External Scientific Report

APPROVED: 4 April 2024
doi: 10.2903/sp.efsa.2024.EN-8761

Inventory of BMR values for BMD analysis

Honoria Ocagli[1], Rebecca Zgheib[2], Dario Gregori[1], Simone Belluco[3], Mauro Dacasto[2], Ileana Baldi[1]

[1]Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova, Italy

[2]Department of Comparative Biomedicine and Food Science, University of Padova, Legnaro, Italy

[3]Department of Food Safety, Istituto Zooprofilattico Sperimentale delle Venezie, Legnaro, Italy

Abstract

This report focuses on creating a repository of Benchmark Response (BMR) values sourced from international databases. The objectives included defining BMRs for specific endpoints and considering various characteristics such as compound characteristics, information on the study (e.g. strains, species and gender), information related to toxicity (target organ, type and mechanism of toxicity) and statistical aspects of the dose-response modelling. The information was retrieved from repositories of risk assessment bodies such as the IRIS-EPA, JECFA, and OpenFoodTox databases. The methodology section details data acquisition, sources pipeline, and software utilized, ensuring transparency and reproducibility. Techniques like web scraping and text mining complemented manual screening to retrieve and interpret JECFA assessments. OpenFoodTox data served as the reference, involving the filtering for Benchmark Dose (BMD) endpoints and mapping variables for a harmonized setting. When filtering IRIS-EPA data, the code selected variables based on mapping information and filtered data from multiple sheets, aligning it with the OpenFoodTox reference. Moreover, the post-processing steps are explained thoroughly, including data import and cleaning. The inventory results in 593 records in a wide format and offers valuable information on BMR values for BMD analysis, facilitating informed decision-making for relevant future applications.

Keywords: Benchmark response, benchmark dose, dose-response model, inventory, OpenFoodTox, IRIS-EPA, JECFA

Question number: EFSA-Q-2024-00197
Correspondence: MESE@efsa.europa.eu

Acknowledgements: We would like to express our gratitude to the secretary staff and trainees who contributed to this project. Their assistance with administrative tasks, data organization, and other supportive roles was invaluable and greatly appreciated. The authors would like to thank the EFSA staff members: Efisio Solazo and José Cortiñas Abrahantes for the support provided to this scientific output

Suggested citation: Ocagli H, Zgheib R, Gregori D, Belluco S, Dacasto M and Baldi I, 2024. Inventory of BMR values for BMD analysis. EFSA supporting publication 2024:EN-8761. 82 pp. doi:10.2903/sp.efsa.2024.EN-8761

ISSN: 2397-8325

Summary

The introduction highlights the importance of defining Benchmark Response (BMR) values in toxicological studies for establishing health-based guidance values (HBGVs). Through the Benchmark Dose (BMD) modelling framework, the report aims to describe the process for creating an inventory of BMRs across various endpoints and substances.

The inventory considers factors influencing BMR selection, such as the substance, target organ, mode of action, type of toxicity, study type, and endpoint used for assessment. It involves mapping evidence from repositories of results provided by risk assessment bodies.

The data and methodologies section outlines the steps taken to create the inventory of BMR values. Firstly, existing information from public databases and reports from risk assessment agencies was retrieved. The primary sources included the OpenFoodTox, IRIS-EPA, JECFA, and CESAR databases, with additional sources like the EFSA Journal, WHO, and FAO consulted for supplementary information.

Data acquisition involved various methods depending on the source type. Reports and PDFs required web scraping, pre-processing, and text mining, while databases were downloaded and filtered for relevant analysis.

The data source pipeline was established to create an inventory of BMR values. It involved simultaneous retrieval of information from various sources, including databases and reports, and filtering based on specific criteria. Missing information was retrieved in two phases: first from designated databases and then from supplemental sources. R software and Docker were used for data management and analysis, while DistillerSR aided in systematic record-keeping and review screening.

The data dictionary provides a comprehensive list of variables and their definitions for the inventory of BMD values, focusing on toxicological information for chemicals. The process involved harmonizing potential variables with existing ones in the EFSA catalogue and matching them among the databases considered. Key variables include information about the source, chemical composition, assessment type, outcomes, study details, and substance toxicity. Moreover, the mapping process aligns variables from original databases with specific variables for the inventory, ensuring a unified dataset.

The following sections detail each of the explored sources. JECFA was established by FAO and WHO and evaluates the safety and functionality of food additives. The data extraction and analysis process from JECFA initially started with web scraping to gather data from JECFA's website due to the absence of accessible APIs. This involved accessing all hosted chemical pages and extracting PDF URLs for toxicological monographs and reports. Next, text mining techniques were applied to extract pertinent information from these PDFs. Records were then filtered based on specific criteria related to BMD assessments.

OpenFoodTox, a database curated by EFSA, is a cornerstone for European food safety risk assessment. The complete dataset was downloaded from the website, with non-genotoxic endpoints undergoing data extraction, while genotoxic endpoints were initially screened at the title/abstract level.

The EPA developed and maintained the IRIS-EPA database, providing valuable insights into the risks associated with these substances.

The CESAR risk assessment database was also consulted. It is managed by Health Canada and holds assessments of existing substances in Canada under CEPA. Accessing CESAR, we filtered substances meeting CEPA's definition of toxicity. Among 192 filtered records, 49 had

BMD assessments, with some overlapping CAS numbers from other sources. Due to limited original content, Health Canada's data weren't further scrutinized.
After the data extraction phase, an expert reviewed the database for accuracy, with some data directly uploaded to DistillerSR and others manually entered. Post-processing efforts addressed variations in variable levels across different databases, conducted in DistillerSR and R. The inventory can be downloaded from DistillerSR. In R, data cleaning and final variable selection were performed, harmonising endpoints to ensure consistency and comparability across studies. We transitioned the dataset from a wide to a long format for toxicity types and target organs to enhance further analyses. In the wide format, toxicity types and organs were separate columns; in the long format the data into each row representing a unique combination of toxicity type and target organ.
The inventory provides a comprehensive information on studies on BMRs for specific endpoints in toxicological studies and consists of 593 records in the wide format. Notable findings include the predominance of the Rat species and the Fischer 344 strain, with variations in gender distribution across databases. Dose-response models and endpoints varied, with "incidence" and "decrease" being standard endpoints. The liver was frequently targeted, and carcinogenicity was the most prevalent toxicity level. Additionally, the conversion of BMR values is provided for continuous endpoints embracing a SD-based definition.

In conclusion, a comprehensive search created an inventory of BMRs for various endpoints, types of studies, assessed substances, and other characterization aspects of the assessment suitable for further quantitative data analysis.

Table of Contents

Abstract 1
Summary 3
1 Introduction 7
1.1 Background and terms of reference as provided by the requestor 7
1.1.1 Inventory of BMR values for BMD analysis 7
1.1.2 Background, description of the activities and methodological support needed 7
1.2 Interpretation of the Terms of Reference 10
1.3 Objectives 11
2 Data and Methodologies 11
2.1 Data 11
2.2 Data Source Pipeline 13
2.2.1 Data acquisition 14
2.2.2 Import into DistillerSR 14
2.2.3 Retrieving missing information 15
2.2.4 Software 15
3 Data Dictionary Definitions 16
3.1 Data Source Information 16
3.2 Chemical information 25
3.3 Assessment Information 25
3.4 Outcomes 25
3.5 Study information 26
3.6 Mapping 27
4 JECFA 28
4.1 Web scraping 31
4.1.1 Phase 1 - FAS 32
4.1.2 Phase 1 - TRS 32
4.2 Text Mining 33
4.2.1 Phase 2 – Text mining on TRS (when a FAS is not available) 33
4.2.2 Phase 2 – Text Mining on FAS 34
4.2.3 Description of JECFA records with text mining on FAS/TRS 34
4.3 Filtering of JECFA records 37
4.3.1 Title/abstract and full-text screening 37
5 OpenFoodTox 40

5.1 Filtering 40
5.2 Recap of the mapping 41
6 IRIS-EPA 42
6.1 Filtering 43
6.2 Recap of the mapping 43
7 CESAR 44
7.1 Access to the source 45
8 Post processing 47
8.1 Download from DistillerSR 47
8.2 Data Import and Cleaning in R 48
8.2.1 Harmonization of endpoints 48
8.2.2 Toxicity 48
9 Inventory 48
9.1 Sources characteristics 48
9.2 Population characteristics 49
9.3 Dose response models 49
9.4 Toxicity and target organs 50
9.5 BMR conversion 51
10 Conclusions 52
References 52
Abbreviations 54
Appendix A – List of variables for the Inventory 56
Appendix B – Cesar 59
Appendix C – Descriptive Statistics of the Inventory 68
Annex A – DB matching 79
Annex B – Original OpenFoodTox 79
Annex C – Original IRIS-EPA 79
Annex D – R markdown file to import the DistilleR extractions 79
Annex E – Inventory cleaned in R 79

1 Introduction

The definition of Benchmark Response (BMR) for specific endpoints in toxicological studies is paramount to establish health-based guidance values. Employing the Benchmark Dose BMD modeling framework, this scientific report endeavors to create a comprehensive inventory of BMR, encompassing various endpoints, types of studies, and substances under assessment. The primary focus of this review centers on the diverse BMR values utilized in BMD analysis, emphasizing their significance in discerning minimal changes relative to responses observed in unexposed groups within regulatory, academic, and industrial frameworks.

Central to this inventory is an exploration of BMR considerations, including their potential influences on risk assessment. Various aspects will be examined, such as the nature of the substance under scrutiny, target organ specificity, mode of action, possible types of toxicity types (e.g., genotoxicity, carcinogenicity), and the definition used for the estimation of BMD (whether relative, additional risk, standardized by σ, or any other). In cases where a specific definition was not used, the report computed and presented the derived BMR, offering a comprehensive overview.

