers, focusing on reporting guidance, reliability for regulatory uptake, and dissemination of data for accessibility and use by assessors. Section 3 presents considerations for assessors, describing approaches to identify research data, evaluate its relevance and reliability, and incorporate research data into weight of evidence (WoE) analyses. Sections 2 and 3 emphasise the importance of clear reporting throughout the lifecycle. Transparency is critical for the uptake of research data in regulatory assessments, trust, acceptance of the assessment outcome, and for the potential reuse of assessments utilising research data across different decision-making contexts. Figure 1.2 graphically summarises the lifecycle of research data and the relationship between researchers and assessors. Section 4 provides recommendations on good practice and needs for improving and harmonising tools and approaches presented in Sections 2 and 3.

This Guidance Document includes Annexes. Annex A presents a non-exhaustive list of resources available to researchers to help design, conduct, and report specific types of research data in order to maximise its consideration in regulatory assessments. Annex B presents a list of software and data repositories for research and regulatory data. Annex C presents examples of regulatory contexts where research data is considered in regulatory assessments. Annex D provides four case studies that illustrate how research data have been integrated into assessment workflows. Case study A examines the potential for reuse across regulatory authorities of a curated compilation of studies (mostly research data) completed by the US EPA on several hundred per- and polyfluoroalkyl substances (PFAS). Case study B describes two examples (glyphosate and bisphenol F) and identifies good practice on how research data can support the identification of endocrine disruptors in the EU regulatory context. Case study C presents a transparent and structured method to evaluate ecotoxicity data used in risk assessment (CRED method). Case study D shows two examples (fenamiphos and imidacloprid) of submission and incorporation of peer reviewed literature for pesticides approval under the EU pesticides legislation.

Figure 1.2. Flow chart of the generation and use of research data from reporting and evaluation perspective. The data in the centre is either reported by researchers (Section 2) or evaluated by assessors (Section 3)



Note: Flow chart of the reporting, evaluation and use of research data. The research data in the centre is either reported by researchers (Section 2) or searched, screened, extracted, evaluated and used by assessors (Section 3). In the publishing or dissemination process, and evaluation of the data and the way is reported is also performed, often in the form of peer review, e.g., by journal article reviewers and editors (see Section 2.3). The assessment framework included the choice of an evaluation tool that could be predefined (see Section 3.2).

1.4 Principles of research data quality

The trustworthiness of scientific research is underpinned by fundamental principles of data quality. However, perspectives and terminology on data quality can vary among the stakeholder groups involved in the lifecycle of research data (Figure 1.1). Data quality is a broad construct and can include reporting quality, reliability/internal validity, and relevance/external validity/generalisability considerations. For this reason, recent trends are to use terminology that is more explicit about the specific aspect of data quality being considered. This Guidance Document addresses data quality through separate discussions of reporting quality, reliability, and relevance. High-quality reporting of both data generated and the underlying methods (Section 2.2) is foundational as it underpins the assessment of reliability and relevance. The concepts of reliability and relevance are introduced below as they are cross-cutting to Sections 2 and 3.

1.4.1 Reliability

A key requirement for the regulatory use of research data from any source is that the data are considered reliable. Reliability refers to how a study was designed, performed, documented, and analysed. This Guidance Document recognises the use of different phrasing across fields and for simplicity uses the term reliability to encompass internal validity and risk of bias (RoB). Reliability considerations depend on the type of data. For data used for human health and (eco)toxicity assessments, reliability considerations depend on whether they are obtained from observational studies (human and environmental), generated experimentally (*in vitro*, *in vivo*, ecotoxicological field studies), or estimated computationally (e.g., (Q)SARs, PBK models), using established or innovative methods. Reliability subsumes the concept of reproducibility. Table 1.1 presents some general reliability considerations common to the majority of study designs. These

are discussed in more detail in Sections 2.3 and 3.4. Annex A includes specific tools used to guide the reliability evaluation process in peer-reviews and regulatory assessments.

Table 1.1. General reliability considerations

Observational*	Experimental**	Computational
• Sample selection • Sample size and statistical power • Exposure measurement • Outcome assessment • Confounding factors • Statistical analysis methods • Complete reporting of results	• Test item identification and characterisation (e.g., analytical verification) • Exposure characterisation • Test system description (e.g., biological model, test conditions) • Experimental setup (e.g., positive and negative controls, randomisation, technical and biological replicates, sample size, and statistical power) • Endpoint/outcome assessment (e.g., measurement techniques, blinded evaluation, cytotoxicity) • Statistical analysis • Complete reporting of results	• Choice of modelling methods • Applicability domain clarity • Quantity, quality and representativeness of training data set • Robustness of the model (insensitivity to changes in the training set) • Identification and where possible quantification of uncertainties • Model verification • Reliability of input parameters • Reproducibility of the model (including access to code) • Consistency of predictions with other models or data sources • Complete reporting of results

Note: *- Human and environmental observational studies investigating chemical exposure and effects. **- Experimental refers to any type of experimental (eco)toxicity study with controlled dose and conditions including *in vivo*, in *vitro,* laboratory to field scale ecotoxicity studies. Human experimental studies are not considered in this Guidance because reliable studies are rarely available. It is unethical to expose people to potentially harmful substances with no perceived benefit. However, there are some exceptions, such as short-term health effect or pharmacokinetic studies or preventive studies aiming at measuring the efficacy of interventions to reduce exposure to chemicals in households (e.g., second-hand smoke, wood smoke) (more examples in (Allen et al., 2015).

In this Guidance Document, reliability focuses mainly on consideration of internal validity of the study. Internal validity evaluates the extent to which limitations in the design, conduct, and analyses of studies may lead to deviation (i.e., bias) of the estimated effect from the true effect, in terms of both magnitude and direction (overestimation/underestimation) (Higgins et al., 2023). In the context of new methods, the 2005 *OECD Guidance Document No. 34 on the Validation and International Acceptance of New or Updated Test Method for Hazard Assessment* (OECD, 2005) defines reliability as the extent of reproducibility of results from a test within and across different laboratories over time and across operators, when performed using the same protocol. The OECD Guidance Document No. 34, however, is currently being revised[2]. Standardisation and validation of new methods is a continuous process that starts from good practices and general quality principles of scientific research and may carry on towards internationally recognised regulatory standards. The use of standardised protocols enhances the reproducibility of results, which supports the reliability of the data. Whereas regulatory standards (e.g., OECD TGs) can enhance the reliability and the reproducibility of results, they should not be a prerequisite for a study to be considered in a regulatory assessment. In EU legislation, reliability refers to *"the inherent quality of a test report or publication relating to preferably standardised methodology and the way the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings. The reliability of data is closely linked to the reliability of the test method used to generate the data"* (ECHA, 2011; EFSA, 2023a).

[2] Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment | OECD

1.4.2 Scientific versus regulatory relevance

When considering relevance, a fundamental distinction exists between scientific relevance and regulatory relevance. Scientific relevance relates to the knowledge advancements in a research field. Regulatory relevance relates to the utility of a given study to provide data for a hazard or risk assessment as defined by legislation. The European Chemicals Agency (ECHA) defines relevance as the extent to which data and tests are appropriate for a particular hazard identification or risk characterisation (ECHA, 2011). Thus, relevance depends on the assessment framework and more specifically on the assessment questions to be addressed within one framework. The 2005 OECD Guidance Document No. 34 defines relevance as the *"relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest"* (OECD, 2005).

Relevance is often associated with the concept of external validity. Regulatory relevance comprises exposure and biological considerations (Rudén et al., 2017). Exposure relevance refers to the representativeness of the substance and the exposure scenario, including doses and concentrations. Issues related to substance composition, purity, and routes of exposure (e.g., studies involving direct injection, such as oral gavage or intradermal administration) influence exposure relevance. Biological relevance is based on the relationship between the results of a study (e.g., a measured biomarker) and the adverse outcome of concern in the species (or population) of interest. The Adverse Outcome Pathway (AOP) framework provides a shared and structured knowledge base which, based on an established sequence of events, can support the relevance assessment. The European Food Safety Authority (EFSA) also defined a biologically relevant effect as an effect considered by expert judgement as important and meaningful for human, animal, plant or environmental health, stressing that a statistically significant effect should not automatically be considered relevant for the outcome of an assessment (EFSA, 2017b).

Regulatory relevance changes as the regulatory framework develops over time. What is not relevant in the current framework may become relevant in the future, as new legislation or scientific guidance is established and *vice versa*. For example, the criterion of a plausible mechanistic link between endocrine activity and adverse outcome, introduced in the ECHA/EFSA guidance for the identification of endocrine disruptors (EFSA/ECHA, 2018), has increased the regulatory relevance of intermediate/mechanistic effects data (Annex D, Case study B). It is likely that research data generated using NAMs to provide mechanistic information will increasingly become regulatory relevant.

Assessing regulatory relevance can be approached in several ways, either by using predefined inclusion and exclusion criteria during the process of identifying studies or using specific tools that probe study relevance. Regardless of the approach, relevance considerations are often similar when the decision-making contexts for the assessments are alike. Differences are typically limited to the step in the process where studies are evaluated for relevance (i.e., early during study screening or later during a deeper analysis of each included study). Using inclusion and exclusion criteria to identify relevant studies is becoming more common with the increasing use of systematic review methods to conduct regulatory assessments (Case study A). In addition to the use of screening (inclusion/exclusion) frameworks, evaluation tools including specific criteria for the evaluation of relevance can be used to assess individual studies. Case study B and Case study C explore some of these approaches. The criteria for identifying relevant studies can differ depending on the focus of the analysis. For example, more stringent relevance criteria are required for quantitative hazard characterisation, when compared to hazard identification, which is typically a qualitative exercise. For dose-response, experimental (eco)toxicology studies should ideally have at least five concentration/dose levels and epidemiological studies would have to include quantitative estimates of exposure (versus qualitative characterisation of exposure).

2 Considerations for researchers generating data

2.1 Targeting regulatory needs

Research on chemical hazard, exposure and risk in general pursue objectives that are often different from regulatory assessments. Consideration of regulatory needs, however, can increase the societal impact of research beyond the scientific relevance. Some research programmes explicitly target regulatory needs. In such cases, research funders may have specific expectations and requirements. Although human health and environmental protection goals, information requirements, and assessment methodologies differ across countries and across sectors, the general considerations regarding regulatory relevance are valid across frameworks. Understanding legislation, including the interplay between different pieces of legislation, regulatory guidance, and regulatory datasets increases researchers' ability to identify regulatory data demands. Regulatory processes often present opportunities for researchers and the wider public to provide input. Participation in expert panels, public data calls, and consultations on draft assessments or dossiers submitted by registrants are direct channels to respond to regulatory needs. These engagements are especially useful to ensure that recent research outcomes from academic or industry research programmes are considered. Such opportunities help researchers to contribute to ongoing or upcoming assessments by placing their research data in the context of the broader evidence base for the assessment (Ågerstrand et al., 2017).

Understanding which regulatory datasets exist for a certain topic is the best first option to identify regulatory data gaps. Researchers should conduct literature searches of the chemical of interest in advance to determine whether, and if so, how it was previously tested and assessed under regulatory programmes. This can inform research study design, including information on tested concentrations or doses. Many of the guideline studies cited in regulatory assessments are not published in scientific journal databases such as PubMed, Web of Science, Scopus, or EMBASE. Key resources to check on guideline testing status include the OECD eChemPortal, the US Environmental Protection Agency (EPA) CompTox Chemical Dashboard ToxVal module, EFSA's homepage and OpenFoodTox database, and ECHA CHEM (Annex B). In the EU, work has started to establish a Common Data Platform on Chemicals, bringing together chemicals data at EU level[3].

The type of regulatory task determines the specificity of the information needed. Where information requirements aimed at excluding properties of concern are based on a defined set of evidence, typically OECD TGs, narrow relevance criteria are applied, limiting the utility of most research studies. On the other hand, the investigation of specific concerns in assessments performed by agencies benefits from any type of research study that adds evidence to raise or remove concerns. In some cases, research studies may not add essential information or are anyway unlikely to change the WoE. However, important knowledge gaps exist for most compounds on the world market. Additional research data often contribute to regulatory

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2023%3A779%3AFIN

conclusions, even for substances with extensive registration dossiers. For example, an analysis of assessments underpinning EU REACH restrictions found that 58% of the key studies were non-standard studies, and 77% of these studies had at least one author affiliated with academia (Borchert et al., 2022). These studies were all based on animals (vertebrates and invertebrates) or included human data. The case studies provide specific examples representative of various regulatory contexts (Annex D).

Studies investigating sensitive endpoints that are not included in guideline studies can result in regulatory endpoints driving the risk assessment. Recent examples include:

1. Non-standard rodent studies of immune system effects which drove EFSA's new tolerable daily intake (TDI) for bisphenol A (EFSA, 2023b).
2. Epidemiological evidence of decreased immunity in children exposed to perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS) which steered the US EPA's new drinking water standards for four polyfluoroalkyl substances (PFAS) (US EPA, 2022).

In ecotoxicity assessments, data coming from non-standard tests may allow the identification of more sensitive species than the ones used in standard tests. For example, in the case of the neonicotinoid insecticide imidacloprid (Annex D, Case study D), EFSA eventually established regulatory acceptable concentrations from the species sensitivity distribution (SSD) derived from a dataset of ten chronic toxicity data, including several non-standard tests on aquatic insects (EFSA, 2014b). There are also numerous examples of studies investigating sensitive endpoints that have had little impact on regulatory decision-making. This is the case for ecotoxicological studies on behavioural changes following exposure to chemicals. Possible reasons for this may include divergent views on the relevance of this type of endpoint at the population level (Ågerstrand et al., 2020).

Regulatory agencies frequently publish overviews of their regulatory needs, sometimes including opportunities for researchers to engage in risk assessment processes, or to obtain dedicated research funding. Research funding bodies have also responded to the need to improve the utility of research data in regulatory assessments. For example, the European Partnership for the Assessment of Risks from Chemicals (PARC) brings the research and the regulatory communities closer together throughout the life cycle of research projects, from co-design to execution and dissemination, to develop next-generation risk assessment strategies(Marx-Stoelting et al., 2023). Researchers can follow the communication channels of the OECD, Health Canada, US EPA, the European Commission, ECHA, EFSA, etc. to learn about the latest updates and opportunities (e.g., ECHA, 2024).

In the context of NAMs, one fundamental challenge is to discriminate between those effects that can be linked to adverse outcomes, and those that are merely indicative of adaptive or homeostatic responses (Rudén et al., 2017). The AOP concept provides a means of establishing the relevance of any type of (eco)toxicity data generated and is especially useful for NAMs. In particular, the AOP Knowledgebase[4] (AOP-KB) provides common ontologies linking molecular initiating events with intermediate effects and adverse outcomes as defined by regulatory endpoints. Common ontologies facilitate the interpretation of study results in regulatory contexts, their integration as part of Integrated Approaches to Testing and Assessment[5] (IATA) (OECD, 2017b; OECD, 2020b), as well as their potential use within Defined Approaches (DA). IATAs follow an iterative approach to answer a defined question in a specific regulatory context, taking into account the acceptable level of uncertainty associated with the decision context (Sakuratani et al., 2018). Several IATA case studies published by the OECD include research data. For complex endpoints, IATAs and DAs provide the methodological basis for animal-free assessments.

[4] https://aopkb.oecd.org/

[5] https://www.oecd.org/en/topics/sub-issues/assessment-of-chemicals/integrated-approaches-to-testing-and-assessment.html

In research that is primarily focused on development and evaluation of new methods, substance selection is usually driven by the need for an evaluation and/or comparison between methods for a given AOP. Hence, data-rich reference substances are needed to determine the reliability of a new method. Nonetheless, inclusion of data-poor substances can increase the regulatory relevance of the data generated, while indirectly raising assessors' interest in the new method.

2.2 Reporting guidance

Many research studies are excluded from regulatory assessments because they lack details in the method/protocol and/or in results presented, impeding the assessor to evaluate relevance and reliability. This section introduces good practices and resources to guide the reporting of published data in a manner that maximises their potential use within the scientific community as well as their utility in regulatory decision-making. Best reporting practices cover not only the reporting of study methods, its performance, statistical analyses, and results but also data provenance ("data lineage"), and transparency regarding sources of funding, who was involved, and their roles in the research. Adhering to reporting guidance aid the implementation of FAIR principles of scientific data management and stewardship (Wilkinson et al., 2016). For researchers, reporting guidance can help to streamline their work and expedite the review process. Table 2.1 presents core reporting information for three types of studies frequently used in regulatory assessments: observational studies, lab-scale experimental *in vivo*, and *in vitro* (eco)toxicity studies, and computational models. A lack of this type of information can make it difficult to reproduce results and can hamper the use of the research in regulatory assessments because its reliability cannot be assessed. Reliability is discussed in more detail in Section 3.5 and the list of considerations overlaps with the core reporting items in Table 2.1. However, the major distinction is that the merely reporting information does not guarantee the reliability of the methodologies used or the data presented. Not reporting the content summarised in Table 2.1 can result in the study not being considered at all or considered "not assignable" in some assessments of reliability. In practice, the studies considered most reliable for regulatory purposes report much more critical detail than presented in Table 2.1. For example, access to raw data, and procedures for handling outliers are often critical to assessors. More comprehensive (minimum or recommended) reporting standards are available for specific types of test methods and technologies used. In the case of emerging test systems (e.g., organ-on-chip, complex *in vitro* models) or measurement analysis (e.g., omics, high-content imaging), specific reporting standards are available or are under development. An overview of more detailed OECD and non-OECD reporting recommendations and tools is included in Annex A. This list covers resources for a broader range of research studies, not limited to those fields covered in Table 2.1. In addition to scientific considerations, animal studies or studies using primary human cells may be subject to ethics approvals and related reporting requirements.

Table 2.1 highlights that providing detail on the test item and exposure characteristics is the first critical reporting element of all studies. Reporting of the test item includes substance identifiers (i.e., IUPAC name, international name, CAS RN), structural identifiers (i.e., SMILES, INCHI(Key), and composition (e.g., formulations/mixtures, extracts, purity, enantiomeric ratios). The unambiguous identification of a substance also makes the research data findable. CAS numbers alone do not always unequivocally identify the substance. Substance identification and naming conventions partly depend on the definition of a "substance" under specific policy areas. An example of detailed guidance on the topic is available for EU REACH and CLP (ECHA, 2023). A special note of caution relates to the identification and characterisation of complex multi-constituent substances, substances with unknown or variable composition, complex reaction products or biological materials (UVCBs), including polymers, natural substances, nanomaterials, and other advanced materials[6]. For instance, guiding principles for measuring and reporting

[6] https://www.oecd.org/en/publications/advanced-materials-working-description_4b5ba38d-en.html

physicochemical parameters of nanomaterials are presented in (OECD, 2019a). If it is unclear what substance(s) or form of a substance was tested, or at what concentration, it is impossible for assessors to confidently determine what caused the observed effect. Detailed analytical characterisation of the test item is often necessary to attribute the observed results in a study to the substance being assessed.

Table 2.1. Core reporting elements for consideration in publications by researchers

	Observational	*In vivo*	*In vitro*	Computational
Exposure and/or test item	• Exposure characterisation (including specification of substance and/or the proxy used) • Confounding factors	• Test item and reagent identifiers (e.g. IUPAC name, CAS number), source, purity, composition of mixture/formulation • Physicochemical characterisation (e.g., solubility, particles properties[7]) of the test item • Administration of test item (e.g., route, dose levels, frequency, duration) • Description of negative (solvent/vehicle) controls; positive controls (if used) • Analytical confirmation of dose (when warranted)	• Test item and reagent identifiers (e.g. IUPAC name, CAS number), source, purity, composition of mixture/formulation • Physicochemical characterisation (e.g., solubility, particle properties[3]) of the test item • Administration of test item (e.g., concentration levels, duration) • Exposure conditions (e.g., temperature, medium composition) • Description of positive, negative/solvent/vehicle controls	• Test item identifier (e.g. IUPAC name, CAS number), chemical structure
Population, organisms, or test system	• Description of the population, including geographic region, sex, and age (life stage)	• Source/supplier/origin, species/strain, sex, and age (or life stage), health status, housing conditions, acclimatisation	• Source/supplier/origin, test system basic information, e.g., cell/tissue type(s), donor characteristics (ideally using Research Resource Identifiers8 and quality control, e.g., purity, mycoplasma testing, genetic stability. • Endogenous metabolic competence/activation of the system (when warranted)	• Description of the conceptual model, including the relevant physicochemical and biological processes and the relationship between them • Identification and where possible quantification of uncertainties • Model verification

[7] Including e.g., particle size, size distribution, shape, stability, surface area, and treatment.

[8] https://www.rrids.org/

	Observational	*In vivo*	*In vitro*	Computational
Study design (or model specification)	• Sample size • Description of endpoint • Methods for endpoint measurements and analysis • Statistical methods	• Sample size, number of organisms (or experimental units) per sex and dose • Description of endpoint • Methods for endpoint measurements and analysis • Statistical methods	• Sample size, technical replicates, biological replicates • Description of endpoint • Methods for endpoint measurements and analysis • Statistical methods	• Source and value of model input parameters related to the substance(s) modelled (identity and properties) and the biological system (e.g., biochemical and physicochemical parameters) • Description of endpoint modelled • Transparent description of the model development workflow and resulting model (e.g. via QMRF*)
Results presentation	• Quantitative (preferred) or qualitative presentation of results, including variability, reasons for data exclusion	• Quantitative presentation of results (e.g., control performance, variability, dose-response, reasons for data exclusion) • In some cases, a qualitative presentation of results is sufficient, e.g., histopathology • Signs of general/systemic toxicity throughout the study (e.g., body weight, mortality, behaviour)	• Quantitative presentation of results (e.g., control performance, variability, concentration-response, reasons for data exclusion) • Consideration of cytotoxicity or other type of interference that can impact the results	• Quantitative or qualitative results (predictions) of results, including estimates of prediction error (e.g. via QPRF** and QAF*** • Access to code

Note: * QMRF= (Q)SAR Model Reporting Format, Annex I at (OECD, 2024); **QPRF= (Q)SAR Prediction Reporting Format, Annex II at (OECD, 2024); ***QAF= (Q)SAR Assessment Framework (OECD, 2024).

Access to detailed descriptions of methods and protocols is essential to ensure reproducibility and build trust in study results. The lack of accessible, detailed methodological descriptions is a major factor behind the reproducibility crises in life science research (Baker, 2016), including (eco)toxicology (Ågerstrand et al., 2014). A 2014 survey by the American Society for Cell Biology showed that incomplete specification of the original protocol is the most prominent reason for unsuccessful replication of published results (American Society for Cell Biology, 2014). For human trials there is a longer history of registering protocols (ICMJE, 2024) and efforts are underway to promote this practice in toxicology and environmental health research (Mellor et al., 2024). For animal studies, for example, the German Animal Study Registry (ASR[9]) was launched in 2019 for preregistration of animal studies worldwide to increase transparency and reproducibility of bioscience research and to promote animal welfare (Bert et al., 2019). Pro-MaP (Promoting Reusable and Open Methods and Protocols) is a multistakeholder initiative led by the European Commission that aims to improve methodological clarity in research publications (European Commission, 2024). Recommendations for researchers, research institutions, publishers, editors, and research funders include (adapted from European Commission, 2024):

- Documenting, sharing, and executing step-by-step study protocols.
- Publishing method descriptions in a user- or reader-friendly way, and with sufficient detail to reproduce the experiment. To facilitate this, some journals consider it to be acceptable to quote exact text describing detailed methods from previously published work with attribution. Plagiarism policies should be updated accordingly.
- Making responsible use of shortcut citations, where the description of the method is to a citation of a previous paper (or papers). Shortcut citations can be effective if authors cite a recent methods paper or protocol that describes exactly what they did. In contrast, shortcut citations hamper reproducibility if the cited resource is inaccessible, does not mention or fully describe the cited method, or cites another resource instead of fully describing the method.
- Sharing of protocols in a format that can be cited and updated, using open access, dynamic repositories that allows protocol versioning and forking (e.g., protocols.io[10]). Protocols should also provide Digital Object Identifiers (DOIs) for citation purposes and have a long-term preservation strategy. Open access ensures that protocols are available to everyone. Versioning and forking allow research groups and the scientific community to track the evolution of protocols within and across research groups, whereas the DOI ensures unique persistent identifiers that can be cited.

2.2.1 OECD resources

Researchers interested in maximising the utility of data for regulatory purposes should be familiar with reporting requirements in the OECD Test Guidelines for the Testing of Chemicals and reporting guidance, available via OECD Harmonised Templates[11] (discussed below). Even if regulatory use of data is not a primary goal or if the test species/conditions are not similar, these guidelines can help researchers because they point to the essential issues to consider and report, also in peer review journals. Knowledge of OECD study designs and reporting standards can shape the design of studies and the presentation of methods and results. By adhering to the concepts in these guidelines, the research will be more applicable to regulatory contexts (i.e., number of treatment groups, clear description of test item, key endpoints, appendices with summarised and individualised results, etc.). Thus, high quality study design and reporting helps reviewers in the peer review process, as well as regulatory assessors (see Sections 2.3 and 3.2).

[9] www.animalstudyregistry.org
[10] www.protocols.io/
[11] www.oecd.org/en/topics/sub-issues/assessment-of-chemicals/harmonised-templates.html

OECD Guidelines for the Testing of Chemicals

The OECD Guidelines for the Testing of Chemicals[12] (OECD TGs) are internationally accepted standard methods for chemicals testing, including many types of organisms (mammals and non-mammalian vertebrates, invertebrates, plants, and algae), model systems (*in vivo*, *in vitro*, and *in silico*), and compartments (water, air, soil, and sediments). The adoption of new OECD TGs is based on rigorous validation procedures to ensure cross-laboratory reproducibility and provide specific validity criteria to confirm reliability of results. Under the OECD Mutual Acceptance of Data (MAD)[13] system, laboratory test results generated in accordance with OECD TGs and OECD Principles of Good Laboratory Practices (GLP)[14] are accepted in all OECD Member Countries and MAD adherent countries, provided that essential and relevant validity criteria of the corresponding OECD TG are met. OECD TGs are used by professionals in industry, contract laboratories, academia and government involved in the testing and assessment of chemicals (industrial chemicals, pesticides, pharmaceuticals, etc.) for the purpose of safety assessment and other uses related to the protection of human health and the environment.

In academic research, it is not always possible to adhere to the test designs set out in OECD TGs and GLP. For example, animal tests, commonly used require significant resources and numbers of animals. Securing funding and ethical approval may be difficult and is not necessarily encouraged, as part of supporting the replacement, reduction, and refinement in use of animals in research. Further, ecotoxicity studies may be performed with species that are not described in any guideline test. Benchmarking against chemicals with known effects in "standard species" and providing information on historical control data (where available) can facilitate use of the data in regulatory assessments. Alignment with standardised test guidelines as much as possible makes it more likely that studies are acceptable in peer reviewed journals (due to the use of accepted methods and clarity in reporting of study conduct and results) and considered in regulatory settings. Reporting guidance (including graphs, statistics) and validity or acceptance criteria provided in OECD TGs are in many cases applicable to research studies. One consideration for researchers performing studies that do not follow standardised guideline methods is to provide a rationale for deviation from the existing methods, focusing especially on the biological rationale and the possible effects of the deviations. This can facilitate decisions on inclusion of the research in an assessment conducted for regulatory purposes (Section 2.4).

OECD Series of Testing and Assessment

The OECD Series on Testing and Assessment[15] includes almost 400 publications related to testing and assessment of chemicals. Some of them support the development of OECD TGs (e.g., validation reports, guidance documents, detailed review papers) and others support best practice in reporting risk assessment methods and data to be used in a regulatory context. A comprehensive method description is a prerequisite to assess and use the corresponding data. Recent examples in the Testing and Assessment Series include guidance on (Annex A):

- Reporting for omics data (OECD, 2023b)
- An assessment framework for (Q)SAR models and predictions (OECD, 2024)
- Characterisation, validation and reporting of PBK models (OECD, 2021)
- The use of AOPs in the development of IATAs (OECD, 2017b)

To cope with the increasing number of non-standard *in vitro* methods and to harmonise their reporting the OECD released guidance for describing non-guideline *in vitro* test methods (OECD, 2017a) and on Good

[12] www.oecd.org/en/topics/sub-issues/testing-of-chemicals/test-guidelines.html
[13] www.oecd.org/en/topics/sub-issues/testing-of-chemicals/mutual-acceptance-of-data-system.html
[14] www.oecd.org/en/topics/sub-issues/testing-of-chemicals/good-laboratory-practice-and-compliance-monitoring.html
[15] www.oecd.org/en/publications/oecd-series-on-testing-and-assessment_20777876.html

in Vitro Method Practices (GIVMP) (OECD, 2018a). Guidance provided in these OECD documents is applicable to research studies and can be especially useful in areas where OECD TGs and OECD Harmonised Templates have not yet been developed, i.e., studies using methods, organisms, or endpoints that are not covered by OECD TGs.

OECD Harmonised Templates (OHTs) and IUCLID

OHTs are standard data formats for reporting information on chemical properties, on their adverse effects on human health and the environment, and on their use and related exposure to workers, consumers, and the environment. Although OHTs are not designed to be used by the research community, they can be used as models for reporting studies and other information on any type of chemicals. To some extent, additional, non-standard information (e.g., new biomarker) can be reported in existing OHTs together with standard fields and endpoints.

The OHTs are regularly updated to cover new or revised OECD TGs, fulfil requests for improvement from users and/or regulators, and extend their functionalities following information technology, new regulatory requirements, and chemical testing developments. Currently, there are over 130 OHTs for reporting chemical safety data used in risk assessment. The templates can be freely downloaded in Word format and xml schema from the OHTs website[16]. Additional reporting materials are available for certain endpoints such as predefined tables and predefined executive summaries.

The OHTs can be used as specifications for data entry screens in regulatory data management systems (e.g., IUCLID). International Uniform Chemical Information Database (IUCLID)[17] is a software application designed to record, store, maintain and exchange data on the intrinsic and hazard properties of chemical substances or mixtures, as well as the uses of these substances and the associated exposure levels. IUCLID is being increasingly used by different jurisdictions and regulatory programmes (Australia, Canada, New Zealand, Switzerland, United Kingdom, United States, European Union entities, and the OECD) (OECD, 2025). Although IUCLID and its OHT specifications were designed mainly to report data from guideline studies, research data are also reported, typically by regulatory assessors (registrants, agencies) extracting data from scientific literature to fit the available format. An ongoing WPHA project aims to adapt OHTs to the reporting needs of research data.

2.2.2 Other reporting guidance and templates

In addition to OECD resources, other tools to assist in the reporting of non-guidelines studies are available (Annex A). These tools are useful to researchers as they provide accessible checklists and guidance for specific items that should be reported to improve scientific publications and make the results applicable to regulatory assessments.

A collection of available reporting guidelines for health-related research can be found on the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) website[18]. The EQUATOR Network initiative was officially launched by the UK National Knowledge Service in 2008, with the aim to improve the reliability of medical publications by promoting transparent and accurate reporting of health research (Altman et al., 2008). Examples of relevant reporting guidelines in the chemical context include the ARRIVE (Animal Research: Reporting of *In Vivo* Experiments) guidelines for *in vivo* animal studies (Percie Du Sert et al., 2020) and STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for observational studies in epidemiology (Von Elm et al., 2007). Currently, resources in EQUATOR are heavily oriented towards the analysis of human health. Evidence and tools focusing on other types of evidence, such as (eco)toxicity and *in vitro*, are underrepresented or absent. One tool

[16] www.oecd.org/en/topics/sub-issues/assessment-of-chemicals/harmonised-templates.html
[17] https://iuclid6.echa.europa.eu/
[18] https://www.equator-network.org/

specifically developed for broad use in regulatory assessments is SciRAP[19]. It is freely available online and includes reporting checklists for epidemiological studies, *in vitro* studies and *in vivo* (eco)toxicity studies, including *in vitro* and ecotoxicity studies on nanomaterials (Beronius et al., 2018; Moermond et al., 2016; Roth et al., 2021; Shao et al., 2023; Hlisníková et al., 2024). A number of efforts are underway to develop tools and question item repositories for *in vitro* studies, e.g., RIVER (Reporting *In Vitro* Experiments Responsibly)[20], Peer Review of *in Vitro* studies Appraisal Tool (PRIVAT)[21], and certain projects under PARC (Svendsen et al., 2023). Reporting tools can build on each other, i.e., the SciRAP ecotoxicity criteria were built on the Criteria for Reporting and Assessing Ecotoxicity Studies (CRED) (Moermond et al., 2016) (see also Box 2.1 and Case study C). These evaluation tools and reporting checklists/guidelines are frequently expanded and refined to improve their scope and applicability to specific types of data and substances.

There are examples where resources have been developed outside of the OECD to help fulfil OECD guidance requirements. For example, to help fulfil the requirements of the *OECD Guidance Document for Describing Non-Guideline In Vitro Test Methods* (OECD, 2017a), an EU funded project (EU-ToxRIsk) developed an annotated toxicity test method template (ToxTemp) to describe cell-based toxicological test methods, facilitating regulatory use of the data (Krebs et al., 2019). The ToxTemp provided the basis for the method description in the OECD document *Initial Recommendations on Evaluation of Data from the Developmental Neurotoxicity (DNT) In-Vitro Testing Battery* (OECD, 2023a).

Templates to capture study metadata are increasingly used in various fields to promote FAIR principles and machine-actionability. For example, the Nanosafety Data Interface[22] has a template wizard to develop aggregated FAIR data where users can enter metadata for physicochemical, ecotoxicity, *in vitro*, and exposure and release data for nanomaterials (Jeliazkova et al., 2021).

Such templates are not yet commonly used by most researchers, perhaps due to a lack of awareness on their existence. In addition, populating the templates requires extra effort for researchers beyond the standard journal publication process. As noted above, the OECD OHTs are not designed to be used by the research community, but they can serve as models for developing templates targeted to researchers. Structured submission of methods and results, as part of the publication process, promotes high quality reporting (Jin et al., 2015; Sim & Detmer, 2005; Swan & Brown, 2008). Support for such an approach may be increasing with more wide-spread use of study registries and protocols. For example, the International Committee of Medical Journal Editors requires prospective registration of clinical trials as a prerequisite for publication, providing formatting guidelines for preparing, sharing and reporting data summarised in tables and results for journal submission (ICMJE, 2024). Assessors have a vested interest in these efforts, as the process of summarising study methods and results (often referred to as “data extraction” or “data abstraction”) is one of the most laborious of the assessment process, with estimates of 0.5–2.5 hours per study (depending on study complexity and type) (National Academies of Sciences, Engineering, 2022). Case study A explored the feasibility of having researchers summarise methods and results of their experimental animal studies using a structured, web-based data extraction model. This task could be implemented during the manuscript submission process (Wilkins et al., 2022). Participants found the process viable and understood the long-term benefits despite the extra effort. The pilot study also suggested that using templates may improve the conduct and completeness of reporting in future research.

Several open science communities and professional societies have initiatives aimed at improving the reporting and collection of research data, in support of the FAIR principles. Some illustrative examples are reported in (Box 2.1).

[19] https://ki.se/en/imm/scirap-science-in-risk-assessment-and-policy
[20] https://nc3rs.org.uk/our-portfolio/river-recommendations
[21] https://osf.io/w4fyp/
[22] https://enanomapper.adma.ai/

Box 2.1. Community resources and professional interest groups

Evidence Based Toxicology Collaboration (EBTC): Founded in 2011 at Johns Hopkins Bloomberg School of Public Health with the vision to make evidence-based methodologies the standard that ensures public health, a healthy environment, and a sustainable future. EBTC is a member-driven organisation, bringing together the international toxicology community to work on adapting and developing evidence-based methods and frameworks that facilitate the use of evidence in informing regulatory, environmental, and public health decisions. Areas of focus include (1) research methods, for the better conduct and reporting of studies, (2) evidence synthesis, to ensure the best use of evidence in policy-making, (3) open science, to support more accessible and reusable research, and (4) evidence and decisions, creating frameworks for transparent use of evidence in policy-making. In 2023, EBTC launched *Evidence-based Toxicology*, an open science journal for the environmental health sciences. EBTC also publishes a newsletter for subscribers.

ELIXIR Toxicology Community: Established in 2020 to support the integration of standards, tools, and resources for toxicological research projects and risk governance at national and international levels (Martens et al., 2021). Goals include developing open community standards to support common interest, including ontologies, application programming interfaces (APIs), data formats, deposition databases, and publication recommendations. The current collection of resources is a mix of meta-information and includes regulatory and scientific databases of diverse scope and specificity.

The Society of Environmental Toxicology and Chemistry (SETAC): A global organisation established in 1979 with the primary objective of promoting environmental sciences. This mission is accomplished through various initiatives, including the organisation of meetings, training programmes, and an active publication agenda. Building on this commitment, SETAC published the Technical Issue Paper in 2019, entitled "Recommended Minimum Reporting Information for Environmental Toxicity Studies". This document provides guidelines to enhance the transparency and reliability of reporting in environmental toxicity studies. In 2024, SETAC supported the development of the Criteria for Reporting and Evaluating Exposure Datasets (CREED) for use in environmental assessments (Merrington et al., 2024).

2.3 Reliability

2.3.1 Reliability and related concepts used in regulatory contexts

While different definitions of reliability exist, they all converge on the concept of internal validity of a study or endpoint, whereas relevance can vary depending on the specific assessment goal (Section 1.4.2). Table 1.1 gives an overview of general reliability considerations for observational studies, experimental studies, and computational models. These considerations are included in domain specific methodological guidance and in study evaluation tools used in regulatory frameworks (Annex A). Researchers, scientific reviewers, and editors can use these considerations, guidance, and tools to improve the quality of research data and hence the applicability for regulatory purposes. Research in (eco)toxicology continuously brings forward a wide range of new types of study designs and technologies, which complicates the reliability assessment. Study reporting and evaluation tools have been developed over time, often by expanding their original scope and/or by increasing their specificity to study designs or substance types. Annex A presents a non-exhaustive list of methodological guidance and reliability assessment tools. Some are discussed in more detail in Section 3.4.

Reliability assessment tools developed in recent years are domain-based tools. These tools break down the appraisal process into core reliability considerations (Table 1.1), allowing for a focused evaluation of

each aspect. Domain-based tools used by the US EPA Integrated Risk Information System (IRIS) Program and US National Toxicology Program (OHAT 2019; US EPA, 2022), consist of key prompting questions developed for a specific assessment, and present a judgement for each reliability consideration supported by a narrative rationale.

Structured evaluation tools can enhance the quality of peer review by systematically identifying study limitations. Unfortunately, awareness of these tools is limited among researchers, reviewers, and editors. While study reliability is fundamental in peer review, no systematic guidance exist, and the assessment of reliability varies widely across reviewers, manuscripts, and journals. Consequently, peer-reviewed studies require additional reliability assessments for regulatory purposes.

In the sections below, specific reliability considerations are discussed for human observational, experimental *in vivo* and *in vitro*, and computational studies.

2.3.2 Human observational studies

Observational studies refer to non-experimental studies. In some cases, data used in regulatory human hazard or risk assessments comes from human epidemiological observational studies, such as cross-sectional, case-control and cohort designs. Among the reliability considerations presented in Table 1.1, issues related to participant selection, confounding factors, and exposure assessment often hinder the use of observational studies in regulatory assessments. These reliability domains, along with a lack of detail on how outcomes were assessed and insufficient presentation of quantitative results, were the sources of most study deficiencies identified in the epidemiological studies included in the US EPA Systematic Evidence Map on PFAS (Annex D, Case study A). Another concept to consider is study sensitivity, here defined as the ability of a study to detect an effect, if present (National Academies of Sciences, Engineering, 2022; Cooper et al., 2016). A study may be well designed and conducted but still have limitations that make it difficult to detect an association. In the PFAS Systematic Evidence Map, insufficient study sensitivity was the most frequent study deficiency (Radke et al., 2022). An overview of other key methodological aspects considered when assessing the quality of epidemiological studies can be found in institutional guidance by EFSA (EFSA, 2024), US EPA's IRIS Handbook (National Academies of Sciences, Engineering, 2022), OHAT Handbook (NIEHS, 2019), NTP Report on Carcinogens Handbook (NTP, 2025), or by IARC (IARC, 2019), and in the reliability assessment tools developed for human epidemiological studies (Shamliyan et al., 2010). For a selection of inventories and reviews on reliability assessment tools, see Appendix D of EFSA Scientific Committee guidance on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments (EFSA, 2024).

Participant selection

Participant selection, if not properly performed, may lead to "selection bias". In case-control studies for example, cases may be more motivated in participating in the study than controls, which may lead to bias if such cases are also those with the greater probability of exposure. In occupational cohort studies, one should consider the "healthy worker effect" (i.e., that people in good health are more likely to join the workforce) and the "healthy worker survivor effect" (i.e., that people in good health and with a low susceptibility/sensitivity to the exposure are more likely to stay in the job), which may potentially attenuate the risk estimate when comparing workers to people that cannot work (National Academies of Sciences, Engineering, 2022).

Confounding factors

The identification and control for potential confounding factors, i.e., factors that are both associated with the outcome and the exposure, but which are not intermediaries on the pathway between the exposure and the outcome, is of utmost importance in observational studies. This can occur if the setting of the study

is not under control of the investigator, as in randomised control trials. Common confounders include demographics like age, gender and race/ethnicity, socio-economic variables like education and income, variables related to health status like body mass index (BMI), or behavioural factors like smoking or alcohol consumption. However, potential confounders depend on the research question and need to be evaluated at the design stage of the study, considering background information on the outcome and the exposure under assessment. Confounding due to co-exposure to multiple chemicals (mixtures) with effects on the same health outcome may affect the interpretability of results, particularly when the chemicals are highly correlated. Directed acyclic graphs (DAGs), which depict graphically prior knowledge about biological and behavioural systems related to the causal research question, may provide a useful tool to help researchers in visualising relationships between variables that could lead to confounding and other types of bias (Digitale et al., 2022). While unidentified factors or limitations in the analysis always introduce "residual confounding" to a certain degree, estimating its likely strength and direction helps the interpretation of study results. Besides confounding, further elements to consider when characterising the causal association and susceptibilities of a defined exposure-outcome relationship include the examination of potential effect modification/interaction (i.e., when a factor modifies the causal effect of another factor on a defined outcome) or mediation (i.e., when a factor is an intermediate along the chain of events between the exposure and the outcome, thus partially, or entirely, accounting for the association between the exposure and the outcome). If needed, these effects should be assessed through appropriate statistical analyses.

Exposure assessment

Exposure assessment is a key source of uncertainty in environmental epidemiology. The exposure metrics should be an acceptable proxy for the true exposure of interest within the relevant population (Arroyave et al., 2021). All relevant exposure routes should be considered with the appropriate time window (National Academies of Sciences, Engineering, 2022) to reduce exposure misclassification as much as possible. In many cases, non-differential exposure misclassification can represent a bias towards the null hypothesis. When reporting exposure through human biomonitoring for example, biomarkers should be chosen considering the reproducibility of measures over time, and factors that could influence measurements related to the chemical composition of the substance and the matrix, e.g., the potential for metabolism of the chemical due to the matrix enzymatic activity (Arroyave et al., 2021; Calafat and Needham, 2007).

A key regulatory task is developing quantitative reference values (e.g., non-cancer reference doses or cancer risk estimates). In such cases, researchers should include effect measures (e.g., relative risk, standardised mortality ratio) based on a comparison group exposed to lower levels (or no exposure/exposure below detection limits), or cases versus controls, or a repeated measures design (Thayer, et al., 2022).