Furthermore, this inventory delved into additional factors influencing BMR selection, including study type (acute, subacute, etc.), the experimental design (in vivo, in vitro, observational), and other relevant characteristics essential for chemical classification. By scrutinizing these multifaceted dimensions, this report aimed to contribute valuable insights into the intricate process of defining BMR and their implications in advancing toxicological risk assessment methodologies.

1.1 Background and terms of reference as provided by the requestor

This report is the result of an EFSA procurement titled "Inventory of BMR values for BMD Analysis" – Contract No. OC/EFSA/MESE/2022/03 – CT 02, awarded by the European Food Safety Authority to the Università degli Studi di Padova as a leading partner in consortium with the Istituto Zooprofilattico Sperimentale delle Venezie as a secondary partner.

1.1.1 Inventory of BMR values for BMD analysis

Defining BMR for specific endpoints in toxicological studies that have been analysed using BMD modelling framework to establish health-based guidance values. Create an inventory of BMRs for endpoints, type of studies considered, substance assessed and other characterization of the assessment.

1.1.2 Background, description of the activities and methodological support needed

The benchmark dose (BD as called by Crump, today the term used is BMD) approach was first proposed by Crump in 1984 (CRUMP, 1984). The methodology proposed make use of a dose-response model that can be defined for continuous or quantal responses. In case of quantal endpoints, the definition of the benchmark dose was given as the dose d corresponding to specific value for the "extra-risk" as:

$$\frac{P(d) - P(0)}{1 - P(0)} = BMR$$ Equation 1

The authors stated that can be interpreted as the probability of an effect at dose d given that no effect would have occurred in the absence of the dose. The specific predetermined value of extra risk is what is now called BMR. The definition of the BD/BMD for continuous endpoint given in his paper is somewhat different, called "extra-response", as:

$$\frac{m(d) - m(0)}{m(0)} = \pm BMR_1$$ Equation 2

Being the dose d corresponding to a specified amount of absolute change in the mean relative to the mean in the absence of the dose. This BMD is defined for a specific predetermined value of extra response that would corresponds to the BMR_1 for continuous endpoints. Also, in this paper Crump pointed towards other definitions for the BMD when dealing with continuous type of responses, but not other definitions were provided. Later in 1995 (Crump, 1995), Crump extended the definitions of benchmark dose for continuous endpoints. In this paper Crump introduces the definition based on change in the mean response relative to the standard deviation (σ) assuming homogeneity across dose groups as:

$$\frac{m(d) - m(0)}{\sigma} = \pm BMR_2$$ Equation 3

Actually, in this paper, the term benchmark response or benchmark risk, being the pre-specified risk that will be used to determine the benchmark dose, was used for the first time. The signs reflect the nature of the response, increasing, or decreasing responses define detrimental responses differently. Other definitions of the benchmark dose are also provided in terms of additional risk or changes in response with respect to the response in the absence of a dose. In the past the different definitions presented above have been used to set a reference point (RP), EFSA has always proposed to use the so-called extra-risk/extra-response. Considering Equation 3, by dividing in both sides by $m(0)$,

$$\frac{m(d) - m(0)}{\sigma \cdot m(0)} = \frac{\pm BMR_2}{m(0)}$$

the relation between both BMR's, leading to the same BMD, can be expressed as:

$$\frac{m(d) - m(0)}{m(0)} = BMR_1 = \pm\sigma \cdot \frac{BMR_2}{m(0)}$$

Furthermore the ratio of BMR's can be expressed in terms of the coefficient of variation of the response at dose zero, see below:

$$\frac{BMR_1}{BMR_2} = \pm\frac{\sigma}{m(0)}$$

Similarly, it can be derived from the additional definitions of the BMD, note that they all can be derived if $m(0)$ and σ used to set the RP are known. The way to set the RP is not part of the scope of this review. The focus of this review is on the continuous type of endpoint, for which setting a BMR value to be able to estimate the associated BMD is not straightforward.

In the recently updated guidance it is stated that *when choosing a BMR for continuous data, EFSA recommends a tiered approach:*

- Tier 1: consider whether a biologically relevant BMR is already established (e.g. internationally agreed, previously used by EFSA, etc.) for the endpoint considered and whether the value is still appropriate. Discussion, including challenges and guiding information, related to the derivation of such BMR values can be found in publications of Dekkers, de Heer & Rennen (2001) (Dekkers et al., 2001) and WHO (2020) (WHO, 2020).
- Tier 2: in the absence of an already established BMR, experts should consider whether it is possible to define quantitatively "biologically relevant" to inform the selection of a BMR for the endpoint considered. The BMR may be defined using any of the methods that are available in the literature (e.g. Expert Knowledge Elicitation, which could be informed by e.g. the effect size theory (Slob, 2017), 1SD of the background response (US EPA, 2012), hybrid approach or other definitions), taking biological relevance into account. This tier assumes that a level of adversity can be identified, even though the minimal degree of adversity may not be known. Thus, biologically relevant BMRs may also be represented by a range rather than by a single point.

If it is not possible to provide an argument for a specific biologically relevant BMR (or range of biologically relevant BMRs) for the endpoint considered, this endpoint should not be used to establish a HBGV (see also (WHO, 2020)). In the absence of endpoints with biologically relevant BMRs, the full set of doses used in the experiment could still be used in a sensitivity analysis to investigate the probability that, for several BMR chosen a priori, the BMD value associated to them would be below or above the doses tested. This information could then be further considered in calculation of a range of MOEs. Another possibility could be to use each of the dose tested and calculate the relative change compared to the background response, and then use these relative changes as BMRs to estimate the BMD distribution. This would aid defining the uncertainty associated to each BMD distribution, which in turn would provide insights on the information contained in the dose-response fitted.

Considering what it is stated in the updated BMD guidance, the review should focus on the BMR values used in the benchmark dose analysis to allow for a creation of an inventory of what have been considered as small change in comparisons to the response observed in the unexposed group in the regulatory framework as well as academy and industry. Also, its related justification and explanation should be provided. The BMR inventory should also consider other aspects that might likely influence the selection of the appropriate BMR for the risk assessment under considerations, such as:

- Substance under scrutiny
- Target organ
- Mode of action of the substance under scrutiny
- The type of toxicity the substance under scrutiny could cause (e.g. genotoxicity, carcinogenic, etc.)
- The definition used to estimate the BMD (relative as in Equation 2, additional risk, standardizing by σ as in Equation 3, or any other definition). In case that the definition in Equation 2 is not used, the derived BMR for this definition should be computed and

provided, together with the derivation in case that is different from the example provided above
- Study type (Acute, subacute, etc.)
- Type of experiment conducted (in vivo, in vitro, observational, etc.)
  other characteristics that might be useful when classifying chemicals as well as the endpoint used for the assessment

A data model allowing to transfer the information extracted in the Data collection Framework of EFSA shall be defined and used to extract the data.

Sources of information that could serve as starting point for such review are the following:

- OpenFoodTox database:
  - The database can be downloaded from EFSA knowledge junction under: https://doi.org/10.5281/zenodo.780543 and can be consulted from a MicroStrategy Dashboard: https://www.efsa.europa.eu/en/microstrategy/openfoodtox
- EFSA journal:
  - It contais risk assessment for which BMD analysis were performed, here a link providing the results retrieved from the EFSA journal when using the terms BMD or BMR
- The CompTox Chemicals Dashboard:
  - It provides information regarding chemical compounds and substances of potential interest to environmental scientists. It contains information about substances, associated chemical structures, experimental and predicted physicochemical and toxicity data, bioassay data, exposure data and additional links to relevant websites and applications
- The Global Portal to Information on Chemical Substances:
  - It contains information of chemical substances relevant to regulatory bodies (https://www.echemportal.org/echemportal/page/1)
- The risk assessment publications from WHO and FAO:
  - Where benchmark dose analysis has been performed
- US-EPA risk assessment performed:
  - In which health-based guidance values or reference points were established (https://iris.epa.gov/AdvancedSearch/)

1.2 Interpretation of the Terms of Reference

Risk assessment includes determining the likelihood of a substance having adverse health effects. These assessments frequently evaluate the dose-response and the relationship between exposure and outcomes that have an impact on health (Munns et al., 2003).

The BMD is an approach that involves dose-response modelling to obtain BMDs, that is, dose levels corresponding to specific response levels near the low end of the observable range of the data. Furthermore, the BMD approach incorporates and communicates more information than the No Observed Adverse Effect Level (NOAEL) or Lowest Observed Adverse Effect Level (LOAEL) process commonly used for non-cancer health effects (Dekkers et al., 2001).

Calculating a BMD is directly determined by selecting the BMR. Selecting BMR involves making judgments about the statistical and biological characteristics of the endpoint considered, type of substance, and about the applications for which the resulting BMDs/BMDLs (benchmark dose lower limit) were used.

Therefore, an inventory of BMR data is important for chemical risk assessments for a variety of health effects that may result from exposure of food to chemical agents.

1.3 Objectives

The scientific literature available on BMD analysis and the associated BMR was used to create a repository of BMR. This repository was used to establish health-based guidance values (HBGV) or reference points (RP), taking into consideration other aspects that might likely influence the selection of the appropriate BMR. The choice of the appropriate BMR could be impacted by factors such as the substance under investigation, the target organ, the mode of action of the substance under investigation, the type of toxicity the substance can cause, the definition used to estimate the BMD, the type of study, the type of experiment conducted, the end point used for the assessment, and other characteristics that could be useful when classifying the chemicals (More et al., 2022).