Study sensitivity

While some of the study features that affect study sensitivity are already included in other reliability domains, such as those already cited above, there could be additional features worth considering when assessing whether a study is able to detect an effect, if present (National Academies of Sciences, Engineering, 2022; Copper et al., 2016). A well-conducted epidemiology study may indeed still have reduced sensitivity due to population characteristics, for example due to a low number of biomonitoring samples with detectable levels of the chemical of interest, limited exposure contrast between groups, or few observed cases of the outcome of interest. Careful consideration by qualified statistical experts should be given during the study design phase to define the appropriate statistical power to detect the expected effect size, considering sample size overall and across subgroups, precision, outcome prevalence, and number of covariates in the model (National Academies of Sciences, Engineering, 2022). It is worth mentioning that while under-powered studies may hinder our ability to interpret null results as a lack of association, they can still be considered when integrating evidence through the use of meta-analyses, statistical approaches that combine the results of multiple scientific studies. The use of such methods,

however, is influenced by the type of studies that are available, which are often impossible to combine due to high levels of heterogeneity (National Academies of Sciences, Engineering, 2022). Study sensitivity can help determine if a null finding indicates a true lack of association. Differences in study sensitivities may explain apparent inconsistencies across studies (i.e., studies with greater sensitivity might be more likely to observe an effect).

2.3.3 Experimental in vivo and in vitro studies

To ensure reliability of experimental studies, it is important to carefully consider study design, test conditions and statistical analyses, and to clearly describe them. Standardised OECD TGs and accompanying guidance documents provide useful guidance, even for research studies that deviate from standard TGs. General reliability considerations are common to experimental *in vivo* and *in vitro* studies (Table 1.1), but some critical considerations are specific to the type of test system. Adequate reporting of methods and results is fundamental to demonstrate reliability based on these considerations. Table 2.1 lists core reporting elements for *in vivo* and *in vitro* studies. Many reliability assessment tools have been developed for experimental animal (eco)toxicity studies (Beronius et al., 2018; Krauth et al., 2013; Moermond et al., 2016; Smith et al., 2018) and *in vitro* studies (Roth et al., 2021; Tran et al., 2021). New ones are in development e.g., for *in vitro* studies (Svendsen et al., 2023). For ecotoxicity studies, an overview of reliability assessment tools and considerations in choosing the best-suited method is given in (Moermond et al., 2017).

The identity, purity and composition of the test item may affect study results, and it is critical that these factors are characterised and clearly described in experimental studies. In case the test item is a formulation or other mixture, it is also important to characterise the composition of constituents. Similarly, the identity of solvents/vehicles, negative controls, and positive control items need to be clearly described. In addition to the minimum reporting elements presented in Table 2.1 for describing the test item, analytical verification of the test item should be conducted at study initiation and termination (and during for longer duration studies) to verify substance identity and stability. For example, in the US National Toxicology Program, about 3% of purchased chemicals were identified as wrong substances during analytical verification. The rate of labelling inaccuracies rises to 10% when inaccurate purity information is included (NIEHS, 2019). For chemically unstable substances, more frequent analyses may be needed. For complex substances, studies lacking analytical verification of chemical identity, purity, and composition may be excluded from use in a regulatory assessment.

The physicochemical properties of the test item should also be known and taken into consideration. For example, volatile or poorly soluble substances need specialised experimental systems to maintain the desired exposure conditions. For nanomaterials (and other substances in particulate form), additional considerations for physicochemical characterisation include particle size, size distribution, shape, degree of aggregation, surface area and charge. In aquatic ecotoxicology, high biomass loading can influence the uptake of chemicals. For ionisable chemicals, the test pH influences the ionisation stage and hence bioavailability (Köhler et al., 2023).

Test conditions must ensure the stability of the test item within the experimental system. Any degradation or formation of new compounds should be minimised (e.g., by storing in the dark to prevent photolysis) or fully characterised (including identification and quantification of degradation products) to obtain reliable results.

In vivo - specific considerations

For *in vivo* studies, it is important that the choice of animal model is justified, that the species, strain, sex, and life stage are clearly described, and that information about the supplier is provided. The animal model used should be reliable and sensitive for investigating the endpoints of interest. Existing OECD TGs and corresponding guidance documents can provide guidance on appropriate animal models, appropriate

timing for dosing as well as considerations for selection of doses and dose spacing, vehicle/solvent, and route of administration of the test item.

The doses administered in an *in vivo* study should preferably be motivated and based on available information, such as existing data on toxicity and toxicokinetics. The achieved concentrations, stability, and homogeneity of the test item (in the prepared solution) should be determined as appropriate for the type of study and test item. In aquatic toxicity studies, including higher tier mesocosm studies, field studies in bees, and other non-target arthropods, test concentrations should be analytically verified at all concentrations and/or doses or at least in the highest and lowest ones. This is to confirm concentrations/doses at the initiation of exposure and throughout the period of exposure.

A concurrent negative control should always be included. Care should be taken that the vehicle/solvent used to solubilise the negative control does not influence study results, e.g., by causing toxicity or affecting how the test item is absorbed. Depending on the solubility of the test item, as well as how it is being administered (orally by gavage or via feed or drinking water, dermally or via inhalation, or via surrounding media such as water, sediment, or soil), different types of solvents or vehicles may be appropriate. The route and method for administration as well as the timing, frequency, and duration of administration should be appropriate for the endpoints being investigated and considering the toxicokinetics of the test item.

Housing conditions and experimental procedures can affect outcome parameters in *in vivo* studies, for example by influencing body weight, stress levels, and hormone levels with potential consequences for the reliability of results (Abidin et al., 2024; Baily, 2018; Schumann et al., 2014; Verwer et al., 2007). Some housing conditions that need to be considered are the number of animals housed together or if individual housing is appropriate, as well as temperature, humidity, light-dark cycle, pH, oxygen content, and feeding regime.

A central objective in regulatory assessments of chemicals is to establish a No Observed Adverse Effect Level (NOAEL) or a Benchmark Dose (BMD) i.e., a dose at which there are no significant (adverse) effects from exposure and that can be used as reference point or a point of departure in quantitative risk assessment. Identification of this reference point is highly influenced by the experimental design and statistical power of a study. In *in vivo* studies, some study design factors that must be especially considered are the number and levels of doses tested and number of animals per sex and treatment group, as well as the methods used for statistical analyses. This includes clear description of the experimental unit, e.g., the individual, litter, or cage/tank of organisms. Randomisation of animals to treatment groups and to different tests should always be carried out, and the method for randomisation should ideally be described. Study reliability may also be compromised by a lack of blinding at the outcome assessment stage, especially when measurements are subjective or not automated. However, blinding is not always best practice. For instance, blinding is important when analysing histopathological data but is generally not recommended during the initial evaluation of tissues because masked evaluation can make the task of separating treatment-related changes from normal variation more difficult and may result in subtle lesions being overlooked (Crissman et al., 2004; OECD, 2010; OECD 2015). Best practice entails initial evaluation with knowledge of treatment group followed by a secondary (blinded) evaluation of tissues. This secondary blinded review may be reserved for cases where a treatment-related finding is observed.

A broad range of endpoints and measurement techniques can be employed to investigate toxicological effects, including observational, physiological, molecular, and biochemical (omics), imaging techniques, each requiring specific reliability considerations (e.g., instrument maintenance, calibration, adherence to standard operational procedures (SOPs). Studies investigating unconventional endpoints that are typically not covered in OECD TG studies often come with a wide variety of designs, posing a challenge to reliability assessment. For example, in the case of ecotoxicity behavioural studies, rapid adoption of emerging reporting and evaluation tools (e.g., EthoCRED, Bertram et al., 2024) is instrumental to build trust for regulatory consideration.

In vitro - specific considerations

For *in vitro* studies, all instruments should be regularly maintained, calibrated and validated (if required) and all measurements should be performed according to SOPs (OECD, 2018a). It is important to note that the performance of the measurement method can also influence the readout reliability and the interpretation of the *in vitro* data.

It is crucial to characterise and report in full all the components of the test method to ensure the reproducibility of results. This includes the test system (e.g., cells, tissue, organ, or sub-cellular fraction(s)), other biological components (e.g., serum, antibodies, proteins), all supporting materials and reagents (e.g., disposables, culture media), and batch references, if relevant.

Information on the components needed to perform the method should include the source (or supplier) information, and for cell-based test systems the species and the sex from which they originate should be recorded. If available, Research Resource Identifier (RRID) should be reported for components, such as cell lines, plasmids, and antibodies, as they provide an easy way to identify which specific component was used. Complete and clear identification of method components helps to clarify possible differences in results obtained from different studies and enables others to reproduce the data using exactly the same test system or other key components. The endogenous metabolic competence of the system, as well as any metabolic activation systems employed, such as the addition of an S9 fraction, should be specified.

Besides conventional 2-D test systems, complex *in vitro* models, such as stem cells, organoids, spheroids, Organ-on-Chip (OoC), Microphysiological Systems (MPS), 3D bioprinting are widely used in the research community. For these models, the biological component of the test system is coupled with supporting materials, used to build a 3D structure and/or to add other physiologically relevant features (e.g., fluid flow, mechanical stretching). These materials can be matrices or scaffolds (e.g., Matrigel, collagen, or fibrin gel) or used for the technical device manufacturing (e.g., polydimethylsiloxane (PDMS) or other polymers).

Physicochemical properties of the test item influence dissolution, sorption of the test item to materials, and cellular uptake. Low solubility of the test item is a common issue in *in vitro* studies. While solubility can be verified by visually inspecting the solution, it is preferable to use more advanced methods such as High Performance Liquid Chromatography (HPLC) and UV spectroscopy. Testing hydrophobic substances in plastic plates may reduce bioavailability because of sorption to the walls (OECD, 2019b). Large, hydrophobic molecules, for example, are easily absorbed into PDMS (Auner et al., 2019) or bound to reagents commonly present in the culture medium (e.g., albumin, serum). To gain a better understanding of the problem, computational models can be used to estimate the amount of compound that the cells are exposed to and the factors influencing it (Proença et al., 2021). Ideally, nominal concentrations are analytically verified. To determine the bioavailability of the test item during the experimental procedure, measuring its effective concentration can help the interpretation of results. For studies on nanomaterials particle size, shape, and surface charge need to be considered, and appropriate measures taken to ensure homogenous dispersion and to avoid particle aggregation (Shao et al., 2023).

An appropriate vehicle/solvent control must be included to account for any effects caused by the vehicle/solvent specifically. The choice of vehicle/solvent is determined by the solubility of the test compound, as well as the test system used. The study should also include an appropriate positive control. Negative control items may also be included to exclude false positives from reagents and test conditions. The negative control is a reference chemical for a specific endpoint and is different from the vehicle/solvent control. Curated lists of reference chemicals developed for specific toxicity mechanisms in the context of international validation frameworks provide an ideal source for the selection of positive and negative controls (e.g. Sund et al., 2021). A good example of a curated data resource is the validation dataset for skin sensitisation including both animal and human data (OECD Series on Testing and Assessment no. 336).

The concentrations and duration of exposure of the test item should be clearly reported and justified, considering solubility as well as cytotoxicity of the test item. Cytotoxicity might affect the reliability of results

in an *in vitro* study and should therefore be measured under the same conditions as the endpoint(s) under investigation, ideally on the same plate or during the same run (Krebs et al., 2019; OECD, 2018a). Conclusions should be drawn under conditions (concentration of test compound and exposure duration) that do not induce significant cytotoxicity.

The statistical design of an *in vitro* study includes consideration of the concentrations tested, including spacing of concentrations, the number of technical and biological replicates, as well as proper methods for statistical analyses. Randomisation of treatments in *in vitro* studies may be applied to control for bias introduced by the position of the sample in a multi-well plate. However, randomisation is not always best practice in *in vitro* studies, especially if dosing is performed manually since randomisation may introduce pipetting errors or data transfer errors (OECD, 2018a).

Additional test conditions that impact the viability of the test system, as well as toxicity of the test item, include incubation temperature, humidity, CO_2 concentration, media used, and control of contamination, as well as seeding density and number of cell passages (OECD, 2018a).

The readout of an *in vitro* study is usually the measurement of one or more (functional) endpoints, through different technologies. Among these, the most common are:

- Microscopy and high-content imaging
- Omics (gene or protein expression, measurement of test items and its metabolites)
- Spectrometric measurements (e.g., liquid chromatography/mass spectrometry, UV, absorbance, luminescence)
- Analytical biochemistry assays (e.g., ELISA)

2.3.4 Computational (in silico) studies

The reliability terminology used in the domain of *in silico* modelling differs in some respects from that used in the experimental literature. Reliability assessment tools are typically embedded within a given computational workflow (e.g., Myatt et al., 2022). Reliability refers to the accuracy of prediction, which should generally be reproducible if the model is adequately described and/or accessible as a software tool. Scientific relevance is thus implicitly assumed to the extent that the model predicts a property or endpoint of toxicological interest, while regulatory relevance refers to whether the property/endpoint predicted corresponds to an information requirement. Typically, information requirements for hazardous properties are expressed in terms of adverse outcomes in standardised studies with different types of organisms or *in vitro* test systems relevant to human health or environmental safety assessment.

Common applications of *in silico* models in toxicology include (Q)SAR and PBK analysis. In ecotoxicology, PBK models may be coupled with toxicodynamic models (TK-TD models) to simulate individual- or population-level effects. Several mathematical models are available to perform (quantitative) *in vitro* to *in vivo* extrapolation (qIVIVE) and translate the data generated with an *in vitro* system to *in vivo* relevant information (Chang et al., 2022). Other types of models, such as quantitative AOP (qAOP) models and system biology models are gaining consideration for potential regulatory application.

The main factors underpinning the reliability of model predictions are listed in Table 1.1. OECD guidance is available for (Q)SAR models (OECD, 2024) and PBK models (OECD, 2021) to assess the validity of models and their predictions for regulatory use. Other types of *in silico* models may require additional guidance. The inclusion of model reporting formats and checklists enhances transparency and reproducibility

An additional layer of complexity with *in silico* models is that they are often built on data (observational or experimental), which ideally should also be evaluated using an appropriate tool. This is necessary, for example, to integrate *in silico* models with Defined Approaches (DAs). (Q)SAR analysis is already part of two of the DAs in OECD TG 497 to assess skin sensitisation. Additionally, *in silico* modelling has been

used to develop DAs based on experimental data. An example is the Genomic Allergen Rapid Detection Test Method for Skin Sensitisation (GARD™skin), described in OECD TG 442E. The supporting document of this method also provides a good example of transparency in the underlying algorithms and computational workflow.

Research studies often combine experimental and computational components. For example, *in vitro* models (e.g., 2-D cultures, spheroids, organoids, microphysiological systems), are often combined with predictive qIVIVE to correlate experimental model parameters to *in vivo* organ and systemic endpoints. In this case, depending on the nature of the experimental system, the predictive model should take into account, medium content and volume, the liquid-to-cell ratio, non-specific binding, organ layout and flow rates and characteristics, transport capacity, and scaling factors (OECD, 2021).

The extent to which a model prediction can be relied upon in regulatory decision-making (i.e., its adequacy or credibility) depends on the context of use, including how consequential the decision is, and the weight of the prediction in reaching the decision. Reliability considerations used in the context of risk of bias are less discussed for computational models compared to observational and experimental studies (Cronin et al., 2019). However, systematic errors can be introduced through the choice of chemicals included in the training set, and the choice of modelling methodology. Good modelling practices can help to identify bias, including the assessment of model robustness by various statistical tests such as Y-scrambling or other intentional perturbations of the training set (Cronin et al., 2019). A robust model is relatively insensitive to changes in the training set and thus unlikely to be a “chance correlation”. There are also statistical tests (such as cross-validation and external validation) to mitigate against overfitting, which gives an inflated measure of model predictivity. The predictive performance of the model also depends on the way in which the applicability domain is defined e.g., a (Q)SAR for neutral molecules is less useful for ionising chemicals. Hence, transparency regarding the applicability domain is crucial contextual information for the interpretation of model performance (this information might not be provided in detail in the case of proprietary software). OECD guidance (OECD, 2024) also addresses the applicability of a (Q)SAR model to individual chemicals. Additional considerations affecting the confidence in *in silico* models, going beyond relevance, reliability and applicability, include a) accessibility to model code, software tools and underlying data (including chemicals and their structures); b) the transparency and interpretability of the model algorithms; and c) whether the model has been peer reviewed (Cronin et al., 2023).

2.4 Publishing data

Research funders, publishers with the associated editorial boards, and repository managers define and implement policies, set requirements, and offer options to researchers to publish research outputs. Generally, these policies have paid more attention to the “data” component of research outputs. There is also an emerging trend to promote the use of method repositories. Within the existing requirements and resources available, it is mostly the responsibility of the researchers to choose what and where to publish, depending on the nature, size, and structure of the data. Researchers have various options for publication of research data, which impact on the ability of assessors to find, screen, and evaluate results efficiently.

2.4.1 Peer reviewed scientific journals

Publication in a scientific journal is the main means for researchers to make the study publicly available. Choosing journals that implement rigorous peer review, open data policies, and structured submission approaches promotes scientific and regulatory (re)use. The peer review process implemented by most scientific journals addresses aspects of data availability, methodological transparency, scientific reliability, and relevance. Therefore, peer review is often a pre-requisite for regulatory consideration of research data. The proliferation of journals and limited availability of reviewers have put the peer review process under stress, making it increasingly difficult to ensure quality of publications. The peer review process, in fact, is

heterogeneous across scientific journals and does not guarantee reliability for regulatory use. Reporting requirements of study design (including protocols, codes) and results for assessors may be more stringent than for journals. Reporting of main results in machine-readable tables comprising relevant endpoints and statistical parameters facilitates interpretation and data extraction. Length limitations implemented by most journals imply that it is often not possible to provide full descriptions of methods and results (including detailed study protocols, codes, and raw data) in the main manuscript (European Commission, 2024). Therefore, information provided in the methods and results sections should be clearly linked to detailed data published in supporting information or, preferably, in dedicated repositories. Clearly, data should always be provided in accordance with the relevant data protection policies. For projects generating large amounts of data, publications presenting database(s) hosting the full results can be very useful to both research and regulatory communities (e.g., Govarts et al., 2023; Richard et al., 2016, 2021).

2.4.2 Data repositories

As humans increasingly rely on computational support to deal with data, the data management dimension has gained prominence in modern research. Integrating research outputs in modern digital infrastructure benefits the findability, accessibility, interoperability, and reuse of data (FAIR principles) for both scientific and regulatory purposes[23]. FAIR principles support reproducibility of results, a fundamental principle in scientific research and regulatory science (Wilkinson et al., 2016).

Data repositories are a fundamental asset in the lifecycle of research data. They provide the means to publish full results, including negative results, implement reporting standards and to integrate new data over time. Research projects, scientific communities, research organisations, and regulatory authorities have developed many data repositories over time to host data from multiple disciplines related to chemical safety science. Data repositories differ in their funding, data integration, data quality control, governance, and sustainability. Some offer desirable features including open access, version control, adoption of internationally recognised reporting standards, and interoperability with data analysis and modelling tools. Different governance models exist to update databases up with newly generated data. Whereas most regulatory databases (e.g., ECHA CHEM[24]) include research data, the focus of this section is on repositories available to researchers for data integration. References and links to the resources introduced in this section are reported in Annex B.

Journals recommend submitters to make data available in the main manuscript, in supporting information or in domain-specific repositories. Some journals refer to inventories of specific repositories, such as Registry of Research Data Repositories (re3data), a global registry of research data repositories. If specific repositories are not available, submitters are referred to generalist repositories (e.g., Zenodo, OSF, Dataverse). In some cases, journals require authors to deposit their data in a repository as part of the manuscript submission. Research funders may also recommend or mandate certain solutions. For example, Horizon Europe researchers are encouraged to publish any large scale environmental and human (bio)monitoring data in the Information Platform on Chemical Monitoring[25] (IPCHEM).

Research programmes often develop data infrastructure and governance to serve needs and objectives of scientific communities, such as project consortia, scientific communities of practice, or research groups involved in federal research programmes. US EPA Toxicity Forecaster (ToxCast) is a well-known example of a large-scale research programme run by regulatory authorities making data available on a dedicated platform[26] (Richard et al., 2016, 2021). Many academic research projects, however, could not secure long-

[23] www.go-fair.org/fair-principles/
[24] https://chem.echa.europa.eu/

[25] https://ipchem.jrc.ec.europa.eu/

[26] https://comptox.epa.gov/dashboard/

term maintenance and continuous data integration beyond the duration of the project. In the EU, the Norman and the Nanosafety Data Interface (the latter implementing the eNanoMapper database systems), are notable exceptions of information systems that successfully secured resources for continuous development and operation over the years (Annex B). Long-term commitment by funders, project leaders, data producers and data managers, is a prerequisite to implement good practice in the management of research data. The focus of data management in research projects has shifted from the development of *ad hoc* solutions to support data producers implementing FAIR principles within existing (meta)data infrastructures, as pursued by PARC (Marx-Stoelting et al., 2023). PARC is developing the FAIR Data Hub[27] to facilitate management of existing, or newly generated data in line with FAIR principles and to support chemical risk assessment. The PARC FAIR Data Hub supports researchers as well as assessors from both research and regulatory sectors. Towards long-term sustainability of the architecture of the PARC FAIR Data Hub, PARC explores collaboration and alignment with research infrastructures within the European Strategy Forum on Research Infrastructures (ESFRI), such as ELIXIR (see Box 2.1. There are also notable examples of public research organisations developing and maintaining data infrastructures and bioinformatics services freely available for storing and analysing large and complex biological data (e.g., NCBI, EMBL-EBI, see Annex B).

In some cases, public authorities have a long-term mandate to develop and manage infrastructure designed to host specific types of research data for regulatory use. In this case, any research group can request data integration, provided they adhere to the adopted reporting standards. This governance model serves end-to-end data providers and final users (assessors). However, it requires resources for infrastructure development and maintenance, scientific coordination including data quality control, as well as commitment from funders, repository managers and data providers. It is therefore resource-intensive, as experience has shown with IPCHEM (Comero et al., 2020).

2.4.3 Importance of publishing all results

Research studies may show effects, sometimes referred as “positive”, or no effects, or “negative”. No-effect results may be due to low study sensitivity or true lack of biological effects or a combination of both. Studies should be analysed considering the results and the uncertainty around them (i.e., effect measures and confidence intervals). Dichotomisation when reporting study results (negative vs positive, or effect vs no effect) only based on statistical significance (P values) can lead to misleading interpretations. Historically, scientific research has been biased towards primarily publishing studies that report effects (or reject the null hypothesis), with an increase of positive conclusions in papers from 63 to 85% in the period 1990-2007 (Fanelli, 2012). A study in France has shown that 81% of researchers have produced negative results, 75% are willing to publish them, and only 12.5% have had the opportunity to do so (Herbet et al., 2022). Negative results are very important because they inform about non-sensitive species or endpoints, prevent duplication of studies (especially animal studies), and provide crucial input to computational model development.

Funders and scientists generally prefer to focus on effects and perceive negative or no-effect results as having limited scientific impact (Bespalov et al., 2019; Echevarriá et al., 2021; Fanelli, 2012). However, in regulatory assessments negative results can be as impactful as results showing effects (Weintraub, 2016). WoE approaches are used in hazard and risk assessments to critically examine, prioritise, and integrate results from different types of studies with similar and different experimental approaches to reach a general conclusion (OECD, 2019; SCHEER, 2018; US EPA, 2022). Selective publication of study results based on the effect sizes and/or their statistical significance ("publication bias") may impact the WoE, biasing the effect size away from the null hypothesis in situations where no effects exist, or skewing the estimate when the effect does exist (EFSA, 2024).

[27] https://www.eu-parc.eu/thematic-areas/tools-resources/parc-fair-data-hub

One example where publishing all results is important is the development of a species sensitivity distribution (SSD) in ecotoxicology, which can be used to derive regulatory endpoints for risk assessment. Inclusion of additional data is necessary, even when no adverse effects are observed at the highest tested concentration, to better represent the distribution of sensitivities within a taxonomic group or between species. There are well accepted approaches to include data in SSDs when no adverse effects are observed at the highest concentration tested (Kon Kam King et al., 2014).

In silico models and machine learning algorithms can be biased, if the training set only contains positive results, reducing their predictive power[28]. Hence, the current machine learning algorithms might have limited application unless there is a noticeable shift in publishing all available results.

Publishing all results supports assay validations, particularly for the validation of NAMs that provide non-animal information for hazard and risk assessment (Browne et al., 2015; Kleinstreuer et al., 2017). During the validation of an assay, it is not only important to test positive substances with low, medium, and high potencies to characterise sensitivity of the assay, it is also critical to include an appropriate number of negative substances to determine the specificity of the assay (Du Pasquier et al., 2024). Finally, a benefit of reporting no-effect results is that rejected hypotheses can help other scientists to avoid flawed concepts, adjust their research plans, and increase their chances of success.

Funders, scientific editors, and reviewers can reduce publication bias by placing more emphasis on publishing negative results. Peer-reviewed journals often reject studies that find no effect, even if they are as scientifically valid and relevant as those showing an effect. This may be because studies that show effects are considered more newsworthy, or because of perceived reduced sensitivity to observe an effect. An approach to increase confidence in negative results is for investigators to simultaneously test well-characterised positive and negative reference chemicals, which will help support the validity of negative as well as positive results (Bespalov et al., 2019; Echevarriá et al., 2021).

Pre-registration of protocols as done in the clinical setting can reduce publication bias. Similar initiatives have been proposed for animal studies (Bert et al., 2019) and may have beneficial effects also for observational epidemiological studies (EFSA, 2024).

[28] https://www.nature.com/articles/d41586-024-01389-7

3 Identification, assessment and use of research data

3.1 General considerations

3.1.1 Regulatory contexts

Research data can be used to address assessment questions across a wide range of regulatory contexts. Regulatory assessments are conducted routinely as part of a regulatory programme, or on an *ad hoc* basis, for example, in the case of a specific question to regulatory agencies or of an emergency incident, such as food contamination. Regulatory programmes differ significantly in terms of:

a. The extent to which information requirements are prescribed and explicitly defined;
b. the ability of authorities to request/generate additional data; and
c. who carries out the assessment and who has the "burden of proof" (legal obligation to demonstrate safe use or that there is an unacceptable risk).

The scope of the search strategy, screening, extraction, evaluation, synthesis, and integration, as well as the extent of reporting at each step of the workflow present similarities and differences across jurisdictions and policy domains. This variability affects the efficiency and effectiveness of the assessment process and requires careful consideration of the appropriate methods and tools to be used.

Many regulatory contexts explicitly require assessors to consider all available scientific evidence, including research data. The practical implementation of such requirement depends on the way it is formulated in legal texts, regulatory guidance, and on the available resources and tools. Annex C presents examples of assessment tasks and regulatory contexts representative of the range of scenarios in scope of this Guidance Document. In some contexts, the collection and evaluation of research data is a preliminary evidence collection step (e.g., fulfilling regulatory information requirements or building systematic evidence maps to support scoping and problem formulation). Preliminary evidence collection potentially feeds into a range of subsequent assessment questions (e.g., Annex D, Case study A). Elsewhere, the collection, screening, extraction, evaluation, synthesis, and integration of research data addresses specific exposure, hazard, or risk assessment questions. The case studies in Annex D explore in detail some of these contexts.

3.1.2 Principles

In regulatory assessments, research data are often considered in weight of evidence (WoE) assessments (Figure 1.2). WoE assessment is *"a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted"* (WHO, 2011). Research data is therefore in the scope of guidance on WoE assessment issued by regulatory authorities (OECD, 2019; SCHEER, 2018; US EPA, 2022). In some recent guidance documents, the phrase evidence integration is used to describe WoE assessment (US EPA, 2022; EFSA 2023).

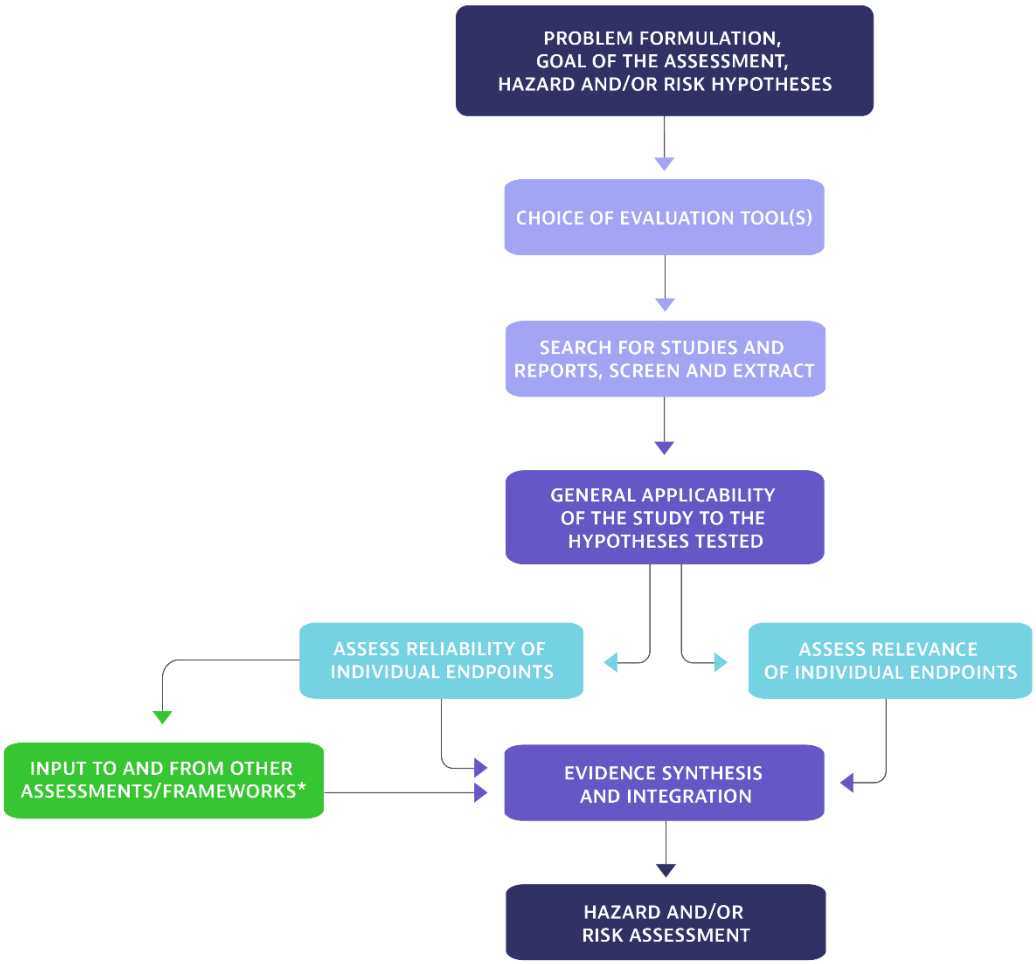
Whereas differences in the regulatory contexts and availability of resources require adaptability in the assessment approaches, guiding principles for retrieving, selecting, extracting, assessing, synthesising, and integrating research data apply independently of specific contexts. Building on those defined for WoE assessments (OECD, 2019b) these principles are:

- **Fitness for purpose**. The chosen approach provides a suitable evidence base to answer a specific regulatory question in an efficient way.
- **Scientific rigour**. Search, screening, extraction, evaluation, synthesis, and integration are based on recognised scientific criteria. These are defined *a priori*, independent of stakeholder interest. Expert judgment is an integral part of any scientific assessment and is necessary for the consideration of research data in regulatory assessments. Intended bias is avoided and unintended bias (variability in expert evaluations) is minimised.
- **Predefined approach**. Implementing predefined protocols improves consistency in the identification, screening, extraction, evaluation, synthesis, and integration of research data. Formal systematic review protocols are one type of predefined approach, but others can be used according to the regulatory context. Deviations from the protocol are acceptable when a sound justification is provided.
- **Transparency and openness**. Clarity and accessibility of assessment methodologies, judgments, and results in all steps of the workflow is fundamental for efficient communication, as well as to build trust and facilitate reuse of assessment outcomes.

3.1.3 Approaches

Systematic review (SR) is the most comprehensive and rigorous approach implementing the guiding principles defined above. SR is a methodology designed to minimise bias and error and maximise transparency when answering a research question via a WoE assessment. SR methodologies were initially developed in the field of health research (Higgins et al., 2023) before being applied to chemicals safety assessments (EFSA, 2010; National Academies of Sciences, Engineering, 2022; NIEHS, 2019; WHO, 2021a). SRs use a structured approach to identify, critically assess, summarise, and analyse data from the studies included in the review. Methods to perform the SR are predefined in a protocol. This includes the identification of the databases where the search is performed, which could cover both published peer-reviewed literature, regulatory and scientific databases, and grey literature. The SR also involves literature search strategies (to be adapted for each source of evidence), inclusion and exclusion criteria for screening the records, a data model for the data extraction, and methodologies for study evaluation and evidence synthesis. The PRISMA reporting guidelines for SRs provide a checklist of 27 items that facilitates the reporting for each step of the SR, and a diagram to present the flow of studies through the screening process (Page et al., 2021). Although designed primarily for health interventions, the PRISMA guidelines are broadly applicable to other disciplines, including chemical assessments. Systematic evidence maps (SEMs) employ SR methods to identify, summarise, and optionally assess the reliability of studies. Unlike traditional SRs, SEMs do not delve into in-depth data analysis to draw definitive assessment conclusions but instead provide a broad overview of the existing evidence landscape. They are used to inform problem formulation, and to guide the development of assessment protocols (Thayer, Shaffer, et al., 2022). Case study A (Annex D) presents a SEM on PFAS.

Figure 3.1. Stepwise approach towards reliability and/or relevance evaluation of research data



Note: *- Generally, reliability assessments are transferable across regulatory assessment contexts, as they focus on the intrinsic quality of the work. In contrast, relevance assessments are specific to the goal of the assessment and may change depending on what is assessed. Thus, they may not be transferrable from across contexts.
Source: Adapted from (C. Moermond et al., 2017).

Systematic reviews consider both reliability and relevance of the evidence (WHO, 2021a). The collection of data by targeted search terms, the inclusion/exclusion search criteria, and the screening of title/abstract and/or full texts constitute a first screen of relevance. When a study is considered in another regulatory context, its relevance must be re-evaluated (Figure 3.1). However, the reliability assessment from previous evaluations may still be applicable. Both relevance and reliability assessments contribute to evidence integration in the subsequent assessments. The case studies illustrate examples of SR steps implemented to address assessment questions (Case study A, Case study B, and Case study D).

Across regulatory frameworks, not all assessments utilise the full SR methodology. The approaches used for identification, use, and integration of data in regulatory assessments should be fit-for-purpose as related to the regulatory framework and assessment question (EFSA, 2017c). Assessors adapt workflows to fit

specific tasks also depending on data availability, time and other resources available, and the acceptable level of uncertainty. Flexible, less resource-intensive approaches can be appropriate in certain contexts. Irrespective of the approach taken, the guiding principles defined above remain valid.

In many frameworks, a review of all available and relevant literature is obligatory for registrants (Case study D). In some frameworks, however, the reporting of a comprehensive screening of research data is not explicitly required. For example, in fulfilling EU REACH information requirements (i.e., as formulated in Annexes VII to X), registrants should consider available research data to assist in identifying the presence or absence of hazardous properties, and to avoid duplication of testing, especially those tests involving animals. In practice, consideration of research data is linked to the obligation to provide guideline studies or waiving them with alternative information, which normally ensures that registrants perform such comprehensive review at least in case guideline studies are not available.

In hazard classification and characterisation, all studies that are critical for a classification or for the derivation of regulatory endpoints need to be considered. However, normally only few studies if not a single one are critical for the assessment. In hazard classification, for example, it is possible that a single reliable and relevant study determines classification. When the evidence at hand is sufficient to reach a conclusion, searching for additional evidence may be unnecessary. In hazard characterisation, there are situations where all relevant and reliable data are equally critical and therefore need to be screened and assessed with the same level of scrutiny. This is the case of species sensitivity distributions (e.g., EFSA, 2014). In other situations, when regulatory reference values are determined by one or a few critical studies, studies reporting results on less sensitive endpoints may not need the same level of scrutiny.

Regulatory risk assessments often follow tiered and iterative processes, with data needs at higher tiers depending on the results obtained at lower tiers. Research data may be used at any tier. Historically, *in vitro* and computational data have been used at lower tiers, while *in vivo* and observational human data has been relied upon to reach conclusions at higher tiers. However, more recently NAM data, supported by mechanistic understanding of AOPs, have been relied upon (at least in part) in reaching a regulatory conclusion (e.g., for the identification of endocrine disruptors; Case study B). Additional questions may arise along a tiered assessment process, triggering new data needs. Targeted searches may be performed to fill very specific information gaps (e.g., a missing parameter in a PBK simulation). In model-driven risk assessments, sensitivity analysis can inform on those parameters that require more extensive screening and evaluation (Dent et al., 2021).

Expert judgment plays a key role in regulatory assessments, as in any scientific activity. The involvement of different assessors may lead to different conclusions on the same topic, even when the same regulatory approach is followed. These differences may also be due to the variety of psychological biases to which expert judgment is subject. Formal approaches for expert knowledge elicitation (EKE) can be used to counter these psychological biases and to manage the sharing and aggregation of judgements between experts. For instance, EFSA has published guidance on the application of these approaches when eliciting judgements for quantitative parameters (EFSA, 2014c).

Assessors can take advantage of previously performed assessments of specific steps of them. This includes recent SRs conducted by other assessors or researchers, although it is important to understand if the existing SR answers the regulatory question at hand. Narrative reviews usually lack a structured methodology to assess reliability and relevance of individual studies. Their conclusions have limited direct utility in regulatory assessments. Narrative review papers could still serve as a starting point for the assessor to obtain references to relevant primary studies and assess these individually. Curated scientific datasets are another type of resource that can feed into assessment workflows. These may implement systematic screening and evaluation, which may satisfy or at least facilitate downstream assessment needs.

3.2 Searching and screening studies

After the assessment goal and/or research questions have been defined (see Figure 3.1), the next step is to identify informative research data. This step includes identifying potential sources, search methods, and the approach to selecting and screening the studies that may contain relevant data. For most regulatory assessments, peer reviewed journal articles are the primary source of research data, which is the type of research data referenced in this section. Clarifying the context and scope is necessary to ensure that the data needs are well defined, and that the search results are relevant and focused. This is to avoid generating an unnecessarily large and unfocused set of literature that will require screening. Through scoping and problem formulation, a more specific topic or question can result in a more narrowly refined literature search and selection process, increasing utility and reducing time and resources.

For the steps described below related to searching for studies and screening studies, a vast number of software tools are available to assist assessors, which can increase efficiency and capacity. This is an active area of development for machine learning. New tools will continue to emerge to help assessors in identifying relevant studies and data. Some of these tools are introduced in the sections below and in Annex B.

3.2.1 Searching for studies

Access to peer reviewed journal articles is most easily attained by searching bibliographic databases such as, but not limited to, Medline, Scopus, EMBASE, SciFinder, or platforms to access multiple databases such as Web of Science, CAS STNext, PubChem, PubMed, Europe PMC (Annex B). It is generally recommended to search at least two literature databases to maximise the coverage and recall of relevant studies (Ewald et al., 2022).

When searching databases and other sources for scientific data, it is beneficial to develop a strategy that is well-suited for the scope and purpose identified and adapted to the different sources of information. Consulting a librarian or information specialist, if available, is helpful in devising the most useful strategy. Librarians or information specialists have expert knowledge on how to better structure searches to capture the research questions, the differences between sources of information, and how to adapt searches accordingly (EFSA, 2010).

Keyword searching is the most common approach to searching these databases. The search keywords are used to identify potentially relevant articles based on terms found in article titles, abstracts, author-identified keywords, and database-controlled vocabularies (e.g., PubMed's Medical Subject Headings, MeSH). To start, searches on specific chemicals should include all chemical names and their synonyms. When designing the search string, it is important to understand how each database interprets keywords. For example, a search string composed for PubMed cannot necessarily be reused to search Web of Science since search string syntax is different between the two resources. The use of logical operators (e.g., "AND", "OR") should be considered and implemented, as appropriate.

Other approaches to identify relevant articles include "forward" and "backward" searches, also referred to as "snowballing". Forward snowball searching is a strategy that collects articles citing a specific article or set of articles. Backward snowball searching is a strategy that gathers the articles that have been cited in an article or set of articles. These types of searches can be supported by AI-powered tools. Both approaches are more topic-specific and can supplement a keyword search.

In addition to searching for peer reviewed journal articles, it is also recommended to consider grey literature, or literature that is not published in traditional peer reviewed sources. Grey literature can include government reports, conference proceedings, graduate dissertations, research, and committee reports, and more. Grey literature may be found by using search engines on the internet, or institutional websites, thesis repositories, websites, or databases of regulatory agencies such as ECHA, EMA, EFSA, US EPA,

etc. The OECD's eChemPortal[29] is also a useful database that links to collections of chemical hazard and risk information from different national, regional, and international organisations. Assessment protocols should identify which grey literature resources are searched.

Searching for grey literature is often more complicated because databases of grey literature are scattered and less comprehensive than for the peer-reviewed literature. The decision to search for grey literature can depend on the amount of peer reviewed journal articles that are identified for the assessment. Some databases of peer reviewed literature, such as Scopus, also catalogue types of grey literature but are not a primary source. Search and evaluation of grey literature can be very time consuming and resource intensive, thus pragmatic considerations may limit the extent of effort.

Lastly, many regulatory assessment processes include public feedback periods, which may incorporate research studies that were not initially identified through systematic searches.

Deduplication of articles is often required when searching across multiple databases because databases have overlapping catalogued information. Deduplication is most easily done by comparing citation information between articles. Ideally, unique identifiers like DOIs can be used for deduplication. For sources that do not have a DOI, title and author can be used, but this is more prone to error since this information may be formatted differently across resources. Most reference management software has deduplication functionality.

Documenting the literature search, including the terms used in the search, the time window of the search, and in- and exclusion criteria, is important for transparency and rigour of the risk assessment. The search process needs to be documented in enough detail so that it can be repeated by others (Higgins et al., 2023), so they can evaluate whether or not the most relevant literature was identified.

In addition to identified published studies in peer reviewed or grey literature, data available in curated databases or repositories may also provide information that is useful to consider in an assessment. There are several examples where regulatory authorities or scientific groups have developed and/or maintain repositories of data that have been screened and extracted in standardised format (e.g., OECD Existing Chemicals Database, HAWC, US EPA Comptox Chemicals Dashboard, EFSA OpenFoodTox, ECHA CHEM, (Q)SAR Toolbox, EASIS, IPCHEM, US ECOTOX, Norman Database System). Annex B provides brief descriptions and links to these resources. Some of these (e.g., EFSA OpenFoodTox, ECHA CHEM) use IUCLID, a software application developed by ECHA to record, store, maintain and exchange data on intrinsic and hazard properties of chemical substances. IUCLID is a key tool for both regulatory bodies and the chemical industry and is used in various regulatory frameworks.

Some publicly managed scientific datasets serve general policy needs covering a defined scientific domain, independently of specific regulatory processes. Examples include the US ECOTOX database of ecotoxicity data and the Endocrine Active Substances Information System (EASIS). US ECOTOX database identifies new ecotoxicity studies on aquatic and terrestrial species predominantly from peer reviewed journal articles, checks the completeness of basic reporting information and updates the public database quarterly (Olker et al., 2022). EASIS implements the OHTs to facilitate the reuse and exchange of the data. It is the first IUCLID installation that implements OHT 201, a template dedicated to reporting mechanistic data (intermediate effects) derived from non-animal methods, mostly from data published in the scientific literature (Carnesecchi et al., 2023).

Taking advantage of these sources of information can inform a risk assessment early in the process, potentially saving time and resources. Assessors, however, should be aware that the level of evaluation of these databases differs. Sometimes this is just on the level of data curation (the correct data in the correct field). In other cases, a rigorous reliability evaluation is performed before data integration.

[29] https://www.echemportal.org/echemportal/

3.2.2 Screening literature

After the search has been conducted, it is necessary to screen the identified information to determine what is, in fact, relevant to the question or objective of the assessment. Often, only a small proportion of identified studies (<5%) are considered relevant. Typically, screening the identified articles is completed in a series of successive steps, first using titles and abstracts, and then acquiring and using the full article. At each step of screening, inclusion and exclusion criteria should be clearly documented. Often, these criteria remain the same for the title and abstract and full-text screening. For the purposes of systematic review, Population, Exposure, Comparator, Outcome (PECO) frameworks are often used. PECO frameworks are adaptable but provide a format that is comprehensive when considering elements of a study to be considered during screening. However, other frameworks do exist. Examples of a PECO framework are provided in Case Study A (Shirke et al., 2024), and in Case study B, example 2 (Table B.2).

A substantial number of articles are often excluded by screening using titles and abstracts, and predefined inclusion and exclusion criteria. Reference management software such as EndNote may be useful in executing this step. Advanced tools are also available that have been built specifically for literature screening and include functionalities such as machine learning to predict relevant references, and options to categorise and annotate the literature. Examples of such tools include SWIFT-ActiveScreener, DistillerSR, Rayyan, ResearchRabbit and HAWC (Annex B). Best practice is to have two assessors independently screen each record and describe approaches for resolving conflicts, e.g., discussion, consultation with a third screener.

Relevant articles identified using titles and abstracts undergo screening again using full-text. Typically, the same criteria used for title and abstract screening are used to confirm relevance based on full-text. In addition, reasons for excluding studies at the full-text level should be documented when an assessment is conducted using systematic review. Specialised screening software applications are helpful during full-text review, as they help structure the workflow (including conflicts among screeners) and have annotation capabilities to inventory or categorise the evidence (e.g., reasons for exclusion, type of evidence, etc.). Typically, full-text review takes longer than title and abstract review, especially when annotation is included in the process. Application of machine learning and automated approaches at full-text review and annotation are still in the exploratory and development phases.

3.3 Data extraction

Extracting data from the included studies entails the systematic collection of information from each study included in the assessment. The data extraction strategy should be tested to ensure its feasibility and effectiveness, and described *a priori* in the protocol to enhance standardisation of the process (EFSA, 2010; National Academies of Sciences, Engineering, 2022; WHO, 2021a).

Data extraction requirements vary according to the context of the assessment and should be tailored to fit the problem formulation and analyses planned in the protocol (EFSA, 2010). The format of data extraction can be narrative, tabular, or graphical. When this step involves data harmonisation, it is essential that any data transformations, such as unit conversions, are accurately accounted for.

Data extraction is a resource-intensive and time-consuming process that requires careful planning and execution. To ensure data consistency it is recommended to use predefined tabular or web-based templates that adhere to reporting templates described in Section 2.2. In principle, data from all the studies considered relevant for the assessment should be extracted. However, study evaluation (Section 3.4) may also be performed prior to or during data extraction (EFSA, 2010; National Academies of Sciences, Engineering, 2022). In this context, to make the best use of the resources available, it may be appropriate not to extract the results of studies that are deemed less informative according to the pre-established protocol (WHO, 2021a).

Box 3.1. Artificial intelligence

Artificial intelligence (AI) is a broad term that refers to technologies and methods that aim to approximate human intelligence capabilities (learning, comprehension, data analysis, decision making) with the intention of replicating human tasks. AI, specifically **machine-learning (ML) methods**, have been investigated for application to tasks described in this Guidance Document. The available technologies continue to advance and enable larger training datasets and complex methods to develop ML models.

ML methods have been applied to screen for relevant references when only titles and abstracts are available. ML tools (e.g., SWIFT-AS, SysRev) reduce the number of references that must be manually reviewed, which reduces the amount of time required to screen many references. Other ML methods have been applied to data extraction as well (e.g., Dextr), reducing the time it takes to extract information from scientific studies.

Generative AI (GenAI) models have seen significant advancements in their capacity and applicability. GenAI refers to ML models that generate content based on different types of inputs like text, images, audio, or video. Gen AI models that handle text are called **Large Language Models (LLMs)**. LLMs, such as GPT-4, generate text by predicting the next word in a sequence based on context. They are trained on vast amounts of text data to learn language patterns. They are useful in tasks such as text generation, translation, and summarisation. ML methods have been applied to screen for relevant references when only titles and abstracts are available. ML tools (e.g., SWIFT-AS, SysRev) may reduce the number of references that must be manually reviewed (which reduces the amount of time required to screen many references), prioritise the references to be reviewed, highlight possible mistakes in the review process. Other ML methods have been applied to data extraction as well (e.g., Dextr), reducing the time it takes to extract information from scientific studies. GenAI also has the potential to be applied for study evaluation (e.g., SciScore). Despite the potential applications of GenAI, research is ongoing and more evaluation and validation of GenAI tools and outputs is needed for these approaches to be confidently applied in scientific assessments.

Moreover, AI has the potential to interrogate, interpret, and integrate various forms of unstructured data, including free text, which were previously inaccessible for regulatory purposes. This has the potential to explain the relevance of research data and to integrate different data sources to address specific assessment needs. For example, AI can significantly facilitate the association of research data with the Key Events (KE) of Adverse Outcome Pathways (AOP), as these KEs explain toxicity in a stepwise approach. Such associations add value to both the data and the AOP knowledge framework.

By providing new tools and methods for analysing and interpreting large amounts of data, AI has the potential to significantly enhance efficiency and streamline risk assessment practice.

3.4 Study evaluation

A rigorous evaluation of data reliability is an essential part of the assessment process. General principles of reliability are outlined in Section 1.4 (Table 1.1) and are described in more detail in Section 2.3 for selected study types. This section details how assessors evaluate study reliability. Reliability evaluation should be performed by endpoint because different endpoints within a study may differ in reliability. Study evaluation is inherently expert judgement based and benefits from inputs from multiple assessors and use of structured evaluation tools.

Evaluation tools help assessors to perform a detailed evaluation of relevance and reliability, following on the considerations implemented at the screening step (Section 3.2). The use of evaluation tools facilitates transparent and structured application of expert judgment, providing a basis for resolving disagreements in cases where multiple assessors are involved. Evaluation tools have been developed for observational (National Academies of Sciences, Engineering, 2022; Shamliyan et al., 2010) *in vivo* (Beronius et al., 2018; Krauth et al., 2013; Moermond et al., 2017; Moermond et al., 2016; National Academies of Sciences, Engineering, 2022), *in vitro* (Roth et al., 2021; Tran et al., 2021), and *in silico* studies (OECD, 2024) (Annex A).

Study evaluation tools are tailored to specific scientific fields and regulatory contexts. They often reflect core principles described in test guidelines. While developed for diverse applications, these tools generally address common aspects of study design and conduct that may influence reliability (Table 1.1). A frequent difficulty experienced by assessors is the impossibility of assessing all evaluation criteria. Most tools envisage this situation and allow choosing “not assignable” to evaluation questions. In some cases, assessors may decide to request access to additional information that is not available in the publication (e.g., raw data) to the study authors (EFSA, 2014a). Since this may add a significant amount of time to the assessment process, it may be unsustainable to systematically implement this practice in workflows of assessment tools. Eventually, it is up to the assessors to decide how gaps and uncertainties impact the overall study evaluation and its consideration in the overall WoE.

Typically, study evaluation tools are used to prioritise studies for subsequent consideration in the assessment. Studies with reliability concerns, or those that lack sufficient data (i.e., non-assignable), can be given less weight in a WoE analysis, not used for quantitative dose-response, or potentially be excluded from further consideration. Generally, high-quality reporting greatly facilitates reliability assessment.

Reliability assessment should identify any concerns with study methods and analyses and not simply identify whether it was reported. Criteria-based tools are commonly encountered in regulatory assessments, for example the SciRAP tools, which list predefined detailed criteria for reporting and methodological quality. Several of the tools provide visualisations of the results of study evaluation and/or are combined with tools that provide visualisation (e.g., in HAWC). Evaluation tools can express expert judgement in a qualitative (descriptive) or quantitative way, i.e., assigning numerical scores to each criteria/domain under evaluation, to then obtain an overall score. Certain numerical scoring tools, such as the ToxRTool (Schneider et al., 2009), have a long history of use in toxicology. However, they are increasingly discouraged, as they make it difficult to capture the source, magnitude and direction of possible biases (Arroyave et al., 2021; Higgins et al., 2023; National Academies of Sciences, Engineering, 2022). Numerical scoring tools can give the impression of an undue level of quantitative precision in an exercise that inherently involves expert judgement. Thus, more recent evaluation systems emphasise presentation of the expert judgement rationale underlying the reliability assessment to foster transparency. Developing guidance to assess reliability often needs to be partially customised for a given assessment, i.e., to the specific exposure/test compound or organism being studied (Moermond et al., 2016). A ring test with assessors has shown that the number of criteria to be met for a study to be found reliable differs per study (Kase et al., 2016).

Study reliability can be evaluated using a tiered approach. Some evaluation tools employ a stopping rule, whereby the identification of critical deficiencies can halt the full assessment. When choosing an approach for reliability assessment, assessors may consider the following aspects:

- Does it fit the type of evidence identified during problem formulation and the purpose of the assessment?
- Is it compatible with the time and expertise available?
- Does it allow for a systematic reporting of the evaluations, including the rationale and justification of expert judgements?

- How does it accommodate multiple inputs and conflict resolutions, possibly using a third-party review?

Specific considerations for the reliability assessment of observational studies, *in vitro* and *in vivo* experimental studies, and *in silico* studies are listed in Annex A and are shortly described below.

3.4.1 Observational studies

During the last several years, various tools have been developed to assess the quality of observational studies and human data (epidemiology, clinical). These tools have been designed by organisations/governmental bodies for application in their own assessments as well as by researchers. Some examples are Cochrane Risk of Bias (RoB) tools, NTP-OHAT, SciRAP Epi, ROBINS-E, BEES-C, as listed in Annex A. A recent list of inventories and reviews of RoB tools was made available by EFSA (EFSA, 2024).

Evaluation tools for observational studies vary widely (EFSA, 2024), e.g., regarding their structure and the purpose for which they can be applied. Consequently, evaluation tools should be chosen based on the type of studies to be appraised as well as the context. With regard to structure, there are tools based on checklists, scoring scales, or domain-based approaches. As for other streams of evidence, current tendency is to be cautious in the use of checklists with overly rigid criteria and algorithmic approaches that imply some sort of quantification (Arroyave et al., 2021). Assessors are moving towards the use of domain-based tools, which better allow to focus on the key domains based on the research questions.

Notably, some tools have been designed for use under rapid timeframes such as short-term requests following incidents. These focus on a subset of the most critical considerations for each domain, to allow for conciseness and usability in multidisciplinary teams (e.g., the RaRob tool).

Evaluation should cover sources of bias and reflect on how such bias may affect the likelihood, degree, and direction of risk estimates. For observational studies, key sources of bias include selection bias, information bias (exposure and outcome misclassification) and bias due to confounding factors. To perform such evaluations, relevant expertise in both methodological and the specific exposure and outcome under assessment is essential, and this should be reflected in the composition of the multi-disciplinary team performing the assessment (Arroyave et al., 2021).

Importantly, studies should not be excluded or downgraded solely based on study design, e.g., considering cross-sectional studies automatically of lower quality than cohort studies, and without focusing on the specific exposure-outcome of interest (Arroyave et al., 2021; EFSA, 2024; Steenland et al., 2020; US EPA, 2022). Studies with different key sources of potential bias can still provide relevant information once evidence is integrated, as recognised by the triangulation approach to causal inference (Arroyave et al., 2021; EFSA, ; Lawlor et al., 2016), which encourages considering the net effect of possible biases (Steenland et al., 2020).

3.4.2 Experimental studies

Several tools are available for evaluating reliability and relevance of *in vitro* and *in vivo* (eco)toxicity data, including for *in vivo* studies SciRAP, CRED, NTP-OHAT RoB tool, and the US EPA's IRIS study evaluation approach, for *in vitro* studies SciRAP and INIVTES-IN, as listed and referenced in Annex A. In addition, guidance such as the OECD GIVIMP (OECD, 2018a) provides additional insight to aspects of study design and conduct that may impact the results and study reliability.

Over the past decades, the Klimisch criteria (Klimisch et al., 1997) have been used for categorising reliability of (eco)toxicity studies in regulatory contexts. However, Klimisch heavily promotes adherence to standardised test guidelines and does not provide specific criteria or much guidance for study evaluation. Several tools have since been developed to facilitate more structured and transparent evaluation of

evidence for hazard and risk assessment, including criteria-based tools such as SciRAP and CRED (see examples in Case study B and Case study C), and domain-based RoB tools such as NTP-OHAT and INVITES-IN. These tools generally include specific questions or criteria, as well as guidance, to help the evaluator consider and evaluate critical aspects of study design and conduct. The RoB tools have been developed from tools that have established and long-standing use in systematic review in the field of epidemiology and clinical medicine (discussed above for observational studies). They have been adapted to the evaluation of experimental animal (*in vivo*) studies or *in vitro* studies. Criteria-based tools have commonly been developed in the field of (eco)toxicology with the specific aim to increase structure and transparency in study evaluation and to facilitate evaluation and use of all relevant evidence, including non-standard research studies, in regulatory hazard and risk assessment of chemicals. For example, the NORMAN CRED sub-module allows for comparison of predefined criteria to metadata for the assessed study stored in the database, requires stating why a criterion failed, and documents expert judgement. The evaluation is stored and available to other experts to facilitate agreement on the reliability (Case study C).

Although available tools have been developed in different contexts and have different structure (criteria-based versus domain-based), they generally address the same overarching aspects of study design and conduct that may influence the reliability of results (e.g., Waspe et al., 2021). In some cases, it may be deemed useful to combine aspects from different tools to achieve a study evaluation that is fit-for-purpose. For example, EFSA has combined specific criteria from SciRAP with the NTP-OHAT RoB tool in several evaluations, including their opinions on glyphosate (Case study B), bisphenol A (EFSA, 2017a, 2023c) and starch sodium octenyl succinate (EFSA, 2020).

3.4.3 In silico studies

OECD guidance documents on (Q)SAR (OECD, 2024) and PBK models (OECD, 2021) constitute the main internationally accepted guidance for *in silico* studies (Annex A). In each case, a model reporting format is provided, which the developer or proponent of the model should compile. Additionally, there is a checklist that the assessor can use to check that the main quality and reporting considerations have been followed.

In the case of both (Q)SAR and PBK models, guidance is given on how to score the overall confidence in the model (high, medium, low). The guidance on (Q)SARs -*(Q)SAR Assessment Framework: Guidance for the Regulatory Assessment of (Quantitative) Structure Activity Relationship Models and Predictions,* (OECD, 2024)- goes a step further in supporting the confidence assessment of individual (chemical-specific) (Q)SAR predictions. The reason for this is that a (Q)SAR model may be considered valid in general terms, but individual predictions may have high uncertainty, particularly if they are outside the applicability domain. ECHA provides practical guidance on how to check whether a substance falls into the applicability domain of a (Q)SAR model (ECHA, 2016).

3.5 Evidence synthesis and integration for decision-making

One of the final steps in conducting a regulatory assessment is to reach conclusions based on identified information that is deemed relevant and reliable. Typically, this is a multi-step process where conclusions are initially reached within a line of evidence, followed by reaching conclusions based on evaluating multiple lines of evidence, which is described as WoE approach. OECD Guidance Document No. 311 on *Guiding Principles and Key Elements for Establishing a Weight of Evidence for Chemical Assessment* (OECD, 2019c) presents this conceptually. It intentionally avoids being prescriptive in methodology since judgements are context-dependent and rules or criteria may differ across individual agencies and scenarios. The key is that the process used should be transparent and document the evaluation of all evidence considered whether it is ultimately used or not.

WoE assessment is also a term commonly used in the EU regulatory setting and is often used to characterise the collection of evidence, evaluation of reliability and relevance, and integration of data/studies within and across lines of evidence to arrive at a conclusion (EFSA, 2017c; SCHEER, 2018). However, terminology varies across regulatory programmes. Structured WoE evidence frameworks are most developed for human and animal evidence. Ongoing efforts are underway to increase the transparency of considering mechanistic (and *in silico*) information. Currently, some WoE approaches consider this type of information as a separate line of evidence on par with human or animal, while others consider it more supportive. For example, the US EPA's IRIS Program uses a structured framework approach where the first step is analysing studies within an evidence stream (i.e., human, animal), referred to as "evidence synthesis". This step is considered analogous to "strength of evidence" used in some other assessment processes. Within IRIS, "evidence integration" is a second step that focuses on the integration of human and animal evidence synthesis judgments to draw an overall conclusion(s). This conclusion considers human relevance of the animal evidence, cross-stream coherence across the human and animal evidence, susceptibility, and biological plausibility/mode of action from mechanistic information. "Evidence integration" is analogous to "weight of evidence" used in some other assessment processes. A similar, but less prescriptive approach, is followed by EFSA. The latter considers evidence synthesis as the process of summarising "similar" evidence (e.g., evidence from similar populations, study designs or evidence streams) and recognises that defining what is similar is subjective and depends on the evaluation of the assessor. Case study B provides an example of using EFSA's approach where *in vitro* mechanistic contributes significantly to the evidence integration and conclusions. Evidence synthesis is often a qualitative analysis (e.g., narrative, tabular format), but can be quantitative when studies are sufficiently similar, i.e., meta-analyses (e.g., EFSA, 2017c). It follows that evidence integration is the process of combining evidence that is "diverse". Evidence integration can also happen within the same evidence stream integrating e.g., observational, experimental and computational studies on the same species (EFSA, 2023c). In a similar methodological concept, Health Canada describes "totality of evidence" as what types and sources of information are to be gathered and considered for subsequent assessment and how it can be influenced by the interpretations of "all" available or relevant evidence to date, allowing a reassessment based on the availability of data at a later date. "Weighting evidence" is defined as how much individual sources of evidence are weighted in and integrated into an overall conclusion or recommendation (Health Canada, 2018).

Over the past decade, the use of structured frameworks for reaching WoE conclusions based on a body of evidence have become more common to increase transparency and consistency of the assessments (EFSA, 2017c; NIEHS, 2019; US EPA, 2022). Structured WoE frameworks systematically evaluate and integrate all elements necessary for establishing causality relationships between chemicals and potential adverse effects, incorporating factors that influence confidence in the evidence. Although specific terminology may vary, the factors can be anchored to the Bradford Hill causality considerations of strength of association, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy (Hill, 1965). For example, the OHAT handbook (NIEHS, 2019) builds on the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework. Certainty in a body of evidence can be rated down for lack of randomisation and other RoB concerns, unexplained inconsistency, indirectness, imprecision, or publication bias, or it can be rated up for the magnitude of the effect, dose–response gradient, direction and impact of residual plausible confounding, and consistency across model systems, study designs, or study design types. This framework is applied separately to animal and human evidence and a matrix approach used to develop overall hazard conclusions based on the within evidence confidence judgements. Consideration of mechanistic data is also incorporated in this matrix approach. The US EPA's IRIS Program (US EPA, 2022) also uses a structured framework, but it is somewhat less anchored to GRADE and considers human relevance of the animal evidence, cross-stream coherence across the human and animal evidence, and biological plausibility/mode of action in determining overall hazard conclusions. Other programmes such as IARC (IARC, 2019) and the NTP Report on Carcinogens (NTP, 2025) present their analyses in a less structured, more narrative format, but have method documents

to describe the analytical approach. Publication bias can be challenging to assess but can be explored using funnel plots, Egger's regression, and trim and fill techniques. Other indications of potential publication bias include identification of abstracts or other types of grey literature that do not appear as full-length articles within a reasonable time frame (NIEHS, 2019).

EFSA does not have a predefined structured approach to integrate evidence (EFSA, 2017c), but in a number of assessments an approach similar to the one described above is followed e.g., re-evaluation of erythritol (E 968) as a food additive (EFSA, 2023c). More recently, the *Scientific Committee guidance on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments* (EFSA, 2024), describes a generic approach for integrating evidence from human studies with other toxicological data around a certain health outcome. This approach allows for flexibility based on the amount of evidence available, which will influence the way in which studies are grouped and described. The approach also encourages the use of existing evidence to build a case for or against causality. For example, in cases where epidemiological evidence is limited, studies on clinical markers, which are intermediate steps or risk factors for the disease, can be considered. The approach to integrating different lines of evidence within EFSA are in line with those already mentioned for US EPA's IRIS Program (US EPA, 2022), IARC (IARC, 2019), OHAT (NIEHS, 2019), NTP (NTP, 2025) and Health Canada (Health Canada, 2018). Reliability and relevance to the risk assessment question are considered to identify the key line(s) of evidence for the effect(s) of interest, and evidence on mechanism of action/biological plausibility is used to provide links between different lines of evidence (EFSA/ECHA, 2018; US EPA, 2022). Integration of evidence using an AOP framework can be particularly useful in cases where there are no existing reliable and relevant data for the adverse outcome of concern (e.g., autism spectrum disorder). Biological plausibility considerations can also help identify situations where animal evidence is not reliable or useful, due to relevant toxicokinetic differences between humans and animals (EFSA, 2015), or not needed, due to the large availability of human data (EFSA, 2022). Where there is non-concordance and similar reliability/relevance of lines of evidence, such uncertainty is taken into consideration, and expert judgement is used to identify the most appropriate studies considering the context.

The structured frameworks are conceptually consistent with OECD Guidance Document No. 311 (OECD, 2019c). The OECD guidance presents an illustrative example where WoE conclusions for a line of evidence are based on relevance and reliability. In systematic review processes, study relevance is largely determined at the outset through the development and application of inclusion and exclusion criteria, i.e., studies with no or very limited relevance would be excluded. Relevance can also be assessed as part of WoE by reaching judgements on the "directness" of assessed outcomes to the goals of the assessment.

The WoE analysis process helps build the foundation of a comprehensive uncertainty assessment, an important part of a risk assessment. Uncertainty assessment is typically qualitative, but it can also be quantitative. Uncertainty analyses are also important because they highlight areas where additional research is most likely to have an impact.

3.6 Reporting

Regulatory assessments should be reported in sufficient detail to allow the reader to understand the scope of the assessment (e.g., regulatory task, substance(s) assessed), the methods used (e.g., literature search, screening criteria, study evaluation tools, approach for evidence synthesis/integration), and expert judgements made. This makes the content transparent and comprehensible for the reader and supports the consistency of assessments. Additionally, it facilitates future reuse and updates either by the original or subsequent assessors. There are however legal and practical limitations in what assessors can report in terms of data. Some research data may be proprietary, and only available to the owner and the receiving agency. These are not for public dissemination and often reported in an assessment as "unpublished report/data". Regarding publication copyrights, assessors cannot report and share substantial portions of

an article in the public domain. In other cases, full-study details may be unavailable because the study was published in a scientific journal that no longer exists. In this case, although the literature had been previously retrieved and used in an assessment reusing the data may be impossible.

A pragmatic, yet comprehensive, set of reporting recommendations is provided in Section 7 of the WHO "Framework for the use of systematic review in chemical risk assessment" (WHO, 2021a). While this document was developed specifically for systematic reviews, these recommendations can be readily adapted and applied to support the use of research data in regulatory assessments which are not always carried out in a systematic review context. The following points are particularly important:

- Rationale or objective of the assessment
- Search strategy
- Selection criteria
- Data collection and extraction
- Approach to evaluation of studies
- Approach to evidence synthesis and integration
- Results of each step and overall interpretation (including overall uncertainty assessment)

Ideally, prior to conducting the assessment, the methods for its implementation are recorded as a protocol and published in a publicly accessible database, and the protocol is then followed with any deviations from planned methods being appropriately justified. However, depending on the resources available, uncertainties on the topic to be assessed, sensitivity of the issue, etc. appropriate, fit-for purpose reporting will vary according to the assessment (EFSA, 2023d).

An adapted version of the Table 7.1 on reporting expectations for systematic review in the WHO Framework (WHO, 2021a), suggested for consideration in scientific assessments is provided below.

Table 3.1. Reporting recommendations with examples from cases studies (Annex D)

Step	Recommendation	Examples
Search studies	• Describe information sources (databases, contact with study authors, grey literature sources, etc.) • Present search strings used in databases	• Case study B – Bisphenol F example • Case study D- Section D.3 – Searching for literature • Case study A- US EPA IRIS Program (Shirke et al., 2024)
Screening for relevance	• State eligibility (inclusion and exclusion) criteria • Describe method for screening • List included studies • List studies excluded ideally at title and abstract and full text (with rationale)	• Case study D- Section D.4 – Selecting studies • Case study A- US EPA IRIS Program (Shirke et al., 2024)
Data extraction	• List all data items • Describe method of extraction • Describe data storage software	• Case study A- US EPA IRIS Program
Study evaluation	• Describe methods for assessing relevance and reliability of individual included studies • Describe how relevance and reliability assessment inform data synthesis and integration • Report evaluation results	• Case study A- US EPA IRIS Program (Shirke et al., 2024) • Case study B- Endocrine Disruptors • Case study C- The CRED evaluation method
Synthesis and integration	• Present the principal summary measures • Describe the statistical and qualitative techniques for combining studies • Describe methods for assessment of characteristics of cumulative evidence relevant to interpreting results (certainty or confidence assessment) • If conducted, describe the methods for integrating multiple streams of evidence	• Case study B– Bisphenol F example

Source: Adapted from Table 7.1 “Systematic review: reporting expectations and explanations” in (WHO, 2021a).

4 Recommendations

❖ Recommendation 1 [Researchers, funding bodies, reviewers, editors, publishers]

Apply the following general principles of data quality in data generation, analysis, reporting and scientific review:

- The study, including its methodological and statistical design, should be reported in enough detail to allow the study to be reproduced and the statistical calculation to be checked.
- FAIR principles should be applied to the data and their methods.
- The study design should be fit for its (scientific) research purpose, e.g., appropriate exposure route and duration, tissues/organisms, models, and endpoints should be chosen (Table 1.1 and Table 2.1). A justification for the design should be provided.
- Exposure should be well characterised, to enable justified conclusions on causality or associations between exposure and effects. This should include test item identification and characterisation (including purity information) and exposure measurements.
- Outcomes or endpoints should be defined and measured in an objective manner, to minimise confounding or bias.
- The statistical design should be fit-for-purpose, including choice of sample size/replicates, dose-response models, reference substances, etc.
- All study results, including positive and negative findings (i.e. effect and no effect results), should be reported, focusing on the endpoint measurements and their associated uncertainties. This improves the overall evidence base and helps prevent unnecessary repetition of research efforts, particularly in the case of animal studies.

It should be acknowledged that scientific studies are not primarily aimed at following regulatory requirements. Innovation beyond standards and creative thinking is necessary to advance the field. In any case, applying general principles of data quality brings benefits to researchers themselves (easier review and more citations), reviewers (easier review process) and users of that data, whether in the scientific or the regulatory domain. Improving the quality of reporting should be one of the main priorities of research funders, publishers, and their editors. In the field of (eco)toxicology, checklists for data quality already exist (Section 2.2) but these are not used in a systematic way in the peer review process. We call upon publishers and editors to make reporting checklists part of the review process and aid authors as well as reviewers in improving study quality, e.g., like in epidemiology with the STROBE statement (see also Recommendation 4).

❖ Recommendation 2 [Regulatory scientists/assessors, researchers]

Adapt existing reporting templates or develop new templates for research data. Reporting templates for research data (including details of methodological and statistical design) should accommodate all elements necessary for regulatory evaluation, while providing flexibility to limit and adapt fields to non-applicable or non-standard elements. They should be based on the general principles of data quality (Recommendation 1, Table 1.1). Structured reporting of core elements should be harmonised (Table 2.1). Fit-for-purpose reporting templates for research data facilitate adoption by researchers and improve the ability of assessors to share data via information tools implementing the standards. Research data

generated using new technologies also benefit from the development of reporting standards, as these facilitate interoperability and reuse in both scientific and regulatory domains. Examples of existing OECD activities include the OECD Omics Reporting Framework (OORF) and updating the OECD Harmonised Templates (OHTs) to support research data. In general, reporting standards should be user-friendly to encourage widespread adoption by researchers and assessors, and flexible to ensure long-term compatibility.

- **Recommendation 3 [Database and software developers, researchers]**

Develop data repositories and software that implements reporting standards. Reporting standards on their own are insufficient without the necessary tools to support their use and understanding. Researchers and repository managers drive the development of software applications (e.g., IUCLID and HAWC) and structured (meta)data repositories that support entry, storage, searching, analysis, and visualisation of research data and their underpinning methods. Currently, many data repositories exist and meet the needs of different research use cases, but, without standards, uptake and use by assessors are limited. Implementing reporting standards in tools enables interoperability (i.e., easily transmitting data between tools), which also increases access to available information. As reporting standards are created or updated, integration into tools must be a focus to continue to facilitate the regulatory use of research data.

- **Recommendation 4 [Researchers, reviewers, editors, publishers, regulatory scientists/ assessors]**

Use recognised reporting templates and data repositories when publishing (or extracting) research data. Structured reporting of research data and methods contributes to scientific quality and open science principles. As such, it facilitates review and reuse for regulatory purposes and builds trust in novel methods (Section 2.2). Scientists and journals implementing structured reporting standards expedite the peer review process, improve long-term efficiency of research activities, and increase the chances of accessing future funding by demonstrating the regulatory fitness and impact of their research. Reporting standards, including those developed by regulatory and scientific organisations, are available to researchers for various types of research data. In some cases, researchers can publish research methods and results in dedicated repositories implementing the standards. Using repositories and reporting standards supported by regulatory authorities further improves findability and trust by assessors. Alignment with existing research infrastructure supports interoperability and long-term access to data (Section 2.4.2).

- **Recommendation 5 [Regulatory scientists/assessors]**

Follow guiding principles for searching, screening, extracting, evaluating, and integrating research data in regulatory assessments. Approaches for the consideration of research data in regulatory assessments should follow a predefined, fit-for-purpose protocol and ensure that scientific rigour and transparent reporting are maintained. Systematic reviews, with its associated guidance and tools provide a comprehensive and rigorous framework that implements the guiding principles defined in Section 3.1.2. Depending on the regulatory context and assessment question(s) addressed, workflows implemented by assessors vary in scope and complexity. Following the above-mentioned guiding principles increases the inter-usability of regulatory assessments between frameworks.

- **Recommendation 6 [Regulatory scientists/assessors, researchers (developers of evaluation tools)]**

Provide evaluation tools and clear guidance covering at least general reliability considerations and core reporting elements. Evaluation tools should be accompanied by clear guidance and practical examples (see Recommendation 9 on training). They should facilitate systematic and transparent reporting of the results of the evaluation. Adhering to the general reliability considerations (Table 1.1) and core reporting elements (Table 2.1) described in this Guidance Document ensures a basic level of functional equivalence between evaluation tools used across regulatory programmes.

❖ **Recommendation 7 [Regulatory scientists/assessors, researchers (users of evaluation tools)]**

Select an evaluation protocol and tool that is appropriate for the data and the assessment needs while maximising potential for future reuse. Generally, qualitative evaluation tools are preferable to quantitative tools using scoring as the latter are easily misinterpreted (Section 3.4). When multiple assessors evaluate studies and data in a specific assessment, an inter-calibration should be conducted between evaluators to ensure consistent application of the evaluation tool i.e., common interpretation of evaluation criteria. Any adaptation of existing tools should be explained. Including rationales to support judgements ensures appropriate interpretation of study evaluations, increases the transparency and credibility of the process, and facilitates the reuse of study evaluation assessments across programmes.

❖ **Recommendation 8 [Regulatory scientists/ assessors]**

Reuse components of completed assessments to the extent possible. The identification, evaluation, integration, and analysis of research data for regulatory use require a considerable number of resources (time, people, and funding). The adoption of reporting standards and interoperable tools to search, screen, extract, and evaluate research data supports potential reuse of components of an assessment, allowing assessors to meet regulatory needs more efficiently. To facilitate future updates, the specific timeframes used for the search, extraction, and evaluation should be stated clearly. Caution, however, must be taken to ensure a component meets the specific needs of an assessment. For example, a literature search from a previous assessment may have been too narrow in scope for reuse in another assessment, as the relevance criteria change with the goal of the assessment. Moreover, interpretation of study results may vary over time. The reuse of study evaluations is easier when assessors use the same evaluation tool, or when tools are at least functionally equivalent. Such equivalence can be verified by general reliability considerations and core reporting elements as described in Table 1.1 and Table 2.1, respectively (see also Recommendation 6).

❖ **Recommendation 9 [Regulatory scientists/assessors, researchers, and reviewers]**

Provide training to researchers, assessors, and reviewers to embrace good practice. Training on the tools and approaches described in this Guidance Document is essential to learn, disseminate, and promote good practice. It lays a foundation of excellence, especially for those early in their career. Training is needed to help assessors choose and apply evaluation tools for hazard and risk assessment. Regulatory authorities should consider opportunities to share experiences. For example, many assessment organisations have internal training resources for staff. These could be made accessible via webinars to promote concise self-paced learning and reuse. It is recommended that the development of new tools such as software applications, data repositories, and evaluation tools is accompanied by training protocols that outline data inputs and outputs to overcome steep learning curves.

References

Abidin, İ., Keser, H., Şahin, E., Öztürk, H., Başoğlu, H., Alver, A., & Aydin-Abidin, S. (2024). Effects of housing conditions on stress, depressive like behavior and sensory-motor performances of C57BL/6 mice. Laboratory Animal Research, 40(1), 1–10. https://doi.org/10.1186/s42826-024-00193-8

Ågerstrand, M., Arnold, K., Balshine, S., Brodin, T., Brooks, B. W., Maack, G., McCallum, E. S., Pyle, G., Saaristo, M., & Ford, A. T. (2020). Emerging investigator series: Use of behavioural endpoints in the regulation of chemicals. Environmental Science: Processes and Impacts, 22(1), 49–65. https://doi.org/10.1039/c9em00463g

Ågerstrand, M., Edvardsson, L., & Rudén, C. (2014). Bad Reporting or Bad Science? Systematic Data Evaluation as a Means to Improve the Use of Peer reviewed Studies in Risk Assessments of Chemicals. Human and Ecological Risk Assessment, 20(6), 1427–1445. https://doi.org/10.1080/10807039.2013.854139

Ågerstrand, M., Sobek, A., Lilja, K., Linderoth, M., Wendt-Rasch, L., Wernersson, A. S., & Rudén, C. (2017). An academic researcher's guide to increased impact on regulatory assessment of chemicals. Environmental Science: Processes and Impacts, 19(5), 644–655. https://doi.org/10.1039/c7em00075h

Allen, R., Barn, P., and Lanphear, B. (2015). "Randomized Controlled Trials in Environmental Health Research: Unethical or Underutilized?", PLoS Medicine, 12(1), p. e1001775, https://doi.org/10.1371/journal.pmed.1001775

Altman, D. G., Simera, I., Hoey, J., Moher, D., & Schulz, K. (2008). EQUATOR: Reporting guidelines for health research. Open Medicine, 2(2), 9–10. https://doi.org/10.1016/S0140-6736(08)60505-X

American Society for Cell Biology. (2014). ASCB Member Survey on Reproducibility. 1–12.

Arroyave, W. D., Mehta, S. S., Guha, N., Schwingl, P., Taylor, K. W., Glenn, B., Radke, E. G., Vilahur, N., Carreón, T., Nachman, R. M., & Lunn, R. M. (2021). Challenges and recommendations on the conduct of systematic reviews of observational epidemiologic studies in environmental and occupational health. In Journal of Exposure Science and Environmental Epidemiology. https://doi.org/10.1038/s41370-020-0228-0

Auner A.W., Tasneem K.M., Markov D.A., McCawley L.J., Hutson M. S. (2019) Chemical-PDMS binding kinetics and implications for bioavailability in microfluidic devices. Lab Chip.;19(5):864-874. https://doi.org/10.1039/C8LC00796A

Baker, M. (2016). Is there a reproducibility crisis in science? Nature, 3–5. https://doi.org/10.1038/d41586-019-00067-3

Bailey, J. (2018). Does the Stress of Laboratory Life and Experimentation on Animals Adversely Affect Research Data? A Critical Review. Alternatives to Laboratory Animals. 46(5):291-305. https://doi.org/10.1177/026119291804600501

Beronius, A., Molander, L., Zilliacus, J., Rudén, C., & Hanberg, A. (2018). Testing and refining the Science in Risk Assessment and Policy (SciRAP) web-based platform for evaluating the reliability and relevance of in vivo toxicity studies. Journal of Applied Toxicology, 38(12), 1460–1470. https://doi.org/10.1002/jat.3648

Bert, B., Heinl, C., Chmielewska, J., Schwarz, F., Grune, B., Hensel, A., Greiner, M., & Schönfelder, G.

(2019). Refining animal research: The animal Study registry. PLoS Biology, 17(10), 1–12. https://doi.org/10.1371/journal.pbio.3000463

Bespalov, A., Steckler, T., & Skolnick, P. (2019). Be positive about negatives–recommendations for the publication of negative (or null) results. European Neuropsychopharmacology, 29(12), 1312–1320. https://doi.org/10.1016/j.euroneuro.2019.10.007

Borchert, F., Beronius, A., & Ågerstrand, M. (2022). Characterisation and analysis of key studies used to restrict substances under REACH. Environmental Sciences Europe, 34(1), 1–15. https://doi.org/10.1186/s12302-022-00662-8

Browne, P., Judson, R. S., Casey, W. M., Kleinstreuer, N. C., & Thomas, R. S. (2015). Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model. Environmental Science and Technology, 49(14), 8804–8814. https://doi.org/10.1021/acs.est.5b02641

Calafat, A. and L. Needham (2007), "Factors affecting the evaluation of biomonitoring data for human exposure assessment", International Journal of Andrology, Vol. 31/2, pp. 139-143, https://doi.org/10.1111/j.1365-2605.2007.00826.x

Carnesecchi, E., Langezaal, I., Browne, P., Batista-Leite, S., Campia, I., Coecke, S., Dagallier, B., Deceuninck, P., Dorne, J. L. C., Tarazona, J. V., Le Goff, F., Leinala, E., Morath, S., Munn, S., Richardson, J., Paini, A., & Wittwehr, C. (2023). OECD harmonised template 201: Structuring and reporting mechanistic information to foster the integration of new approach methodologies for hazard and risk assessment of chemicals. Regulatory Toxicology and Pharmacology, 142(May), 105426. https://doi.org/10.1016/j.yrtph.2023.105426

Chang X, Tan Y-M, Allen DG, Bell S, Brown PC, Browning L, Ceger P, Gearhart J, Hakkinen PJ, Kabadi SV, et al. (2022). IVIVE: Facilitating the Use of In Vitro Toxicity Data in Risk Assessment and Decision Making. Toxics.10(5), 232. https://doi.org/10.3390/toxics10050232

Cooper, G. S. et al. (2016). Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures. Environment International, 92-93, pp. 605-610. https://doi.org/10.1016/j.envint.2016.03.017

Comero, S., Dalla Costa, S., Cusinato, A., Korytar, P., Kephalopoulos, S., Bopp, S., & Gawlik, B. M. (2020). A conceptual data quality framework for IPCHEM – The European Commission Information Platform for chemical monitoring. TrAC - Trends in Analytical Chemistry, 127, 115879. https://doi.org/10.1016/j.trac.2020.115879

Crissman, J. W., Goodman, D. G., Hildebrandt, P. K., Maronpot, R. R., Prater, D. A., Riley, J. H., Seaman, W. J., & Thake, D. C. (2004). Best Practices Guideline: Toxicologic Histopathology. In Toxicologic Pathology. https://doi.org/10.1080/01926230490268756

Cronin, M. T. D., Belfield, S. J., Briggs, K. A., Enoch, S. J., Firman, J. W., Frericks, M., Garrard, C., Maccallum, P. H., Madden, J. C., Pastor, M., Sanz, F., Soininen, I., & Sousoni, D. (2023). Making in silico predictive models for toxicology FAIR. Regulatory Toxicology and Pharmacology, 140(April), 105385. https://doi.org/10.1016/j.yrtph.2023.105385

Cronin, M. T. D., Richarz, A. N., & Schultz, T. W. (2019). Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction. Regulatory Toxicology and Pharmacology. https://doi.org/10.1016/j.yrtph.2019.04.007

Dent, M. P., Vaillancourt, E., Thomas, R. S., Carmichael, P. L., Ouedraogo, G., Kojima, H., Barroso, J., Ansell, J., Barton-Maclaren, T. S., Bennekou, S. H., Boekelheide, K., Ezendam, J., Field, J., Fitzpatrick, S., Hatao, M., Kreiling, R., Lorencini, M., Mahony, C., Montemayor, B., Yang, C. (2021). Paving the way for application of next generation risk assessment to safety decision-making for cosmetic ingredients. Regulatory Toxicology and Pharmacology, 125, 105026. https://doi.org/10.1016/j.yrtph.2021.105026

Digitale, J. C., Martin, J. N., & Glymour, M. M. (2022). Tutorial on directed acyclic graphs. Journal of Clinical Epidemiology, 142, 264–267. https://doi.org/10.1016/j.jclinepi.2021.08.001

Du Pasquier, D., Salinier, B., Coady, K. K., Jones, A., Körner, O., LaRocca, J., Lemkine, G., Robin-Duchesne, B., Weltje, L., Wheeler, J. R., & Lagadic, L. (2024). How the Xenopus eleutheroembryonic thyroid assay compares to the amphibian metamorphosis assay for detecting thyroid active chemicals. Regulatory Toxicology and Pharmacology, 149(February). https://doi.org/10.1016/j.yrtph.2024.105619

ECHA. (2011). Guidance on information requirements and chemical safety assessment. European Chemicals Agency, version 2.1, 1–23. https://www.echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment

ECHA. (2016) Practical Guide. How to use and report (Q)SARs. European Chemicals Agency. https://echa.europa.eu/documents/10162/17250/pg_report_qsars_en.pdf/407dff11-aa4a-4eef-a1ce-9300f8460099

ECHA. (2023). Guidance for identification and naming of substances under REACH and CLP. European Chemicals Agency. https://echa.europa.eu/documents/10162/2324906/substance_id_en.pdf/ee696bad-49f6-4fec-b8b7-2c3706113c7d

ECHA. (2024). Key areas of regulatory challenge. European Chemicals Agency. https://data.europa.eu/doi/10.2823/858284

Echevarriá, L., Malerba, A., & Arechavala-Gomeza, V. (2021). Researcher's Perceptions on Publishing "negative" Results and Open Access. Nucleic Acid Therapeutics, 31(3), 185–189. https://doi.org/10.1089/nat.2020.0865

EFSA/ECHA. (2018). Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009. EFSA Journal, 16(6), 1–135. https://doi.org/10.2903/j.efsa.2018.5311

EFSA. (2010). Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal, 8(6). https://doi.org/10.2903/j.efsa.2010.1637

EFSA. (2014a). Conclusion on the peer review of the pesticide human health risk assessment of the active substance chlorpyrifos. EFSA Journal, 12(4), 1–34. https://doi.org/10.2903/j.efsa.2014.3640

EFSA. (2014b). Conclusion on the peer review of the pesticide risk assessment for aquatic organisms for the active substance imidacloprid. EFSA Journal, 11(1), 1–49. https://doi.org/10.2903/j.efsa.2013.3066

EFSA. (2014c). Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal, 12(6). https://doi.org/10.2903/j.efsa.2014.3734

EFSA. (2015). Scientific Opinion on the safety of caffeine. EFSA Journal, 13(5 https://doi.org/10.2903/j.efsa.2015.4102

EFSA. (2017a). Bisphenol A (BPA) hazard assessment protocol. EFSA Supporting Publications, 14(12), 1354E. https://doi.org/10.2903/sp.efsa.2017.en-1354

EFSA. (2017b). Guidance on the assessment of the biological relevance of data in scientific assessments. EFSA Journal, 15(8), e04970. https://doi.org/10.2903/j.efsa.2017.4970

EFSA. (2017c). Guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal, 15(8), e04971. https://doi.org/10.2903/j.efsa.2017.4971

EFSA. (2017d). Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms. EFSA Journal, 15(2), e04690. https://doi.org/10.2903/j.efsa.2017.4690

EFSA. (2019). Guidance on harmonised methodologies for human health, animal health and ecological risk assessment of combined exposure to multiple chemicals. EFSA Journal, 17(3), e05634. https://doi.org/10.2903/j.efsa.2019.5634

EFSA. (2020). Opinion on the re-evaluation of starch sodium octenyl succinate (E 1450) as a food

additive in foods for infants below 16 weeks of age and the follow-up of its re-evaluation as a food additive for uses in foods for all population groups. EFSA Journal, 18(8), e05874. https://doi.org/10.2903/j.efsa.2020.5874

EFSA. (2022). Tolerable upper intake level for dietary sugars. EFSA Journal, 20(2), e07074. https://doi.org/10.2903/j.efsa.2022.7074

EFSA (2023a). Harmonised approach for reporting reliability and relevance of genotoxicity studies, EFSA Supporting Publications, 20(9). https://doi.org/10.2903/sp.efsa.2023.en-8270

EFSA. (2023b). Re-evaluation of the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs. EFSA Journal, 21(4), e06857. https://doi.org/10.2903/j.efsa.2023.6857

EFSA. (2023c). Re-evaluation of erythritol (E 968) as a food additive. EFSA Journal, 21(12), 1–95. https://doi.org/10.2903/j.efsa.2023.8430

EFSA. (2023d). Guidance on protocol development for EFSA generic scientific assessments. In EFSA Journal 21(10), e08312. https://doi.org/10.2903/j.efsa.2023.8312

EFSA. (2024). Scientific Committee guidance on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments. EFSA Journal, 22(7), e8866. https://doi.org/10.2903/j.efsa.2024.8866

European Commission. (2017). H2020 Programme Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. European Commission, Directorat (21 March), 1–10. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

European Commission (2023). Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs and Cronin, M., Report of the European Commission workshop on “The Roadmap Towards Phasing Out Animal Testing for Chemical Safety Assessments” – Brussels, Publications Office of the European Union, 2024, https://data.europa.eu/doi/10.2873/34576

European Commission (2024). Joint Research Centre, Batista Leite, S., Brooke, M., Carusi, A., Collings, A. et al., Promoting reusable and open methods and protocols (PRO-MaP) – Recommendations to improve methodological clarity in life sciences publications, Publications Office of the European Union, 2024, https://data.europa.eu/doi/10.2760/46124

Ewald, H., Klerings, I., Wagner, G., Heise, T. L., Stratil, J. M., Lhachimi, S. K., Hemkens, L. G., Gartlehner, G., Armijo-Olivo, S., & Nussbaumer-Streit, B. (2022). Searching two or more databases decreased the risk of missing relevant studies: a metaresearch study. Journal of Clinical Epidemiology, 149, 154–164. https://doi.org/10.1016/j.jclinepi.2022.05.022

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. Scientometrics, 90(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Govarts, E., Gilles, L., Rodriguez Martin, L., Santonen, T., Apel, P., Alvito, P., Anastasi, E., Andersen, H. R., Andersson, A. M., Andryskova, L., Antignac, J. P., Appenzeller, B., Barbone, F., Barnett-Itzhaki, Z., Barouki, R., Berman, T., Bil, W., Borges, T., Buekers, J., Schoeters, G. (2023). Harmonized human biomonitoring in European children, teenagers and adults: EU-wide exposure data of 11 chemical substance groups from the HBM4EU Aligned Studies (2014–2021). International Journal of Hygiene and Environmental Health, 249. https://doi.org/10.1016/j.ijheh.2023.114119

Guyatt, G. H., Oxman, A. D., Kunz, R., Atkins, D., Brozek, J., Vist, G., Alderson, P., Glasziou, P., Falck-Ytter, Y., & Schünemann, H. J. (2011). GRADE guidelines: 2. Framing the question and deciding on important outcomes. Journal of Clinical Epidemiology, 64(4), 395–400. https://doi.org/10.1016/j.jclinepi.2010.09.012

Health Canada. (2018). Weight of Evidence: General Principles and Current Applications at Health Canada. Prepared for: Task Force on Scientific Risk Assessment; by: Weight of Evidence Working

Group. Canada.Ca. https://www.canada.ca/en/health-canada/services/publications/science-research-data/weight-evidence-general-principles-current-applications.html

Herbet, M. E., Leonard, J., Santangelo, M. G., & Albaret, L. (2022). Dissimulate or disseminate? A survey on the fate of negative results. Learned Publishing, 35(1), 16–29. https://doi.org/10.1002/leap.1438

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, W. V. (editors). (2023). Cochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023) Available from www.cochrane-handbook.org. The Cochrane Collaboration.