Defining BMR for specific endpoints in toxicological studies analysed using BMD modelling framework to establish HBGVs or RP.

This work aimed to create an inventory of BMR for endpoints, type of studies considered, substance assessed, and other characterization of the assessment. It has covered mapping and categorizing the available evidence from risk assessment bodies' repositories of results.

2 Data and Methodologies

To create the inventory, we have followed the following steps:

- Inventories/Databases:
    - Retrieving and merging existing information in public databases available online through their websites
- Reports/Articles:
    - Retrieving risk assessment information in reports from risk assessment agencies that were not present in the databases

2.1 Data

The starting points for retrieving relevant databases are those specified in the contract and shown in

Table 1: OpenFoodTox database, Integrated Risk Information System-United States Environmental Protection Agency (IRIS-EPA), Joint FAO/WHO Expert Committee on Food Additives (JECFA), and Canada's Existing Substances Assessment Repository (CESAR).

Table 1: Table representing information sources

Information source	Type	Access/reachability	URL
OpenFoodTox	Database	Can be downloaded from EFSA knowledge junction	https://www.efsa.europa.eu/en/microstrategy/openfoodtox
IRIS-EPA	Database	Can be downloaded	https://iris.epa.gov/AdvancedSearch/
JECFA	Reports/PDFs	Website searchable by partial name or CAS number contains summaries of all the evaluations	https://apps.who.int/food-additives-contaminants-jecfa-database/
CESAR	Reports/PDFs	Website searchable by partial name or CAS	https://www.canada.ca/en/sr/srb/sra.html

JECFA, Joint FAO/WHO Expert Committee on Food Additive; CESAR, Canada's Existing Substances Assessment Repository

The CESAR information source was considered to corroborate that if dose response analysis were considered by Health Canada, they would be, in general, retrieved as well from the primary information sources (OpenFoodTox, IRIS-EPA, and JECFA). The idea of including such a source was to explore the coverage of the three primary sources listed above when dose-response analysis is of interest.

The supplemental sources of information mentioned in Table 2, such as the European Food Safety Authority (EFSA) journal, the World Health Organization (WHO), the Food and Agriculture Organization (FAO), were utilized as supplementary sources in OpenFoodTox and JECFA, respectively. These resources were consulted to retrieve missing data and to stay up to date with updates not yet incorporated into the respective databases. When dealing with searchable textual sources, keywords such as BMD and BMR were used, among others, while for textual sources in PDF format, R software v.4.3.2 was used for pre-processing and conducting text mining tasks.

Table 2. Table representing supplemental information sources to consult

		Access

Information source	Type	Reachability	URL
EFSA Journal	Searchable database	Open access online scientific journal	https://efsa.onlinelibrary.wiley.com/journal/18314732
WHO	Searchable database	Open access website providing reports	https://www.who.int/
FAO	Searchable database	Open access website providing reports	https://www.fao.org/home/en

IRIS-EPA, Integrated Risk Information System-Environmental Protection Agency; JECFA, Joint FAO/WHO Expert Committee on Food Additives; CESAR, Canada's Existing Substances Assessment Repository; EFSA, European Food Safety Authority; CAS number, Chemical Abstracts Service Number

2.2 Data Source Pipeline

The pipeline approach outlined in **Error! Reference source not found.** has served as a guide for creating the inventory. Sources were simultaneously consulted, with a focus on BMD (BMDL), benchmark concentration (BMC) or benchmark concentration lower limit (BMCL) assessments. This means applying specific filters, such as checking if variables (e.g., "reference point": present in the OoenFoodTox database, sheet ENDPOINTSTUDY_KJ and in the IRIS-EPA database, sheet RfD Toxicity Values/RfC Toxicity Values) are equal to BMD or BMDL, for sources already in a database format. For textual sources, a list of specific keywords was employed.

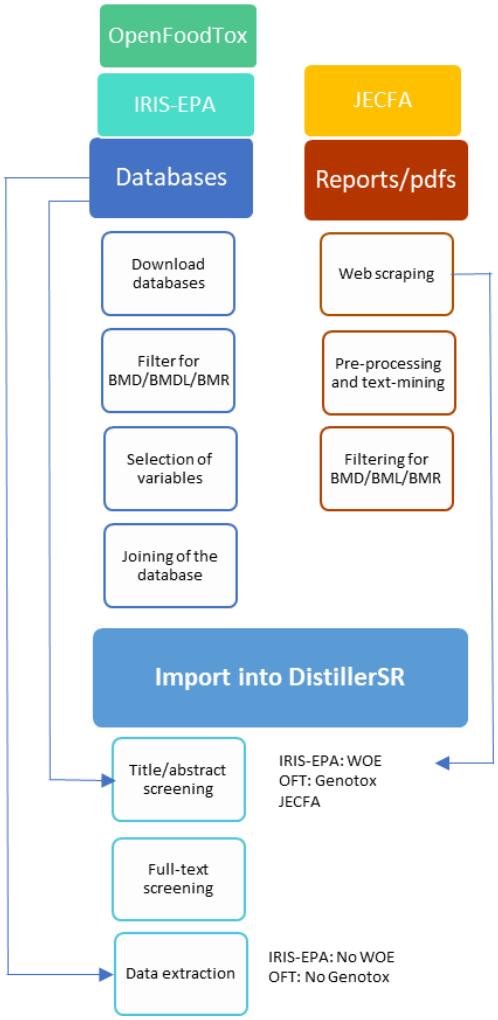

Figure 1. Data source pipeline

Potential variations were expected in the substances covered by the four sources. However, despite these disparities, all substances were retained after the required filters were applied. This entails the inclusion of not only substances appearing in multiple sources but also those unique to each source. Consequently, when identical substances were identified in multiple sources but associated with different evaluations, they were recorded as distinct entries.

Moreover, the records were entered into different phases according to the information available: databases filtered for BMD/BMDL/BMR were entered directly in the data extraction phase, whereas unfiltered databases and records from reports underwent the title/abstract phase.

2.2.1 Data acquisition

As shown in Figure 1, the information export varies according to the type of source. Sources were checked to retrieve a downloadable data version considering BMR/BMD information. When needed, information was retrieved by web scraping according to the website's policy.

Reports/PDFs: When considering sources containing reports, the export and filtering phase required a web scraping part with pre-processing and text mining.

Databases: databases, when available, were downloaded. The relevant information was then maintained and, whenever possible, the entire database was filtered for BMD analysis. The variables of interest are those mentioned in the contract and a group of other variables that were discussed with toxicology experts.

A preliminary analysis has been performed on the OpenFoodTox chemical hazards database of the EFSA and the IRIS-EPA dataset by extracting and consolidating data on a single set of data. Consequently, variables that respond to the main questions of EFSA have been identified.

The variables provided specific information about:

- The chemical compound to distinguish it from other chemicals
- The principal toxicological study conducted to derive the risk assessment values
- The biological response involved in the development of toxic effects and the key endpoint for which the BMD analysis was conducted
- The organ system most sensitive to the toxic effect will be reported alongside the mechanism involved in the development of the toxic effects
- The risk assessment values with their appropriate units derived from the toxicological studies
- The source from which the information was retrieved, and the specific URL

2.2.2 Import into DistillerSR

The output of the different sources was imported into DistillerSR to keep track of each record (**Error! Reference source not found.**). Each source has been imported into a specific phase of the screening according to the available information. In the data extraction phase, the missing information was manually filled in. Reports and PDFs were imported in the Title/Abstract phase, and in the Title/Abstract, reports were imported from JECFA and part of OpenFoodTox and IRIS-EPA, as previously mentioned. The JECFA reports were derived from Web scraping and text mining and filtered according to BMD/BMR. An expert manually checked each record to evaluate its eligibility to go through the full-text and data extraction phase.

2.2.3 Retrieving missing information

When all the documents collected from the different sources were in the data extraction phase, all missing information was filled in. Filling in missing information was carried out in two distinct phases to ensure comprehensive data coverage. In Phase 1, the team focused on extracting relevant information from reports provided by designated databases. In Phase 2, the team broadened their search beyond the primary databases and explored supplemental sources to address any remaining gaps further.

- IRIS-EPA:

- Missing information in the IRIS-EPA database was first filled in by consulting reports provided by IRIS (Integrated Risk Information System) assessments. IRIS stands for the Integrated Risk Information System, a U.S. Environmental Protection Agency (EPA) program. IRIS assessments are a critical component of this program, and they involve the systematic review and analysis of scientific information on the potential human health effects of exposure to various chemical substances. These reports serve as data sources and contain valuable insights related to the variables of interest.

- OpenFoodTox:
  - The EFSA Journal was consulted to fill in the additional information for the OpenFoodTox datasets. By expanding the scope of the search, the team aims to access alternative data reservoirs that may contain valuable information.
- JECFA:
  - The WHO website was consulted to provide additional information for the JECFA dataset.

Each record was reopened to fill in the missing information when available for the three different information sources (IRIS-EPA, OpenFoodTox, and JECFA). Each record is inspected to guarantee the development of a comprehensive and high-quality dataset. This involves a review of each imported record, where any gaps, discrepancies, or missing information is identified. Each PDF document and record will undergo an examination to locate and extract the required information whenever it is available. A search was carried out for supplementary data sources and materials to bridge these information gaps.

Initially, each PDF and record will be carefully prepared for analysis. A clear understanding of the structure and layout of the documents will be established. This involves identifying headings, subheadings, tables, figures, and other relevant markers. The objective is to locate and extract data points, facts, or references that match the specific missing information in our datasets. Once the relevant information is located and verified, it will be carefully extracted and recorded in a structured format.