Hill, A. (1965). President' s Address The Environment and Disease: Association or causation? Proc R Soc Med, 58, 295–300.

Hlisníková, H., & Beronius, A. (2024), P20-02 Developing the SciRAPepi tool for assessment of reliability and relevance of observational epidemiological studies, Toxicology Letters, 399, S293. https://doi.org/10.1016/j.toxlet.2024.07.705

IARC. (2019). List of Classifications – IARC Monographs on the Identification of Carcinogenic Hazards to Humans. January. https://monographs.iarc.fr/list-of-classifications

ICMJE. (2024). Recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals. 22(10), 781–791. https://doi.org/10.1016/j.cnre.2016.01.001

Jeliazkova, N. et al., 2021. Towards FAIR nanosafety data. Nature Nanotechnology, 16(6), pp. 644-654. https://doi.org/10.1038/s41565-021-00911-6

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and Challenges of Big Data Research. Big Data Research, 2(2), 59–64. https://doi.org/10.1016/j.bdr.2015.01.006

Kase, R., Korkaric, M., Werner, I., & Ågerstrand, M. (2016). Criteria for Reporting and Evaluating ecotoxicity Data (CRED): comparison and perception of the Klimisch and CRED methods for evaluating reliability and relevance of ecotoxicity studies. Environmental Sciences Europe, 28(1), 1–14. https://doi.org/10.1186/s12302-016-0073-x

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. Chemical Research in Toxicology, 30(4), 946–964. https://doi.org/10.1021/acs.chemrestox.6b00347

Klimisch, H. J., Andreae, M., & Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regulatory Toxicology and Pharmacology. https://doi.org/10.1006/rtph.1996.1076

Köhler, H. R., Gräff, T., Schweizer, M., Blumhardt, J., Burkhardt, J., Ehmann, L., Hebel, J., Heid, C., Kundy, L., Kuttler, J., Malusova, M., Moroff, F. M., Schlösinger, A. F., Schulze-Berge, P., Panagopoulou, E. I., Damalas, D. E., Thomaidis, N. S., Triebskorn, R., Maletzki, D., von der Ohe, P. C. (2023). LogD-based modelling and ΔlogD as a proxy for pH-dependent action of ionizable chemicals reveal the relevance of both neutral and ionic species for fish embryotoxicity and possess great potential for practical application in the regulation of chemicals. Water Research, 235(February). https://doi.org/10.1016/j.watres.2023.119864

Kon Kam King, G., Veber, P., Charles, S., & Delignette-Muller, M. L. (2014). MOSAIC_SSD: A new web tool for species sensitivity distribution to include censored data by maximum likelihood. Environmental Toxicology and Chemistry, 33(9), 2133–2139. https://doi.org/10.1002/etc.2644

Krauth, D., Woodruff, T. J., & Bero, L. (2013). Instruments for assessing risk of bias and other methodological criteria of published animal studies: A systematic review. In Environmental Health Perspectives. https://doi.org/10.1289/ehp.1206389

Krebs, A., Waldmann, T., Wilks, M. F., van Vugt-Lussenburg, B. M. A., van der Burg, B., Terron, A., Steger-Hartmann, T., Ruegg, J., Rovida, C., Pedersen, E., Pallocca, G., Luijten, M., Leite, S. B.,

Kustermann, S., Kamp, H., Hoeng, J., Hewitt, P., Herzler, M., Hengstler, J. G., Leist, M. (2019). Template for the description of cell-based toxicological test methods to allow evaluation and regulatory use of the data. Altex, 36(4), 682–699. https://doi.org/10.14573/altex.1909271

Lawlor, D. A., Tilling, K., & Smith, G. D. (2016). Triangulation in aetiological epidemiology. International Journal of Epidemiology, 45(6), 1866–1886. https://doi.org/10.1093/ije/dyw314

Martens, M., Stierum, R., Schymanski, E. L., Evelo, C. T., Aalizadeh, R., Aladjov, H., Arturi, K., Audouze, K., Babica, P., Berka, K., Bessems, J., Blaha, L., Bolton, E. E., Cases, M., Damalas, D., Dave, K., Dilger, M., Exner, T., Geerke, D. P., Willighagen, E. L. (2021). ELIXIR and Toxicology: a community in development. F1000Research, 10, 1–24. https://doi.org/10.12688/f1000research.74502.2

Marx-Stoelting, P., Rivière, · G, Luijten, · M, Aiello-Holden, · K, Bandow, · N, Baken, · K, Cañas, · A, Castano, · A, Denys, · S, Fillol, · C, Herzler, · M, Iavicoli, · I, Karakitsios, · S, Klanova, · J, Kolossa-Gehring, · M, Koutsodimou, · A, Lobo Vicente, · J, Lynch, · I, Namorado, · S, Sanders, · P. (2023). A walk in the PARC: developing and implementing 21st century chemical risk assessment in Europe. Archives of Toxicology, 97, 893–908. https://doi.org/10.1007/s00204-022-03435-7

Mellor, D., Corker, K., & Whaley, P. (2024), Preregistration templates as a new addition to the evidence-based toxicology toolbox, Evidence-Based Toxicology, 2(1). https://doi.org/10.1080/2833373x.2024.2314303

Merrington, G., Nowell, L. H., & Peck, C. (2024). An introduction to Criteria for Reporting and Evaluating Exposure Datasets (CREED) for use in environmental assessments. Integrated Environmental Assessment and Management, 00(00), 1–6. https://doi.org/10.1002/ieam.4899

Moermond, C., Beasley, A., Breton, R., Junghans, M., Laskowski, R., Solomon, K., & Zahner, H. (2017). Assessing the reliability of ecotoxicological studies: An overview of current needs and approaches. Integrated Environmental Assessment and Management, 13(4), 640–651. https://doi.org/10.1002/ieam.1870

Moermond, C. T. A., Kase, R., Korkaric, M., & Ågerstrand, M. (2016). CRED: Criteria for reporting and evaluating ecotoxicity data. Environmental Toxicology and Chemistry, 35(5), 1297–1309. https://doi.org/10.1002/etc.3259

Myatt, G. J., Bassan, A., Bower, D., Johnson, C., Miller, S., Pavan, M., & Cross, K. P. (2022). Implementation of in silico toxicology protocols within a visual and interactive hazard assessment platform. Computational Toxicology, 21(October 2021), 100201. https://doi.org/10.1016/j.comtox.2021.100201

National Academies of Sciences, Engineering, and Medicine. (2022). Review of U.S. EPA's ORD Staff Handbook for Developing IRIS Assessments: 2020 Version (2022). https://doi.org/10.17226/26289

NIEHS. (2019). Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. National Toxicology Program, 1–98. https://ntp.niehs.nih.gov/whatwestudy/assessments/noncancer/handbook/index.html

NTP. (2025). NTP Report on Carcinogens Handbook on Methods for Conducting Cancer Hazard Evaluations. National Toxicology Program. https://doi.org/10.22427/NTP-OTHER-1008

OECD. (2005). Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment, OECD Series on Testing and Assessment, No. 34, OECD Publishing, Paris. https://doi.org/10.1787/e1f1244b-en

OECD. (2010). Guidance Document for the Diagnosis of Endocrine-related Histopatology of Fish Gonads. OECD Series on Testing and Assessment, No. 123, OECD Publishing, Paris. https://doi.org/10.1787/8f7cf3b5-en

OECD. (2015). Guidance on the GLP Requirements for Peer Review of Histopatology. OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring, No. 16, OECD Publishing, Paris. https://doi.org/10.1787/9789264228306-en

OECD. (2017a). Guidance Document for Describing Non-Guideline In Vitro Test Methods, OECD Series on Testing and Assessment, No. 211, OECD Publishing, Paris. https://doi.org/10.1787/9789264274730-en

OECD. (2017b). Guidance Document for the Use of Adverse Outcome Pathways in Developing Integrated Approaches to Testing and Assessment (IATA), OECD Series on Testing and Assessment, No. 260, OECD Publishing, Paris. https://doi.org/10.1787/44bb06c1-en

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

OECD. (2018b). Recommendation of the Council Concerning Access to Research Data from Public Funding. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347

OECD (2019a), Guiding Principles for Measurements and Reporting for Nanomaterials: Physical Chemical Parameters, OECD Series on the Safety of Manufactured Nanomaterials and other Advanced Materials, OECD Publishing, Paris. https://doi.org/10.1787/70c84148-en

OECD. (2019b). Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures, OECD Series on Testing and Assessment, OECD Publishing, Paris, https://doi.org/10.1787/0ed2f88e-en

OECD. (2019c). Guiding Principles and Key Elements for Establishing a Weight of Evidence for Chemical Assessment, OECD Series on Testing and Assessment, No. 311, OECD Publishing, Paris. https://doi.org/10.1787/f11597f6-en

OECD. (2020a). Enhanced Access to Publicly Funded Data for Science, Technology and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/947717bc-en

OECD. (2020b). Overview of Concepts and Available Guidance related to Integrated Approaches to Testing and Assessment (IATA), OECD Series on Testing and Assessment, No. 329, OECD Publishing, Paris. https://doi.org/10.1787/cd920ca4-en

OECD. (2021). Guidance Document on the Characterisation, Validation and Reporting of Physiologically Based Kinetic (PBK) Models for Regulatory Purposes, OECD Series on Testing and Assessment, No. 331, OECD Publishing, Paris. https://doi.org/10.1787/d0de241f-en

OECD. (2023a). Initial Recommendations on Evaluation of Data from the Developmental Neurotoxicity (DNT) In-Vitro Testing Battery, OECD Series on Testing and Assessment, No. 377, OECD Publishing, Paris. https://doi.org/10.1787/91964ef3-en

OECD. (2023b). OECD Omics Reporting Framework (OORF): Guidance on Reporting Elements for the Regulatory Use of Omics data from Laboratory-based Toxicology Studies, OECD Series on Testing and Assessment, No. 390, OECD Publishing, Paris. https://doi.org/10.1787/6bb2e6ce-en

OECD. (2024). (Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions, Second Edition, OECD Series on Testing and Assessment, No. 405, OECD Publishing, Paris. https://doi.org/10.1787/bbdac345-en

OECD. (2025). Customisation Opportunities of IUCLID for the Management of Chemical Data – 4th edition, OECD Series on Testing and Assessment, OECD Publishing, Paris, https://doi.org/10.1787/d8db13f7-en

Olker, J. H., Elonen, C. M., Pilli, A., Anderson, A., Kinziger, B., Erickson, S., Skopinski, M., Pomplun, A., LaLone, C. A., Russom, C. L., & Hoff, D. (2022). The ECOTOXicology Knowledgebase: A Curated Database of Ecologically Relevant Toxicity Tests to Support Environmental Research and Risk Assessment. *Environmental Toxicology and Chemistry*. https://doi.org/10.1002/etc.5324

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A.,

Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *In The BMJ*. https://doi.org/10.1136/bmj.n71

Percie Du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., MacLeod, M., Wuerbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *BMJ Open Science*, 4(1), 1–7. https://doi.org/10.1136/bmjos-2020-100115

Proença, S., Escher, B.I., Fischer, F.C., Fisher, C.P., Grégoire, S., Hewitt, N.J., Nicol, B., Paini, A., & Kramer, N.I. (2021). Effective exposure of chemicals in in vitro cell systems: A review of chemical distribution models. *Toxicology in vitro*, 105133. https://doi.org/10.1016/j.tiv.2021.105133

Radke, E. G., Wright, J. M., Christensen, K., Lin, C. J., Goldstone, A. E., Lemeris, C., & Thayer, K. A. (2022). Epidemiology Evidence for Health Effects of 150 per-and Polyfluoroalkyl Substances: A Systematic Evidence Map. *Environmental Health Perspectives*, 130(9), 1–10. https://doi.org/10.1289/EHP11185

Richard, A. M., Huang, R., Waidyanatha, S., Shinn, P., Collins, B. J., Thillainadarajah, I., Grulke, C. M., Williams, A. J., Lougee, R. R., Judson, R. S., Houck, K. A., Shobair, M., Yang, C., Rathman, J. F., Yasgar, A., Fitzpatrick, S. C., Simeonov, A., Thomas, R. S., Crofton, K. M., Tice, R. R. (2021). The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chemical Research in Toxicology*, 34(2), 189–216. https://doi.org/10.1021/acs.chemrestox.0c00264

Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M. T., Wambaugh, J. F., Knudsen, T. B., Kancherla, J., Mansouri, K., Patlewicz, G., Williams, A. J., Little, S. B., Crofton, K. M., & Thomas, R. S. (2016). ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chemical Research in Toxicology*, 29(8), 1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135

Roth, N., Zilliacus, J., & Beronius, A. (2021). Development of the SciRAP Approach for Evaluating the Reliability and Relevance of in vitro Toxicity Data. *Frontiers in Toxicology*, 3(October), 1–13. https://doi.org/10.3389/ftox.2021.746430

Rudén, C., Adams, J., Ågerstrand, M., Brock, T. C., Poulsen, V., Schleka, C. E., Wheeler, J. R., & Henry, T. R. (2017). Assessing the relevance of ecotoxicological studies for regulatory decision making. *Integrated Environmental Assessment and Management*, 13(4), 652–663. https://doi.org/10.1002/ieam.1846

Sakuratani, Y., Horie, M., & Leinala, E. (2018). Integrated Approaches to Testing and Assessment: OECD Activities on the Development and Use of Adverse Outcome Pathways and Case Studies, *Basic & Clinical Pharmacology & Toxicology*, 123(S5), 20-28. https://doi.org/10.1111/bcpt.12955

SCHEER. (2018). Memorandum on weight of evidence and uncertianties. *In Gender and Development* (Vol. 120, Issue 1). https://doi.org/10.2875/386011

Schneider, K., Schwarz, M., Burkholder, I., Kopp-Schneider, A., Edler, L., Kinsner-Ovaskainen, A., Hartung, T., & Hoffmann, S. (2009). "ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicology Letters*, 189(2), 138–144. https://doi.org/10.1016/j.toxlet.2009.05.013

Schumann, K., Guenther, A., Jewgenow, K., & Trillmich, F. (2014). Animal Housing and Welfare: Effects of Housing Conditions on Body Weight and Cortisol in a Medium-Sized Rodent (Cavia aperea). *Journal of Applied Animal Welfare Science*, 17(2), 111–124. https://doi.org/10.1080/10888705.2014.884407

Shao, G., Beronius, A., Nymark, P. (2023). SciRAPnano: a pragmatic and harmonized approach for quality evaluation of in vitro toxicity data to support risk assessment of nanomaterials. *Front Toxicol*. 5:1319985. https://doi.org/10.3389/ftox.2023.1319985

Shamliyan, T., Kane, R. L., & Dickinson, S. (2010). A systematic review of tools used to assess the

quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2010.04.014

Shirke, A. V., Radke, E. G., Lin, C., Blain, R., Vetter, N., Lemeris, C., Hartman, P., Hubbard, H., Angrish, M., Arzuaga, X., Congleton, J., Davis, A., Dishaw, L. V., Jones, R., Judson, R., Kaiser, J. P., Kraft, A., Lizarraga, L., Noyes, P. D., Carlson, L. M. (2024). Expanded Systematic Evidence Map for Hundreds of Per-and Polyfluoroalkyl Substances (PFAS) and Comprehensive PFAS Human Health Dashboard. *Environmental Health Perspectives*, 132(2), 1–24. https://doi.org/10.1289/EHP13423

Sim, I., & Detmer, D. E. (2005). Beyond trial registration: A global trial bank for clinical trial reporting. *PLoS Medicine*, 2(11), 1090–1092. https://doi.org/10.1371/journal.pmed.0020365

Smith A.J., Clutton RE, Lilley E, Hansen KEA, Brattelid T. (2018). PREPARE: guidelines for planning animal research and testing. *Laboratory Animals*. 2018;52(2):135-141. https://doi.org/10.1177/0023677217724823

Steenland, K., Schubauer-Berigan, M. K., Vermeulen, R., Lunn, R. M., Straif, K., Zahm, S., Stewart, P., Arroyave, W. D., Mehta, S. S., & Pearce, N. (2020). Risk of bias assessments and evidence syntheses for observational epidemiologic studies of environmental and occupational exposures: Strengths and limitations. *Environmental Health Perspectives*, 128(9), 1–10. https://doi.org/10.1289/EHP6980

Sund, J., Deceuninck, P. (2021). EURL ECVAM library of reference chemicals. European Commission, Joint Research Centre (JRC) PID: http://data.europa.eu/89h/92614229-d020-4d96-941c-c9604e525c9e

Sung JH. (2021). Multi-organ-on-a-chip for pharmacokinetics and toxicokinetic study of drugs. *Expert Opin Drug Metab Toxicol.*, 17(8):969-986. https://doi.org/10.1080/17425255.2021.1908996

Svendsen, C., Whaley, P., Vist, G. E., Husøy, T., Beronius, A., Di Consiglio, E., Druwe, I., Hartung, T., Hatzi, V. I., Hoffmann, S., Hooijmans, C. R., Machera, K., Robinson, J. F., Roggen, E., Rooney, A. A., Roth, N., Spilioti, E., Spyropoulou, A., Tcheremenskaia, O., Mathisen, G. H. (2023). Protocol for designing INVITES-IN, a tool for assessing the internal validity of in vitro studies. *Evidence-Based Toxicology*, 1(1). https://doi.org/10.1080/2833373x.2023.2232415

Swan, A., & Brown, S. (2008). To share or not to share: Publication and quality assurance of research data outputs. Report Commissioned by the Research Information Network (RIN). Annex Detailed Findings for the Eight Research Areas (June 2008), June, 56. https://eprints.soton.ac.uk/266742/

Thayer, K. A., Angrish, M., Arzuaga, X., Carlson, L. M., Davis, A., Dishaw, L., Druwe, I., Gibbons, C., Glenn, B., Jones, R., Phillip Kaiser, J., Keshava, C., Keshava, N., Kraft, A., Lizarraga, L., Persad, A., Radke, E. G., Rice, G., Schulz, B., Vetter, N. (2022). Systematic evidence map (SEM) template: Report format and methods used for the US EPA Integrated Risk Information System (IRIS) program, Provisional Peer Reviewed Toxicity Value (PPRTV) program, and other “fit for purpose” literature-based human health. *Environment International*. https://doi.org/10.1016/j.envint.2022.107468

Thayer, K. A., Shaffer, R. M., Angrish, M., Arzuaga, X., Carlson, L. M., Davis, A., Dishaw, L., Druwe, I., Gibbons, C., Glenn, B., Jones, R., Kaiser, J. P., Keshava, C., Keshava, N., Kraft, A., Lizarraga, L., Markey, K., Persad, A., Radke, E. G., Yost, E. (2022). Use of systematic evidence maps within the US environmental protection agency (EPA) integrated risk information system (IRIS) program: Advancements to date and looking ahead. *Environment International*, 169(August), 107363. https://doi.org/10.1016/j.envint.2022.107363

Tran, L., Tam, D. N. H., Elshafay, A., Dang, T., Hirayama, K., & Huy, N. T. (2021). Quality assessment tools used in systematic reviews of in vitro studies: A systematic review. *BMC Medical Research Methodology*. https://doi.org/10.1186/s12874-021-01295-w

US EPA. (2018). Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the TSCA Program. https://www.epa.gov/sites/default/files/2018-06/documents/epa_alt_strat_plan_6-20-18_clean_final.pdf

US EPA. (2022). ORD Staff Handbook for Developing IRIS Assessments. https://iris.epa.gov/Document/&deid%3D356370

US EPA. (2022). Technical Fact Sheet: Drinking Water Health Advisories for Four PFAS (PFOA, PFOS, GenX chemicals, and PFBS). Drinking Water Health Advisories, June 2022, 1–7. https://www.epa.gov/system/files/documents/2022-06/technical-factsheet-four-PFAS.pdf%0Ahttps://www.epa.gov/ground-water-and-drinking-water/drinking-water-health-advisories-pfoa-and-pfos

Verwer, C. M., van der Ven, L. T. M., Bos, R. van den, & Hendriksen, C. F. M. (2007). Effects of housing condition on experimental outcome in a reproduction toxicity study. *Regulatory Toxicology and Pharmacology*, 48(2), 184–193. https://doi.org/10.1016/j.yrtph.2007.03.004

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbrouckef, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. *Bulletin of the World Health Organization*, 85(11), 867–872. https://doi.org/10.2471/BLT.07.045120

Waspe, J., Bui, T., Dishaw, L., Kraft, A., Luke, A., & Beronius, A. (2021). Evaluating reliability and risk of bias of in vivo animal data for risk assessment of chemicals – Exploring the use of the SciRAP tool in a systematic review context. *Environment International*, 146(August 2020), 106103. https://doi.org/10.1016/j.envint.2020.106103

Weintraub, P. G. (2016). The importance of publishing negative results. *Journal of Insect Science*, 16(1), 1–2. https://doi.org/10.1093/jisesa/iew092

WHO, (2011). Principles and Methods for the Risk Assessment of Chemicals in Food. *International Journal of Environmental Studies*, 68(2), 251–252. https://doi.org/10.1080/00207233.2010.549617

WHO, (2021a). Framework for the use of systematic review in chemical risk assessment. https://www.who.int/publications/i/item/9789240034488

WHO, (2021b). World Health Organization Human Health Risk Assessment Toolkit: Chemical Hazards. (Issue 8). https://www.who.int/publications/i/item/9789240035720

Wilkins, A. A., Whaley, P., Persad, A. S., Druwe, I. L., Lee, J. S., Taylor, M. M., Shapiro, A. J., Blanton Southard, N., Lemeris, C., & Thayer, K. A. (2022). Assessing author willingness to enter study information into structured data templates as part of the manuscript submission process: A pilot study. Heliyon, 8(3), e09095. https://doi.org/10.1016/j.heliyon.2022.e09095

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.,Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 1–9. https://doi.org/10.1038/sdata.2016.18

Yang Y., Chen Y., Wang L., Xu S., Fang G., Guo X., Chen Z., Gu Z. (2022). PBPK Modeling on Organs-on-Chips: An Overview of Recent Advancements. *Front Bioeng Biotechnol*. 14(10),900481. https://doi.org/10.3389/fbioe.2022.900481

Annex A. Available resources supporting the design, conduct, and report of specific types of research data

The list of selected resources provides researchers with some relevant references for consideration. Records providing both reporting standards and good practice/evaluation tools are listed in the column reflecting their main aim.

The list of reporting standards is not exhaustive, nor is it necessarily reflective of national/international regulatory endorsement. Especially for specific scientific domains, substance types and technologies more detailed standards and guidance exist. In several cases, general guidance presented in this Guidance and in this Annex includes links to more specific resources.

Table A A.1. Available resources supporting the design, conduct, and report of specific types of research data

Evidence type	Resources to promote good practice	
Data type	**Reporting standards**	**Methodological quality**
Human data (epidemiology, clinical)	- OHTs 79-83 - Reporting guidelines for randomised trials (CONSORT) - Reporting guidelines for observational studies (STROBE) - SciRAP Reporting checklist for epidemiological data including cross-sectional, case-control, nested case-control, and cohort studies (SciRAP)	- Cochrane RoB tools (RoB-2 for randomised trials, ROBINS-I for non-randomised/observational studies of interventions, ROBINS-E for Risk Of Bias In Non-randomised Studies of Exposure) - NTP-OHAT Risk of Bias Rating Tool for Human and Animal Studies - SciRAP tool for evaluation of epidemiological data covering cross-sectional, case-control, nested case-control, and cohort studies - Biomonitoring, Environmental Epidemiology, and Short-lived Chemicals (BEES-C) instrument for evaluating the quality of research proposals and studies that incorporate biomonitoring data on short-lived chemicals - RoB instrument for non-randomised studies of exposures - Inventories and reviews of critical appraisal tools (EFSA, 2024a, Appendix D)
In vivo ecotoxicology / toxicology	- OHTs 41-54, OHTs 60-(66-2), OHTs 67-(69-2), OHT 71-(75-1), OHT (75-3)-77 and OHT 84 - OECD Harmonised Endpoint Summaries - Guidelines for reporting animal research (ARRIVE) - CRED criteria for reporting and evaluating (aquatic) ecotoxicity studies, and its adaptation for sediments and soil, for nanomaterials and behavioural studies - SciRAP reporting checklists for *in vivo* toxicity studies (and for ecotoxicity studies, based on CRED)	- Norecopa Planning Research and Experimental Procedures on Animals: recommendations for Excellence (PREPARE) - NTP-OHAT Risk of Bias Rating Tool for Human and Animal Studies - SciRAP tool for evaluation of in vivo toxicity data - CRED criteria for reporting and evaluating (aquatic) ecotoxicity studies, and its adaptation for sediments and soil and for nanomaterials and behavioural studies(available on the SciRAP platform)
In vitro	- OHT 66-3, OHT 70, OHT 75-2, OHT 201	- OECD Guidance Document on Good In Vitro Method Practices

Evidence type	Resources to promote good practice	
Data type	**Reporting standards**	**Methodological quality**
ecotoxicity / toxicity	(intermediate effects) - OECD Harmonised Endpoint Summaries - NC3Rs Reporting recommendations for in-vitro experiments (RIVER) - Template to implement GD No 211and GIVIMP guidance (ToxTemp) - SciRAP reporting checklists for in vitro toxicity studies, including a separate checklist for *in vitro* studies on nanomaterials	(GIVIMP) (No 286) - OECD Guidance Document for Describing Non-guideline In Vitro methods (No. 211) - Guidance Document on Good Cell and Tissue Culture Practice 2.0 (GCCP 2.0) (Pamies et al 2022) - SciRAP tool for evaluation of in vitro toxicity data, including a separate tool for the evaluation of *in vitro* studies on nanomaterials - Peer review of *in vitro* studies Appraisal Tool (PRIVAT) - A protocol for designing INVITES-IN, a tool for assessing the internal validity of *in vitro* studies has recently been published (Svendsen et al 2023) - Standards developed by ISO Technical Committee TC 276 Biotechnology. For instance: - ISO 21709:2020(en) Biobanking — Process and quality requirements for establishment, maintenance and characterization of mammalian cell lines; - ISO/TS 23511:2023 (en) - General requirements and considerations for cell line authentication. - Stem cells - ISSCR Guidelines for Stem Cell Research and Clinical Translation - Standards developed by ISO Technical Committee TC 276Biotechnology. For instance: - ISO 24603:2022(en)— Biobanking — Requirements for human and mouse pluripotent stem cells; - Quality standards on human stem cells (Ludwig et al 2023 and Pistollato et al 2022) - Microphysiological systems (MPS) - Recommendations on fit-for-purpose criteria to establish quality management for microphysiological systems and for monitoring their reproducibility (Pamies et al 2024) - Technical framework for enabling high quality measurements in new approach methodologies (NAMs). (Petersen et al 2023)
In silico ecotoxicity / toxicity – (Q)SAR	- OHTs 41-57, OHTs 60-78, OHT 86, OHT 201 (intermediate effects) - OECD Harmonised Endpoint Summaries - OECD (Q)SAR Model Reporting Format (ENV/CBC/MONO(2023)32/ANN1) - OECD (Q)SAR Prediction Reporting Format (ENV/CBC/MONO(2023)32/ANN2)	- (Q)SAR Assessment Framework: Guidance for the Regulatory Assessment of (Quantitative) Structure - Activity Relationship Models, Predictions, and Results Based on Multiple Prediction (ENV/CBC/MONO(2023)32)
Omics	- OECD Omics Reporting Framework (OORF): Guidance Document (ENV/CBC/MONO(2023)41) and associated reporting template	- ISO standards: Biotechnology - Massively parallel sequencing - Part 1: Nucleic acid and library preparation (ISO 20397-1:2022) and Part 2: Quality evaluation of sequencing data (ISO 20397-2:2021); Molecular *in vitro* diagnostic examinations – Specifications for pre-examination processes in metabolomics in urine, venous blood serum and plasma (ISO 23118:2021) - Genomics informatics - Reliability assessment criteria for high-throughput gene-expression data - (ISO/TS 22690:2021) - Genomics informatics – Omics Markup Language (OML) (ISO 21393:2021) - ICH– Guideline on genomic sampling and management of genomic data E18 - Use cases, best practice and reporting standards for metabolomics in regulatory toxicology (Viant et al. 2019)
In silico - toxicokinetic and toxicodynamic modelling	- OHT 58, OHT 59 - OECD Guidance Document on Characterisation, Validation and Reporting of Physiologically Based	- OECD Guidance Document on Characterisation, Validation and Reporting of Physiologically Based Kinetic (PBK) Models for Regulatory Purposes (No 331) - US EPA Quality Assurance Project Plan (QAPP) for PBPK

Evidence type	Resources to promote good practice	
Data type	**Reporting standards**	**Methodological quality**
	Kinetic (PBK) Models for Regulatory Purposes (No 331) - US EPA PBPK model templates	models - EFSA Scientific Opinion on the state of the art of Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms (EFSA 2018) - EFSA Scientific Opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products (EFSA 2014)
(Bio)monitoring	- CREED Template for reporting environmental exposure datasets - Reporting standards defined in the Information Platform for Chemical Monitoring data (IPCHEM)	- Criteria for Reporting and Evaluating environmental Exposure Datasets (CREED) - OECD Occupational Biomonitoring Guidance (No 370) - Quality framework for chemical biomonitoring under the National Health and Nutrition Examination Survey (NHANES) - HBM4EU Quality Assurance/Quality Control (QA/QC) program
Environmental fate and behaviour	- OHTs 24-40, OHT 401 - OECD Harmonised Endpoint Summaries	-
Non target analysis	- NTA Study reporting tool	- Best Practice for Non-target analysis (BP4NTA)

Annex B. Data repositories and software for storing, sharing, searching, and screening research data

The following table is a (non-comprehensive) list of data repositories and software for research and regulatory data. This table provides easy access to resources that were mentioned or related to the topic addressed in this Guidance Document.

Table A B.1. Data repositories and software for storing, sharing, searching, and screening research data

Data repositories and software for research and regulatory data	Description
General purpose scientific data management tools	
Open Science Framework	Open source platform designed to support researchers through the life cycle of research projects, providing data management functionalities
Registry of Research Data Repositories (re3data)	Registry of research data repositories to store and share research data
Dataverse	Open source application to share, store, and analyse research data
Zenodo	Open-access repository for research outputs. It allows researchers to share and preserve their research data, software, publications, presentations, and other digital assets
Databases, platforms, and tools of bibliographic and chemical information	
Web of Science (WoS)	Commercial web platform that offers various features to find and access research publications. WoS is commonly used as a literature database that is searched to identify relevant information
US National Centre for Biotechnology Information (NCBI)	Information platform hosting a large number of biomedical resources including PubMed, PubChem, Gene Expression Omnibus, and many other research database
CAS SciFinder	Information platform produced by the Chemicals Abstracts Service to providing access to a chemical and bibliographic information
CAS STNext	Platform providing access to global databases in the field of chemistry, biomedicine and pharmaceuticals
Europe PMC	Web-based platform providing access to multiple life science bibliographic databases
PubChem	The largest publicly available repository of chemical information. Information includes assay response data and links to other resources
PubMed (Medline)	Publicly available literature repository for research publications within medical and related life sciences fields
Scopus	Bibliographic database of scientific publications from a wide range of scientific disciplines
EMBASE	Bibliographic databases of peer-reviewed scientific journal articles in the field of medical science
ResearchRabbit	AI-tool to help the discovery of relevant scientific literature
Chemical properties databases managed by OECD/ regulatory authorities	
OECD eChemPortal	OECD information platform bringing together collections of chemical hazard and risk information prepared for government chemical programmes at national, regional, and international levels
OECD Existing Chemicals DB	Resource listing all OECD High Production Volume Chemicals together with any annotations provided by Member countries. For assessed chemicals, links to download completed assessments are provided
QSAR Toolbox	Toolbox containing a large collection of chemical properties databases. It also provides computational workflows for grouping chemicals and filling data gaps by read-across
US EPA CompTox	Web application providing information on over 1 million chemicals. Information includes hazard data from

Data repositories and software for research and regulatory data	Description
Chemicals Dashboard (CCD)	*in vitro* and *in vivo* sources, physicochemical data, and links to other resources
US EPA ECOTOX Knowledgebase	Publicly available database of toxicity information on aquatic and terrestrial species
ECHA CHEM	ECHA's public chemical database including data submitted by companies in REACH registrations
EFSA OpenFoodTox	EFSA's database of chemical and toxicological information on chemicals assessed by the agency and included in published scientific opinions
Specialised scientific databases and knowledge platforms	
Adverse Outcome Pathway Knowledgebase (AOP KB)	Platform bringing together all knowledge on how chemicals can induce adverse effects, using the Adverse Outcome Pathways analytical construct and ontologies
Endocrine Active Substances Information System (EASIS)	Database providing information on endocrine active properties of chemical substances
Nanosafety Data Interface and eNanoMapper database system	The Nanosafety Data Interface is a platform providing aggregated data to support the safety assessment of nanomaterials, including data generated by EU funded projects and the US cancer Nontechnology Laboratory portal. The platform implements the structured framework of the eNanoMapper database system
Gene Expression Omnibus	Public repository of high-throughput gene expression and other functional genomics data sets
EMBL's European Bioinformatics Institute (EBI)	Global resource for biological data, providing information on DNA and protein sequences, structures, genomes, gene expression, molecular interactions, and pathways
Norman Database System	Information platform operated by a network of European research organizations providing access to a range of environmental data, focusing on chemicals and their impact on the environment
Information Platform on Chemical Monitoring (IPCHEM)	EU's information platform for searching, accessing, and retrieving chemical (bio)monitoring data collected and managed in Europe
Software application supporting regulatory use of research data	
IUCLID	Software application used by different jurisdictions and regulatory programmes to record, store, maintain and exchange data on the intrinsic and hazard properties of chemical substances or mixtures, as well as the uses of these substances and the associated exposure levels
US EPA Health and Environmental Research Online (HERO)	Citation management web application used by the US EPA. All references used within a chemical assessment are made publicly available through HERO
US EPA Health Assessment Workspace Collaborative (HAWC)	Open-source web application. HAWC has a collection of features to support actions like data extraction and study evaluation with built-in visualizations. Linked is the version US EPA uses for several assessment programmes like Integrated Risk Information System (IRIS)
DistillerSR	Commercial application that supports workflow management and systematic review steps like relevance screening and data extraction.
Rayyan	Software tool designed to assist researchers in conducting systematic literature reviews
SWIFT-Active Screener (SWIFT-AS)	Commercial application that supports relevance screening using reference metadata like titles and abstracts. This application uses an active learning machine-learning method to reduce the total number of references that need to be manually screened
SWIFT-Review	Commercial application to further categorise references identified through literature searches. For example, identify the non-human animal toxicology studies returned in the broader literature search

Annex C. Examples of regulatory contexts where research data has been considered in regulatory assessments

Table A C.1. Examples of regulatory contexts where research data has been considered in regulatory assessments (not exhaustive)

Assessment domain	Assessment tasks	Examples
Collection of evidence and prioritisation	Collection of (eco)toxicological data to fulfil information requirements	• (Eco)toxicity endpoints in EU REACH registration dossiers (e.g., bisphenol A[30])
	Systematic evidence maps to understand availability and summarise evidence on potential health effects	• Literature inventory heat map (evidence map) defining the scope of the evaluation of diethylhexyl phthalate by US EPA under the Toxic Substances Control Act[31] • PFAS systematic evidence maps developed by the US EPA (Case study A)
	Prioritisation of substances for risk assessment and/or management	• Data landscaping as part of a working approach to identify potential candidate chemicals for prioritisation for risk evaluation under the Toxic Substances Control Act in the US EPA Office of Chemical Safety and Pollution Prevention[32]
	Assessment of occurrence	• Section 3.1 of Scientific Opinion on the safety of caffeine[33]

[30] https://chem.echa.europa.eu/100.001.133/dossier-view/8d9de292-990f-403c-82a8-096416da9af0/376807a6-1d87-48c3-ac94-f40c8f167a81_376807a6-1d87-48c3-ac94-f40c8f167a81?searchText=80-05-7

[31] https://www.epa.gov/sites/default/files/2020-09/documents/casrn_117-81-7_di-ethylhexyl_phthalate_final_scope.pdf

[32] https://www.epa.gov/sites/default/files/2018-09/documents/preprioritization_white_paper_9272018.pdf

[33] EFSA NDA Panel, (2015). Scientific Opinion on the safety of caffeine. EFSA Journal 2015; 13(5):4102, 120 pp. https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/j.efsa.2015.4102

Assessment domain	Assessment tasks	Examples
Exposure assessment		• Section 3.3.2 of Scientific Opinion on update of risk assessment of phthalates in food contact materials[34]
	Assessment of non-dietary exposure	• Section 1.6.1.2 of Scientific Opinion on risk to public health related to bisphenol A in foodstuff [35]
	Environmental exposure assessment	• US EPA Risk Evaluation for Cyclic Aliphatic Bromide Cluster (HBCD) under the Toxic Substances Control Act (Section 2.3, Non-scenario Specific Approach)[36]
Hazard identification/ classification of substances	Hazard classification of substances	• Evaluation of environmental hazard of bisphenol A for harmonised classification and labelling under EU CLP [37] • Additional lines of evidence from research data for harmonised classification and labelling (Case study C) • Identification of 4-MBC as substance of very high concern for endocrine disrupting properties in EU REACH[38]
Hazard characterisation, including establishment of	Causality determination on health effects	• Safety assessment of titanium dioxide (E171) as a food additive[39] • Re-evaluation of erythritol (E968) as a food additive[40] • US EPA Integrated Science Assessment for Lead - Causality determinations on health effects related to ambient exposures (Table IS-1)[41]

[34] EFSA CEP Panel, (2019). Scientific Opinion on the update of the risk assessment of di-butylphthalate (DBP), butyl-benzyl-phthalate (BBP), bis(2-ethylhexyl)phthalate (DEHP), di-isononylphthalate (DINP) and di-isodecylphthalate (DIDP) for use in food contact materials. EFSA Journal 2019;17(12):5838, 85 pp. https://doi.org/10.2903/j.efsa.2019.5838