2.2.4 Software

In the workflow of IRIS-EPA and OpenFoodTox, R software version 4.3.2 (R Core Team, s.d.) is integral to several stages of data management, particularly the initial data acquisition and pre-processing phase. Notably, R is adept at efficiently preparing raw data for subsequent analysis. Moreover, in the specific context of text mining from PDFs, Docker is employed to streamline and containerize the webscraping of JECFA and text mining processes. This approach ensures a consistent and reproducible environment for extracting valuable information from documents, thereby enhancing the overall comprehensiveness of the dataset. While R is the driving force behind data management, Docker serves as a specialized tool, optimizing the workflow.

To ensure systematic record keeping, the output from various sources was imported into DistillerSR (Systematic Review and Literature Review Software by DistillerSR, s.d.). During the data extraction phase, databases were directly imported, and any incomplete information

was manually supplemented. In Annex E, the data extraction preview form. As for reports and PDFs, they were imported during the Title/Abstract phase. Only documents containing BMD/BMR-related information were included in the import process. The records were imported into different levels according to the completeness of the information. Each record was subjected to a manual evaluation by an expert to determine its eligibility. DistillerSR assisted in the management of duplicate records. As different sources may yield multiple documents for the same substance, DistillerSR helped identify and manage these variations. Once all the records are imported, they go through several stages of review screening, including title/abstract screening, full-text screening, and data extraction.

3 Data Dictionary Definitions

As a first step, the list of potential variables suggested by the team needed to be harmonized with those already existing in the EFSA catalogue to verify for similarity or overlap with terminologies (

3.1 Data Source Information

The most relevant variables have been defined to provide toxicological information for chemicals.

The variable 'Year' indicated the publication year of the information source. 'BMD_adopted' specifies whether the JECFA database benchmark dose analysis was adopted.

The database source from which the information was retrieved is reported under the variable 'db_source' and its source of retrieval is under the variable 'information_source_URL'. For this variable, it indicated whether the information was retrieved from OpenFoodTox, IRIS-EPA or JECFA.

Table 3). A process of matching and mapping was carried out to ensure the alignment of variables suggested by the team with the EFSA catalogue. The team reviewed the EFSA catalogue, which encompassed a comprehensive range of standardized classifications and terminologies relevant to food safety. Each suggested variable was examined and compared with the entries in the EFSA catalogue to identify the most appropriate match. When a corresponding variable was found, it was replaced with that name in the matching table among the different sources. For some variables, there were numerous levels. In this case, we arranged the levels to reflect the most frequently encountered situations. When the suggested variables did not find an exact match in the EFSA catalogue, the new variables indicated by the team were used. As a second step, the potential variables were matched among the list of databases considered for the inventory.

3.2 Data Source Information

The most relevant variables have been defined to provide toxicological information for chemicals.

The variable 'Year' indicated the publication year of the information source. 'BMD_adopted' specifies whether the JECFA database benchmark dose analysis was adopted.

The database source from which the information was retrieved is reported under the variable 'db_source' and its source of retrieval is under the variable 'information_source_URL'. For this variable, it indicated whether the information was retrieved from OpenFoodTox, IRIS-EPA or JECFA.

Table 3. List of variables for the Inventory

Variable name	Description	Levels
Year	Information source year of publication	- free text -
BMD adopted	This variable was dedicated to the JECFA database source to indicate whether the BMD analysis performed was adopted or not	YES NO
db_source	Specification of the database used for the retrieval of information	OpenFoodTox IRIS-EPA JECFA
information_source_URL	The specific URL dedicated to the source of information	- free text -
com_name	Substance name as defined in the opinions. If multiple names for the same substances are reported, the most specific (or more common) is reported	- free text -
substance_CASRN_number	Chemical Abstracts Service (CAS) Number: Standardised identifier used to identify chemical substances uniquely	- free text -
molecular_formula	Molecular formula	- free text -
substance_description	Brief description of the chemical substance	- free text -
assessment_type	Type of assessment carried out to derive toxicological values, whether non-cancer assessment or cancer assessment	Cancer Non-cancer
exposure_routes	It refers to the pathway through which a toxic substance enters the body (e.g., inhalation, dermal, oral)	oral: feed oral: gavage oral: unspecified oral: drinking water oral: capsule aereo: inhalation intraperitoneal intragastric other
principal_toxicological_study	Specific study that is considered the primary source of toxicity data for a particular chemical or substance	- free text -
principal_toxicological_study_DOI/URL	Digital Object Identifier is a permanent character string (a "digital identifier") used to uniquely identify an object such as an electronic	- free text -

Variable name	Description	Levels
	document, source the study report or publication, URL or EFSA Journal	
endpoint	Key endpoint for which BMD analysis was conducted	- free text -
Factor_Endpoint	Variable indicating the variations that occurred at the key endpoint	Increase Decrease Induction Incidence Prevalence Change
Factor_Endpoint2	Variable allowing the categorization of the key endpoint	Liver haemangiosarcoma Tail muscular atrophy Anorectic effects Hepatocellular carcinoma Liver weight Kidney weight Body weight Erythrocytes, leucocytes and platelet Serum plasma globulin Other
key_adverse_response	Key adverse response refers to the adverse effect or health outcome	- free text -
dose_response_model	Models that can be used for BMD analysis	Dichotomous Hill Exponential Gamma Hill Inverse Exponential LMS-two stage Logistic Log-logistic Log-Normal Log-Probit Multistage Nested Log-Logistic Probit

Variable name	Description	Levels
		Polynomial Power Weibull Other
selected_BMR_value	Values for BMR used for BMD calculation, selected BMR value could be reported as a percentage, probability, standard deviation, etc.	1% 5% 10% 1 SD 0.5% 3% 50% 200%
definition_for_bmd_estimation	- Continuous endpoints metrics: Relative definition, standard deviation, absolute, or other definitions If the relative definition is not used, a conversion should be provided. - Quantal endpoints metrics: "extra risk" or "additional risk"	Extra risk Relative definition Standard deviation Summed risk
reference_point	It refers to a specific dose or concentration of a toxic substance that is used as a basis for assessing the potential adverse effects on human health or the environment. For the purpose of this inventory, the reference points included are BMD, BMDL, BMC, BMCL	BMC BMCL BMD BMDL
reference_point_value	Numerical value of the reference point	- number -
reference_point_unit	mg/kg body weight per day, unless otherwise specified	mg/m3 mg/kg-day mg/kg bw/day mg/L µg/kg µg/kg bw/day µg/kg bw µg/kg bw/week fiber/cc ng/L

Variable name	Description	Levels
uncertainty_factor	Safety factor/Uncertainty factor used to derive the health-based guidance values	- number -
health_based_guidance_values	It refers to a numerical threshold or concentration level of a substance in a specific environmental medium (such as air, water, soil, or food) that is deemed safe for human health. Health based guidance value, type of value set in the report such as: an ADI, Rfd, Rfc, or TDI, cancer CSF, Unit risk value etc. …	ARfD ADI RfD RfC MOE MSI/FC Slope TDI TWI Other
HBGV_value	Numerical value of the HBGV	- number -
HBGV_unit	mg/kg body weight per day, unless otherwise specified	fiber/cc mg/day mg/kg mg/kg bw mg/kg-day mg/kg bw/day mg/m3 ng/kg bw/week µg/kg bw Other
study_duration_type	Length of time over which the study is conducted. For example, acute toxicity studies chronic toxicity studies, etc.	Chronic/long term toxicity Short term dietary toxicity Subacute Subchronic Acute toxicity Short-term toxicity Long-term toxicity
study_species	Species used in the principal toxicological study	American mink Bobwhite quail Cat Cattle Dog

Variable name	Description	Levels
		Hamster Human Mallard duck Monkey Mouse Pig Rat Unspecified Other
study_strains	Strains used in the principal toxicological study	B6C3F1 Fischer 344 Sprague-Dawley Wistar Wistar Han Other
sex_of_test_subjects	Sex of the tested animals in vivo toxicity study	Female Male Male/Female Not specified
study_experimental_model	Specific method used to investigate the toxic effects of a chemical substance on living organisms. These models can include in vitro (cell-based) or in vivo (animal-based) systems	In vitro In vivo
type_of_substance_toxicity	It refers to the specific mechanism or mode of action by which a substance exerts its harmful effects on living organisms. Can be categorised into genotoxicity, immunotoxicity, neurotoxicity, carcinogenicity, reproductive toxicity, developmental toxicity, etc.	Acute toxicity Carcinogenicity Mutagenicity Genotoxicity Developmental and Reproductive Hepatotoxicity Cardiotoxicity Respiratory Toxicity Endocrine Disruption Irritation and Corrosion Ecotoxicity Clinical Toxicity Systemic toxicity Neurotoxicity

Variable name	Description	Levels
		Nephrotoxicity Immunotoxicity
target_organ	Organ system that is most sensitive to the toxic effects of a chemical substance	Bladder Developmental Duodenum Endocrine Esophagus Gastrointestinal Kidney Liver Heart Blood Immune Lung Musclo-skeletal Nervous Ocular Skin Reproductive Spleen Thyroid Urinary Other
mechanism_of_toxicity	Description of the specific biological mechanisms involved in the development of toxic effects following exposure to a chemical substance	- free text -
single_grouped_chem_assess	Suggestions to consider whether the chemical substance is a single chemical entity or part of a group assessment	part of group assessment single metabolite unknown
specify_group_chem_assessment	If the chemical substance was part of a group assessment, this variable was dedicated to specifying which group it belonged to	- free text -
type_of_data	It provides a framework for understanding the nature of the measurements or observations collected in a study. Data can be broadly classified into two types: continuous and quantal	Continuous Quantal

Variable name	Description	Levels
statistical_paradigm	Some common statistical paradigms include frequentist and Bayesian	Bayesian Frequentist
m(0)	Modeled mean response at control dose	- number -
sigma	Modeled standard deviation	- number -
transformed_bmr_value	If the BMD is not expressed in relative definition, it should be transformed to be expressed as a relative change with respect to the background response, and then the derivation of the BMR is added to the inventory	- number -

JECFA, Joint FAO/WHO Expert Committee on Food Additives; CAS number, Chemical Abstracts Service Number; EFSA Journal, European Food Safety Authority Journal; BMD, benchmark dose; BMR, benchmark response; BMDL, benchmark dose lower limit; BMC, benchmark concentration; BMCL, benchmark concentration lower limit; ADI, acceptable daily intakwue; Rfd, oral reference dose; Rfc, inhalatory reference concentration; TDI, tolerable daily intake; CSF, cancer slope factor; HBGV, health-based guidance value

3.3 Chemical information

The variable 'com_name' referred to the unique name or identifier given to a specific chemical substance to distinguish it from other chemicals.

The variable 'substance_CASRN_number' (chemical abstract service (cas) number) variable is the standardised identifier used to uniquely identify chemical substances. The symbolic representation of a chemical compound that indicates the number and types of atoms present in a molecule is reported under the variable 'molecular_formula'.

The variable 'substance_description' represented a brief description of the chemical substance.

3.4 Assessment Information

The variable 'assessment_type' defined the type of evaluation performed on a specific chemical compound to derive toxicological values that differentiate between non-cancer dose-response assessment and cancer dose-response assessment.

The route of exposure (e.g., inhalation, dermal, oral) to a certain chemical compound was reported under the variable 'exposure_routes'.

The variable 'principal_toxicological_study' reported a specific study that was considered the primary source of toxicity data for a particular chemical or substance. The study was identified by the variable 'principal_toxicological_study_DOI/URL'.

3.5 Outcomes

The variable 'endpoint' refers to the key endpoint for which the BMD analysis was performed. The variable 'Factor_Endpoint' indicated the variations that occurred at the key endpoint. The "Factor_Endpoint2" allowed the categorization of the key endpoint.

The variable 'key_adverse_response' was used to describe the biological response or effect that is most responsive or sensitive to the toxic effects of a chemical substance.

The variable 'dose_response_model' reported the models that were used for BMD analysis.

The variable 'selected_BMR_value' provided us with information related to the BMR choice. BMR should be reported in terms of SD, relative definition, or absolute value (in the contract).

The variable 'definition_for_bmd_estimation' represented the definition used to estimate the BMD. Continuous endpoint metrics: relative definition or standard deviation.

The variable 'reference_point' refers to the type of experimental dose in the study. Included in the inventory are BMD, BMDL, BMC, and BMCL. The variables 'reference_point_value' and 'reference_point_unit' refer respectively to the numerical value of the reference point and the unit in mg/kg body weight per day unless otherwise specified.

The 'uncertainty_factor' was used to derive the health-based guidance values.

The variable 'health_based_guidance_values' is a level of exposure to the substance that is considered safe for human consumption over a lifetime such as: an ADI, oral reference dose (Rfd), inhalatory reference concentration (Rfc), or TDI, cancer slope factor. The variables 'HBGV_value' and 'HBGV_unit' reported, respectively, the numerical value and the units of the HBGV.

3.6 Study information

The type of duration of the toxicological study was reported under 'study_duration_type'.

The variables 'study_species', 'study_strains' and 'sex_of_test_subjects' referred to the characteristics of the study population in the main toxicological study.

The 'study_experimental_model' referred to the specific method used to investigate the toxic effects of a chemical substance on living organisms.

The type of substance toxicity was reported under 'type_of_substance_toxicity'. A framework derived from Toxric, a comprehensive database of toxicological data and benchmarks (Wu et al., 2022) was employed. In the original classification, the following categories were present: acute toxicity, carcinogenicity, mutagenicity, developmental and reproductive, hepatotoxicity, cardiotoxicity, reproductive toxicity, disruption of the endocrine system, irritation and corrosion, ecotoxicity, clinical toxicity, toxcast&tox21 assay, and cyp 450. In our adaptation, we expanded the classification by incorporating additional levels to improve the specificity. These augmented levels encompass genotoxicity, systemic toxicity, neurotoxicity, nephrotoxicity, and immunotoxicity. Some levels were not adopted, including toxcast&tox21 assay, and CYP450. By incorporating these subcategories, our classification system aimed to provide a more nuanced and detailed understanding of the diverse toxicological effects associated with the substances under investigation.

The variable 'target_organ' referred to the system most sensitive to toxic effects of a chemical substance, and the variable 'mechanism_of_toxicity' described the specific biological mechanisms involved in the development of toxic effects.

The variable 'single_group_chem_assessment' was used to suggest the consideration if the chemical substance is a single chemical entity or part of the group assessment. If the chemical substance was part of a group assessment, this variable was dedicated to specifying to which group it belonged “specify_group_chem_ssessment”.

Continuous or quantal data were defined under the variable 'type_of_data'. The variable 'statistical_paradigm' in this data set was related to the statistical approach used in toxicology assessments, described in dose_response_model as provided in the benchmark dose technical guidance (Hogan, s.d.) And the guidance on the use of the benchmark dose approach in risk assessment (more et al., 2022). It highlighted the choice between common statistical paradigms, with examples being the frequentist and Bayesian methodologies. 'm (0)' represented the mean response of the model at the control dose. 'Sigma' represented the standard deviation (under homoscedasticity).

In case the relative definition is not used, conversion elements are provided under m(0) and sigma for recalculation as BMR*sigma/m(0) "transformed_BMR_value".

3.7 Mapping

Appendix A – List of variables for the Inventory presents how the matching works among IRIS-EPA and OpenFoodTox sources.
For example, the "Exposure routes" variable in the EFSA catalogue is reported in the library of the catalogue as ROUTEXP, and 9 levels were considered.
In order to integrate data from the original databases, we embarked on a comprehensive mapping and matching of the available variables. This process involved aligning and correlating the variables present in both original databases with the specific variables created for our inventory. The purpose was to align disparate variables and ensure a unified dataset for more robust and comprehensive analyses.

As shown in Annex A, the information on 'Exposure routes' in OpenFoodTox is available on the ENDPOINTSTUDY_KJ/ variable sheet: ROUTE or on the sheet GENOTOX_KJ/ variable: route.
Whereas in IRIS-EPA the information related to the route can be retrieved in the Sheet: chemical details/ variable: ORAL EXPOSURE, or sheet: chemical details/ variable: INHAL EXPOSURE STATUS, or sheet: WOE Toxicity Values/ variable: STUDY ROUTE.

The sample case on Benzo[a]pyrene is reported as an example of how the dataset appeared before importing into DistillerSR. Each row refers to a substance. Some substances, for example, Benzo[a]pyrene (BaP), often have multiple rows in databases due to the different types of chemical assessment used to derive toxicological values. The different types of chemical assessment for BaP ensure a comprehensive understanding of non-cancer and cancer-related toxicological values, enabling appropriate risk management strategies (Table 4).

Table 4: Example of mapping for Benzo[a]pyrene in OpenFoodTox and IRIS-EPA sources

Db_source	Com_name	Assessment_type	Route_exp
IRIS-EPA	Tributyltin oxide (TBTO)	Non-cancer	Oral
IRIS-EPA	Benzo[a]pyrene (BaP)	Non-cancer	Oral
IRIS-EPA	Benzo[a]pyrene (BaP)	Non-cancer	Oral
IRIS-EPA	Benzo[a]pyrene (BaP)	Non-cancer	Intragastric
IRIS-EPA	Benzo[a]pyrene (BaP)	Cancer	Oral
IRIS-EPA	Benzo[a]pyrene (BaP)	Cancer	Inhalatory
OpenFoodTox	Benzo[a]pyrene	Non-cancer	Oral:feed
JECFA	1,3-dichloro-2-propanol	Cancer	Oral

IRIS-EPA, Integrated Risk Information System-Environmental Protection Agency; JECFA, Joint FAO/WHO Expert Committee on Food Additives
The orange-coloured cells contain information from the original source that was matched and imported. The cells represented in blue colour are the ones for which the data must be retrieved from the additional information sources.

4 JECFA

JECFA stands as a preeminent international scientific body, recognized for its pivotal role in evaluating and assessing the safety and functionality of food additives. Established by the FAO and WHO, it operates at the intersection of scientific expertise and global health policy. The JECFA site was accessed on 15 June 2023.
JECFA serves as an independent scientific expert committee that performs risk assessments and provides advice to FAO, WHO, the member countries of both organizations, and the Codex Alimentarius Commission (CAC). This searchable database contains summaries of all the evaluations of flavours, food additives, contaminants, toxicants, and veterinary drugs JECFA has performed. Each summary contains basic chemical information, acceptable daily intakes (ADIs)/ tolerable daily intakes (TDIs), links to the most recent reports and monographs, the specification database, and a history of JECFA evaluations. The database is searchable by partial name or CAS number, first character (letter or symbol), or functional class (WHO, 2023).

WHO is part of the JECFA committee. It publishes monographs containing detailed descriptions of the biological and toxicological data considered in the evaluation and the intake assessment. FAO is also part of the JECFA committee. It is a specialized agency that leads international efforts to defeat hunger (Bhalla et al., 2019).

The Food Safety and Quality Unit supports the strengthening of systems of food safety and quality control at national, regional, and international levels. FAO supports science-based food safety governance and decisions by providing sound scientific advice (through the JECFA and Joint FAO/WHO Expert Meetings on Microbiological Risk Assessment (JEMRA) expert bodies) to underpin food safety standards at national, regional, and international levels (Bhalla et al., 2019).