[35] EFSA CEF Panel, (2015). Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs: Executive summary. *EFSA Journal* 2015; 13 (1):3978, 23 pp. https://doi.org/10.2903/j.efsa.2015.3978

[36] Risk Evaluation for Cyclic Bromides Aliphatic Cluster CASRN: 25637-99-4, 3194-55-6, 3194-57-8 (epa.gov)

[37] https://www.echa.europa.eu/documents/10162/b8a9b144-33c0-064f-bedc-39032a59e0dc

[38] https://www.echa.europa.eu/documents/10162/41008a30-53db-84bd-6d4e-7f31d9aa78dc

[39] EFSA FAF Panel, (2021). Scientific Opinion on the safety assessment of titanium dioxide (E171) as a food additive. EFSA Journal 2021;19(5):6585, 130 pp. https://doi.org/10.2903/j.efsa.2021.6585

[40] EFSA FAF Panel, (2023). Re-evaluation of erythritol (E 968) as a food additive. EFSA Journal, 21(12), e8430. https://doi.org/10.2903/j.efsa.2023.8430

[41] https://assessments.epa.gov/isa/document/&deid=359536#downloads

Assessment domain	Assessment tasks	Examples
toxicity or regulatory reference values	Hazard characterisation in safety evaluations for the approval, renewal, restrictions or bans of substances	•Peer review of the scientific literature supporting safety assessments of plant protection products under the EU Plant Protection Products Regulation (Case study D) •Restriction of substances under the EU REACH Regulation. Examples include DecaBDE, Formaldehyde, Lead, 4-Nonylphenol, and PFAS[42] •Section 3.2.4. of Re-evaluation of the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs[43] •US EPA IRIS Toxicological Review of Formaldehyde-Inhalation[44] •US EPA Provisional Peer reviewed Toxicity Values for Methylnaphthalene[45]
	Establishment of environmental quality standards (EQS)	•Draft Environmental Quality Standards for Priority Substances under the Water Framework Directive – Diclofenac[46]
Health and environmental risk assessments	Risk characterisation underpinning approvals or risk management	•US EPA Risk Evaluation for Methylene Chloride (Risk characterisation, Section 4) under the Toxic Substances Control Act [47] •Health risk assessment of aldehydes group by Health Canada[48] •Comparison of environmental risks of pharmaceuticals (pain killers), informing Dutch stakeholders how to reduce the use diclofenac and ibuprofen, priority substances in the EU Water Framework Directive[49]

[42] https://echa.europa.eu/registry-of-restriction-intentions/-/dislist/details/0b0236e18663449b, see also (Borchert et al., 2022)

[43] EFSA CEP Panel, (2023). Scientific Opinion on the re-evaluation of the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs. EFSA Journal 2023; 21(4):6857, 392 pp. https://doi.org/10.2903/j.efsa.2023.6857

[44] US Environmental Protection Agency, Washington, DC, EPA/635/R-22/039, 2022 (External Review Draft, 2022). https://iris.epa.gov/Document/&deid=248150

[45] US Environmental Protection Agency, Washington, DC, EPA/690/R-24/017F, 2024 https://cfpub.epa.gov/ncea/pprtv/recordisplay.cfm?deid=361053

[46] https://health.ec.europa.eu/publications/scheer-scientific-opinion-draft-environmental-quality-standards-priority-substances-under-water-0_en

[47] https://www.epa.gov/sites/default/files/2020-06/documents/1_mecl_risk_evaluation_final.pdf

[48] https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-substances/assessment-aldehydes-group.html#toc5

[49] https://www.rivm.nl/publicaties/risicos-van-pijnstillers-in-het-oppervlaktewater#abstract_en

Annex D. Case studies

Case study A. Reuse of curated analysis of research data: Polyfluoroalkyl Substances (PFAS) Systematic Evidence Maps (SEMs)

Developed by the United States Environmental Protection Agency (US EPA) Office of Research and Development (ORD)
Case study authors and contributors: Kristina Thayer*, Sean Watford*, Laura Carlson, Avanti Shirke, Michelle Angrish (US EPA/ORD). *- Affiliation listed reflects the author's institution at the time this work was conducted.
Participants: French National Agency of Food Safety, Environment and Work- ANSES (Nawel Bemrah, Geraldine Carne, Isabelle Maniere, Aurélie Mathieu), French School of Public Health- EHESP (Pauline Rousseau-Guetin), European Commission (Veerle Vanheusden), European Food Safety Authority- EFSA (Fulvio Barizzone, Chantra Eskes, Maria Anastassiadou), EFSA Panel on Contaminants in the Food Chain- EFSA CONTAM Panel (Ron Hoogenboom – Wageningen University & Research, Christer Hogstrand – King's College London), ICAPO (Scott Belcher, North Carolina State University representing the Endocrine Society), Dutch National Institute for Public Health and the Environment- RIVM (Astrid Bulder), Swedish National Food Agency (Irina Gyllenhammar), Maastricht University (Dick T. H. M. Sijm, Victor Amstutz), US EPA (Jennifer Nichols)

A.1. Background

Systematic evidence maps (SEMs) are increasingly used as a problem formulation tool to refine the focus of scientific issues that are evaluated in subsequent assessments and expedite assessment development (Thayer et al., 2022a). SEMs can be defined as "*A comprehensive summary of the characteristics and availability of evidence as it relates to broader themes of policy or decision-making relevance* (Wolffe et al., 2019). *SEMs do not seek to synthesise evidence but instead to catalogue it, utilising systematic search, selection, and coding strategies to produce searchable databases of studies. These databases are accompanied by descriptive information that helps the reader use and evaluate the evidence map and interpret its contents*."[50]

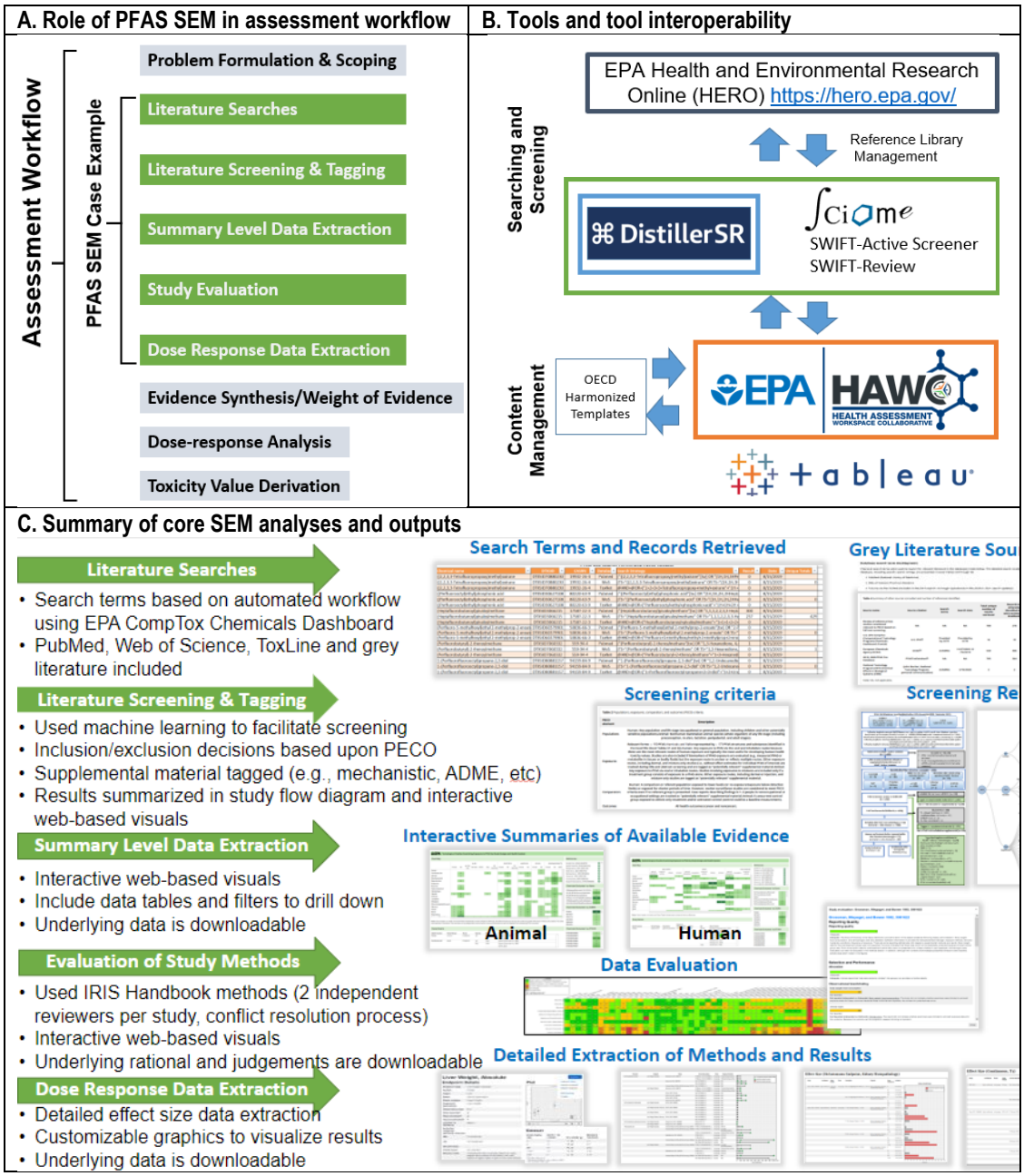
Most studies included in SEMs are considered research data as defined in this OECD guidance document. SEMs have been used within US EPA for various purposes, including to understand data gaps for research prioritisation, determine the need for updated assessments, inform assessment priorities and refine scope, inform development of analysis plans for mechanistic information, catalogue ADME (absorption, distribution, metabolism, elimination) and similar evidence, and inform development of study evaluation considerations. Increased utilisation of SEMs across the environmental health field has the potential to increase transparency and efficiency for data gathering, problem formulation, and evidence surveillance. The US EPA/ORD Health and Environmental Risk Assessment (HERA) National Research Program has been using SEMs to inform and facilitate the development of human health toxicity assessments for environmental chemicals (Thayer et al., 2022a), including for polyfluoroalkyl substances (PFAS) (Carlson et al., 2022; Carlson et al., 2024; Shirke et al., 2024; Radke et al., 2022).The PFAS SEMs include detailed

[50] Environment International Policies and Guidelines, July 20, 2023. https://www.sciencedirect.com/journal/environment-international/about/policies-and-guidelines

descriptions of study methods, study results, and study evaluation (potential bias and sensitivity[51]), but they do not present conclusions on potential human hazard(s) or present toxicity values based on dose-response analysis. A SEM template has also been developed for chemical human health assessments to foster consistency within the HERA portfolio of assessment products and expedite development of SEMs (Thayer et al., 2022b). Template availability can also promote harmonisation in the environmental health community and create more opportunities for sharing extracted content. Figure A.1 shows where the PFAS SEMs integrate into the assessment workflow, the tools (and tool interoperability) used, and a schematic summary of core SEM analyses and outputs.

[51] Potential bias (factors that affect the magnitude or direction of an effect in either direction) and insensitivity (factors that limit the ability of a study to detect a true effect; low sensitivity is a bias toward the null when an effect exists). Additional details available in the Integrated Risk Information System (IRIS) Handbook (US EPA, 2022a)

Figure A.1. Overview of the PFAS SEM case example



Note: Panel A describes the role of PFAS SEM in problem formulation and scoping. Panel B shows the tools that are utilised for literature searching and library management (HERO, https://hero.epa.gov/), literature screening (DistillerSR®, https://www.distillersr.com/; Sciome's SWIFT-Review https://www.sciome.com/swift-review/ and SWIFT-Active Screener https://www.sciome.com/swift-activescreener/), and data extraction/visualisation (US EPA's Health Assessment Workspace Collaborative (HAWC), https://hawc.epa.gov/ or Tableau, https://www.tableau.com/). Panel C summarises core SEM analyses and outputs. Access to the graphics (publications and URLs) in Panel C are presented in Table A.1.

A.2. Main Goals of the Case Study

In the context of this OECD guidance document, the overall goal of this case study is to promote the reuse of analysis of research data conducted by one agency (e.g., identification of studies, relevance and reliability assessment, and data extraction) to support the assessment work of other agencies. SEMs can potentially serve as an evidentiary foundation for conducting assessments of the catalogued information, even when assessments are required for different specific regulatory purposes[52]. The PFAS SEMs serve as an ideal case example since the SEMs make extensive use of Health Assessment Workspace Collaborative (HAWC) (Shapiro et al., 2018)[53], a free and open-source web-based software application designed to manage and facilitate the process of conducting health assessments and provide online access to their associated data and analyses. More specifically, HAWC is a modular, web-based content management system designed to store, display, and synthesise multiple data sources for the purpose of supporting the development of human health and environmental risk assessments of pollutants. Key HAWC modules include screening, study quality evaluation, data extraction (human epidemiology, animal bioassay, and *in vitro*), and evidence synthesis. Data extraction can be downloaded to support dose-response analysis conducted in other platforms. Intended for human health and environmental risk assessors, HAWC allows collaboration within assessment teams comprised of managers, team-members, and reviewers to synthesise this information. HAWC supports systematic review methodology to increase scientific rigour and transparency of chemical assessments by using a predefined, multi-step process to identify, and critically evaluate the underlying evidence. It serves as a repository for study quality decisions and extracted data used to support an assessment and provides interactive visuals of the results both within individual studies and across the entire evidence base.

A.3. Methods

In order to evaluate the potential feasibility of using the existing PFAS SEMs for subsequent analyses as a case study, a risk-assessment related analysis would need to be conducted. Full evaluation of this case study would entail using the SEM content to conduct a risk assessment-related analysis (e.g., problem formulation, hazard characterisation, or risk evaluation). However, no specific analyses were planned by participants during the timeframe of developing this OECD guidance document that would overlap with the PFAS SEMs content (Spring 2023-Spring 2024). Therefore, as feasible alternative conceptual feedback was sought via sharing electronic resources and online meeting discussions to summarise the SEM methods and content (Table A.1). Several meetings with the same agenda were held to accommodate schedules. At these meetings, a slide set was used to overview the SEMs with demos to display the interactive components (Table A.1). Most participants were from European government agencies that have

[52] The existing SEMs do not include studies that may contain confidential business information (CBI) for all the chemicals. This is because the search processes used to explore CBI information are highly manual and do not lend themselves to being applied to hundreds of chemicals at a time. When US EPA's HERA Program uses the SEMs to facilitate conducting a chemical assessment, a targeted search of CBI is conducted at that time.

[53] US EPA's deployment of HAWC was used for the PFAS SEMs, but a freely available deployment for the public with the same features is also available at https://hawcproject.org/. The US EPA supports development of HAWC, which is an MIT-licensed open-source application. US EPA maintains EPA HAWC (https://hawc.epa.gov), which is used as a compendium for US EPA assessments. Many public assessments demonstrating HAWC's capabilities are available on this website. However, the assessment development portion of US EPA HAWC is not available to the wider public. Since the application is open-source, there are other deployments available that allow the public to develop assessments, including https://hawcproject.org/ (not affiliated with US EPA). Mention of or referral to commercial products or services, and/or links to non-US EPA sites does not imply official US EPA endorsement of or responsibility for the opinions, ideas, data, or products presented at those locations, or guarantee the validity of the information provided.

responsibilities for conducting regulatory assessments (or from scientific panels that support these agencies). Soliciting feedback was facilitated with the discussion prompts below. Feedback was summarised through a draft version of this annex, which was shared with participants for review of accuracy and completeness of the discussion.

- Conceptually, is this product type helpful for problem formulation?
- Are the data structured in a way that makes them easy to access and reuse?
- What are barriers to using this structured format?
- This project utilised the US EPA Integrated Risk Information System (IRIS) study evaluation method (US EPA, 2022a). Can this be potentially reused, or would your group need to use a specific study evaluation methodology to match your given need and context?

A.4. Results

Conceptual feedback was sought from case study participants in the form of discussion during webinars, because none of the case study participants were in a position to fully explore usage of the SEMs for problem formulation and assessment analyses during the timeframe of conducting the case studies. Overall, participants expressed a high level of support for the PFAS SEMs. There was an appreciation for the large amount of work involved and transparent organisation of the materials. In principle, the structured information appeared to lend itself to reuse. One participant noted the importance of advertising the availability of these materials as potentially duplicative work may be underway to support EU-based analyses of PFAS (e.g., EFSA CONTAM panel). One potential barrier may be users acclimating to the newness of a digital format (versus paper/narrative format).

With respect to the ability to utilise study evaluations conducted using IRIS methods (US EPA, 2022a) some participants expressed a need to better understand the IRIS methods, but one noted that in principle there is the possibility for reuse since the methods have been reviewed. The IRIS study evaluation methods underwent public comment and peer-review by the National Academy of Science, Engineering, and Medicine (NASEM) before being finalised in the 2022 IRIS Handbook (NASEM, 2022a; US EPA, 2022a). It is worth noting that the HAWC study evaluation module was designed to be flexible. The study evaluation domain items can be customised to accommodate other study evaluation frameworks, including the Office of Health Assessment and Translation (OHAT) risk of bias framework. The format is domain based and does not develop numerical scores because such scoring is discouraged in systematic review. A variety of judgement rating approaches are also available: none, yes/no, a continuum expressed in context of high/low risk of bias, a continuum expressed in context of good/deficient, a continuum expressed in context of high/low confidence, and a continuum expressed in context of minor/critical concerns. Users can also decide whether to develop an overall study evaluation judgment or not. The platform also allows documentation of different judgements in the same study, i.e., for different health endpoints or different exposure characterisation scenarios.

Several participants asked whether the SEM content could be updated. Although the US EPA PFAS SEM project cannot be updated by another organisation or group, the content could be copied from US EPA HAWC (https://hawc.epa.gov/) to a project on public HAWC (https://hawcproject.org/) or any other deployment of HAWC. This would leave the published work by US EPA intact and time-stamped and allow users to conduct the update, potentially drawing from resources at multiple organizations if desired. Copying references and study tags into another project can be done using HAWC's bulk import and bulk tagging features. However, copying other types of content including study evaluations and data extractions can only be done by using the available HAWC application programming interface (API). Using the API requires programming skills that may not be available within some organisations or groups.

US EPA has several examples of using the structured SEM development processes to promote working across organisations or different programmes within US EPA. In 2022, a NASEM report *"Guidance on PFAS Exposure, Testing, and Clinical Follow-Up* (NASEM, 2022b) utilised data on some epidemiologic studies that had been abstracted by US EPA's Office of Water (OW) and Office of Research and Development (ORD) in their literature review. Within US EPA, the data abstraction conducted by OW is being incorporated by ORD as they assemble a consolidated PFAS dashboard that includes the chemicals examined in the SEMs, as well as data abstraction conducted in OW's human health assessment of PFOS and PFOA. The PFAS SEM work conducted by US EPA's ORD is being used by certain programmes within US EPA's Office of Chemical Safety and Pollution Prevention (OCSPP). Outside of PFAS, US EPA has also conducted a joint SEM with the Agency for Toxic Substances and Disease Registry (ATSDR) to get feedback on SEM workflows and identify opportunities to harmonise methods (Smith et al., 2022). Subsequently, ATSDR used the SEM methods to conclude that an update of the Toxicological Profile for Methylene Chloride was not needed because no new studies had been published that would impact existing inhalation or oral minimal risk levels (ATSDR, 2022). These examples demonstrate the possibilities of data re-use across programmes.

In addition to providing feedback on the discussion points, some participants expressed interest in HAWC more generally, potentially for use in their workflows on other topics. While the PFAS SEMs were developed in US EPA's deployment of HAWC where access is only for US EPA users and collaborators, a public version of HAWC with the same functionality is available at https://hawcproject.org/. Because HAWC is open source, it is possible for users to develop and maintain their own versions although maintenance of these versions would not be supported by US EPA staff. US EPA clarified that HAWC is not a vehicle to conducting quantitative analyses, such as dose-response modelling and meta-analysis. This is by design to minimise duplicating functionality of software platforms that already exist and are widely used by research and regulatory communities (e.g., US EPA Benchmark Dose Software). The data in HAWC can be readily downloaded for quantitative analyses conducted outside of HAWC.

A.5. Summary and future directions

Overall, participants expressed a high level of support for the potential feasibility of reuse of published SEMs. It is possible that additional questions and feedback will arise if the PFAS SEMs are used in further risk assessment-related case study analyses. US EPA indicated willingness to provide support for follow-up use of the PFAS SEMs. If the level of effort in providing support becomes challenging and interest seems high, then a "train the trainer" approach can be used where US EPA trains a point of contact within certain organisations or panels.

Moving forward, US EPA's HERA program (US EPA, 2022b) is utilising the structured collection of information assembled in SEMs to support longer term follow-up research endeavours. A major area of focus is to expand HERA's current use of machine-learning (ML)/artificial intelligence (AI) to semi-automate the processes of data labelling (computationally auto-labelling studies), extraction (summarising study methods and results), study evaluation, and data standardisation (ontologies/controlled vocabularies) using cloud services (Beebe et al., 2022; US EPA, 2022a). Automation of full text screening, study evaluation, and data extraction steps, each with user verification (i.e., human-in-the-loop) remain the pinch points in operationalising ML-assisted steps. There is a need to move away from currently used costly and complex infrastructures toward modern data stacks and workflows fit for AI. However, the establishment of automated approaches has languished in part due to a lag in the development of training data needed to develop successful natural language processing models and, to a larger extent, due to a lack of stackable software applications with the flexibility to test new technology and evolve over time.

Semi-automated ML and AI is already being used at US EPA and elsewhere to reduce the cost and time associated with screening studies for inclusion in assessments (Howard et al., 2020; US EPA, 2022a). Several research initiatives within the HERA program at US EPA are focused on expanding use of ML/AI

to other phases of assessment development. One is a consolidated semi-automated workflow to screen and extract data from grey and published literature in support of developing SEMs. The workflow incorporates the use of multiple tools with ML features and a new automated data extraction tool with a human-in-the-loop structure. More specifically, the workflow includes use of Dextr[54], a web-based data extraction tool that provides a user-verification workflow of ML predictions for data entities pertinent to conducting a human health assessment (Walker et al., 2022). US EPA is conducting a pilot study to integrate use of Dextr into the SEM and assessment development processes (Angrish et. al., 2025). The workflow describes data transfer from one step to the next with the goal of data integrity, visibility, and control while operationalising efficiency through modernisation of processes fit for AI and content experts. Another is the labelling or tagging of included studies during the screening process. Labels can be applied manually or based upon classifiers (aka search strategies) that are specified by key words, e.g., mechanistic studies pertinent to evaluation of carcinogenesis. Being able to refine search strategies developed by human information specialists or develop new strategies with ML/AI could reduce the time and costs of tagging studies. In the context of HAWC, imported studies could be automatically labelled. Feedback from HERA human health assessment teams indicate that the ability to auto label references would be very useful during database search and screening, promising a considerable cost and time savings that has been evidenced through AI/ML improvements to the ECOTOX database (Olker et al., 2022). US EPA is currently focusing on labels for mechanistic evidence.

Data extracted into HAWC are controlled through the use of terminology resource standards (such as picklists and controlled vocabularies) to facilitate standardisation of the author reported data. Endpoints extracted from the experimental animal data are mapped to the Environmental Health Vocabulary (EHV) to promote interoperability and consistency across assessments (Angrish et. al., 2025). The EHV, which can be accessed from HAWC[55], is an organised collection of words and phrases that includes preferred terms that are non-redundant, unambiguous, can be indexed, are machine-readable and can be used to search a content management system. In the EHV, endpoint terms are placed in five-level hierarchy: organ system, organ, effect, sub-effect, and endpoint. In addition, the EHV includes various alphanumerical identifiers, term definitions, source information, and other metadata in support of FAIR data principles (Wilkinson et al., 2016). The EHV is built into HAWC as its controlled vocabulary to support assimilation, visualisation, interaction, and data accessibility to the human and animal finding and studies information for use in chemical assessments (Angrish et. al., 2025). Although the EHV is not integrated into the OECD AOP knowledge base (AOPKB) this is an area of future exploration as the fully digital EHV could be easily included as an additional resource among existing ontologies and picklists currently integrated into the AOPKB (Ives et al., 2017). Currently, the data extraction in HAWC is most developed for findings from observational human and experimental animal studies. The EHV is most developed for phenotypic ("apical") findings for experimental animal studies. There is a data extraction module for *in vitro* studies in HAWC, but it is not regularly used and does not connect well to the current EHV. Plans are underway to refine the *in vitro* module in HAWC to make data extraction more efficient[56], approached with a goal of compatibility with non-apical findings covered in OHT 201 for intermediate effects/mechanistic information. Thus, it may be possible to map the data extraction fields in HAWC to the data extraction fields in the OECD Omics Reporting Framework (OECD, 2023), AOP database management systems, and/or related OECD Harmonised Templates such as OHT 201. This would set the stage for enhanced interoperability between OECD and US EPA managed data resources by mapping terminology and associated data managed by OECD reporting frameworks (e.g., OORFs, AOP-KB) with HAWC and vice versa.

HERA is also interested in higher level conversations on adjusting approaches used to disseminate primary research in journal articles to a more structured format and has conducted a pilot exercise in this area

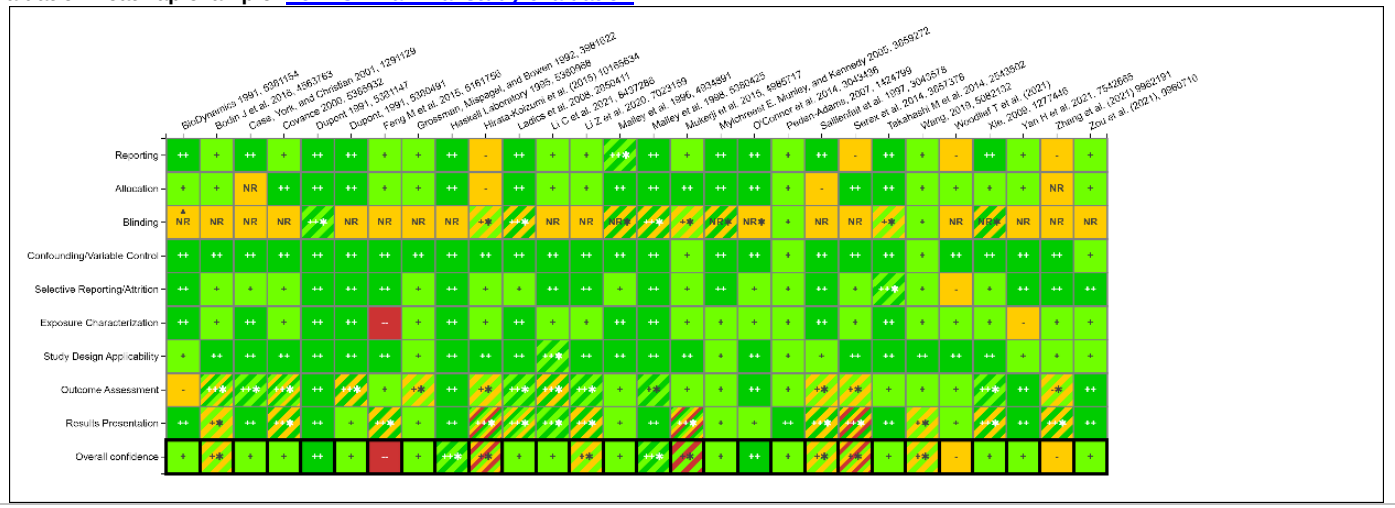
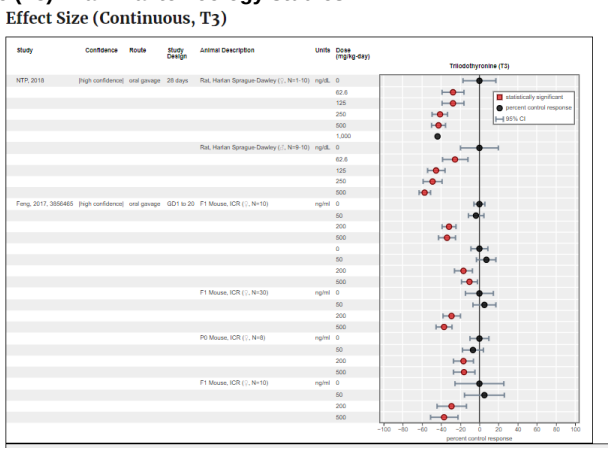
[54] Dextr is a customized version of the Laser AI tool (https://laser.ai)
[55] https://hawc.epa.gov/vocab/ehv/

[56] The target timeframe for updating the HAWC *in vitro* data extraction module is by the end of 2025

(Wilkins et al., 2022). The pilot study was conducted to assess the feasibility of asking participants to summarise study methods and results for experimental animal studies using a structured, web-based data extraction model (HAWC) as an illustration of a potential workflow that could be implemented during the manuscript submission process. Participants were also asked to conduct study evaluation (risk of bias and sensitivity) using IRIS methodology to explore whether awareness of study evaluation methods would impact how participants might approach the conduct and reporting of future research. Having journals disseminate data in structured machine-readable formats would mitigate the expensive and time-consuming process of developing ML/AI approaches to extract content from individual pdfs where content is presented in text and complex tables. It follows that ML/AI approaches could then focus on tasks that entail looking at findings across studies to facilitate evidence synthesis. The data gathered by structured data entry has exponential value as it can be used as training data in existing and developing AI/ML models currently in use. Further, structured data entry supports interoperability between data management systems so that the data can be easily exchanged, addressing a core FAIR principle. Findings from the (Wilkins et al., 2022) pilot study suggested that asking authors to provide data via structured templates may be a viable process. Participants understood the long-term positive implications and did not find the overall process prohibitively arduous. The pilot study also found some support for the hypothesis that use of study templates may have “halo” benefits in improving the conduct and completeness of reporting of future research.

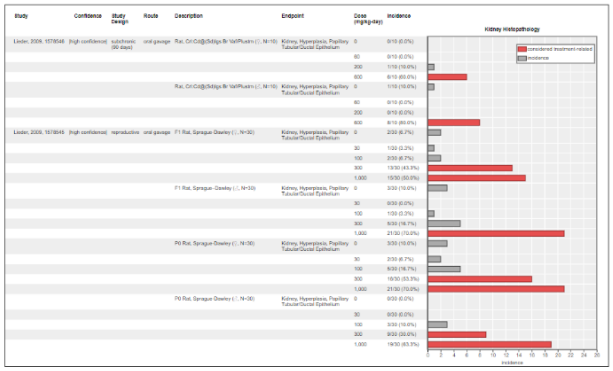
Table A.1. PFAS SEM resources

Resource	Citation
Slide presentation	(Please see Annex to Case Studies supporting document)
Peer-reviewed journal publications	• Carlson, LA et al. (2022). Systematic evidence map for 150+ per- and polyfluoroalkyl substances (PFAS). EHP 130(5):56001. https://doi.org/10.1289/EHP10343 • Carlson, LA et al. (2023).Erratum: Systematic evidence map for 150+ per and polyfluoroalkyl substances (PFAS) • Radke, EG et al. (2022) Epidemiology evidence for health effects of 150 per- and polyfluoroalkyl substances: A systematic evidence map. EHP 130:9. https://doi.org/10.1289/EHP11185 • Shirke, A., et al. (under review) Expanded Systematic Evidence Map for Hundreds of Per- and Polyfluoroalkyl Substances (PFAS) and Comprehensive PFAS Human Health Dashboard https://ehp.niehs.nih.gov/doi/10.1289/EHP13423 **Background** • Patlewicz G et al. (2019) A chemical category-based prioritization approach for selecting 75 per- and polyfluoroalkyl substances (PFAS) for tiered toxicity and toxicokinetic testing. EHP 127(1):14501. https://doi.org/10.1289/EHP4555 • Patlewciz G et al. (2022) Towards reproducible structure-based chemical categories for PFAS to inform and evaluate toxicity and toxicokinetic testing Comp Tox 24:100250 https://doi.org/10.1016/j.comtox.2022.100250 • Thayer, KA et al. (2022). Use of systematic evidence maps within the U.S. environmental protection agency (EPA) integrated risk information system (IRIS) program: advancements to date and looking ahead. https://doi.org/10.1016/j.envint.2022.107363 • Thayer, KA et al. (2022). Template Systematic Evidence Map (SEM) template: Report format and methods used for the U.S. EPA Integrated Risk Information System (IRIS) program, Provisional Peer-Reviewed Toxicity Value (PPRTV) program, and other "fit for purpose" literature-based human health analyses. https://doi.org/10.1016/j.envint.2022.107468 • Williams et al. (2022). Assembly and Curation of Lists of Per- and Polyfluoroalkyl Substances (PFAS) to Support Environmental Science Research. Front Environ. Sci. Apr 5; 10:1-13 10.3389/fenvs.2022.850019
Interactive dashboards and HAWC projects	• Carlson, LA et al. (2022). Systematic evidence map for 150+ per- and polyfluoroalkyl substances (PFAS). • Interactive Overview of Available Animal Evidence: https://hawc.epa.gov/summary/visual/assessment/100500085/Figure-6-Survey-of-animal-studies/ • Interactive Overview of Available Human Evidence: https://hawc.epa.gov/summary/visual/assessment/100500085/Figure-5-Survey-of-human-studies/ • HAWC project: https://hawc.epa.gov/assessment/100500085/ • Radke, EG et al. (2022) Epidemiology evidence for health effects of 150 per- and polyfluoroalkyl substances: A systematic evidence map. EHP 130:9 • Interactive Summary of Available Evidence: https://hawc.epa.gov/summary/visual/assessment/100500085/Epidemiological-Studies-and-Study-Confidence/ • Shirke, A et al (under review) Expanded Systematic Evidence Map for Hundreds of Per- and Polyfluoroalkyl Substances (PFAS) and Comprehensive PFAS Human Health Dashboard • Tableau Interactive Dashboards: • Expanded PFAS SEM: https://public.tableau.com/app/profile/literature.inventory/viz/ExpandedPFASEvidenceMapVisualizations/ReadMe • Comprehensive PFAS Dashboard: https://public.tableau.com/app/profile/literature.inventory/viz/ComprehensivePFASEvidenceMapVisualizations/ReadMe • HAWC Project: https://hawc.epa.gov/assessment/100500256/
Example US EPA PFAS toxicity value assessment that was based on the SEMs:	• U.S. EPA. ORD Human Health Toxicity Value for Perfluoropropanoic Acid (PFPrA) (CASRN 422-64-0 | DTXSID8059970). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-22-042F, July 2023.

Resource	Citation
Example figure that can be used to display study quality information across studies. The figure is interactive and can be used to view more detailed study quality information for a given study and evaluation domain	**Study evaluation heatmap example: [nonECHA animal study evaluation]** 
Example figures that can be created to evaluate dose gradient, magnitude of effect, and patterns across studies. The figures are interactive and underlying data are downloadable	**Continuous endpoint example: triiodothyronine (T3) in animal toxicology studies**  **Continuous endpoint example: Kidney histopathology in animal toxicology studies**

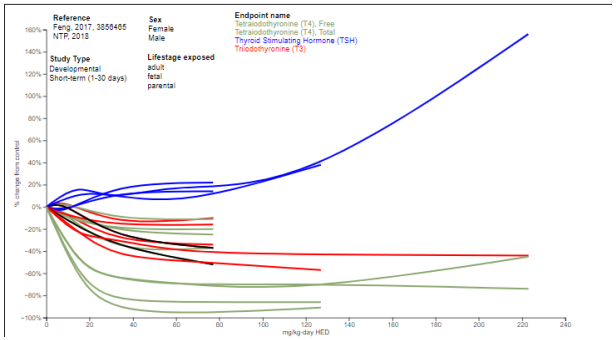
Resource	Citation

Effect Size (Dichotomous Endpoint, Kidney Histopathology)



Continuous endpoint example: Thyroid hormones in animal toxicology studies

Crossview (Thyroid Hormone)



References

Angrish et. al. (2025), "An environmental health vocabulary and its semi-automated curation workflow", *Evidence Based Toxicology*, Vol. 3(1), https://doi.org/10.1080/2833373X.2025.2485111

ATSDR (2022), Systematic Evidence Map (SEM) for Metlylene Chloride, US Department of Health and Human Services, https://www.atsdr.cdc.gov/ToxProfiles/SEM-for-Methylene-chloride.pdf

Beebe, J. et al. (eds.) (2022), *Artificial Intelligence Tools and Open Data Practices for EPA Chemical Hazard Assessments*, National Academies Press, Washington, D.C., https://doi.org/10.17226/26540.

Carlson, L. et al. (2022), "Systematic Evidence Map for Over One Hundred and Fifty Per- and Polyfluoroalkyl Substances (PFAS)", *Environmental Health Perspectives*, Vol. 130/5, https://doi.org/10.1289/ehp10343.

Carlson, L. et al. (2024), "Erratum: "Systematic Evidence Map for over One Hundred and Fifty Per- and Polyfluoroalkyl Substances (PFAS)"", *Environmental Health Perspectives*, Vol. 132/1, https://doi.org/10.1289/ehp14191.

Howard, B. et al. (2020), "SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation", *Environment International*, Vol. 138, p. 105623, https://doi.org/10.1016/j.envint.2020.105623

Ives, C. et al. (2017), "Creating a Structured Adverse Outcome Pathway Knowledgebase via Ontology-Based Annotations", *Applied In Vitro Toxicology*, Vol. 3/4, pp. 298-311, https://doi.org/10.1089/aivt.2017.0017.

NASEM. (2022a), Committee to Review EPA'S IRIS Assessment Handbook et al. (2022), *Review of U.S. EPA's ORD Staff Handbook for Developing IRIS Assessments*, National Academies Press, Washington, D.C., https://doi.org/10.17226/26289.

NASEM. (2022b), Committee on the Guidance on PFAS Testing and Health Outcomes, *Guidance on PFAS Exposure, Testing, and Clinical Follow-Up*, National Academies Press, Washington, D.C., https://doi.org/10.17226/26156.

OECD. (2023), OECD Omics Reporting Framework (OORF): Guidance on Reporting Elements for the Regulatory Use of Omics data from Laboratory-based Toxicology Studies, OECD Series on Testing and Assessment, No. 390, OECD Publishing, Paris. https://doi.org/10.1787/6bb2e6ce-en

Olker, J. et al. (2022), "The ECOTOXicology Knowledgebase: A Curated Database of Ecologically Relevant Toxicity Tests to Support Environmental Research and Risk Assessment", *Environmental Toxicology and Chemistry*, Vol. 41/6, pp. 1520-1539, https://doi.org/10.1002/etc.5324.

Radke, E. et al. (2022), "Epidemiology Evidence for Health Effects of 150 per- and Polyfluoroalkyl Substances: A Systematic Evidence Map", *Environmental Health Perspectives*, Vol. 130/9, https://doi.org/10.1289/ehp11185.

Shapiro, A. et al. (2018), "Software Tools to Facilitate Systematic Review Used for Cancer Hazard Identification", *Environmental Health Perspectives*, Vol. 126/10, https://doi.org/10.1289/ehp4224.

Shirke, A. et al. (2024), "Expanded Systematic Evidence Map for Hundreds of Per- and Polyfluoroalkyl Substances (PFAS) and Comprehensive PFAS Human Health Dashboard", *Environmental Health Perspectives*, Vol. 132/2, https://doi.org/10.1289/ehp13423.

Smith, M. et al. (2022), "Evidence Mapping: Exploring Feasibility for ATSDR's Minimal Risk Level (MRL) Determination", *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.4279746.

Thayer, K. et al. (2022a), "Use of systematic evidence maps within the US Environmental Protection Agency (EPA) Integrated Risk Information System (IRIS) program: Advancements to date and looking ahead", *Environment International*, Vol. 169, p. 107363, https://doi.org/10.1016/j.envint.2022.107363.

Thayer, K. et al. (2022b), "Systematic evidence map (SEM) template: Report format and methods used

for the US EPA Integrated Risk Information System (IRIS) program, Provisional Peer Reviewed Toxicity Value (PPRTV) program, and other "fit for purpose" literature-based human health analyses", *Environment International*, Vol. 169, p. 107468, https://doi.org/10.1016/j.envint.2022.107468.

US EPA (2022a), *Ord Staff Handbook for Developing IRIS Assessments*, US Environmental Protection Agency, Washington, DC.

US EPA (2022b), *Health and environmental risk assessment (HERA) strategic research action plan. Fiscal years 2023-2026*, US Environmental Protection Agency, Office of Research and Development.

Walker, V. et al. (2022), "Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr", *Environment International*, Vol. 159, p. 107025, https://doi.org/10.1016/j.envint.2021.107025.

Wilkins, A. et al. (2022), "Assessing author willingness to enter study information into structured data templates as part of the manuscript submission process: A pilot study", *Heliyon*, Vol. 8/3, p. e09095, https://doi.org/10.1016/j.heliyon.2022.e09095.

Wilkinson, M. et al. (2016), "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, Vol. 3/1, https://doi.org/10.1038/sdata.2016.18.

Wolffe, T. et al. (2019), "Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management", *Environment International*, Vol. 130, p. 104871, https://doi.org/10.1016/j.envint.2019.05.065.

Case study B. Identification of Endocrine Disruptors in the EU regulatory context. Identifying best practices on how research data can assist the regulatory assessment of Endocrine Disruptors

Developed by Swedish Karolinska Institutet (KI), European Food Safety Authority (EFSA), and European Joint Research Centre (EU-JRC)
Case study authors and contributors: Anna Beronius (KI); Iris Mangas, Andrea Terron, Maria Arena, Simone Rizzuto (EFSA); Effrosyni Katsanou, Antonio Franco*, Sharon Munn (EU-JRC); Tanja Burgdorf, Johanna Kaltenhäuser, Carsten Kneuer, Lars Niemann (German Federal Institute for Risk Assessment- BfR); Laurent Lagadic, Steven Levine (BIAC); Scott M. Belcher (Endocrine Society). *- Affiliation listed reflects the author's institution at the time this work was conducted.

B.1. Regulatory context

The current case study aims to explore best practices on the use of research data to assist the regulatory identification of Endocrine Disruptors (EDs) in accordance with the ED criteria laid down in Commission Delegated Regulation (EU) No 2017/2100[57] and Commission Regulation (EU) No2018/605[58] for biocidal products (BPs) and plant protection products (PPPs). According to the International Programme on Chemical Safety (IPCS) of the World Health Organization (WHO)[59], an endocrine disruptor is defined as an *"exogenous substance or mixture that alters function(s) of the endocrine system and consequently causes adverse health effects in an intact organism, or its progeny, or (sub)populations"*. According to the EU criteria for ED identification defined by the relevant EU Regulations[60], a substance shall be considered as having ED properties if it meets all the following criteria:

a. It shows endocrine activity;
b. It shows an adverse effect in an intact organism or its offspring or future generations; and
c. There is a biologically plausible link between the endocrine activity and the adverse effect.