WHO and FAO are the bodies that prepare the toxicological monographs Food Additives Series (FAS) and Technical Report Series (TRS) reports associated with the chemicals under investigation (and are publicly available on their official websites (Figure 2). To retrieve these documents, a web scraping strategy was employed.

Figure 2. Overview of the JECFA portal for the substance Benzyl Alcohol

All the project is within a Docker following this link: docker run -d --rm --name jecfa -e PASSWORD=jecfa -p 18787:8787 corradolanera/jecfa:v0.1. Figure 3, Figure 4 and Figure 5 report the infrastructure.

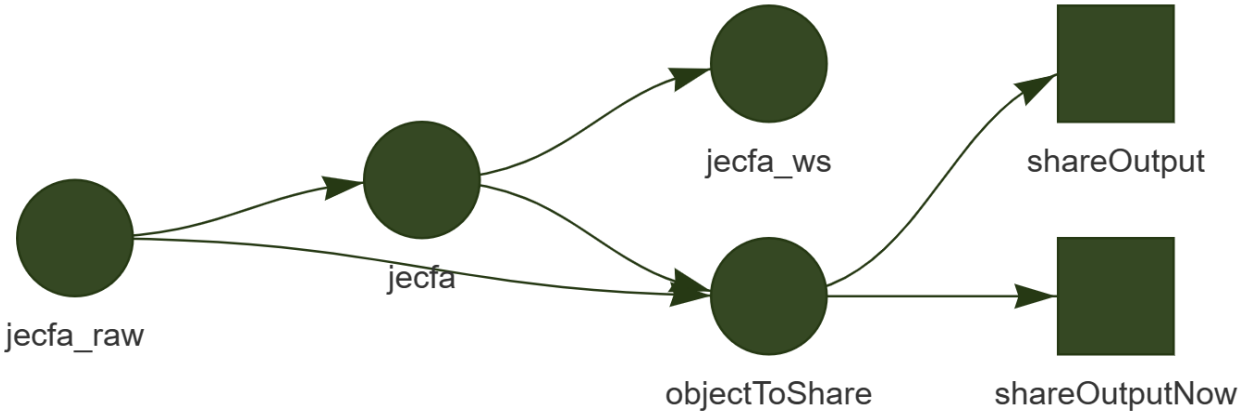

Figure 3. Overview of JECFA pipeline to get the webscraped dataset

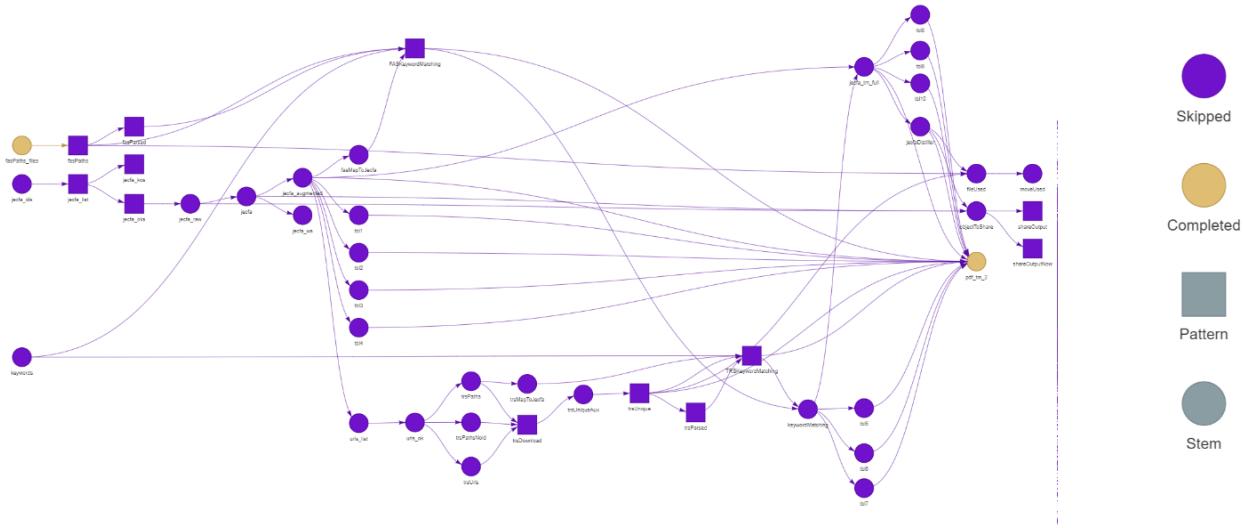

Figure 4. Overview of JECFA pipeline including the text-mining on PDFs

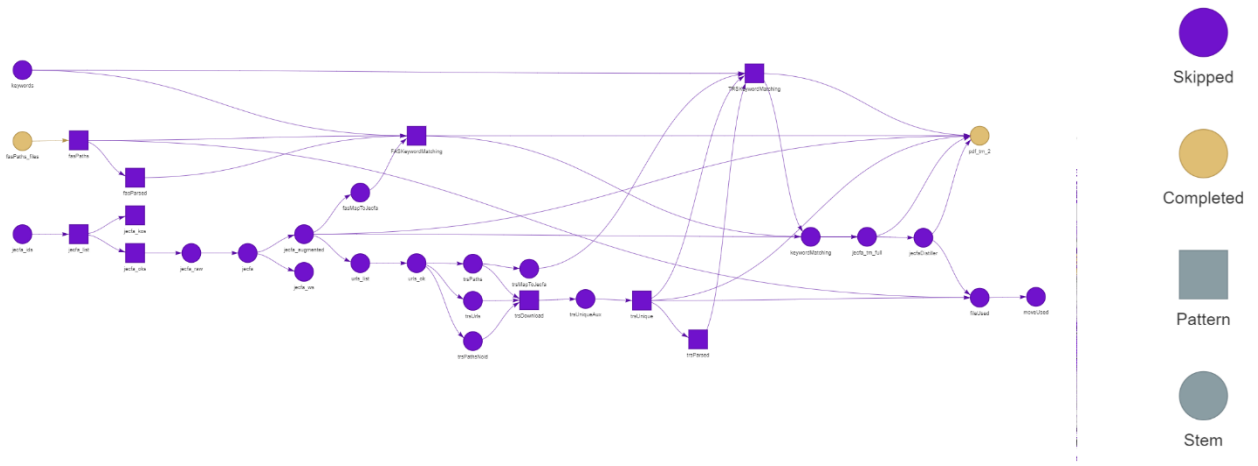

Figure 5. Overview of the JECFA simplified pipeline including the text-mining on PDFs (without tables and exported objects)

4.1 Web scraping

The process begins with the scraping of web data from the JECFA portal, resulting in a csv format data set. Subsequently, data pre-processing is initiated to extract valuable information from both URLs and strings. In particular, experts have singled out the field of the 'Tox.Monograph' as the most reliable source of information (Figure 2).

A web-scraping approach was applied because JECFA did not expose the APIs to query their database. The JECFA site was iteratively explored by accessing all of the pages of hosted chemicals using the R packages `httr` and `rvest`.

The records usually contain two links, but not necessarily to the PDF: one is the toxicological monograph and the other is the report. The toxicological monograph usually starts with 'FAS', while reports start with 'TRS'. Sometimes, an addendum was available. For some substances, the toxicological monograph was not available. The information related to our scope was reported in the toxicological monograph, so only this document was searched for relevant information on the BMD/BMR assessment.

The first step involves addressing concerns related to handling chemicals in cases where the monographs are marked as 'NOT PREPARED' or are missing, constituting 8% of the instances. This process is elucidated in the flowchart depicted in Figure 3. A suggested approach to tackle this issue is to utilize the WHO TRS. In contrast to the FAS initiative, TRS documents can be conveniently downloaded in PDF format directly from the links accessible in the prepared database.

The second step deals with the challenge of dealing with chemicals that lack both monograph and TRS documentation, making up 3% of the cases. A key question that arises in this context

is whether the absence of documentation indicates a deficiency in the risk assessment output, which is the primary focus of this project.

In an initial effort, an approach was taken to directly access documents through the 'Tox.Monograph' links. Unfortunately, this method posed significant challenges due to the inherent heterogeneity of the data. These links led primarily to HTML pages, lacking direct access to the desired PDF documents, as elucidated in Table 5.

Table 5. Description of the link to the monograph

Characteristic host	.htm N 1,060[(a)]	.pdf N 1,587[(a)]	html N 2,751[(a)]	Other N 1,127[(a)]
www.inchem.org	1,060 (100)	0 (0)	2,751 (100)	4 (1.4)
apps.who.int	0 (0)	1,489 (94)	0 (0)	232 (80)
www.who.int	0 (0)	87 (5.5)	0 (0)	55 (19)
whqlibdoc.who.int	0 (0)	11 (0.7)	0 (0)	0 (0)
Unknown[(b)]	0	0	0	836

(a): n (%)
(b): corrupt link such as /food-additives-contaminants-jecfa-database/Document/Index/989

4.1.1 Phase 1 - FAS

As a strategic shift, the code centres its efforts on discerning the names of the 'Tox.Monograph' within the context of FAS, as exemplified in Table 5. This approach is designed to facilitate efficient access to crucial information within the field of 'Tox.Monograph', improving its usability for subsequent steps. Although almost all 85 FAS codes are referenced, the ones collectively accounting for 50% of the records are 40, 42, 44, 46, 48, 50, and 52. We access and store all FAS from https://www.who.int/groups/joint-fao-who-expert-committee-on-food-additives-(jecfa)/publications/toxicological-monographs in PDF format, whenever available, or HTML as the first, fourth, fifth, sixth, eighth, 10th, and 12th through 52nd series of FAS monographs are available in HTML format only. WHO monographs beginning with the 51st series are also available in PDF format.