For this purpose, EFSA's regulatory procedures on the assessment of pesticide active substances with regard to their endocrine disruption potential for both human health and environment are explored. According to the Regulation (EC) No 1107/200961 of the European Parliament and of the Council concerning the placing of PPPs on the market, the applicants are required to present a dossier containing a set of mandatory safety studies. They are also required to carry out a literature review according to Art. 8 Par. 5 which states that "Scientific peer-reviewed open literature, as determined by the Authority, on the

[57] Commission Delegated Regulation (EU) 2017/2100 of 4 September 2017 setting out scientific criteria for the determination of endocrine-disrupting properties pursuant to Regulation (EU) No 528/2012 of the European Parliament and Council. OJ L 301,17.11.2017, p. 1–5. Available online: https://eur-lex.europa.eu/eli/reg_del/2017/2100/oj/eng

[58] Commission Regulation (EU) 2018/605 of 19 April 2018 setting out scientific criteria for the determination of endocrine-disrupting and amending Annex II to Regulation (EC) 1107/2009. OJ L 101, 20.4.2018, p. 33–36. Available online: http://data.europa.eu/eli/reg/2018/605/oj

[59] WHO/IPCS (World Health Organization/International Programme on Chemical Safety), 2012. Global Assessment of the State-of-the Science of Endocrine Disruptors. WHO/PCS/EDC/02.2, publicly available at https://www.who.int/publications/i/item/9789241505031

[60] Commission Delegated Regulation (EU) 2023/707 of 19 December 2022 amending Regulation (EC) No 1272/2008 as regards hazard classes and criteria for the classification, labelling and packaging of substances and mixtures. https://eur-lex.europa.eu/eli/reg_del/2023/707/oj/eng

[61] Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. https://eur-lex.europa.eu/eli/reg/2009/1107/oj/eng

active substance and its relevant *metabolites dealing with side-effects on health, the environment and non-target species and published within the last 10 years before the date of submission of the dossier shall be added by the applicant to the dossier"*. Typically, the dossier comprises original studies on the hazards or other properties of the substance that are relevant for the risk assessment, as well as original studies and meta-analyses of epidemiological evaluations.

The ED assessment includes the integration of regulatory studies with public literature studies and different types of evidence including *in vivo*, *in vitro*, *in silico*, as well as the Mode of Action analysis. The Weight of Evidence (WoE) and systematic literature review approaches are in line with the unconditional requirements of the current EFSA guidance on WoE document[62]. In the identification of ED properties, a formalised WoE assessment is needed, and a specific guidance exists[63].

EU Commission Regulation No 2018/605 states that scientific data, other than those generated in regulatory toxicity tests according to internationally agreed study protocols, shall be selected using systematic review methodology. The ECHA/EFSA guidance document (ECHA/EFSA/JRC 2018) for the identification of EDs in the context of Regulations (EU) No 528/2012[64] and (EC) No 1107/2009[5] describes the stepwise process for ED assessment.

B.2. Main goals of the case study

The case study aims to provide two illustrative examples of the tools and processes for inclusion (gathering and evaluating the quality) of research data from the open literature (referring to scientific papers in this case) in the regulatory process following the ECHA/EFSA guidance on ED identification. Example 1 is a real case of a pesticide for which a data-rich dossier is available and describes EFSA's Critical Appraisal Tool for evaluating the internal validity of non-guideline studies before they could be taken into consideration in the regulatory assessment. Example 2 is an academic research study of a chemical not regulated under PPPR and BPR in which non-guideline studies were evaluated with the SciRAP tool. The regulatory use of information from public databases is not in the scope of this case study.

- Glyphosate active substance (Plant Protection Product): EFSA's already finalised ED assessment (Alvarez et al., 2023) following the ECHA/EFSA ED guidance was used to demonstrate the process followed to evaluate the internal validity of non-guideline research data. Glyphosate is a molecule for which a data rich dossier is available, and two thorough assessments took place recently by EFSA using all tools and processes currently available. Considering the large number of studies evaluated in the assessment, detailed statistics about evaluation outcomes provide valuable information about common shortcomings hindering full regulatory consideration.
- Bisphenol F: ED assessment was conducted according to the ED guidance as part of an academic research study. The example illustrates how systematic selection, evaluation and integration of non-guideline research data could be performed for a chemical not regulated under PPPR and BPR.

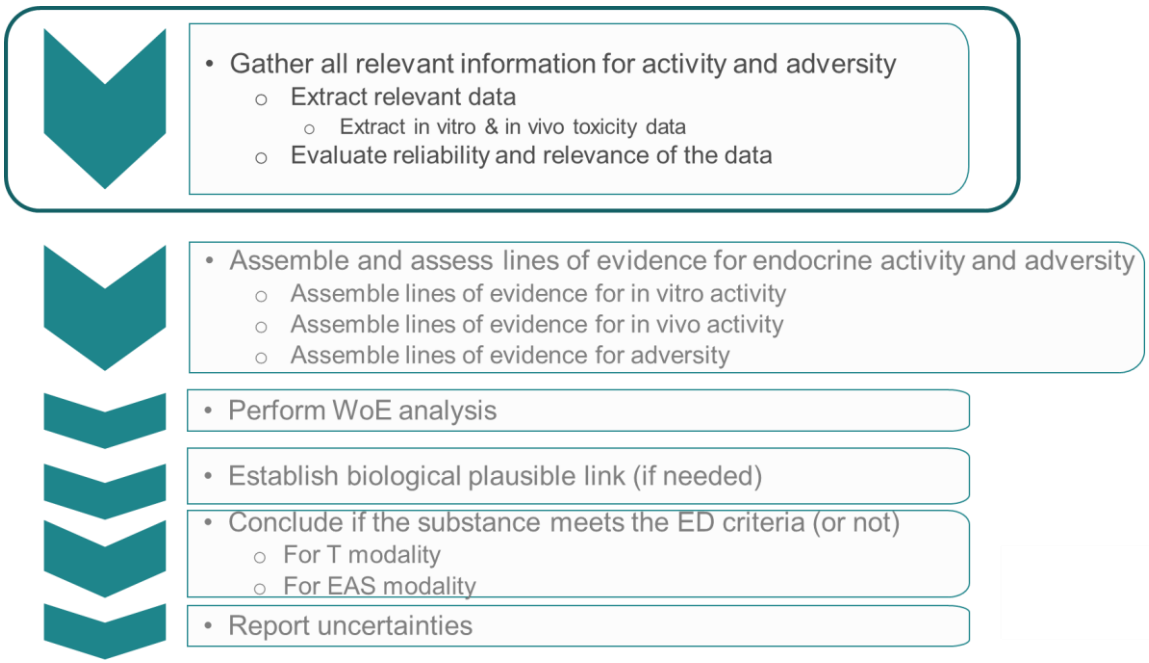
The processes described in the examples provide recommendations on best practices for the use of research data to be included in this Guidance Document.

[62] https://www.efsa.europa.eu/en/efsajournal/pub/4971

[63] Guidance for the identification of endocrine disruptors in biocides and pesticides https://www.efsa.europa.eu/en/efsajournal/pub/5311

[64] Regulation (EU) No 528/2012 of the European Parliament and of the Council of 22 May 2012 concerning the making available on the market and use of biocidal products. https://eur-lex.europa.eu/eli/reg/2012/528/oj/eng

Figure B.1. Workflow for ED identification for human health and environment in line with the ECHA/EFSA ED Guidance, 2018



Note: The first example (glyphosate) focuses on the specific parts of the process indicated by the box in the figure. The second example (bisphenol F) implements the whole workflow. It should be noted here, for clarification, that this workflow, in principle, applies to both guideline and research studies. All relevant data from both guideline and non-guideline studies are incorporated in the workflow and are used for ED identification.

Specific aims of the case study

- Identify common shortcomings of research studies, based on the reliability assessments of research data for use in regulatory assessments from the two examples.
- Contribute to the establishment of minimum/common methodological and reporting standards for non-standard research studies before they can be used in the regulatory decision-making process.
- Identify needs and opportunities to harmonise and share the outcome of evaluation processes. Harmonisation of the evaluation process helps to identify the reasons for divergent views on the scientific reliability among experts for a given research study.

Expert judgement is an inherent characteristic of regulatory evaluations and part of a WoE assessment since a great deal of expertise is required for evaluating the different types of data to inform on the different endpoints. However, there can be differences among experts in their evaluation of the design, analysis and/or interpretation of the results from a study. For example, the regulatory reviewer(s) may consider that a study's methodology is valid but that the conclusion proposed by the study author(s) is not substantiated by the findings. In such cases, the regulatory reviewer(s) will use a different interpretation of the study results in their assessment than that proposed by the author(s). Setting of common data quality standards with regard to the reporting and evaluation of data generated in a research study is critical to improve the transparency and quality of evaluations as well as minimise bias and contribute to harmonisation of the evaluation process. This contributes towards building trust and confidence in the process and avoid duplication of work among different players or in different regulatory contexts.

B.3. Example 1

EFSA's approach for ED assessment of glyphosate for human health

The example provides an overview of the key elements of the detailed ED assessment of glyphosate available at Open EFSA[65].

Extraction of relevant data

According to EFSA's procedure, the extraction of the relevant data i.e., from human observational studies, *in vitro*, *in vivo* experimental toxicity studies was facilitated by DistillerSR® (Evidence Partners, Ottawa, Canada) using predefined forms by an EFSA Working Group of independent experts. The predefined forms allow for the structured collection of data on the characteristics of the studies (e.g., study design, funding source, test system, species), the concentration/dose/exposure characteristics, the endpoints, and methods for measuring them, and the results. Data from DistillerSR® were then transferred to Excel. A two-step approach involving two independent reviewers was followed. The first reviewer performed the data extraction in DistillerSR® which was then transferred into Excel and then the second reviewer performed an independent quality check of the data populated in Excel versus the original publications.

Assess quality of data – Risk of Bias analysis

The internal validity (or Risk of Bias, RoB) of each research study was appraised using a Critical Appraisal Tool (CAT), a customised version of the OHAT/NTP RoB assessment tool[66]. Moreover, for *in vitro* studies, the OHAT/NTP tool developed for the Monograph on PFAS (NTS, 2016) and integrated with some items of the SciRAP tool[67] was used. The following documents have also been integrated and considered: OECD Good In Vitro Method Practices (GIVIMP), 2018; OECD Harmonised Templates 201 (OHT201), and OECD Detailed Review Paper (DRP) No 178 (2012) on methods and endpoints for evaluating EDs. Potential sources of bias are assessed with a set of 6 questions or "domains" and an additional category to consider "other potential threats to internal validity". There are several aspects of the tool that go beyond RoB/internal validity. This is because the customisation of the OHAT included consideration of other aspects of study design that go beyond RoB and capture core features of suitability of the study for use in risk assessment. The tool also allows to categorise based on expert judgement which RoB domains are most relevant depending on the regulatory problem formulation (i.e., key questions).

For each research study, the appraisal was performed for each specific endpoint or group of endpoints because, for the same study, the design and conduct may have affected the RoB differently depending on the endpoints measured.

The following 4-level rating scale was used as shown in Figure B.2:

[65] https://open.efsa.europa.eu/study-inventory/EFSA-Q-2020-00140

[66] The OHAT/NTP tool was developed based on guidance from the Agency for Healthcare Research and Quality (Viswanathan et al., 2012), the Cochrane RoB tool for non-randomised studies of interventions (Sterne et al., 2014), the Cochrane Handbook (Higgins and Green, 2011), CLARITY Group at McMaster University (CLARITY, 2013) and other sources. Available at: https://ntp.niehs.nih.gov/whatwestudy/assessments/noncancer/riskbias

[67] http://scirap.org/

Figure B.2. OHAT 4-level rating scale

++	***Definitely Low*** *risk of bias*: There is direct evidence of low risk of bias practices (May include specific examples of relevant low risk of bias practices)
+	***Probably Low*** *risk of bias*: There is indirect evidence of low risk of bias practices **OR** it is deemed that deviations from low risk of bias practices for these criteria during the study would not appreciably bias results, including consideration of direction and magnitude of bias
– NR	***Probably High*** *risk of bias*: There is indirect evidence of high risk of bias practices **OR** there is insufficient information (e.g., not reported or "NR") provided about relevant risk of bias practices
––	***Definitely High*** *risk of bias*: There is direct evidence of high risk of bias practices (May include specific examples of relevant high risk of bias practices)

Source: OHAT/NTP RoB tool

Risk of bias analysis – In vivo studies

Figure B.3. Critical Appraisal Tool (CAT) for endpoints assessed in *in vivo* studies

Key Questions (Key Q) are highlighted in yellow

Appraisal questions for *IN VIVO* studies
1. Was administered dose or exposure level adequately randomised?
2. Were experimental conditions identical across study groups?
3. Were outcome data complete without attrition or exclusion from analysis?
4. Can we be confident in the exposure characterisation?
5. Can we be confident in the outcome assessment?
6. Were all measured outcomes reported?
7. Were there other potential threats to internal validity? – systemic toxicity

Note: The tool is based on the OHAT/NTP RoB tool [66]

Examples to help to critically review endpoints from in vivo studies in relation to each question of the CAT

Q1. All animals were allocated to any study group including controls using a method with a random component, e.g. both parents and pups. The method for randomisation must be specified.

Q2. The same vehicle and amount were used in control and experimental animals, same housing conditions in both control and experimental animals. Were there large temperature deviations across treatment groups that could confound interpretation of the results? Were all the organisms the same age/developmental stage and source across treatment groups?

Q3. Were excluded animals or missing values properly identified for a given endpoint? If data were censored or rejected from an analysis was a valid rationale provided in the paper?

Q4. Is it clear in the paper what the test substance was (e.g., active ingredient or formulation and if a formulation what formulation was tested)? Was the purity, stability, homogeneity, and exposure levels of the active substance adequately characterised within the study? Was the number of doses tested enough

to perform a dose-response analysis (at least 3 doses plus control)? Was the duration of exposure suitable for the investigated endpoint(s)?

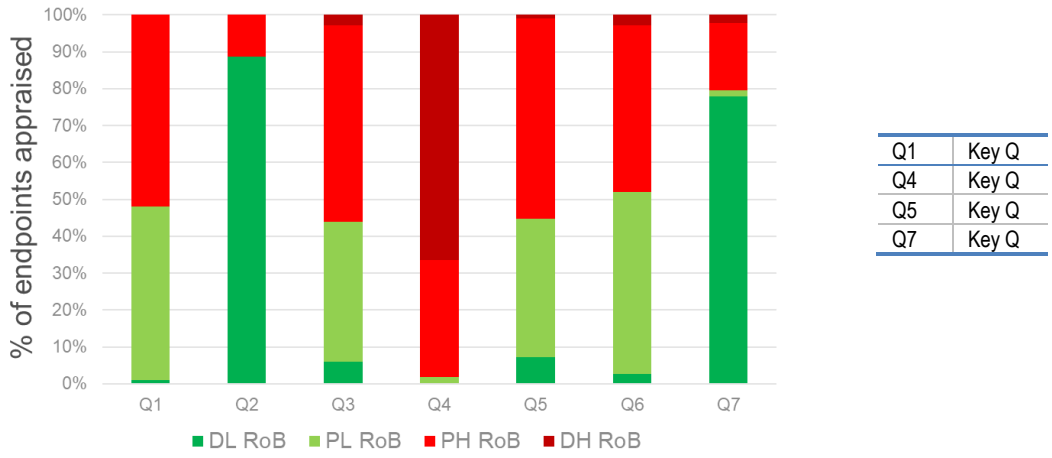
Q5. Were reliable and relevant methods used? Was an indication of their validation status provided? Were assessors adequately trained? Was the study truly replicated or was their pseudo replication (i.e., replicates are not statistically independent)? Was the study statistically powered to determine a biologically significant effect? Did the data for a given endpoint have a valid statistical analysis? Were variance terms for endpoints reported (e.g., standard error, 95% confidence intervals)?

Q6. Did the study use a concurrent control group? Was there a positive control to demonstrate the sensitivity of the test system or a negative control to demonstrate the specificity of the test system? All the study's measured outcomes should be reported. Ideally the raw data (individual measurement data) that allow independent statistical analysis should be included.

Q7. Did overt or systemic toxicity confound the interpretation of the results of assessing a potential adverse effect resulting from an endocrine mechanism? Consult ECHA/EFSA ED guidance (ECHA, EFSA, 2018) on thresholds for overt and systemic toxicity that may confound an assessment of an adverse effect for a given endpoint through an endocrine mechanism. As a minimum, the following parameters should be considered: survival or body weight and body weight gain or food/water consumption or clinical signs. However, overt toxicity assessment requires expert judgement.

Table i in the *Annex to Case Studies* supporting provides a very detailed rationale for scoring each question of the CAT tool for *in vivo* studies.

Figure B.4. Percentage of endpoints from the *in vivo* studies appraised in the different levels of risk of bias (RoB)



Note: Definitely Low (DL), Probably Low (PL), Probably High (PH), Definitely High (DH) RoB for each question of the developed CAT. A total of 221 endpoints were appraised from a total of 24 *in vivo* studies.
Source: peer review report of glyphosate ED assessment humans. Open EFSA [65]

Table B.1 Percentage of endpoints from *in vivo* studies with DH and PH RoB for each of the Key Questions (Key Q) and rationale for their appraisal

Key Qs (*in vivo* studies)	% of endpoints with High (DH and PH) RoB in the different studies	Reasons

Q4 - Exposure characterisation	98.2	• Formulation was used • The doses tested were insufficient for adequate dose-response analysis • Duration of exposure was not suitable for measuring certain endpoints
Q5 - Outcome assessment	55.2	• Blinding was not conducted, and the outcome methodology could be subject to subjective interpretation • Lack or inappropriate statistical analysis
Q1 - Randomisation	52	• No information on how the animals were randomised to be included in control or treated group
Q7 – Systemic Toxicity	20.4	• Systemic toxicity data were not reported or measured e.g., body weight or body weight gain

Note: Probably High (PH), Definitely High (DH) Risk of Bias (RoB).
Source: Peer review report of glyphosate ED assessment humans. Available at Open EFSA [65]

A high risk of bias was identified as well for attrition (Q3) and for outcomes reporting (Q6) (56.1% and 48%) of the endpoints in the different studies with the main reasons being:

For Q3: insufficient information provided about loss/exclusion of animals or measurements. No information about the final number of animals at the end of the study. Not clear why the number of animals was different depending on the endpoint.

For Q6: no adequate reporting e.g., only figures presented, no tables.

Risk of bias analysis – In vitro studies

Figure B.5. Critical Appraisal Tool (CAT) for endpoints assessed in *in vitro* studies

Key Questions (Key Q) are highlighted in yellow

Appraisal questions for *IN VITRO* studies
1. Was administered dose or exposure level adequately randomised?
2. Were experimental conditions identical across study groups?
3. Can we be confident in the exposure characterisation?
4. Can we be confident in the outcome assessment?
5. Were all measured outcomes reported?
6. Were there other potential threats to internal validity? – Cytotoxicity
7. Were there other potential threats to internal validity? – Replicates/repetitions

Note: The tool is based on the OHAT/NTP tool developed for the Monograph on PFAS (NTP, 2016) and integrated with some items of the SciRAP tool (http://scirap.org/). The following documents have also been integrated and considered: OECD GIVIMP., 2018; OHT 201 and OECD, 2012

Examples to help to critically review endpoints from in vitro studies in relation for each question of the CAT

Q1. Did all cells in culture come from a homogeneous cell suspension? Did the study include a concurrent control group to indicate that randomisation covered all study groups?

Q2. If a solvent was used to administer the test substance, was the level of solvent equivalent across all treatments? Were culture conditions the same across all treatments?

Q3. Did the concentration of the test item exceed its solubility? Was the purity of the test item known and was there adequate number of concentrations tested?

Q4. Were positive and negative controls included in assays to demonstrate the sensitivity and specificity? Was the concentration-response adequately characterised to determine the endpoint of interest (e.g., IC50)? Were the data statistically analysed correctly (e.g., the right types of statistical test for categorical data or continuous data). Were variance terms for endpoints reported (e.g., standard error, 95% confidence intervals)?

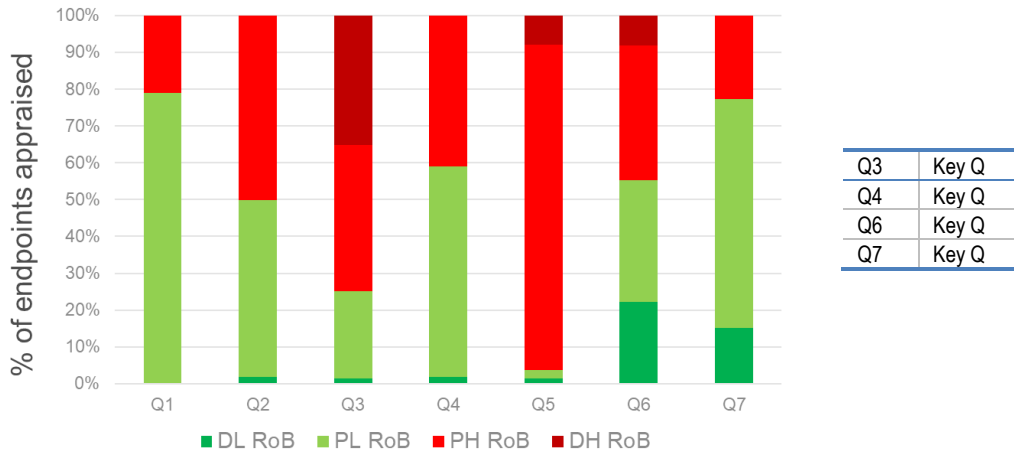
Q5. If protein or mRNA measurements were made by western and northern/slot blots, respectively, were representative blots shown? Have all measured outcomes of the study indicated in the protocol been reported?

Q6. Was cytotoxicity evaluated and if so, was an appropriate cytotoxicity assessment conducted in relation to the endpoint evaluated (e.g., mitochondrial toxicity if an assessment of effects on steroid production is evaluated). Typically, cytotoxicity that exceeds 20% is considered to have confounded a treatment. Could a pH or ionic effect have confounded the results in cell or cell free systems?

Q7. Were assays sufficiently replicated (i.e., within an assay day and number of times the assay was replicated)?

Table ii in the Annex to Case Studies supporting document provides a very detailed rationale for scoring each question of the CAT tool for *in vitro* studies.

Figure B.6. Percentage of endpoints from the *in vitro* studies appraised in the different levels of risk of bias (RoB)



Note: Definitely Low (DL), Probably Low (PL), Probably High (PH), Definitely High (DH) RoB for each question of the developed CAT. A total of 375 endpoints were appraised from a total of 31 *in vitro* studies.
Source: Peer review report of glyphosate ED assessment humans. Available at Open EFSA [65]

Table B.2. Percentage of endpoints from *in vitro* studies with DH and PH RoB for each of the Key Questions (Key Q) and rationale for their appraisal

Key Qs (*in vitro* studies)	% of endpoints with High (DH and PH) RoB in the	Reasons

	different studies	
Q3 - Exposure characterisation	74.9	• The purity of the test item was unknown or too low or that the formulation was used instead of the active substance • Solubility of the test substance was not assessed • The concentrations tested were insufficient for adequate concentration-response analysis. At least three different concentrations and control is required to perform a proper concentration-response analysis
Q4 – Outcome assessment	41.1	• The outcome assessment method was inappropriate • The test system was not appropriate • No or incomplete blinding
Q6 – Cytotoxicity (or other interference)	44.8	• Cytotoxicity was not measured or reported
Q7 – Replicates/ repetitions	22.7	• Not clear how many independent studies and/or how many technical replicates were included

Note: Probably High (PH), Definitely High (DH) Risk of Bias (RoB)
Source: Peer review report of glyphosate ED assessment humans. Available at Open EFSA [65]

A high risk of bias was identified for reporting (Q5) as well for 96.3% of the endpoints in the different studies. The main reasons were the following:

- Data were presented only as summary data in figures; no values of the individual experiments were reported.
- No information provided on the number of experiments/replicates.
- No information in Results section of number of biological independent studies/technical replicates.
- Only data of single concentration (highest) presented in text.

B.4. Example 2

The example presents the workflow followed for the assessment of the endocrine potential of bisphenol F using the ECHA/EFSA ED guidance (ECHA, EFSA, 2018), based on:

Wiklund L, Beronius A. Systematic evaluation of the evidence for identification of endocrine disrupting properties of Bisphenol F. Toxicology. 2022 Jun 30;476:153255. doi: 10.1016/j.tox.2022.153255. Epub 2022 Jul 8. PMID: 35811010.

Purpose and aim, assessment question

The aim of the assessment was to collect and evaluate evidence relevant for evaluating ED properties for bisphenol F (BPF). In addition, the purpose was to explore the application of the ED criteria and assessment process set up for PPPs and biocidal products in the EU on a data-poor non-pesticide. BPF is commonly detected in urine, blood, and breast milk samples in European countries. However, it is not registered under EU REACH and therefore no regulatory toxicity data according to standardised test guidelines are available.

The specific question was “What evidence is available to support an ED evaluation of BPF and does an initial analysis of the evidence indicate ED potential?” This could be further divided into sub-questions:

A. Is there evidence supporting that BPF causes Estrogen-Androgen-Thyroid-Steroidogenesis (EATS)-mediated adverse effects?

B. Is there evidence supporting that BPF has endocrine activity?

Gather all available information for endocrine activity and adversity

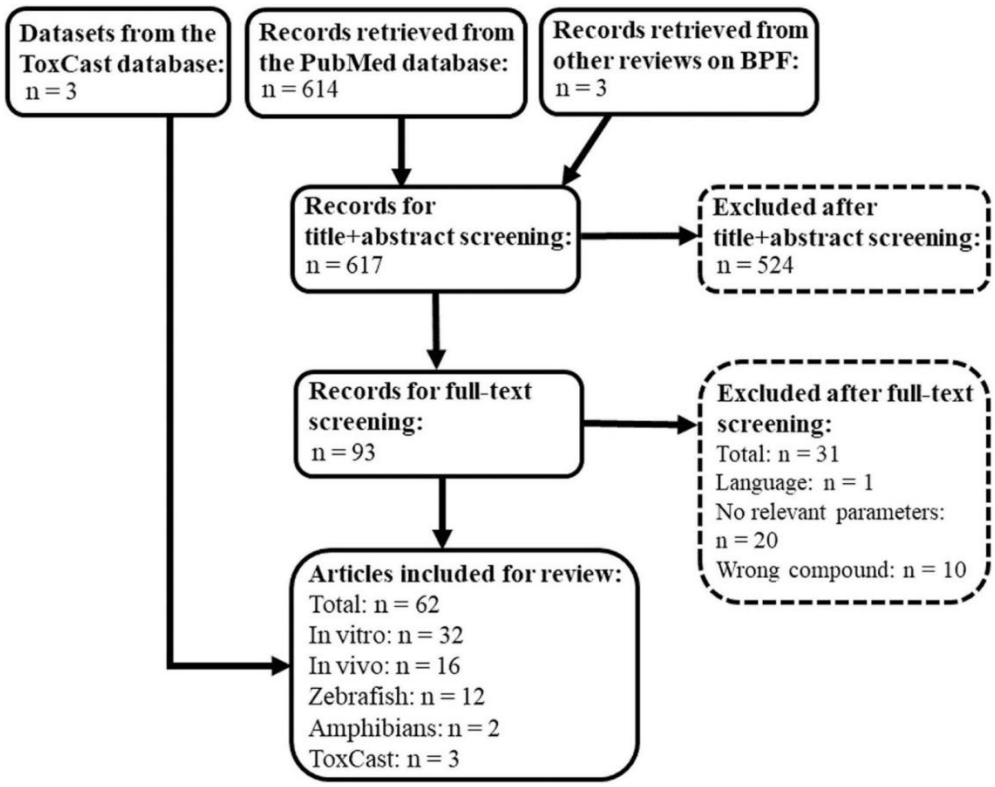
A PECO statement was developed based on the review question to provide a basis for the literature search and for formulating inclusion/exclusion criteria for the screening and selection of studies.

Table B.2. PECO statement constructed for the purpose of evidence collection in the BPF case

PECO statement	
Populations	Animal/human cell lines, primary cells, tissues/organ cultures and embryo Animals (mammals, fish, and amphibians) Humans
Exposure	Bisphenol F
Comparator	Control versus exposed (Experimental data) Different exposure levels (Epidemiological data)
Outcome	Any of the parameters mentioned in table 12 (*in vitro mechanistic*), table 13 (*in vivo* mechanistic) or table 14 (*in vivo* mechanistic, EATS-mediated and 'sensitive to, but not diagnostic of, EATS') in the ECHA/EFSA ED guidance document (2018), as well as non-EATS endocrine-related parameters

Figure B.7 provides the study flow diagram depicting the flow of information through the different steps of gathering the information.

Figure B.7. Information flow through the process of gathering information for the assessment of BPF



Source: (Wiklund L et. al., 2022)

Systematic literature search

A single concept search using the compound name, synonyms, and identifiers for the three BPF isomers was used to search PubMed. The search was performed on January 13th, 2020, using the following search terms:

"bisphenol F"[Supplementary Concept] OR "bisphenol F"[All Fields] OR “bisphenol-F”[All Fields] OR "4,4'-methylenediphenol"[All Fields] OR “2,4'-methylenediphenol”[All Fields] OR "2,2'-methylenediphenol"[All Fields] OR "620-92-8"[All Fields] OR "2467-02-9"[All Fields] OR "2467-03-0"[All Fields] OR “1333-16-0”[All Fields] OR “Reaction mass of 2,2'-methylenediphenol and 4,4'-methylenediphenol and o-[(4-hydroxyphenyl)methyl]phenol” OR "4,4'-methylenebisphenol"[All Fields] OR “2,4'-methylenebisphenol”[All Fields] OR “2,2'-methylenebisphenol”[All Fields] OR "Bis(4-hydroxyphenyl)methane"[All Fields] OR "Phenol, 4,4'-methylenebis-"[All Fields] OR "4,4'-Methylenebis[phenol]"[All Fields] OR "Phenol, 2,4'-methylenebis-"[All Fields] OR "2,4'-Methylenebis[phenol]"[All Fields] OR "Phenol, 2,2'-methylenebis-"[All Fields] OR "2,2'-Methylenebis[phenol]"[All Fields] OR "4,4'-dihydroxydiphenylmethane"[All Fields] OR "2,4'-dihydroxydiphenylmethane"[All Fields] OR "2,2'-dihydroxydiphenylmethane"[All Fields] OR "4,4'-bisphenol F"[All Fields] OR "2,4'-bisphenol F"[All Fields] OR "2,2'-bisphenol F"[All Fields] OR "4,4'-BPF"[All Fields] OR "2,4'-BPF"[All Fields] OR "2,2'-BPF"[All Fields] OR "o-[(4-Hydroxyphenyl)methyl]phenol"[All Fields] OR "o,p'-Bis(hydroxyphenyl)methane"[All Fields] ”.

Retrieved articles were imported into Mendeley[68] reference management software for screening. A backward citations search of the reference lists of retrieved articles was conducted to identify relevant articles not found in the literature search. This strategy did not result in an excess of irrelevant records in this case, and it was therefore not considered necessary to refine the search using targeted search strings. However, a search filter to facilitate targeted searches in the scientific literature for evidence relevant for ED assessment has been developed and validated and published separately (Escrivá et al., 2020).

Searches were also conducted in eChemPortal and in ToxCast[69] to retrieve grey literature and any additional relevant datasets.

Screening and selection of the studies

Screening was conducted using Mendeley. A total of 618 records were retrieved in the search, of which 524 were removed in the screening of titles and abstracts and another 31 removed in the full-text screening (Figure B.7). The following exclusion criteria were applied in the screening process:

- Ineligible exposure (articles only investigating mixtures)
- Studies that are not in silico, in vitro, in vivo (vertebrate), or human data investigating the endocrine-related parameters stated in the PECO statement
- Environmental studies not assessing effects (e.g., exposure data, environmental fate, prevalence in the environment, foods, or water)
- Studies in languages other than English or Swedish

Only one reviewer screened the articles in the BPF case. In a similar study assessing the evidence for ED potential of bisphenol AF (BPAF), Web of Science and EMBASE were searched in addition to PubMed (Escrivá et al., 2021). In that case, titles and abstracts of retrieved articles were screened independently by two reviewers using the RAYYAN tool[70] under 'blind on' mode. Conflicts between the reviewers were resolved by discussion.

Extract relevant data

Information was extracted from the included studies and systematically reported into the Excel template provided as Appendix E to the ECHA/EFSA ED guidance (ECHA, EFSA, 2018). Mechanistic data from both mammals and non-mammalian vertebrates were extracted. Endocrine pathways are well conserved across vertebrate species, and mechanistic data from non-mammalian vertebrates (fish and amphibians) were therefore considered to be relevant also for ED assessment for human health. For the assessment of adversity, only data from studies in mammals were extracted. According to the principles of the Excel template, each parameter investigated in a study was reported in a separate row, generating multiple rows for each study. Both positive and negative data were extracted. In total, 164 parameters (rows) were extracted.

Assemble lines of evidence

Data were organised according to the principles set out in the ECHA/EFSA ED guidance (ECHA, EFSA, 2018). The 164 parameters were organised into 62 lines of evidence collecting data on similar or related

[68] https://www.mendeley.com/

[69] https://comptox.epa.gov/dashboard

[70] https://www.rayyan.ai/

endpoints. Examples of lines of evidence were, for example, data on hormone levels, gene expression, or organ-specific effects. The lines of evidence were then grouped into:

- In vitro mechanistic
- In vivo (mammalian) mechanistic
- Zebrafish and amphibian mechanistic data
- EATS-mediated parameters
- Parameters sensitive to but not diagnostic of EATS

The groups *in vitro* mechanistic, *in vivo* mechanistic, and zebrafish and amphibian mechanistic data were combined as evidence for endocrine activity, while EATS-mediated parameters and parameters sensitive to but not diagnostic of EATS were combined as evidence for EATS-mediated adversity.

Assess quality of data

The quality of the extracted data was evaluated using the Science in Risk Assessment and Policy (SciRAP) tools for *in vivo* and *in vitro* data. SciRAP is a criteria-based tool that allows for evaluating reporting quality, methodological quality and relevance of a study or dataset, separately. The output is a colour profile summarising which criteria are fulfilled, partially fulfilled or not fulfilled. The tool also provides a numerical score corresponding to the % fulfilled criteria. The numerical score should be used with caution, as quantitative measurements of reliability may be misleading, and should be used together with a qualitative analysis of the SciRAP outcome.

In this case, the SciRAP assessment was used to categorise individual datasets on different endpoints (can be several within one study) into reliability categories corresponding to the Klimisch categories "reliable without restrictions", "reliable with restrictions", "not reliable", and "not assignable" (Klimisch et al., 1997). The principles for translating the SciRAP assessment into the reliability categories are given in Table B.3.

Table B.3. Principles for translating SciRAP assessment output into reliability categories for each dataset in the extracted data

Reliability Category	Principles
1.Reliable without restriction	SciRAP methodological quality score > 80 and all key criteria[(a)] are "Fulfilled" and there are no deficiencies in the non-key criteria that might affect study reliability
2. Reliable with restriction	SciRAP methodological quality score > 65 and one or several of the key criteria are "Partially Fulfilled" or there are minor deficiencies in the non-key criteria that might affect study reliability
3. Not reliable	SciRAP methodological quality score < 65 or one or several of the key criteria are "Not Fulfilled" or there are major deficiencies in the non-key criteria that affect study reliability
4. Not assignable	Two or more of the key criteria are "Not Determined"[(b)]

Note: (a) Some SciRAP criteria were considered especially critical for this case and were identified as key criteria *a priori* (b) Criteria that were judged as" Not Determined" in SciRAP were not reported or were considered too poorly reported to confidently classify as" Fulfilled"," Partially Fulfilled" or" Not Fulfilled"

Other examples of the use and interpretation of SciRAP evaluations can be found in (Escrivá et al., 2021), (Ingre-Khans et al., 2020), and (Röhl et al., 2022).

Integrating evidence within and between lines of evidence

To integrate the data within lines of evidence, a structured WoE assessment approach was developed based on the ED guidance as well as guidance for WoE evaluation from the European Commission's

Scientific Committee on Health, Environmental and Emerging Risks (EC, 2018). Specific principles for categorising the confidence of each line of evidence as “strong”, “moderate” or “weak” were applied (Table B.4), based on the quality (as assessed using the SciRAP tool), as well as consistency among studies and species.

Table B.4. Principles for categorising the confidence in lines of evidence as “strong”, “moderate”, or “weak”

Category	Principle for categorisation
Strong	• Effects were observed in one or more studies judged as reliable without restriction; there are no conflicting results
Moderate	• Effects were observed in one or more studies judged as reliable with restriction; there are no conflicting results, or • Effects were observed in one or more studies judged as reliable (with or without restriction) but with conflicting results, i.e., no, or opposite effects were observed in other studies. However, conflicts of results can be explained by differences in study design, for example different exposure periods, doses or animal species or cell models
Weak	• Effects were observed in one or more studies judged as reliable (with or without restriction) but with conflicting results, i.e., no, or opposite effects were observed in other studies. Conflicts of results cannot be explained by differences in study design, for example different exposure periods, doses or animal species or cell models, or • Effects were only observed in one or more studies judged as not reliable or not assignable

It can be noted that the principles applied in this case were relatively strict, i.e. confidence in the evidence was only judged as “strong” if effects were observed in datasets judged as reliable without restrictions and there were no conflicting data. In later case studies, updated principles have been used where evidence can be judged as “strong” also based on datasets judged as reliable with restrictions, as well as when there are conflicting data, if conflicts can be explained by differences in study design such as different exposure periods, doses or animal species or cell models (Holmer et al., 2024). It is important to note that different principles for data integration may be applied in different cases, and such principles should always be fit for purpose and transparently described.

The lines of evidence were then integrated to provide an overall conclusion for adversity and endocrine activity, for the EAS- and T-modalities, respectively. This was done based on the WoE assessment of the empirical evidence and using expert judgment, as described in the ECHA/EFSA ED guidance (ECHA, EFSA, 2018).

Table B.5 and Table B.6 summarise the conclusions regarding the confidence in the lines of evidence for adversity and endocrine activity, respectively, based on the quality, as well as consistency among studies and species available for this example. For more details see (Wiklund L., Beronius A. 2022).

Table B.5. Summary of lines of evidence for EATS-mediated adversity

EATS-modality	EATS-mediated parameters	‘Sensitive to, but not diagnostic of’ EATS parameters
E, A, S	**Moderate – Strong:** • Uterus weight increased,IDs: 1,5,9	**Moderate – Strong:** • Brain organ weight increased, ID: 2

EATS-modality	EATS-mediated parameters	'Sensitive to, but not diagnostic of' EATS parameters
	• Cowpers gland weight increased, ID: 8 • Altered estrous cyclicity, ID: 5 **Weak:** • Testis organ weight increased, IDs: 2,3,6 • Epididymis organ weight decreased, ID: 3 • Seminal vesicle organ weight decreased, IDs: 3,8 • Altered ovary histopathology, ID: 1 • Altered testis histopathology, IDs: 3, 6, 7 • Altered epididymis histopathology, ID: 3 • Decreased sperm parameters, IDs: 3,10 • Seminal vesicle organ weight in offspring decreased, ID: 4 • Prostate organ weight in offspring increased, ID: 4 • Altered testis histopathology in offspring, ID: 4 • Altered epididymis histopathology in offspring, ID: 4 • Decreased sperm parameters in offspring, ID: 4	**Weak:** • Adrenals organ weight in offspring increased, ID: 4
T	**Moderate – Strong:** • Absolute thyroid organ weight increased, ID: 2	
No evidence of effect	• Ovary organ weight, ID: 2 • Testis organ weight in offspring, ID: 4 • Epididymis organ weight in offspring, ID: 2 • Prostate organ weight, IDs: 3, 8 • LABC organ weight, ID: 8 • Glans penis organ weight, ID: 8 • Ano-genital distance, ID: 4 • Nipple development, ID: 2	• Litter size, ID: 4

Note: ID numbers refer to study IDs included in the assessment, for more information see the published article

Table B.6. Summary of lines of evidence for endocrine activity

EATS-modality	*In vitro* mechanistic data	*In vivo* mechanistic data	Zebrafish and amphibian mechanistic data
E, A, S	**Moderate – Strong:** • ER Binding and Activation, ID: 14, 16, 17, 18, 21, 23, 29, 32, 33, 34 • ER-dependent cell proliferation, ID: 16, 19, 22, 27, 28, 29, 30, 31, 34, 35 • Increased ER-dependent gene expression, ID: 21, 27, 37 • AR Binding and Inhibition, ID: 14, 21, 23, 26, 30, 33, 34 • Altered steroidogenesis, ID: 15, 20, 23, 26, 52	**Moderate – Strong:** • Aromatase levels decreased in offspring, ID: 12 **Weak:** • Testosterone levels decreased, IDs: 1, 2, 4, 6, 7 • Oestradiol levels increased, IDs: 1, 3, 4 • Progesterone levels decreased, ID: 1 • LH levels decreased in adults and offspring, IDs: 1, 3, 4, 6 • FSH levels decreased in adults and offspring, ID: 4	**Moderate – Strong:** • Oestradiol levels increased, IDs: 40, 43 • Testosterone levels decreased, ID: 43 • LH levels increased, ID: 49 • FSH levels increased, ID: 49 • Vitellogenin levels increased, ID: 46, 49 • Aromatase mRNA and protein levels increased, ID: 46, 49
T	**Moderate – Strong:** • TR Binding and Activation, ID: 39 • TR-dependent cell proliferation, ID: 39	**Moderate – Strong:** • T3/T4 hormone measurements, ID: 2	**Moderate – Strong:** • Altered T3/T4 hormone measurements, IDs: 44, 45 • TSH levels increased, ID: 44 • TR transcriptional activity, ID: 25

Note: ID numbers refer to study IDs included in the assessment, for more information see the published article

In conclusion, it was found that EATS-mediated adversity was not sufficiently investigated, for any of the modalities, due to lack of data or limited quality of data. However, EAS-mediated adversity in the form of effects on male and female reproductive systems could be inferred, although no strong conclusions could be made. In females, increased uterine weight and altered estrous cyclicity were observed in studies assessed as reliable with restrictions. In males, effects on several sperm parameters were observed together with histopathological changes in testis and epididymis, as well as testis and seminal vesicle weight. However, there was a lack of reported general toxicity data and low study quality in the studies investigating these endpoints.

Endocrine activity was considered to have been sufficiently investigated in regard to the E-, A-, and S-modalities. There was considered to be moderate to strong evidence for ER activation as well as inhibition of AR activity. This was supported by *in vivo* mechanistic evidence from both mammals and non-mammal vertebrates.

B.5. Considerations to enhance the role of open literature in regulatory ecotoxicological assessment

The experience gained by EFSA in appraising evidence from open literature can result in useful tips for researchers who wish to see the outcome of their research taken more into account in the regulatory environmental risk assessment (ERA) process. An analysis of the most frequent issues hindering the reliability of literature studies was conducted over several systematic literature reviews of different active substances/formulations, in order to offer possible ways to produce outcomes that are useful for the regulatory process, while maintaining the freedom to investigate specific and independent research questions.

Several peer-reviewed open literature studies were gathered from authorisation dossiers of PPPs. An appraisal was conducted to identify which issues were most frequently responsible for affecting the reliability (internal validity) on a sample of peer reviewed studies (n=85). Different critical appraisal tools (CAT) were applied depending on the different non-target organisms (NTOs). Such approach presented considerable similarities with Criteria for Reporting and Evaluating Ecotoxicity Data (CRED) developed for aquatic toxicity studies (Moermond et al., 2016).

Among the most frequent parameters hindering the reliability of a literature study, the complete absence of analytical verification, followed by the poor reporting (absence or incomplete information) of age/sex/origin of tested species and tested conditions were frequently identified, together with the low number of replicates (compared to guideline studies), the reduced number of treatment groups (less than 3 concentrations), and the unjustified selection of the concentration range (Table B.7).

Table B.7. Frequency of un/mis/underreported experimental parameters hampering reliability (internal validity) in a sample (n=85) of the appraised peer-review open literature studies in regulatory ERA

Taxa	Age/sex/origin	Blank control	Test conditions	Test item	Analytical verification	Statistics	Number of samples	Number of treatments	Concentration range unjustified
Fish (n= 31)	29.76%	9.92%	24.09%	4.45%	62.35%	9.31%	34.21%	44.94%	32.59%
Amphibians (n= 25)	3.02%	13.49%	14.65%	0.47%	67.21%	3.49%	71.63%	40.23%	56.05%
Aquatic invertebrates (n=16)	72.93%	9.02%	14.29%	3.01%	29.32%	9.02%	13.53%	69.17%	59.40%
Avian (n=10)	0.00%	6.85%	42.47%	23.29%	39.73%	10.96%	28.77%	47.95%	84.93%
Reptiles (n=3)	0.00%	0.00%	0.00%	0.00%	15.22%	0.00%	0.00%	84.78%	15.22%

The issues hindering the reliability can be relatively easy to fix. For instance, a good recommendation for researchers would be to include as much information as possible regarding the tested species (i.e., age/sex/origin) and experimental conditions in the supplementary information, since in the majority of cases, insufficient information on these parameters is due to space constraints in journals. The analytical verifications are instead usually missing due to associated high costs. While having daily analytical verification would be optimal, a good cost-effective solution would be to include the analysis of the stock solutions and/or of the first period of exposure (as a bare minimum). Another good tip would be to include a power analysis to support the selection of a replication number lower than recommended by internationally agreed guidelines (i.e., OECD TGs), and to extend the number of treatments to at least three concentrations plus the control.

B.6. Conclusions

The aim of this case study is to describe the processes for performing the quality assessment of the literature and its integration into the risk assessment using a transparent approach. The two examples in the case study have different but complementary added value. The first example (glyphosate) shows the common parameters missing from the research studies which compromise the validity of the research studies and hinders their integration into the body of evidence used in the risk assessment of pesticides. The second example (BPF) demonstrates how the ECHA/EFSA ED guidance (ECHA, EFSA, 2018) could be implemented for a data-poor non-pesticide substance by applying the EU criteria for the identification of endocrine disruptors by using only research data, in the context of an academic exercise.

The goal of the case study is to improve the use of research data in the regulatory decision-making process to meet scientific and societal needs. Several key parameters that are currently missing in the design, conduct and reporting of most non-standard studies have been identified and have been proposed as *General Reliability Considerations* and *Core Reporting Elements for consideration in publications by researcher* in Table 1.1 and Table 2.1 of the OECD Guidance Document on the Regulatory Use of Research Data.

The impact of the processes proposed in this case study could facilitate a shared interpretation and usage of the scientific peer reviewed literature data between European agencies such as EFSA and ECHA in important areas of collaboration such as the harmonised classification and labelling (CLH) of pesticide active substances. They could also help in the interoperability of data between EU agencies and international organisations (see also Annex B). At the end of the day, the common goal is to use all available data that result in more informed and accurate regulatory assessments.

References

Álvarez, F., Arena, M., Auteri, D., Binaglia, M., Castoldi, A. F., Chiusolo, A., Crivellente, F., Egsmose, M., Fait, G., Ferilli, F., Gouliarmou, V., Nogareda, L. H., Ippolito, A., Istace, F., Jarrah, S., Kardassi, D., Kienzler, A., Lanzoni, A., Villamar-Bouza, L. (2023). Peer review of the pesticide risk assessment of the active substance glyphosate. *EFSA Journal*, 21(7), 1–52. https://doi.org/10.2903/j.efsa.2023.8164

ECHA (European Chemicals Agency) and EFSA (European Food Safety Authority) with the technical support of the Joint Research Centre (JRC), Andersson N, Arena M, Auteri D, Barmaz S, Grignard E, Kienzler A, Lepper P, Lostia AM, Munn S, Parra Morte JM, Pellizzato F, Tarazona J,Terron A and Van der Linden S, 2018. Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009. EFSA Journal 2018; 16(6):5311,135 pp. https://doi.org/10.2903/j.efsa.2018.5311. ECHA-18-G-01-EN

Escrivá L, Hessel E, Gustafsson S, van Spronsen R, Svanberg M, Beronius A. (2020). A validated search filter for the identification of endocrine disruptors based on the ECHA/EFSA guidance recommendations. Environ Int. 142:105828. https://doi.org/10.1016/j.envint.2020.105828

Escrivá L, Zilliacus J, Hessel E, Beronius A. (2021). Assessment of the endocrine disrupting properties of bisphenol AF: a case study applying the European regulatory criteria and guidance. Environ Health. 20(1):48. https://doi.org/10.1186/s12940-021-00731-0

European Commission, Directorate-General for Health and Food Safety, Memorandum on weight of evidence and uncertainties – Revision 2018, Publications Office, 2018, https://data.europa.eu/doi/10.2875/386011

Holmer, M. L. et al., (2024). Methodology for developing data-rich Key Event Relationships for Adverse Outcome Pathways exemplified by linking decreased androgen receptor activity with decreased anogenital distance. Reproductive Toxicology, Volume 128, p. 108662. https://doi.org/10.1016/j.reprotox.2024.108662

Ingre-Khans E, Ågerstrand M, Rudén C, Beronius A. (2020). Improving structure and transparency in reliability evaluations of data under REACH: suggestions for a systematic method, Human and Ecological Risk Assessment: An International Journal. 2020 26:1, 212-241, https://doi.org/10.1080/10807039.2018.1504275

Klimisch HJ, Andreae M, Tillmann U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol. 25(1):1-5. https://doi.org/10.1006/rtph.1996.1076

Moermond CT, Kase R, Korkaric M, Ågerstrand M. (2016), CRED: Criteria for reporting and evaluating ecotoxicity data. Environ Toxicol Chem. 35(5):1297-1309. https://doi.org/10.1002/etc.3259

Rizzuto, S., Arena, M., Auteri, D., Alvarez, F., Ferilli, F., Kienlzer, A., Ippolito, A., Linguadoca, A., Sharp, R., Szentes, C., Villamar, L. (2023). Enhancing the Role of Open Literature in Regulatory Environmental Risk Assessment. Poster Presentation. SETAC Europe 2023, Dublin.

Röhl C, Batke M, Damm G, Freyberger A, Gebel T, Gundert-Remy U, Hengstler JG, Mangerich A, Matthiessen A, Partosch F, Schupp T, Wollin KM, Foth H. (2022). New aspects in deriving health-based guidance values for bromate in swimming pool water. Arch Toxicol. 96(6):1623-1659. https://doi.org/10.1007/s00204-022-03255-9

Wiklund L, Beronius A. Systematic evaluation of the evidence for identification of endocrine disrupting properties of Bisphenol F. Toxicology. 2022 Jun 30; 476:153255. https://doi.org/10.1016/j.tox.2022.153255

Case study C. The CRED Method: A transparent and structured method for evaluation of ecotoxicity data used in risk assessment

Developed by the Swiss Federal Office for the Environment (FOEN) and the German Environment Agency (UBA).
Case study authors and contributors: Mireia Martí-Roura, Muris Korkaric (FOEN); Franziska Kaßner, Peter von der Ohe (UBA); Francisco Sanchez-Bayo (Australian Department of Climate Change, Energy, the Environment and Water); Caroline Moermond (RIVM); Marlene Ågerstrand (Stockholm University); Maria Arena, Fulvio Barizzone, Simone Rizzuto (EFSA); Laurent Lagadic (BIAC); Marion Junghans (Swiss Centre for Applied Ecotoxicology)

C.1. Introduction to the CRED evaluation method

Ecotoxicological studies are used in chemical risk assessment under different regulatory frameworks and for various purposes. These studies come primarily from manufacturers and importers of chemicals, following regulatory testing requirements, but may also come from scientific literature. The increasing number of research data and non-standard tests, with different test designs and endpoints, can make it difficult for regulators to assess their overall reliability and relevance. For a transparent and structured assessment of such research data, while being adaptable to the broad field of ecotoxicology, some guidance is needed.

Over time, several approaches have been proposed for the evaluation of the reliability of ecotoxicity data (Moermond et al., 2017). The “Criteria for Reporting and Evaluating Ecotoxicity Data” (CRED) evaluation method (Moermond et al., 2016) was created to provide a systematic framework for reporting and evaluating the reliability and relevance of ecotoxicity data. It aims to ensure a structured methodology to increase consistency and transparency of evaluations, based on science-based criteria that assist assessors in the evaluation process. The CRED evaluation method was developed to accommodate the use of studies in the context of regulatory frameworks as well as from scientific literature, including studies that do not follow test guidelines.

The starting point for the development of the CRED evaluation method was the reporting requirements of the chronic aquatic OECD TGs No. 201, 210, and 211 (OECD, 2011, 2012, 2013), the evaluation methods already available in the scientific literature, and the expertise of the authors on the subject. This was then combined with the expertise of risk assessors from different sectors through a ring test (Kase et al., 2016).

The CRED evaluation method was recommended for use in the setting of Environmental Quality Standards (EQS) under the EU Water Framework Directive (European Commission, 2018). Under this framework, the evaluation of additional toxicity studies that are published in peer-reviewed literature is crucial, since all relevant and reliable research data should be included, and not just the data from marketing authorisation. Conducting and presenting a systematic literature review with all relevant literature studies is required in most regulatory frameworks (e.g., for plant protection products or biocidal products) and several evaluation methodologies can be used to evaluate the reliability of the studies. More recently, for several marketing authorisation frameworks, such as for medicinal products for human use (European Medicines Agency, 2024), the CRED evaluation method has been recommended.

Even though the CRED evaluation method was developed from the perspective of aquatic ecotoxicity studies, it can be adapted for use in other types of ecotoxicity studies, as the general principles underlying the development of the CRED evaluation method apply to all studies. To improve the reporting of test conditions of ecotoxicity studies, especially in the open literature, the CRED reporting recommendations have also been developed for use by researchers and editors involved in the publication process to allow for a sound evaluation.

C.2. Workflow: use of the CRED evaluation method to evaluate relevant aquatic ecotoxicological studies

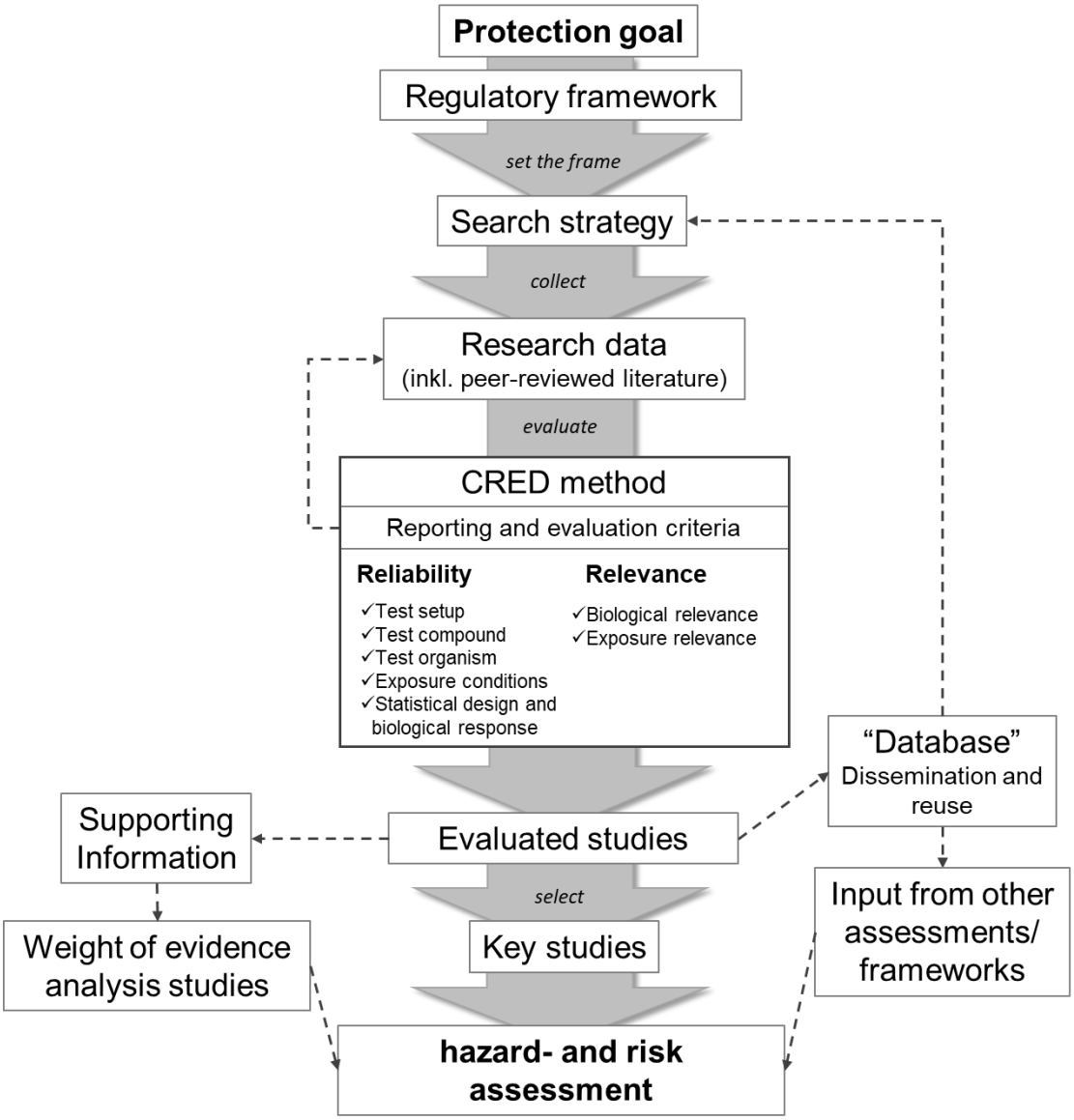
Search strategy and evaluation of studies

Problem formulation is the first step of a risk assessment which allows the risk assessor to identify the potential exposure pathways and hazards, formulate risk hypotheses, and identify the proper risk assessment methodology. The problem formulation sets the boundaries for risk assessment for making it fit-for-purpose. Once the problem formulation is defined, the next step involves the identification of the receptors at risk in the relevant compartments (e.g., sediment organisms). After that, a search strategy needs to be defined to identify key data necessary for the assessment. The CRED evaluation method does not provide specific guidance on this step, but a systematic and transparent exploration of databases and the scientific literature using predefined search terms and clear inclusion and exclusion criteria is recommended. Examples of guidance on how to search and identify publicly available scientific ecotoxicity data can be found in the Technical Guidance for Deriving Environmental Quality Standards (European Commission, 2018), in the EFSA Guidance “Submission of scientific peer-reviewed open literature for the approval of pesticide under Regulation (EC) No 1107/2009” (EFSA, 2011) (see also Case study D); as well as the ECHA Guidance on information requirements and chemical safety assessment - Chapter R.3 and R.4 (ECHA, 2011a and 2011b).

In practice, it is a time-efficient procedure to first screen studies for general relevance (e.g., whether the test organism is relevant for the environmental compartment) and, if necessary, to sort those studies out in a fast-track procedure. All remaining studies are then evaluated for reliability and relevance according to the CRED evaluation method. Figure C.1 gives a schematic overview of the steps for applying the CRED evaluation method to research data.

To facilitate the use of the CRED evaluation method, an Excel file was provided in the original publication's supplemental information. Further tools for reporting evaluation results have been added to later developments of the CRED evaluation method see Section C.4, e.g., NORMAN CRED Tool https://www.normandata.eu/nds/ecotox/credIndex.php, and the SciRAP tool https://www.scirap.org/.

Figure C.1. Schematic representation of a workflow using the CRED evaluation method to report and evaluate the relevance and reliability of ecotoxicological studies for use in hazard and risk assessments



Evaluation of relevance

Relevance refers to the extent to which data and tests are appropriate for a particular hazard identification or risk characterisation (Moermond et al., 2016) and differs for each assessment/protection goal. Thus, the relevance can change over time due to the constant developments in the regulatory field (esp. new/shifting protection goals) and new scientific findings. Examples of the latter are developments in endocrine disruptors and behavioural studies (Ford et al., 2021, EC 2023). The relevance of a study may also change based on the regulatory framework as there may be different exposure regimes that are not all covered in the same publication. Thus, relevance should be evaluated the first time a study or endpoint is used in a framework and re-evaluated over time. The CRED evaluation method provides 13 relevance criteria.

- The first ten CRED criteria concern the biological relevance of a study or endpoint. The primary question here is whether the study design can provide the endpoints relevant to the regulatory

question. For example, results based on recovery might not be relevant for regulatory frameworks other than the EU and the Great Britain authorisation of plant protection products.

- The last three CRED criteria concern the exposure relevance. The main question is whether the exposure scenario is relevant for the test substance or the test organisms (e.g., is the relevant life-stage tested) and if the product tested is representative and relevant for the substance being assessed.

Evaluation of reliability

The evaluation of reliability involves a close examination of critical elements in the design, execution, reporting, and statistical analysis of ecotoxicity studies. To guide the assessor, the CRED evaluation method provides 20 reliability criteria, presented as questions. In the original approach, some criteria can be answered unambiguously, while in other criteria several aspects come together, and the assessor must evaluate their overall compliance. In practice, this has led to the use of response categories “fulfilled”, “partially fulfilled” and “not fulfilled”, alongside an option to comment on each aspect of a criterion. No absolute weighing of criteria is provided in the CRED evaluation method, as the importance of criteria may differ between compounds and organisms. For example, for a dissipating compound, multiple analytical measurements are much more important than for a stable compound. Thus, expert judgement is always needed, and a box-ticking exercise is not recommended.

- The first four CRED criteria are focused on the test design. Importantly, the absence or the non-compliance of a specific guideline and/or GLP is regarded of minor importance for study reliability. On the other hand, critical aspects, such as the lack of proper controls will most likely disqualify the dataset for regulatory use.
- CRED criteria 5 to 7 concern the test substance and aim to ensure that it can be proven beyond reasonable doubt which exact chemical component is responsible for an observed biological effect, e.g., by asking for known impurities or other components in a formulation.
- CRED criteria 8 and 9 pertain to the test organisms and the suitability of the pre-exposure conditions (e.g., health status and acclimatisation to test conditions) and the thoroughness of the organism description.
- CRED criteria 10 to 16 aim to evaluate if the exposure conditions are suitable for both the test organism and the test substance (e.g., appropriate test medium and exposure concentrations below the solubility limit) and if the derived endpoint is analytically verified.
- The last criteria, 17 to 20, concern statistical design and analysis of the biological response and touch on the basis necessary to prove statistically robust endpoints and ideally to be able to re-evaluate the data.

Overall categorisation of reliability and relevance

After the evaluation, the study/endpoint is categorised in four reliability (R1-R4) and relevance categories (C1-C4), depending on whether the assessor concludes the endpoint to be reliable or relevant without restriction (R1 and C1), with restriction (R2 or C2), or not reliable or relevant (R3 or C3). The fourth category is given to studies/endpoint for which reliability or relevance could not be assigned due to insufficient details in the study report (R4 or C4). In principle, when more information becomes available (e.g., through a request to the author), R4 studies may become R1, R2 or R3 studies. Combinations of reliability and relevance categories are possible for the study's assessment (e.g., R2 C3). Overall, the categorisation is not meant to be a bookkeeping exercise and depends on the expert judgement and, for relevance, the specific framework for which the endpoint is intended to be used. While this obviously leaves room for variability in reliability evaluations, the approach supports informed decision making due to the transparency of the evaluation that opens the possibility for a targeted discussion amongst experts. For

the latter, the consideration of multiple CRED analyses is beneficial, pointing to the potential flaws in the study identified by each expert. As an example, in the NORMAN CRED-Tool[71], these analyses are recorded and made available to the public. For this specific tool and related to the nature of the NORMAN database, a fifth category (R5 and C5) was introduced for studies that have not yet been evaluated.

C.3. Example of a study evaluation using CRED

The current exercise aims to show the use of the CRED method as a transparent evaluation tool that can be applied to both standard and non-standard studies and with a potential use under several regulatory frameworks. The following publication was selected to exemplify the use of the CRED evaluation method:

Perillon C., Feibicke M., Sahm R., Kusebauch B., Hönemann L., Mohr S. 2021. The auxin herbicide mecoprop-P in new light: Filling the data gap for dicotyledonous macrophytes. Environmental Pollution 272, 116405.

The study was selected because the tested substance, Mecoprop-P, is registered under both EU REACH and the EU Plant Protection Products (PPP) Regulation and is listed as “substance subject to review for possible identification as priority substance or priority hazardous substance” in the Directive 2008/105/EC on environmental quality standards in the field of water policy. The publication reports an experimental microcosm study with several macrophyte species exposed to the herbicide Mecoprop-P. The study provides new insights into the effects of this herbicide on aquatic dicotyledonous macrophytes, which are the most sensitive taxonomic group.

The results from Perillon et al., 2021 became very relevant for the discussions on the harmonised classification for Mecoprop-P within ECHA’s RAC (Risk Assessment Committee). The original proposal for a harmonised classification was to change the Aquatic Chronic 2 classification into an Aquatic Chronic 3 one. For the most sensitive species, the dicotyledonous plant *Myriophyllum spicatum*, only a test with a preparation was available (Gonsior, 2015). The contribution of the co-formulants of this preparation to toxicity was not known. This study was therefore not used as a basis for the harmonised classification. The results of Perillon et al., 2021 showed that the co-formulants had no major influence on the result. Therefore, after the discussion, the study with the preparation was used in the RAC opinion as the basis for the harmonised classification - as Perillon et al., 2021 was not published at this time - and resulted in an Aquatic Chronic 1 harmonised classification instead of the originally proposed Aquatic chronic 3. The new data from Perillon et al., 2021 also provided new insights on the effects of aquatic macrophytes and allowed the update of the acute and chronic EQS for Mecoprop-P in Switzerland (Kroll et al., 2023).

To illustrate the use of the CRED evaluation method, the original study, together with supporting information and the CRED template, was sent to scientific and regulatory experts working in different regulatory frameworks with the question to assess the study for use in the framework(s) in which they were experts. Four assessors provided a full assessment (Table C.1). Although the study describes multiple species and multiple endpoints per species, for simplicity, the participants applied the CRED evaluation to one species (*Callitriche palustris*), one effect parameter (dry weight) and two toxicity endpoints (EC10 and EC50). The summary of the results of this evaluation and the framework chosen for the relevance evaluation are shown in Table C.1 (for further information about the detailed CRED evaluation of each participant see Appendix C.1 of the *Annex to Case Studies* supporting document. Please note that the current exercise has been carried out with the goal of showing the use of the CRED evaluation method when being applied under different regulatory frameworks. However, it has been used only with a few participants, so conclusions should be treated with caution.

[71] https://www.normandata.eu/nds/ecotox/credIndex.php

Table C.1. *Summary of the results obtained during the CRED exercise*

Assessor	*CRED evaluation*	*Framework*	*Comment from the assessor*
A 1	EC10 & EC50: R2 C1 (Reliable with restriction and relevant without restriction)	EC10 relevant for long-term assessment in (harmonised) classification in the EU CLP Regulation and to fulfil information requirements under the EU REACH Regulation EC50 relevant for acute assessment in (harmonised) classification in the EU CLP Regulation, to fulfil information requirements under the EU REACH Regulation, and authorisation under the EU Plant Protection Product Regulation	Before approving the regulatory use, the assessor would like to re-assess the data to clarify why the NOEC is much higher than the EC10. For this, raw data from the authors is needed
A 2	EC10 & EC50: R4 C4 (Not assignable)	Authorisation under the EU Plant Protection Product Regulation	The assessor concluded that the EC50 of the study could become evaluated as "Relevant without restrictions" (C1) and "Reliable with restrictions" (R2) if the missing information is provided
A 3	EC10 & EC50: R3 C2 (Not reliable and relevant with restriction)	Authorisation under the EU Plant Protection Product Regulation	The assessor concluded that the information on the intended uses and/or application rate of the test substance, which is relevant for the risk assessment of PPP, is missing. Limitations in the reliability of the study (e.g., inadequate analytical verification in case of substances with stereoisomers) were also observed
A 4	EC10 & EC50: R2 C1 (Reliable with restriction and relevant without restriction)	Development of EQS values under the EU Water Framework Directive	The assessor concludes that the communication with the author for the clarification of the test substance is needed. The NOEC is higher than EC10 because of control variability of 28.1%. This is quite high but lower than the validity criteria that the authors have defined. Thus, the assessor concludes that the EC10 is preferred over the NOEC

Note: Further information about the individual CRED assessments can be found in the *Annex to Case Studies* supporting document.

Differences in the evaluation of studies under different frameworks can be observed (Table C.1). Reliability and relevance were differently assessed even when they were evaluated under the same framework. Differences in relevance (C) can occur as relevance depends partly on the framework. For example, some toxicity parameters might be relevant for some frameworks, but not relevant for others. The relevance of the EC10 has been weighted differently in this exercise. In case of the effect assessment of herbicides, for some frameworks, the EC10 and EC50 are considered relevant toxicity endpoints for chronic and acute effect assessment, respectively, e.g., for the assessment in (harmonised) classification and labelling (CLH)-ECHA, under EU REACH and under the EU Water Framework Directive. However, under the EU Plant Protection Product authorisation, although the toxicity endpoint selected for risk assessment is the EC50, the tests with algae and macrophytes are placed under the chronic risk assessment since these tests comprise the complete life cycle, or a large part of the life cycle, of these organisms (EFSA PPR

Panel, 2013). Thus, different data requirements in different frameworks can result in different relevance assessments for individual studies but also individual endpoints within a study.

Differences in the reliability assessment (R) within the same study should not occur, regardless of the framework, since reliability criteria are used to evaluate the inherent quality of a test report or publication. In this example, the same study and toxicity endpoints (EC10 and EC50) have been rated as reliable with restrictions (Assessors 1 and 4: R2), not assignable (Assessor 2: R4) and not reliable (Assessor 3: R3). Different assessments are mostly based on how the participants addressed the uncertainties and the lack of information in the study. One example of this different assessment was related to the chemical verification of the test substance. Mecoprop-P is a racemate with two isomers. While for the Assessor 3 the criterion related to the chemical verification was not fulfilled and incorrectly reported, since only information of the racemate was provided, for other assessors the information was only missing and could be reported with a communication with the authors (Assessor 2 and 4). The ability to obtain the necessary data might depend on various factors, e.g., when the study was performed (i.e., it might be more difficult to obtain additional unpublished data, especially for older studies), the language, or the accessibility of the authors. Overall, the results of this case study, in which differences in the assessments from the same study are shown, argues strongly for the usefulness of a globally accessible database for sharing evaluation results (e.g., NORMAN ECOTOX database) and globally acceptable reporting requirements for peer-reviewed publications.

This case study shows that the CRED evaluation method does not automatically produce consistent assessments across different frameworks and between experts, highlighting the importance of the peer review process and the regulatory context in which the latter is applied. This is, to some degree, expected, given the different data requirements in the different frameworks and the overall reliance on expert judgement. The assessment does make it clear on which arguments the assessments were based, though. When the original CRED method was developed, it was tested in a two-stage ring-test (Kase et al., 2016) against the method to evaluate the reliability of studies established by Klimisch and colleagues (Klimisch et al., 1997). In the ring-test, all risk assessors were informed before the evaluation that studies should be evaluated for their potential use in EQS-derivation under the EU Water Framework Directive, thus excluding the differences that might arise from the different regulatory contexts. This study showed that the number of criteria that must be met for a reliable study differs per study and thus, no general cut-off can be set. In general, it was shown that the CRED evaluation method provided a more detailed and transparent evaluation of reliability and relevance compared to the Klimisch method. The ring-test participants found CRED to be less dependent on expert judgement, more accurate and consistent, and practical regarding the use of criteria and time needed for performing an evaluation. Thus, the results from the present case study do not contradict the results from the ring test. Therefore, the use of the CRED evaluation method is expected to produce transparent and more consistent evaluations compared to less-detailed evaluation methods such as the Klimisch method. This can facilitate focused discussions in the respective expert groups that perform regulatory risk assessments. However, since a comprehensive assessment and the associated recording of criteria using CRED is a relatively time-consuming process, it was considered most appropriate for the assessment of potential key studies (Kase et al., 2016).

C.4. Further developments and new fields of application

The CRED evaluation method represents a useful system allowing assessors to systematically and transparently evaluate studies in the remits of different regulatory frameworks. The advantages of using such method for the assessment of peer-reviewed studies are large, due to: i) its systematic and transparent structure, ii) the possibility of evaluating both reliability and relevance, iii) its regulatory-based development specific for ecotoxicology, and iv) its use of generally accepted categories for appraisal of criteria, compared to others (e.g., Klimisch et al., 1997). As the CRED evaluation method focuses on

aquatic ecotoxicity studies, further developments may be needed for use for other types of toxicity studies. Some recent initiatives, or potential initiatives, are listed below.

Applicability of CRED to other in vivo non-aquatic and/or higher-tier ecotoxicity studies

One of the possibilities to expand the applicability of the CRED evaluation method would be to apply it to non-aquatic ecotoxicity studies such as soil organisms, bees, and non-target arthropods. Initial work has already been carried out for some of them, for example, CRED has been used for the assessment of sediment studies in the context of EQS derivation[72] and for the retrospective soil hazard assessment in Switzerland. For those assessments, the sediment and soil specific aspects have been incorporated in specific CRED criteria (Casado-Martinez et al., 2024). Another example is the development of Critical Appraisal Tools (CATs) (Lahr et al., 2023) for the assessment of non-standard higher-tier studies for aquatic and terrestrial organisms based on the CRED method. The EthoCRED has been developed to support the evaluation of behavioural changes in ecotoxicity studies, as these studies have been underrepresented in hazard and risk assessments (Ågerstrand et al., 2020; Bertram et al., 2024). Moreover, difficulties in the evaluation of toxicity studies are not only linked to the study design, the observed effects and derived endpoints or the organisms tested but can also be linked to the tested compounds. Nanomaterials, for example, behave very differently in ecotoxicity tests compared to conventional soluble chemicals. Thus, to accommodate the CRED approach to nanomaterials, the NanoCRED method has been proposed (Hartmann et al., 2017).

Applicability to in silico studies

The current trend in risk and hazard assessment is pushing towards New Approach Methodologies (NAM), which includes *in vitro*, *in silico*, and other non-animal approaches. This is supported in the EU Chemical Strategy for Sustainability (EC 2020). As a consequence of limiting animal testing, the number of modelling approaches to support the regulatory risk assessment of plant protection products have indeed increased in recent years (Hommen et al., 2016; EFSA PPR Panel, 2018). However, mechanistic effect models are rarely used in a regulatory context (Larras et al., 2022). Trust in these results could also benefit from a systematic and transparent CRED-like evaluation. Further work to elaborate a method to support the evaluation of literature-based mechanistic models could be initiated.

Quantity and/or quality of reported information

As observed in this case study and in Kase et al. (2016), limitations to the applicability of the CRED evaluation method can arise due to the quantity and/or quality of information reported by the authors. Literature studies may indeed represent an invaluable source of additional information on effects not covered by standard data requirements. However, their CRED evaluation (and consequently their regulatory use) can be drastically hindered by data under-reporting. A systematic evaluation may lead to lower reliability scores, as flaws in the study setup and performance are more easily observed (Kase et al., 2016). It is acknowledged that some information must be excluded from peer-reviewed articles, e.g., due to strict word limits. However, from a regulatory perspective, it is strongly encouraged to report all crucial information at least in supplementary information, where generally no limitations on word count are set. In this context, the EU authorities are currently working on providing guidance to authors who wish to see their work considered in the regulatory environment (Rizzuto et al., 2023). Better reporting of methodological aspects also improves peer review and science in general, since replication of studies also becomes easier. When peer-reviewed studies improve in reporting, this may also prevent unnecessary repetition of animal studies, since more studies could be used for regulatory purposes.

Expert judgement and (semi-)quantitative scoring systems

[72] https://www.ecotoxcentre.ch/expert-service/quality-criteria/sediment-quality-criteria

CRED relies to some extent on expert judgement since many criteria are not simply binary yes/no decisions. The CRED evaluation method does not provide guidance on which criteria are critical or non-critical for the assessment, as the criteria may vary depending on, for example, the study type, and the substance or organism studied. The flexibility in the assessment can be seen as positive for some regulatory frameworks since the evaluation can be adapted to the specific regulatory question. For others, the lack of clear instructions on how to assess the criteria can be seen as a limitation. Thus, the use of tools like the CATs or the NORMAN CRED Tool (https://www.norman-network.com/nds/ecotox), that both report a (semi-) quantitative scoring system to guide the user while allowing for expert judgement, could be helpful for those situations.

Fully automated scoring systems have also been proposed; however, care should be taken when using these systems. Especially with less experienced assessors, which may be reluctant to deviate from the automatically applied score, believing that 'the system' will always be better than their own judgement.

Dissemination and reuse of CRED assessments

Recording the study assessment with the CRED evaluation method requires a certain investment of time. However, when these assessments can be disseminated and reused by other risk assessors and risk managers, an efficient use of the resources invested in the assessments, especially for potential key studies, can be ensured. It also contributes to the overall aim of reducing the use of animal studies. Although some work to exchange data and CRED evaluations have been done on a small scale, the lack of dissemination of CRED evaluations requires attention. The NORMAN Network has been working for some time on a database for ecotoxicological studies (https://www.norman-network.com/nds/ecotox), and more recently developed a tool linked to the existing database to evaluate the reliability of ecotoxicological studies, based on CRED. This NORMAN CRED tool can open the possibility for global exchange and reuse of CRED assessments and, at the same time, increase the transparency of the assessments.

Some modifications have been added to the NORMAN CRED-tool, e.g., splitting one criterion into several questions to increase transparency and clarity and remove the ambiguity of a single decision for that criterion. An example of this is the use of appropriate test conditions, where test conditions like temperature and hardness might be suitable for the test organism under investigation, but the test pH might be unsuitable for the ionisable test compound under investigation (Köhler et al., 2023, Kroll e al., 2024).

Structured evaluation and reporting of exposure data

Data of measured concentrations of chemicals in environmental matrices (exposure data) are crucial components of risk assessment and management. However, the evaluation of such data can be challenging, due to lacking reporting guidelines and variable data quality. Very few examples of structured evaluation and reporting schemes exist. One example is the Criteria for Reporting and Evaluating Exposure Datasets (CREED) system, developed as an outcome of a technical workshop of the Society of Environmental Toxicology and Chemistry (SETAC). CREED offers systematic evaluation criteria to enhance the reliability and relevance of exposure data for diverse environmental assessments (Merrington et al., 2024).

C.5. Conclusions

The CRED evaluation method is a transparent and structured system that allows chemical risk assessors to assess ecotoxicological studies in the context of different frameworks and to cross-validate assessments within the same regulatory framework. Adopting common appraisal tools will enhance harmonisation and transparency of study evaluations performed by experts in ecotoxicology.

The case study has shown that the information (metadata) reported in the research studies is crucial for the evaluation with the CRED method and that the use of research data for regulatory purposes depends largely on the quality of this reporting.

Overall, the CRED method has been proven to be very valuable for the evaluation of ecotoxicological studies. Based on the scope/framework of the assessment, adaptations of the CRED method have been already proposed and implemented to make it even more fit-for-purpose and to increase its application across different regulatory frameworks and assessment areas.

References

Ågerstrand, M., Arnold, K., Balshine, S., Brodin, T., Brooks, B.W., Maack, G., McCallum, E. S., Pyle, G., Saaristo, M., Ford, A.T. (2020). Emerging investigator series: use of behavioural endpoints in the regulation of chemicals. *Environ. Sci.: Processes Impacts*, 22, 49-65. https://doi.org/10.1039/C9EM00463G

Bertram, M.G., Ågerstrand, M., Allen, J., Brooks, B.W., Dang, Z., Duquesne, S., Ford, A.T., Hoffmann, F., Hollert, H., Jacob, S., Kloas, W., Klüver, N., Lazorchak, J., Ledesma, M., Maack, G., Melvin, S.D., Mohr, S., Padilla, S., Pyle, G., Saaristo, M., Sahm, R., Smit, E., Steevens, J.A., van den Berg, S., Wong, B.B.M., Ziegler, M., Brodin, T. (2024) . EthoCRED: A framework to guide reporting and evaluation of the reliability and relevance of behavioural ecotoxicity studies. *Biological Reviews*, 100, 556-585. https://doi.org/10.1111/brv.13154

Casado-Martinez, M., Dell'Ambrogio, G., Campiche, S., Kroll, A., Lauber, E., Marti-Roura, M., Mendez-Fernandez, L., Renaud, M., Tierbach, A., Wildi, M., Wong, J.W.Y., Werner, I., Junghans, M. and Ferrari, B.J.D. (2024), Incorporation of sediment- and soil-specific aspects in the Criteria for Reporting and Evaluating Ecotoxicity Data (CRED). *Integr Environ Assess Manag*. https://doi.org/10.1002/ieam.4948

EC. (2020). European Commission—chemicals strategy for sustainability: towards a toxic-free environment. COM(2020) 667 final. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52020DC0667

EC. (2023). Commission Delegated Regulation (EU) 2023/707 of 19 December 2022 amending Regulation (EC) No 1272/2008 as regards hazard classes and criteria for the classification, labelling and packaging of substances and mixtures (Text with EEA relevance). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32023R0707

ECHA. (2011a). Guidance on information requirements and chemical safety assessment. Chapter R.3: Information gathering.https://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment

ECHA. (2011b). Guidance on information requirements and chemical safety assessment. Chapter R.4: Evaluation of available information. https://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment

EFSA PPR Panel. (2013). Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. EFSA Journal 2013;11(7):3290, 268 pp. https://doi.org/10.2903/j.efsa.2013.3290

EFSA PPR Panel. (2018). Scientific Opinion on the state of the art of Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. EFSA Journal 2018; 16(8):5377, 188 pp. https://doi.org/10.2903/j.efsa.2018.5377

EFSA. (2011). Submission of scientific peer-reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009 (OJ L 309, 24.11.2009, p. 1-50). EFSA Journal ;9(2): 2092. (49 pp.). https://doi.org/10.2903/j.efsa.2011.2092

European Commission. (2018). Technical Guidance for Deriving Environmental Quality Standards. Guidance Document No. 27. Updated version 2018. https://circabc.europa.eu/ui/group/9ab5926d-bed4-4322-9aa7-9964bbe8312d/library/7573707d-410b-4ea0-aa84-6ef762e40ecd/details

European Medicines Agency. (2024). Guideline on the environmental risk assessment of medicinal products for human use. EMEA/CHMP/SWP/4447/00 Rev. 1 Committee for Medicinal Products for Human Use (CHMP). https://www.ema.europa.eu/en/environmental-risk-assessment-medicinal-products-human-use-scientific-guideline

Ford, A.T., Ågerstrand, M., Brooks, B.W., Allen, J., Bertram, M.G., Brodin, T., Dang, Z., Duquesne, S., Sahm, R., Hoffmann, F., Hollert, Jacob, S., Klüver, N., Lazorchak, J.M, Ledesma, M., Melvin, S.D.,

Mohr, S., Padilla, S., Pyle, G.G., Scholz, S., Saaristo, M., Smit, E., Steevens, J.A., van den Berg, S., Kloas, W., Wong B.B.M., Ziegler, M., Maack, G. (2021). The Role of Behavioral Ecotoxicology in Environmental Protection. Environ. Sci. Technol. 55, 9, 5620–5628 https://doi.org/10.1021/acs.est.0c06493

Gonsoir, G. 2015. Mecoprop-p K, 600 g/L: Growth inhibition of Myriophyllum spicatum in a water/sediment system. Study Number: S13-04889. Eurofins Agroscience Services EcoChem GmbH. Not published.

Hartmann, N. B., Ågerstrand, M., Lützhøft, H-C. H., & Baun, A. (2017). NanoCRED: A transparent framework to assess the regulatory adequacy of ecotoxicity data for nanomaterials – relevance and reliability revisited. NanoImpact, 6, 81-89. https://doi.org/10.1016/j.impact.2017.03.004

Hommen, U., Forbes, V., Grimm, V., Preuss, T.G., Thorbek, P., Ducrot, V. (2016). How to use mechanistic effect models in environmental risk assessment of pesticides: case studies and recommendations from the SETAC workshop MODELINK Integr. Environ. Assess. Manag., 12 (1), pp. 21-31, https://doi.org/10.1002/ieam.1704

Kase, R., Korkaric, M., Werner, I., & Ågerstrand, M. (2016). Criteria for Reporting and Evaluating ecotoxicity Data (CRED): comparison and perception of the Klimisch and CRED methods for evaluating reliability and relevance of ecotoxicity studies. Environmental Sciences Europe, 28(1), 7 (14 pp.). https://enveurope.springeropen.com/articles/10.1186/s12302-016-0073-x

Klimisch, H-J., Andreae, M., Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol 25:1–5, https://doi.org/10.1006/rtph.1996.1076

Köhler, H. R., Gräff, T., Schweizer, M., Blumhardt, J., Burkhardt, J., Ehmann, L., Hebel, J., Heid, C., Kundy, L., Kuttler, J., Malusova, M., Moroff, F. M., Schlösinger, A. F., Schulze-Berge, P., Panagopoulou, E. I., Damalas, D. E., Thomaidis, N. S., Triebskorn, R., Maletzki, D., von der Ohe, P. C. (2023). LogD-based modelling and ΔlogD as a proxy for pH-dependent action of ionizable chemicals reveal the relevance of both neutral and ionic species for fish embryotoxicity and possess great potential for practical application in the regulation of chemicals. Water Research, 235(February). https://doi.org/10.1016/j.watres.2023.119864

Kroll, A. Casado-Martinez, C., Junghans, M. (2023). EQS - Vorschlag des Oekotoxzentrums für: Mecoprop-P. Dübendorf (CH): Swiss Centre for Applied Ecotoxicology; 50 pp. https://www.ecotoxcentre.ch/media/nprb2nvk/mecoprop-p_eqs-dossier-2023.pdf

Kroll, A. et al., (2024). Aquatic thresholds for ionisable substances, such as diclofenac, should consider pH-specific differences in uptake and toxicity. Science of The Total Environment, 908, p. 168222. https://doi.org/10.1016/j.scitotenv.2023.168222

Lahr, J., Arts, G., Duquesne, S., Mazerolles, V., de Jong, F., Moermond, C., van der Steen, J., Alalouni, U., Baujard, E., van den Berg, S., Buddendorf, B., Faber, M., Mahieu, K., Montforts, M., Smit, E., van Spronsen, R., Swarowsky, K., Chaton, P.F., Foldrin, J., Lambin, S., Pieper S. (2023) Proposal of critical appraisal tools for the evaluation of ecotoxicology studies. EFSA Supporting Publications. Volume 20, Issue 3, https://doi.org/10.2903/sp.efsa.2023.EN-7787

Larras, F., Beaudouin, R., Berny, P., Charles, S., Chaumot, A., Corio-Costet, M.-F., Doussan, I., Pelosi, C., Leenhardt, S., Mamy, L. (2022) A meta-analysis of ecotoxicological models used for plant protection product risk assessment before their placing on the market. Sci Total Environ., 844, Oct, 157003. https://doi.org/10.1016/j.scitotenv.2022.157003

Merrington, G., Nowell, L.H., Peck, C. (2024) An introduction to Criteria for Reporting and Evaluating Exposure Datasets (CREED) for use in environmental assessments. Integrated Environmental Assessment and Management, https://doi.org/10.1002/ieam.4899

Moermond, C.T.A., Kase, R., Korkaric, M., & Ågerstrand, M. (2016). CRED: criteria for reporting and

evaluating ecotoxicity data. Environmental Toxicology and Chemistry, 35(5), 1297-1309. https://doi.org/10.1002/etc.3259

Moermond, C., Beasley, A., Breton, R., Junghans, M., Laskowski, R., Solomon, K., & Zahner, H. (2017). Assessing the reliability of ecotoxicological studies: an overview of current needs and approaches. Integrated Environmental Assessment and Management, 13(4), 640-651. https://doi.org/10.1002/ieam.1870

OECD. (2011). Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test. OECD Guidelines for the Testing of Chemicals, Section 2. OECD Publishing, Paris, https://doi.org/10.1787/9789264069923-en

OECD. (2012). Test No. 211: Daphnia magna Reproduction Test. OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, https://doi.org/10.1787/9789264185203-en

OECD. (2013). Test No. 210: Fish, Early-life Stage Toxicity Test. OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, https://doi.org/10.1787/9789264203785-en

Perillon C., Feibicke M., Sahm R., Kusebauch B., Hönemann L., Mohr S. (2021). The auxin herbicide mecoprop-P in new light: Filling the data gap for dicotyledonous macrophytes. Environmental Pollution 272, 116405. https://doi.org/10.1016/j.envpol.2020.116405

Case study D. Submission and incorporation of peer reviewed literature for pesticide approval

Developed by the European Food Safety Authority (EFSA)
Case study authors: Maria Arena, Fulvio Barizzone, Anna Federica Castoldi, and Simone Rizzuto (EFSA)

D.1. Introduction to the European Food Safety Authority (EFSA) Guidance

Under Regulation (EC) No. 1107/2009[73], all Applicants submitting dossiers for the approval (or re-approval) of active substances of Plant Protection Products (PPPs) must submit a systematic review of the scientific peer-reviewed open literature on the active substance, its relevant metabolites or Plant Protection Product (PPP) containing the active substance, dealing with side-effects on health, the environment and non-target species and published within the last ten years before the date of submission of the dossier. The literature search should be updated within 6 months before the date of submission of the dossier.