4.1.2 Phase 1 - TRS

To obtain the WHO, TRS in cases where a FAS is unavailable, the process involves downloading from the "Report_sourcelink" column in the pre-processed jecfa.csv file. This operation is performed on a dataset consisting of 662 records, and it ensures that TRS is accessed as an alternative source when FAS information is not present.

Here is a breakdown of what the R code does:

- Cleaning the 'Report' column in the data set:
  - The code cleans the column 'Report' in the dataset (Report_clean) by removing unwanted characters at the beginning of the string. If the string starts with

'">', it removes the first two characters. This step aims to handle formatting issues in the data

- Sanitizing file names:
  - The code defines a function that replaces characters not allowed in file names with underscores. It also removes leading and trailing underscores from the file name
- Creating a folder named 'TRS':
  - If the 'TRS' folder does not exist in the current directory, the code creates it. This folder is intended to store downloaded PDF files
- Extracting PDF URLs:
  - The code extracts a vector of PDF URLs from the column 'Report_sourcelink' in the data set and assigns it to the variable PDF_urls
- Downloading PDFs:
  - The code enters a loop to process each URL in the PDF_urls vector. It checks various conditions before attempting to download a PDF: It skips URLs associated with specific conditions based on values in the dataset (e.g., Monograph in a FAS format). Check if the URL is missing or empty and skips such cases. It verifies that the URL has a valid format (starting with "http://" or "https://"). For URLs that meet the conditions, the code constructs a file name for the PDF file. The file name includes a prefix based on the position in the original dataset and a sanitized version of the column 'Report_clean'. Check if the PDF file exists in the 'TRS' folder. If not, it downloads the PDF file using the GET function from the `httr` library. If the download is successful (status code 200), it saves the PDF file in the 'TRS' folder with the constructed file name. It logs the appropriate messages if the download fails or the file already exists.
- Identify unique PDFs within the TRS folder t (n=54)

4.2 Text Mining

In general, the R code automates the extraction of specific information from a collection of PDF files, making it easier to identify which PDFs contain relevant content based on predefined keywords.

4.2.1 Phase 2 – Text mining on TRS (when a FAS is not available)

This code will create a data frame with one row per PDF file, and each column corresponds to a keyword and its list of page matches. The column 'File' contains the file name of the PDF. In detail:

- Extract text from PDFs:
  - The code uses the pdf_ocr_text function from the `pdftools` library to extract text from each PDF file. The extracted text is converted to lowercase for a consistent search.
- Keyword search:
  - For each PDF file, the code searches for predefined keywords related to the content of interest, such as "dose-response," "modelling," "BMD", "BMR"

"benchmark-dose", "dose response", "modeling", "benchmark dose". It records the pages within the PDF on which these keywords are found.

- Store Results:
  - The results for each PDF file, including the file name and the list of pages where keywords are found, are stored in a list. The list is indexed by the PDF file name.
- Final Data Frame:
  - Lastly, the code combines the individual results of each PDF file into a final data frame.

It follows a match between JECFA refID (i.e., row names) and the TRS PDF.

4.2.2 Phase 2 – Text Mining on FAS

The text mining process is applied to the FAS dataset, which comprises 73 unique FAS instances. The same procedure described for text mining TRS is applied to FAS. The only difference is in the way PDFs are accessed. It follows a match between the JECFA refID and the FAS PDF.

4.2.3 Description of JECFA records with text mining on FAS/TRS

Table 6 compares the words retrieved between FAS and TRS with respective sample sizes of 5,671 and 663 (191 records were excluded due to the absence of FAS and TRS). A higher percentage of FAS records (89%) involve dose-response information compared to TRS records (33%), with similar trends observed across other characteristics such as modeling (75% vs. 20%), BMD (23% vs. 15%), BMR (24% vs. 1.7%), and benchmark dose (28% vs. 11%). The note clarifies that both the American and British spellings for certain terms were considered during the searches.

Table 6. Comparison of the retrieved words between FAS and TRS

Characteristic	**FAS, N = 5,671[(a)]**	**TRS, N = 663[(a)]**
dose response[(b)]	5,073 (89%)	218 (33%)
Modelling[(c)]	4,263 (75%)	132 (20%)
BMD	1,277 (23%)	98 (15%)
BMR	1,379 (24%)	11 (1.7%)
benchmark dose[(d)]	1,561 (28%)	71 (11%)

FAS, Food additive status; TRS, Technical Report Series; BMD, benchmark dose; BMR, benchmark response
(a): n (%)
(b): both dose-response and dose response terms were searched for
(c): both modelling and modeling terms were searched for
(d): both benchmark-dose and benchmark dose terms were searched for

Table 7 and Table 8 report the distribution of keywords in each document source, stratified by the co-presence of "bmd" or "dose response", respectively. It appears that the term 'bmd' is always accompanied by 'dose response' and 'modelling', and this suggested considering this as a trustworthy keyword.

Table 7. Distribution of keywords by the co-presence of 'bmd' and the source of the document

Characteristics	FAS		TRS	
	bmd no, N = 4,394[(a)]	**bmd yes**, N = 1,277[(a)]	**bmd no**, N = 565[(a)]	**bmd yes**, N = 98[(a)]
dose response[(b)]	3,796 (86%)	1,277 (100%)	120 (21%)	98 (100%)
modelling[(c)]	2,986 (68%)	1,277 (100%)	34 (6.0%)	98 (100%)
BMR	832 (19%)	547 (43%)	0 (0%)	11 (11%)
benchmark dose[(d)]	637 (14%)	924 (72%)	39 (6.9%)	32 (33%)

FAS, Food additive status; TRS, Technical Report Series; BMD, benchmark dose; BMR, benchmark response
(a): n (%)
(b): both dose-response and dose response terms were searched for
(c): both modelling and modeling terms were searched for
(d): both benchmark-dose and benchmark dose terms were searched for

Table 8. Distribution of keywords by co-presence of "dose response" and document source

Characteristics	FAS		TRS	
	dose-re no, N = 598[(a)]	**dose-re yes**, N = 5,073[(a)]	**dose-re no**, N = 445[(a)]	**dose-re yes**, N = 218[(a)]
modelling[(b)]	477 (80%)	3,786 (75%)	3 (0.7%)	129 (59%)
BMD	0 (0%)	1,277 (25%)	0 (0%)	98 (45%)
BMR	0 (0%)	1,379 (27%)	0 (0%)	11 (5.0%)
benchmark dose[(c)]	0 (0%)	1,561 (31%)	3 (0.7%)	68 (31%)

FAS, Food additive status; TRS, Technical Report Series; BMD, benchmark dose; BMR, benchmark response
(a): n (%)
(b): both dose-response and dose response terms were searched for
(c): both modelling and modeling terms were searched for

Table 9 compares the characteristics of FAS and WHO TRS data, showing that all FAS records (100%) have monograph information, while only 18% of TRS records and 3.7% of records with unknown sources include monograph details; similarly, all FAS records (100%) and 100% of TRS records have report information.

Table 9. Distribution of retrieved documents

Characteristic	FAS, N = 5,671[(a)]	TRS, N = 663[(a)]	Unknown, N = 190[(a)]
Monograph	5,671 (100%)	118 (18%)	7 (3.7%)
Report	5,668 (100%)	663 (100%)	14 (100%)

FAS, Food additive status; TRS, Technical Report Series
(a): n (%)

Table 10 presents data on unique combinations of Functional.Class, CAS.number, Evaluation.year, COE.number, and FEMA.number. Overall, among 6,334 records, 47% exhibit

non-unique combinations, with varying distributions among FAS, TRS, and records with unknown sources across different uniqueness criteria.

Table 10. Distribution of unique records depending on the set of variables considered

Characteristic	Overall, N = 6,334[a]	FAS, N = 5,671[a]	TRS, N = 663[a]
unique1[b]	2,981 (47%)	2,629 (46%)	352 (53%)
unique2[c]	2,520 (40%)	2,169 (38%)	351 (53%)
unique3[d]	1,462 (23%)	1,123 (20%)	339 (51%)
unique4[e]	1,296 (20%)	986 (17%)	310 (47%)

FAS, Food additive status; TRS, Technical Report Series
(a): n (%)
(b): unique combination of Functional.Class, CAS.number, Evaluation.year, COE.number, FEMA.number
(c): unique combination of Functional.Class, CAS.number, Evaluation.year, COE.number
(d): unique combination of Functional.Class, CAS.number, and Evaluation.year
(e): unique combination of CAS.number and evaluation.year

4.3 Filtering of JECFA records

The process involves filtering JECFA records based on BMD, BMR, or benchmark dose criteria, resulting in the identification of 1122 unique records.

4.3.1 Title/abstract and full-text screening

After title/abstract and full-text screening, 10 records were included.

Thirty-five different PDFs containing keywords such as 'BMD,' 'BMR,' or 'benchmark dose' cover all the 1,122 records. However, 10 records passed through the full-text screening process, indicating a true assessment of BMD. The positive predictive value of these keywords in identifying BMD assessments is 21/35. The reasons for false positives are shown in

The split of records when several endpoints were evaluated resulted in 27 records.

In the database, each record corresponds to a distinct chemical substance, typically featuring a singular endpoint under study. However, in instances where a particular chemical substance undergoes evaluation for multiple endpoints, the record is systematically divided to accommodate each endpoint separately. By segregating records based on individual endpoints, the database facilitates a comprehensive and nuanced exploration of the various facets of each chemical's impact.