The EFSA Guidance on "Submission of scientific peer-reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009" (EFSA, 2011) provides instructions on how to identify and select publicly available scientific literature and how to report it in a dossier. The Guidance is consistent with the fundamental principles of systematic review and aims at minimising bias in the identification, selection, and inclusion of peer-reviewed open literature in dossiers.

The EFSA Guidance is based on the three initial steps of the systematic review process, namely:

1. Clarification of the objective of the review of the scientific literature and setting the criteria for study relevance to the dossier
2. Use of searching tools to find scientific literature on the subject
3. Selection of relevant scientific literature for inclusion in the dossier

It also requires the clear and systematic reporting of the searching and study selection processes followed by an Applicant and it is compatible with existing OECD Guidance documents for the preparation of active substances dossiers (OECD, 2005, 2006).

The intended users of the EFSA Guidance are:

- Applicants submitting dossiers, under Regulation (EC) No 1107/2009, for the approval of new active substances of PPPs and the re-approval of active substances already authorised in PPPs on the European Union (EU) market
- Competent Authorities of the EU Member States in charge of evaluating the submitted dossiers and of drafting either the Draft Assessment Reports (DARs) for new active substances or the Renewal Assessment Reports (RARs) for already authorised active substances
- EFSA, responsible for drawing conclusions on the safety of the active substances and PPPs

This case study aims at describing the workflow followed to submit peer-reviewed literature in accordance with Regulation (EC) No 1107/2009 and the EFSA 2011 Guidance (EFSA, 2011) for the (re-)approval of active substances of PPPs. The focus of this case study is on the process of literature search, review, and

[73] Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. OJ L 309, 24.11.2009, p. 1. Avalilable online: https://eur-lex.europa.eu/eli/reg/2009/1107/oj/eng

reporting, whereas the evaluation of the review outcome and of the study appraisals is beyond the scope of this work.

Examples of submission of scientific peer-reviewed open literature as performed by the Applicant will be provided by using the RARs of the active substances Fenamiphos[74] (EFSA et al., 2019) and Imidacloprid[75] (EFSA, 2014). These examples will showcase the Applicants' interpretations of the EFSA Guidance (EFSA, 2011) and Regulation (EC) No. 1107/2009. It should be noted that the presented case studies are included for illustrative purposes only. They are based on literature reviews, which were conducted in 2013 and may not necessarily reflect the best interpretation of the EFSA Guidance and Regulation (EC) No. 1107/2009. The methodology applied and the results obtained through the systematic literature search reflect only the Applicant's work and not the outcome of the EFSA and EU Member States peer-review process.

D.2. Clarification of the objective of the review of the scientific literature and setting the criteria for study relevance to the dossier

The scope of the literature review is defined *a priori* by the data requirements set in Regulation (EC) No 1107/2009 which represents the review questions. The Regulation establishes data requirements for chemical and microbial active substances as well as for PPPs based on chemical or microbiological preparations.

The list of data requirements given in the Regulation on chemical active substances is reported below:

a. Toxicological and toxicokinetic studies (OECD code: IIA 5)
b. Residues in or on treated products, food, and feed (metabolism and residues data) (OECD code: IIA 6)
c. Fate and behaviour in the environment (OECD code: IIA 7)
d. Ecotoxicological studies (OECD code: IIA 8)
e. Other data requirements for which information may have a direct or indirect effect on overall risk assessment (OECD code: IIA 1- IIA2 -IIA 3 - IIA 4) (only data requirements under these points having a direct impact on the risk assessment need to be considered)

Considering the data requirements, a list of relevance criteria is drawn. Studies relevant for the inclusion into the dossier are those that inform one or more data requirement(s), including hazard identification, hazard characterisation and exposure assessment of the active substance under assessment, its relevant metabolites, or PPP containing the active substance. It should be considered that the selection of relevance criteria is generally an iterative process. It should begin with a clear analysis of the different components characterising the data requirements to set the characteristics of the relevant studies. A useful strategy is to carry out a preliminary search of the literature to test the applicability of the relevance criteria on a sub-set of summary records or full text documents and then refine if necessary. Examples of fundamental components for (eco)toxicological data are the test species, the test material, the use of different doses/concentrations, and the specific endpoints of interest. Relevant studies can be considered those appropriately addressing these components, i.e., studies that present a well-defined test material (including its purity and impurity profile); tests relevant for the mammalian toxicological/environmental assessment; a number of animals/organisms per group sufficient to establish statistical significance; several dose/treatment levels tested (e.g., at least 3) and a negative control to establish a dose-response relationship; a relevant route of administration in terms of risk assessment (e.g., oral, dermal or by

[74] https://open.efsa.europa.eu/study-inventory/EFSA-Q-2016-00278

[75] https://open.efsa.europa.eu/study-inventory/EFSA-Q-2014-00028

inhalation for mammalian toxicology, uptake from water for aquatic ecotoxicology), and a description of the observations, examinations, analysis performed, or necropsy/histopathology data.

In cases where a study has not been conducted in accordance with Good Laboratory Practice (GLP), this does not automatically imply that the study is not relevant.

The EFSA Guidance (EFSA, 2011) requires screening of the identified publications at two levels, applying relevance criteria which have been previously defined: first using the Title and Abstract to exclude summary records which are obviously irrelevant (called Rapid Assessment) and then using the Full Text (called Detailed Assessment).

Fenamiphos

Criteria for study relevance for the dossier

a. In the case of fenamiphos, the Applicant considered as relevant the studies informing the following data requirements:
b. *"Toxicological and metabolism studies on the active substance*
c. *Residues in or on treated products, food, and feed*
d. *Fate and behaviour in the environment*
e. *Other data requirements for which information may have a direct or indirect effect on overall risk assessment (only data requirements under these points having a direct impact on the risk assessment need to be considered)"*

Moreover, they clarified that "Data requirements on ecotoxicology were deemed non relevant as the ecotoxicology section has been waived (no exposure to non-target organisms under the conditions of use – permanent greenhouses and application by drip irrigation in Southern Europe only)."

Rapid assessment criteria

To perform the Rapid assessment, the Applicant used the following criteria to classify references as being non-relevant:

- *"Efficacy*
- *Analytical method*
- *Ecotoxicity*
- *Studies on a molecular level, which cannot be related to risk assessments*
- *Non-EU monitoring studies*
- *Publications in non-EU language without English abstract*
- *Abstract refers to a conference contribution and does not contain data, full text not available*
- *Target organisms*
- *Soil remediation and pollutants*
- *Stereochemistry (as EU guidance is not yet agreed)"*

Detailed assessment criteria

To perform the Detailed assessment, the Applicant used the following criteria to classify references as being non-relevant:

- *"Test substance is not fenamiphos*

- *Study design/test system not adequate*
- *Study design/test system not relevant to EU data requirements*
- *Test system not relevant to representative uses/GAPs*
- *No endpoint can be derived*
- *Observations (e.g., toxicological) are not attributable to a specific substance*
- *Observations cannot be transferred into an endpoint*
- *The information is already available in other peer reviewed articles"*

Imidacloprid

Criteria for study relevance for the dossier

Imidacloprid was included in Annex I to Directive 91/414/EEC on 1 August 2009 by Commission Directive 2008/116/EC[76], and has been deemed to be approved under Regulation (EC) No 1107/2009. A systematic literature review had been conducted by the Applicant at the time of the submission of the approval dossier in 2003.

In January 2014, the European Commission requested EFSA to perform a re-evaluation of imidacloprid and provide conclusions as regards the risk to aquatic organisms following consideration of a new study on the toxicity of imidacloprid on aquatic organisms (Roessink et al., 2013). This publication reported on the acute and chronic toxicity of imidacloprid to non-standard invertebrate species, some of them, namely mayflies, being found more sensitive than standard invertebrate species.

EFSA requested the Applicant to conduct a systematic literature review in accordance with the EFSA guidance on the submission of scientific peer-reviewed open literature (EFSA, 2011). EFSA specified the data needed to support its mandate for the aquatic risk assessment as follows:

"A systematic review of scientific literature on all studies concerning the risk assessment on aquatic organisms, conducted in accordance with the EFSA guidance on the submission of scientific peer-reviewed open literature and the EFSA guidance on application of systematic review methodology to food and feed safety assessment to support decision making".

The systematic review of scientific literature on studies concerning the risk assessment of imidacloprid and its metabolites for aquatic organisms was performed by the Applicant on the following review question:

"What are the acute and/or chronic effects of imidacloprid and/or its metabolites (imidacloprid-desnitro (M09), imidacloprid-urea (M12), 6-chloronicotinic acid (M14) and imidacloprid-desnitro-olefine (M23)), in aquatic organisms such as fish, amphibians, aquatic invertebrates, aquatic plants and/or sediment dwelling invertebrates?"

Rapid assessment criteria

To perform the Rapid assessment, the Applicant decided to screen the references with a single reviewer on the basis of relevant terms in the titles and abstracts. Manual selection was preferred over search by electronic key terms.

Relevant terms focused on aquatic species to meet the specific request from EFSA on aquatic risk assessment and included, but were not limited to, expressions such as:

[76] Commission Directive 2008/116/EC of 15 December 2008 amending Council Directive 91/414/EEC to include aclonifen, imidacloprid and metazachlor as active substances.

"Americamysis, amoeba, amphibian, amphipod, aquatic, aquaticus, Asellus, bacterium, bahia, batrachus, benthic, benthos, bioaccumulation, bioconcentration, biomonitoring, Brachydanio, Branchiopoda, brine, Bufo, caddisfly, Callinectes, carp, carpio, catfish, Ceriodaphnia, Channa, Cheumatopsyche, Chironimidae, Chironomus, cladoceran, Clarias, Copepoda, Copera, crabs, crustacea, crustacean, Cyprinus, Danio, Daphnia, Desmodesmus, Dictyostelium, dubia, ecologic, ecological, eco-risk, ecosystem, ecosystems, ecotoxicity, ecotoxicological, embryo, embryogenesis, emergent, environment, environmental, fish, fossarum, freshwater, frog, frogs, Gammarus, Hyalella, Hydropsychidae, immobilization, invertebrate, invertebrates, Labeo, larvae, larval, latipes, lentic, lethal, lethality, Libellulidae, limnocharis, lotic, Lumbriculus, macroinvertebrate, Macro-invertebrate, macrozoobenthos, magna, mayflies, mayfly, medaka, mesocosm, mesocosms, microalgae, microcosm, microcosms, microcrustacean, microorganisms, model, modeling, models, mollusc, monitoring, mortality, mossambicus, nontarget, non-target, Odonata, Oligochaete, oligochaetes, Oncorhynchus, Oreochromis, Oryzias, Ostracoda, paddy, Palaemonetes, phytotoxicity, pond, population, predators, Prosobranchia, pugio, pulex, pulse-exposure, punctatus, Rana, reproduction, rerio, riparian, riparius, risk, riverine, roeseli, rohita, runoff, Salmo, Salmon, Salmonids, Salmons, sediment, shrimp, Simulium, snail, snails, stream, sublethal, sub-lethal, subspicatus, survival, tadpole, tadpoles, tentans, tilapia, toxicity, Trichoptera, Tubifex, variegatus, water, watershed, waterways, xenobiotic, Zebrafish, zoocenoses, zooplankton, Zygoptera."

Detailed assessment criteria

To perform the Detailed assessment, the Applicant applied a two-step approach.

In the first step, the Applicant identified the specific criteria to assess the "adequacy" of the identified literature in relation to the specific data request from EFSA on aquatic risk assessment, and the methodology used for this selection.

The following criteria were used to classify the literature as "adequate", and thus subject to further review, or "inadequate":

- *"Method development without unique endpoints*
- *Studies on molecular level, which cannot be related to risk assessment*
- *Monitoring studies*
- *Abstracts refers to a conference contribution and does not contain data, full text not available*
- *Not relevant due to missing information: studies with target organisms"*

In the second step, the Applicant carried out the final assessment for adequacy, based on the principle that the available literature should provide comparable information requirements as the standard regulatory tests. The following criteria were used to classify references as being non-relevant:

- *"Target substance not a test item*
- *Conversion into units useful for risk assessment not possible*
- *Study design/test system not sufficiently described*
- *Study design /test system not adequate*
- *Study design/test system not relevant to EU data requirements*
- *Test method does not cover the right targets*
- *Findings not related to a certain test system*
- *No endpoint can be derived*
- *Observations are not attributable (i.e., ecotox) to a specific substance*
- *Observations cannot be transferred into an endpoint.*
- *The information is already available in other peer reviewed articles."*

D.3. Searching for scientific literature

This step involves the development of a search strategy (combinations of search terms) and identification of information sources that must be searched to retrieve as many relevant studies as possible.

According to the EFSA Guidance (EFSA, 2011), the Applicants are requested to perform an extensive literature search and to report it in detail, following a template provided in the Guidance, the following information:

- The bibliographic databases used in the literature review
- The justification for choosing the databases
- The date of the search – online search service used when applicable (e.g., Scientific Technical Network or PROQUEST)
- The time window of the literature search and the frequency of updates
- The search strings for each engine/tool/database. All the search terms should be reported. For all the search terms, the fields that were searched in a database must be indicated. For example [All fields], [MeSH terms], [Title and Abstract]
- The possible filters that were applied to the search and, in case they were, a justification for their application.

Fenamiphos

An example of information reported on bibliographic databases used in the case of fenamiphos is available below:

Table D.1. Databases searched from 01/01/2002 to 30/12/2013

DATABASES	Frequency of updates
MEDLINE	Daily and annual reload
AGRICOLA	Monthly
PASCAL	Closed file (1/2/2015)
CABA	Daily
BIOSIS	Weekly
TOXCENTER	Weekly
CHEMICAL ABSTRACTS (HCAPLUS)	Daily
PQSCITECH	Weekly
EMBASE	Daily

The search strategy was based on a single concept search strategy, where the input parameters reported for the literature search were as follows:

Table D.2. Input parameters for the database search for fenamiphos and its metabolites

Substance name:	Fenamiphos
Known synonyms:	(IUPAC) Ethyl 4-methylthio-m-tolyl isopropylphoshoroamidate (CA) Phosphoramidic acid, (1-methylethyl)-, ethyl 3-methyl-4-methylthio)phenyl ester
EC number:	244-848-1

CAS number:	222224-92-6
Comments:	active substance
Substance name:	**Fenamiphos sulphone**
Known synonyms:	(IUPAC)N-[ethoxy-(3-methyl-4-methylsulfonylphenoxy)phosphoryl]propan-2-amine (CA) Phosphoramidic acid, (1-methylethyl)-, ethyl 3-methyl-4-(methylsulfonyl)phenyl ester
EC number:	
CAS number:	31972-44-8
Comments:	relevant metabolite
Substance name:	**Fenamiphos sulfoxide**
Known synonyms:	(IUPAC) N-[ethoxy-(3-methyl-4-methylsulfinylphenoxy)phosphoryl]propan-2-amine (CA)Phosphoramidic acid, (1-methylethyl)-, ethyl 3-methyl-4-(methylsulfinyl)phenyl ester
EC number:	
CAS number:	31972-43-7
Comments:	relevant metabolite

As regards to the filter, the Applicant declared:

"For the search of fenamiphos and its metabolites fenamiphos sulphone and fenamiphos sulfoxide, no keyword filter was used."

Imidacloprid

A time scale of 10 years prior to the date of the request from EFSA on aquatic risk assessment was initially considered by the Applicant (*i.e.*, 2003-2014) in accordance with the requirements of Regulation (EC) No 1107/2009. EFSA however requested the Applicant to cover the period from 1993 to 2003 due to submission of the imidacloprid Annex I inclusion dossier in 2003. Database access was obtained via the STN online database.

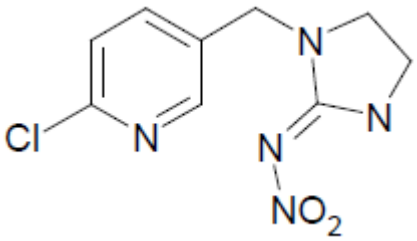
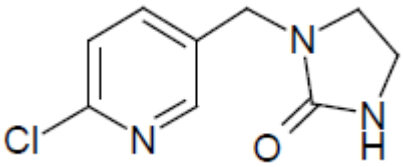
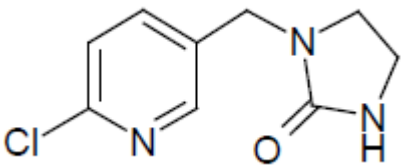
Table D.3. Databases searched from 01/1993 to 01/2014

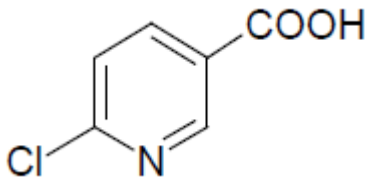
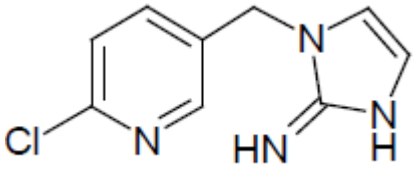
DATABASES	Date of literature search
AGRICOLA	08/01/2014
BIOSIS	22/01/2014
CABA	22/01/2014
CHEMICAL ABSTRACT	27/01/2014
DERWENT DRUG FILE (DRUGU)	22/01/2014
EMBASE	27/01/2014
ESBIOBASE	27/01/2014
IPA	24/01/2014
MEDLINE	22/01/2014
PASCAL	27/01/2014
PQSISCITECH	13/01/2014
REGISTRY	27/01/2014

SCISEARCH	27/01/2014

The search strategy was based on a single concept search strategy, where the input parameters reported for the literature search were as follows:

Table D.4. Input parameters for the database search for imidacloprid and its metabolites

Substance name:	**Imidacloprid**
IUPAC name:	1-(2-Chloro-5-pyridylmethyl)-2-(N-nitroimino)imidazolidene
CAS number:	138261-41-3
Reason for inclusion:	Parent substance
STN query:	(138261-41-3 OR 105827-78-9 OR "1-(2-CHLORO-5-PYRIDYLMETHYL)-2-(NNITROIMINO) IMIDAZOLIDINE" OR "1-(6-CHLORO-3-PYRIDYLMETHYL)-NNITROIMIDAZOLIDIN-2-YLIDENEAMINE" OR "1-[(6-CHLORO-3-PYRIDINYL)METHYL]-4,5-DIHYDRO-N-NITRO-1H-IMIDAZOL-2-AMINE" OR AE-F 106464 OR AEF 106464 OR AEF106464 OR (ADMIRE OR GAUCHO OR MARATHON OR GENESIS OR COMMANDO OR PREMISE OR ALIAS OR MERIT OR PASADA)(W)(RTM OR TM OR R) OR CONFIDOR OR IMIDACLOPRID OR TRIMAX OR PROVADO) AND PY>1992 NOT P/DT AND any keyword listed in appendix B
Molecular structure:	
Substance name:	**Imidacloprid Desnitro (M09)**
IUPAC name:	1-[(6-Chloropyridin-3-yl)methyl]imidazolidin-2-imine
CAS number:	1155875-74-3
Reason for inclusion:	Relevant metabolite
STN query:	115970-17-7 OR IMIDACLOPRID(A)DESNITRO OR "1-[(6-CHLOROPYRIDIN-3-YL)METHYL]IMIDAZOLIDIN-2-IMINE") and PY>1992 NOT P/DT
Molecular structure:	
Substance name:	**Imidacloprid Urea (M12)**
IUPAC name:	1-[(6-Chloropyridin-3-yl)methyl]imidazolidin-2-one
CAS number:	120868-66-8
Reason for inclusion:	Relevant metabolite
STN query:	(120868-66-8 OR IMIDACLOPRID-UREA OR "1-[(6-CHLOROPYRIDIN-3-YL)METHYL]IMIDAZOLIDIN-2-ONE") and PY>1992 NOT P/DT
Molecular structure:	
Substance name:	**6-Chloronicotinic acid (6-CNA, M14)**
IUPAC name:	6-Chloronicotinic acid

CAS number:	5326-23-8
Reason for inclusion:	Relevant metabolite
STN query:	(5326-23-8 OR NSC 277 OR 6-CHLORONICOTINIC ACID OR 2-CHLORO-5-PYRIDINECARBOXYLIC ACID OR 3-CARBOXY- 6-CHLOROPYRIDINE OR 6-CHLOROPYRIDIN-3-CARBOXYLIC ACID) AND PY>1992 NOT P/DT
Molecular structure:	
Substance name:	**Imidacloprid-desnitro-olefine (M23)**
IUPAC name:	1-[(6-Chloropyridin-3-yl)methyl]-1,3-dihydro-2H-imidazol-2-imine
CAS number:	187022-17-9
Reason for inclusion:	Relevant metabolite
STN query:	(187022-17-9 OR "1-[(6-CHLOROPYRIDIN-3-YL)METHYL]-1,3-DIHYDRO-2HIMIDAZOL-2-IMINE") and PY>1992 NOT P/DT
Molecular structure:	

Note: PY identifies the year of publication, e.g., > 1992 articles published from 1992 onwards. In addition, in order to exclude patents (document types not considered to be subjected to a peer-review process) search terms were combined with the search order “NOTP/DT”.

D.4. Selecting relevant scientific studies and reporting the results

According to the EFSA Guidance, following the initial removal of any duplicate reference retrieved, the remaining references should be assessed for relevance by applying the relevance criteria that have been previously defined in the initial step (see Section D.2).

Finally, the document prescribes the reporting of the following information concerning the selection of studies according to specific templates:

1. The results of the selection process for each data requirement or group of data requirements searched
2. A list of the bibliographic references, in a format exportable to reference management software, for all relevant studies and for studies whose relevance remains unclear (i.e., the studies which were not excluded after the detailed assessment of the full-text documents), ordered by data requirement
3. A list of the bibliographic references, in a format exportable to reference management software, for all relevant studies and for studies whose relevance remains unclear (i.e., the studies which were not excluded after the detailed assessment of the full-text documents), ordered by first author
4. A list of the bibliographic references, in a format exportable to reference management software, for all studies excluded from the dossier after detailed assessment of full-text documents for relevance, with justification for their exclusion

Examples of information provided by the Applicants are reported in the following sections.

Fenamiphos

Results of the selection process

In the case of fenamiphos the following information was reported.

Table D.5. Results of the study selection process for fenamiphos

Summary of the review	n
Total number of summary records retrieved from both searches	1102
Total number of summary records retrieved after removing duplicates from all database searches	1000
Number of summary records excluded after rapid assessment for relevance (by title/abstract)	917
Number of studies excluded from the risk assessment after detailed assessment of full-text documents	62*
Number of studies not excluded for relevance after detailed assessment (i.e., relevant studies and studies of unclear relevance)	10
Number of studies included in the dossier as supporting information	7
Number of relevant and reliable studies (Klimisch criteria 1-2) identified by the literature search and appraisal process	0

Note: *According to the number of not excluded studies (10) this number should be 73.

The following information was given to complement the reporting of the study selection process:

"This process identified a total of 10 relevant studies. One out of the ten studies was considered relevant and possibly of use in the risk assessment but was found to be unreliable after detailed evaluation (Klimisch score of 3). Nine studies were considered relevant and included in the dossier but not assessed for reliability as they were not standard studies.

These references are used as supplementary information to EU Chemical Active (CA) and Chemical Product (CP) data points as presented in Table 4."

Reliability assessment was carried out by applying (Klimisch et al., 1997) criteria only on studies that were considered clearly relevant to the risk assessment, that is a single study that was assigned a Klimisch Code 3 (not reliable). Two additional studies out the 10 studies selected for either relevance or unclear relevance were excluded from the RAR, as they were concluded as not relevant. For the remaining 7 relevant studies, reliability was instead not assessed, because of their "non-standard" status. These results were used as supplementary information and included in the form of a narrative summary in the RAR's sections they specifically referred to (e.g., 'studies on absorption, distribution, metabolism and excretion in mammals'; 'endocrine disrupting properties', etc.).

Specific limitations related to the use of Klimisch criteria and more general considerations on the appraisal of reliability of non-standard studies are reported in Section D.5.

List of the bibliographic references included in the dossier

Table D.6. Examples of bibliographic references for relevant and unclear studies related to the fenamiphos application

Data requirement (indicated by the corresponding CA[a] and CP[b] data point)	Author(s)	Year	Title	Source
5.1.1[c]	Moser VC, Padilla S	2011	Esterase metabolism of cholinesterase inhibitors using rat liver *in vitro*	Toxicology 281 (2011) 56– 62
5.8.3[d]	Kojima H, Katsura E, Takeuchi S, Niiyama K, Kobayashi K	2004	Screening for estrogen and androgen receptor activities in 200 pesticides by *in vitro* reporter gene assays using chinese hamster ovary cells	Environmental Health Perspectives, Volume 112, Number 5 April 2004

Note: (a) CA= Chemical Active (b) CP= Chemical Product (c) Data point of 'studies on absorption, distribution, metabolism and excretion in mammals'/'absorption, distribution, metabolism and excretion by oral route' (d) Data point of 'endocrine disrupting properties'.

List of the bibliographic references excluded from the dossier after detailed assessment with justification for their exclusion

Table D.7. Examples of the publications excluded from the risk assessment after detailed assessment of full-text documents related to the fenamiphos application

Author(s)	Year	Title	Source	Reason(s) for not including publication in dossier
Baun A, Ledin A, Reitzel LA,Bjerg PL,Christensen TH*[a]	2012	Fenamiphos and related organophosphorus pesticides: environmental fate and toxicology	Water Research Volume 38, Issue 18, November 2004, pages 3845–3858	Limit of detection for fenamiphos analyzed for 0.1µg/L, but not found in the ten leachates
Bjørling-Poulsen M,Raun Andersen Hand Grandjean P[b]	2008	*In vitro* study of the neuropathic potential of the organophosphorus compounds fenamiphos and profenofos: Comparison with mipafox and paraoxon	Environmental Health 2008, 7:50 doi: 10.1186/1476-069X-7-50	Only one mention of fenamiphos in table. Overview of neurotoxicity linked to pesticide exposure

Note: (a) There was a typo in the bibliographic information reported. The correct year is 2004 and the title is "Xenobiotic organic compounds in leachates from 10 Danish MSW landfills-chemical analysis and toxicity tests" (b) There was a typo in the bibliographic information reported. The correct title is *"Potential developmental neurotoxicity of pesticides used in Europe"*

Imidacloprid

Results of the selection process

In the case of imidacloprid the following information was reported.

Table D.8. Results of the study selection process for imidacloprid

Data requirement(s) captured in the search	n
Total number of summary records retrieved after all searches of peer-reviewed literature (excluding duplicates)	6512
Number of summary records excluded from the search results after rapid assessment of relevance	6367
Total number of full-text documents assessed in detail*	145
Number of studies excluded from further consideration at step 2	63
Number of studies excluded from further consideration after detailed assessment for relevance	31
Number of studies not excluded for relevance after detailed assessment (i.e., relevant studies and studies of unclear relevance)	42
Number of studies which could not be evaluated (full text or translation not received before 14 March 2014)	9

Note: * Excluding articles not received after prior cut-off date.

The reliability of information obtained from a report was evaluated using a reliability score. The Applicant applied the criteria of (Klimisch et al., 1997) by using a score system similar to TOXRTool introduced by (Schneider et al., 2009), but adapted for literature on ecotoxicity. Standardised questions assist in the evaluation process by forcing yes/no answers and allocating points accordingly. The overall score suggests then whether the article may be considered as "reliable" (Klimisch Code 1), "reliable with restrictions" (Klimisch Code 2) or "non-reliable" (Klimisch Code 3). Articles recognised as secondary literature are assigned the Klimisch Code 4 ("not assignable"). Articles assigned Klimisch Code 3 may also be used as supportive "weight of evidence" literature.

As a result of the selection process the Applicant listed 42 peer-reviewed studies as part of the body of evidence to be considered for the assessment.

Specific limitations related to the use of Klimisch criteria and more general considerations on the appraisal of reliability of non-standard studies are reported in Section D.5.

List of the bibliographic references included in the dossier

Table D.9. Examples of bibliographic references for relevant and unclear studies related to the imidacloprid application

Annex Point / Reference Number	Author(s)	Year	Title Source (where different from company) Company name, Report No., Date, GLP/GEP status (where relevant), published or not
K IIA 8.2.1 /01	Chen, A.-M.; Wang, J.-H.; Xia, X.-M.; Wang, J.; Zhu, L.-S.; Fan Y.-Y.	2014	Acute toxicity of imidacloprid with different formulation on earthworm and zebrafish. Location: doi:10.11654/jaes.2013.09.008, Journal:Journal of Agro-Environment Science 1758-1763, Volume: 32, Issue: 9, Pages:1758-1763, Year: 2013, Report No.: M-479153-01-2, Edition Number: M-479153-01-2

Annex Point / Reference Number	Author(s)	Year	Title Source (where different from company) Company name, Report No., Date, GLP/GEP status (where relevant), published or not
			GLP/GEP: n.a., published
K IIA 8.2.1 /02	Tisler,T.; Jemec, A.; Mozetic, B.; Trebse, P.	2009	Hazard identification of imidacloprid to aquatic environment. Publisher:Elsevier, Location: doi:10.1016/j.chemosphere.2009.05.002, Journal: Chemosphere, Volume: 76, Issue: 7, Pages: 907-914, Year: 2009, Report No.: M-479105-01-1, Edition Number: M-479105-01-1 GLP/GEP: n.a., published

List of the bibliographic references excluded from the dossier after detailed assessment with justification for their exclusion

Table D.10. Examples of the publications excluded from the risk assessment after detailed assessment of full-text documents related to the imidacloprid application

Author(s)	Year	Title	Source	Reason(s) for not including publication in dossier
Chang, Xiaoli; Zhai, Baoping [Reprint Author]; Wang, Beixin; Sun, Changhai	2009	Effects of the mixture of avermectin and imidacloprid on mortality and developmental stability of Copera annulata (*Odonata: Zygoptera*) larvae.	Biological Journal of the Linnean Society, (JAN 2009) Vol. 96, No. 1, pp. 44-50.	Mixture tested
Chang, Xiao-Li; Zhai, Bao-Ping; Wang, Bei-Xin; Zhou, Yu.	2008	Acute toxicity of four new types of insecticides to the fourth instar larvae of Chironomus flaviplumus Tokunage (Diptera: Chironimidae).	Shengtai Yu Nongcun Huanjing Xuebao, Volume 24, Issue 1, Page 47-50, Publication Year 2008	Test medium was "running water" (not further described/measured); marginal description of results; no chemical analysis performed; the test concentrations and the observed effects cannot be related to imidacloprid as a formulation was tested

D.5. Considerations on reliability assessment

In section 5.4.2, the EFSA Guidance defines reliability as "*the extent to which a study is free from bias and its findings reflect true facts*". Moreover, it highlights that the reliability of studies available in the literature is likely to vary and that, in addition, the reliability of a study depends on the nature of the assessment the study needs to inform. In fact, the same study may be considered unreliable to establish a deterministic endpoint for human toxicity but reliable in the context of a probabilistic assessment in the ecotoxicological field.

The EFSA Guidance does not establish/adopt specific Critical Appraisal Tools (CATs) to assess the reliability of the studies. Rather, it provides a list of possible resources that can be used for that purpose.

In addition, it provides specific considerations related to the assessment of the methodological quality of the studies and it warns against considering compliance with good laboratory practice (GLP) as a guarantee of reliability. Actually, the latter should be assessed only based on the scientific validity of a study. In that respect, GLP studies have strict requirements for the recording and archiving of the raw data, which may be made available to regulators, facilitating the assessment of a study. At the same time, GLP is not synonymous of reliability. On the other hand, deciding on the reliability of a study on the basis of adherence to testing guideline (sometimes in combination with GLP compliance) - as it is recommended when evaluating studies using the Klimisch method (Klimisch et al., 1997) - may exclude a number of papers.

It has to be noted that at the time of the submission of the dossiers used as examples in this case study (i.e., fenamiphos and imidacloprid) both the Criteria for Reporting and Evaluating Ecotoxicity Data (CRED) (Moermond et al., 2016) and the CATs on non-standard ecotoxicity studies (Lahr et al., 2023) had not been published.

Specific examples on how to assess the reliability (and relevance) of studies are reported in Case study B (*Identification of an endocrine disruptor in the EU regulatory context*) and Case study C (*The CRED Method: A transparent and structured method for evaluation of ecotoxicity data used in risk assessment*) in this Guidance Document.

D.6. Final considerations

Although not reporting specific instructions for study appraisal, since its introduction in 2011, the EFSA Guidance has promoted the use of research data in regulatory assessments and their integration with "standard studies" in a WoE approach. The Guidance promotes a systematic approach to search and select studies and report the results of the process. Moreover, in line with the EFSA Guidance on "*Application of systematic review methodology to food and feed safety assessments to support decision making*" (EFSA, 2010), the EFSA Guidance on "*Submission of scientific peer-reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009*" (EFSA, 2011) has triggered at EFSA developments in the appraisal of "non-standard studies" that promoted the use of CATs. For instance, to support the challenging evaluation of non-standard ecotoxicity studies specific CATs were developed (Lahr et al., 2023). Such CATs were developed on the basis of the Criteria for Reporting and Evaluating Ecotoxicity Data (CRED) for assessing reliability and relevance of studies (Moermond et al., 2016). Developments on the same line were also done in areas beyond the one of pesticides e.g., (EFSA, 2015), (EFSA Scientific Committee (SC) et al., 2020).

References

EFSA (European Food Safety Authority). (2010). Application of systematic review methodology to food and feed safety assessments to support decision making. 90 pp. http://www.efsa.europa.eu/en/efsajournal/pub/1637.htm

EFSA (European Food Safety Authority). (2011). Submission of scientific peer-reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009. 1831-4732. 2092 pp. https://doi.org/10.2903/j.efsa.2011.2092

EFSA (European Food Safety Authority). (2014). Conclusion on the peer review of the pesticide risk assessment for aquatic organisms for the active substance imidacloprid. 1831-4732. 3835 pp. https://efsa.onlinelibrary.wiley.com/doi/abs/10.2903/j.efsa.2014.3835

EFSA (European Food Safety Authority), 2015. Tools for critically appraising different study designs, systematic review and literature searches. 65 pp. https://www.efsa.europa.eu/it/supporting/pub/836

EFSA, Abdourahime H, Anastassiadou M, Arena M, Auteri D, Barmaz S, Brancato A, Brocca D, Bura L, Carrasco Cabrera L, Chiusolo A, Civitella C, Court Marques D, Crivellente F, Ctverackova L, De Lentdecker C, Egsmose M, Fait G, Ferreira L, Gatto V, Greco L, Ippolito A, Istace F, Jarrah S, Kardassi D, Leuschner R, Lostia A, Lythgo C, Magrans JO, Medina P, Messinetti S, Mineo D, Miron I, Nave S, Molnar T, Padovani L, Parra Morte JM, Pedersen R, Raczyk M, Reich H, Ruocco S, Saari KE, Sacchi A, Santos M, Serafimova R, Sharp R, Stanek A, Streissl F, Sturma J, Szentes C, Tarazona J, Terron A, Theobald A, Vagenende B, Vainovska P, Van Dijk J, Verani A and Villamar-Bouza L (European Food Safety Authority), (2019). Peer review of the pesticide risk assessment of the active substance fenamiphos. 1831-4732. e05557 pp. https://doi.org/10.2903/j.efsa.2019.5557

EFSA Scientific Committee (SC), More S, Bambidis V, Benford D, Bragard C, Hernandez-Jerez A, Bennekou SH, Koutsoumanis K, Machera K, Naegeli H, Nielsen SS, Schlatter JR, Schrenk D, Silano V, Turck D, Younes M, Fletcher T, Greiner M, Ntzani E, Pearce N, Vinceti M, Ciccolallo L, Georgiadis M, Gervelmeyer A and Halldorsson TI (EFSA Scientific Committee), (2020). Draft for internal testing Scientific Committee guidance on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments. 1831-4732. e06221 pp. https://efsa.onlinelibrary.wiley.com/doi/abs/10.2903/j.efsa.2020.6221

Klimisch HJ, Andreae M and Tillmann U, (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol, 25:1-5. https://doi.org/10.1006/rtph.1996.1076

Lahr J, Arts G, Duquesne S, Mazerolles V, de Jong F, Moermond C, van der Steen J, Alalouni U, Baujard E, van den Berg S, Buddendorf B, Faber M, Mahieu K, Montforts M, Smit E, van Spronsen R, Swarowsky K, Chaton PF, Foldrin J, Lambin S and Pieper S, (2023). Proposal of critical appraisal tools for the evaluation of ecotoxicology studies. EFSA Supporting Publications, 20:7787E. https://doi.org/10.2903/sp.efsa.2023.EN-7787

Moermond CTA, Kase R, Korkaric M and Ågerstrand M, 2016. CRED: Criteria for reporting and evaluating ecotoxicity data. 35:1297-1309. https://doi.org/10.1002/etc.3259

OECD (Organisation for Economic Co-operation and Development), (2005). OECD Guidance for Industry Data Submissions on Plant Protection Products and their Active Substances (Dossier Guidance). Revision 2, May 2005. OECD Environment Directorate. https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/pesticides-and-biocides/guidance-for-industry-data-submissions-on-plant-protection-products-and-their-active-substances.pdf

OECD (Organisation for Economic Co-operation and Development), (2006). OECD Guidance for Industry Data Submissions for Microbial Pest Control Products and their Microbial Pest Control Agents. August 2006. OECD Environment Directorate. https://web-archive.oecd.org/2012-06-

14/115378-43435253.pdf

Roessink I, Merga LB, Zweers HJ and Van den Brink PJ, (2013). The neonicotinoid Imidacloprid shows high chronic toxicity to mayfly nymphs. Environmental Toxicology and Chemistry, 32:1096-1100. https://doi.org/10.1002/etc.2201

Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, Hartung T and Hoffmann S, (2009). "ToxRTool", a new tool to assess the reliability of toxicological data. Toxicol Lett, 189:138-144. https://doi.org/10.1016/j.toxlet.2009.05.013

Glossary of selected terms

The glossary introduces important concepts and terms for the correct interpretation of this Guidance Document and a reference to the section(s) where these are introduced. The intention of the authors was to align terminology used in this Guidance with definitions introduced in existing OECD guidance (e.g., OECD Guidance Document No. 34, (OECD, 2005)). For general terminology not defined here, this Guidance refers to definitions commonly used in existing OECD guidance and other international frameworks (e.g., (EFSA, 2019; WHO, 2021b)).

Assessor: any individual or organisational entity, including public regulatory authorities, registrants and consultants that performs one or more steps of a regulatory assessment workflow (identification, screening extraction, evaluation, and synthesis of available evidence, including research data, (Section 1.2).

Chemical: any substance subject to regulatory assessment, including e.g., natural and man-made, multi-constituents, nanomaterials, as defined in chemical legislation. Please note that there might be some variability across jurisdictions in the definition of substance.

Evaluation tool: tool used by assessors to evaluate reliability and relevance of research data in the form of structured checklists, criteria, or domain-based questions. In some frameworks, evaluation tools are also referred to as “critical appraisal tools”. Evaluation tools designed to assess reliability are also called “Risk of Bias (RoB)” or “reliability assessment tools”, (Sections 2.3 and 3.4).

Data repository: any database or information technology system that supports storage of data and/or metadata associated with research or regulatory activities. Data repositories include bibliographic repositories (e.g., Medline, Scopus, etc.), as well as structured content repositories. These include repositories designed to host data from regulatory submissions or regulatory programmes (e.g., CompTox, ECHA CHEM), domain specific repositories for defined hazard categories/endpoints (e.g., US EPA ECOTOX, EASIS, IPCHEM), or generalist data repositories (e.g., Zenodo, Re3data). Data repositories are often associated with software for interacting with the data. For example, software applications like IUCLID and Health Assessment Workspace Collaborative (HAWC) allow users to enter and retrieve data, (Section 2.4.2, and Annex B).

Expert judgement: the application of knowledge and experience from experts in the evaluation, interpretation, synthesis, and integration of (research) data to reach conclusions, (Section 3.4).

Guideline study: a study that follows a protocol (or protocols) established by a national or international regulatory authority or standardisation body. Examples include OECD Test Guidelines, EU Test Methods, US EPA and FDA Test Guidelines.

Regulatory relevance: core attribute in the consideration of research data in regulatory assessments. It relates to the utility of a given study to provide data for a specific hazard or risk assessment task, in the context of a regulatory framework/process, (Sections 1.4.2 and 1.2).

Reliability: core quality attribute of research data in regulatory assessments. Reliability refers to how a study is designed, performed, and analysed. Assessing reliability requires sufficient reporting of study methods and results, (Sections 1.3.1 and 2.1)

Reporting guidance: refers to guidance used to promote best reporting practices. Reporting guidance can also be referred to as reporting standards or reporting quality tools. Reporting guidance varies

according to the evidence type e.g., epidemiological, animal, *in vitro*. Reporting guidance covers not only reporting of study methods, performance, statistical analyses, and results but also data provenance ("data lineage") and transparency regarding sources of funding, who was involved, and their roles in the research, (Section 2.2).

Reporting template: a structured layout in the form of a document or table designed to report information in specific fields (e.g., OECD Harmonised Templates, OHTs), (Section 2.2).

Reproducibility: The ability of independent researchers or assessors to reach consistent results when repeating a given task. In the context of this Guidance, the term "reproducibility" is used to refer to a study (e.g. epidemiological, experimental, computational), while "consistency" is used for a literature search, a study evaluation, or a regulatory assessment (Sections 2.3, 2.4, and 3.5).

Research data: any scientific data generated in a research context that could potentially inform hazard, exposure, and/or risk assessments of chemicals. The focus of this Guidance is mostly on data that may be used for human health and (eco)toxicity assessments, (Section 1.1 with full definition).

Scientific relevance: scientific relevance relates to the advancement of scientific knowledge on a subject matter. It refers to the extent that a study advances the knowledge about a property or endpoint of interest in a scientific domain e.g., (eco)toxicology, (Section 1.4.2).