Table 11. Among them, a toxicological assessment with other references, such as LD50 and NOEL (n = 11) is the most common.

The split of records when several endpoints were evaluated resulted in 27 records.

In the database, each record corresponds to a distinct chemical substance, typically featuring a singular endpoint under study. However, in instances where a particular chemical substance undergoes evaluation for multiple endpoints, the record is systematically divided to accommodate each endpoint separately. By segregating records based on individual endpoints, the database facilitates a comprehensive and nuanced exploration of the various facets of each chemical's impact.

Table 11. JECFA reasons for exclusion based on the source document (FAS or TRS)

Source	Presence of BMD analysis	Compound	Remarks	Included
FAS 5	NO		Toxicological assessment with LD50	NO
FAS 10	NO		Toxicological assessment with LD50	NO
FAS 21	NO		Toxicological assessment with LD50	NO
FAS 22	NO		Toxicological assessment with LD50	NO
FAS 23	NO		Toxicological assessment with LD50	NO
FAS 48	NO		Toxicological assessment with LD50	NO
FAS 52	YES	• Mercury • Methylmercury		YES
FAS 53	NO		Toxicological assessment with LD50	NO
FAS 55	YES	• Benzo[a]pyrene • Acrylamide		YES
FAS 58	YES	• 1,3-Dichloro-2-propanol • Methylmercury		YES
FAS 59	YES	• Acidified sodium chlorite (chlorate)		YES
FAS 60	NO		Toxicological assessment with NOAEL	NO
FAS 64	YES	• Lead		YES
FAS 65	YES	• Fumonisins	BMD analysis conducted but not adopted -> BMD analysis for pure Fumonisin B1 for cancer endpoint	YES
FAS 66	YES	• Derquantel	BMD analysis present but not adopted	YES
FAS 69	YES	• Gentian violet	BMD analysis present but not adopted	YES
FAS 72	YES	• Teflubenzuron		YES
		• Sisapronil	BMD analysis present but not adopted	YES
FAS 74	YES	• Sterigmatocystin • 3-monochloro-1,2-propanediol • Fumonisins		YES
FAS 77	NO		"Benchmark dose" words were present as part of the text	NO
FAS 82	YES	• Ergotamine	BMD analysis present but not adopted	YES

Source	Presence of BMD analysis	Compound	Remarks	Included
	YES	• Sec-butanol	Compound not present in the JECFA list	NO
FAS 83	YES	• N-butanol	Compound not present in the JECFA list	NO
TRS_909	NO		"Benchmark dose" words were present as part of the text	NO
TRS_913	NO		Toxicological assessment with NOAEL	NO
TRS_930	YES	• Benzo[a]pyrene • Acrylamide		YES
TRS_940	YES	• 1,3-Dichloro-2-propanol • Methylmercury		YES
TRS_947	YES	• Acidified sodium chlorite (chlorate)		YES
TRS_954	YES	• Malachite green		YES
TRS_959	YES	• Furan • Acrylamide • Arsenic • Deoxynivalenol • Mercury • Perchlorate		YES
TRS_960	NO		Toxicological assessment with NOAEL	NO
TRS_966	YES	• Fumonisin	BMD analysis conducted but not adopted -> have BMD analysis for pure Fumonisin B1 for cancer endpoint	YES
TRS_969	YES		BMD analysis present but not adopted	YES
TRS_988	YES	• Gentian violet	BMD analysis present but not adopted	YES
TRS_997	YES	• Teflubenzuron		YES
TRS_1008	NO		"Benchmark dose" words were present as part of the text	NO
TRS_1014	NO		Toxicological assessment with NOAEL	NO

FAS, Food additive status; TRS, Technical Report Series; LD50, lethal dose 50%; NOAEL, no observed adverse effect level; BMD, benchmark dose; JECFA; Joint FAO/WHO Expert Committee on Food Additives

5 OpenFoodTox

EFSA is the keystone of European risk assessment regarding food safety (Rebeca Fernández, 2021). EFSA's chemical hazards database, OpenFoodTox, is updated on a regular basis.

OpenFoodTox supplies EFSA's systematically organized hazard information for more than 5000 chemicals present in food and feed. These data have undergone rigorous peer review by a diverse group of scientific experts, exceeding 30 individuals, who are part of EFSA's various working groups and panels.
The database was downloaded on 24 March 2023. The database can be downloaded from the EFSA knowledge junction under https://doi.org/10.5281/zenodo.780543 and can be consulted from a MicroStrategy Dashboard. Two sets of data can be downloaded to spread OpenFoodTox to a wider community: five individual spreadsheets extracted from the EFSA microstrategy tool that provide for all compounds: a. substance characterisation, b.EFSA outputs, c. reference points, d. reference values and e.genotoxicity or the entire database (Dorne et al., 2021). In Annex B – Original OpenFoodTox is provided.
OpenFoodTox provides information on substance characterization, links to EFSA related output, European background legislation, and a summary of critical toxicological endpoints and reference values (Dorne et al., 2021).
It provides information about a) substance characterisation, b) EFSA output, c) reference points, d) reference values, and e) genotoxicity. It is a structured database summarising the outcomes of hazard identification and characterisation for human health, animal health, and the environment.

5.1 Filtering

A notable feature of the OpenFoodTox database was the inclusion of direct links to specific articles corresponding to each substance of interest. These links were thoughtfully provided in the original OpenFoodTox database, specifically in the 'OPINION' sheet.

The code accomplishes the following tasks:

- Read Data:
  - Read information from the OpenFoodTox and mapping Excel files.
- Create Intermediate Data Frames:
  - Divide the data into separate data frames for each unique source based on mapping entries
- Filter and Modify Data:
  - Filter data in the ENDPOINTSTUDY_KJ data frame for rows containing 'BMD' in the ENDPOINT column. Adds a new column (ASSESSMENT_TYPE) based on the presence of cancer-related terms in the EFFECT_DESC column. Modifies data in the STUDY and GENOTOX_KJ data frames by adding a new column (TOXICITY) based on certain conditions outlined in the mapping (add reference or link)
- Check Missing Values:
  - Check for missing values in specific columns (TOX_ID, HAZARD_ID, SUB_COM_ID) in the STUDY, CHEM_ASSESS, and COMPONENT data frames.
- Join Data Frames:
  - Combine data from different data frames (ENDPOINTSTUDY_KJ, STUDY, CHEM_ASSESS, COMPONENT, GENOTOX_KJ) using left joins based on specific identifier columns (TOX_ID, HAZARD_ID, SUB_COM_ID). Creates two separate result data frames by joining and selecting specific columns
- Data Transformation and Cleaning:

- Transform and clean the resulting data frames by adding new columns, removing duplicates, and selecting specific columns to prepare a database compliant with the variables present in the mapping
- Final Output:
  - Two datasets are uploaded in DistillerSR
- The records referring to non-genotoxic endpoints:
  - They entered the data extraction phase in DistillerSR, while the genotoxic endpoints were screened at the title/abstract level. This difference arises because preliminary filtering for BMD is unavailable, unlike the process for the previous phase

5.2 Recap of the mapping

We examine how we combined information from the original OpenFoodTox database. To make this happen, we did a detailed matching exercise, aligning the variables in both databases with the ones we needed for the inventory. The goal was to have a unified dataset for reliable and more thorough analyses. For example, the details about "COM_NAME" in OpenFoodTox are in the sheet COMPONENT under the word COM_NAME (Table 12).

Table 12. Recap of the mapping in OpenFoodTox

Variable name in the original source	Sheet name in the source	Variable name output
COM_NAME / COM_CASNUMBER / MOLECULARFORMULA / SUB_DESCRIPTION / QUALIFIER / SUB_COM_ID	COMPONENT	COM_NAME / COM_CASNUMBER / MOLECULAR_FORMULA / SUBSTANCE_DESC / SINGLE_GROUPED_CHEM_ASSESS / SUB_COM_ID
ROUTE / EFFECT_DESC / ENDPOINT / VALUE / DOSEUNIT / TESTTYPE / SPECIES / STRAIN / SEX / TOXICITY / TARGETTISSUE / TOX_ID	ENDPOINTSTUDY_KJ	ROUTE_EXP / KEY_ADVERSE_RESPONSE / REFERENCE_POINT / REFERENCE_POINT_VALUE / REFERENCE_POINT_UNIT / STUDY_TYPE / SPECIES / STRAINS / GENDER / TOXICITY / TARGET_TISSUE / TOX_ID
ROUTE / SPECIES / STRAIN / SEX / IS_GENOTOXIC / GENOTOX_ID	GENOTOX_KJ	ROUTE_EXP / SPECIES / STRAINS / GENDER / TOXICITY / GENOTOX_ID
SAFETY_FACTOR / ASSESSMENTTYPE / RISKVALUE / RISKUNIT / HAZARD_ID	CHEM_ASSESS	UNCERTAINTY_FACTOR / HBGV / HBGV_VALUE / HBGV_UNIT / HAZARD_ID
IS_GENOTOXIC / SUB_COM_ID / HAZARD_ID / TOX_ID / GENOTOX_ID	STUDY	TOXICITY / SUB_COM_ID / HAZARD_ID / TOX_ID / GENOTOX_ID

Table 13. Example of connecting sheets for the Benzo[a]pyrene chemical substance

SHEET/IDENTIFIER	GENOTOX_ID	TOX_ID	HAZARD_ID	SUB_COM_ID	N. of records
GENOTOX_KJ	-	-	-	-	0
STUDY	-	124, 125, 126, 127	81, 82, 83, 84, 85, 86, 87,88	142, 143, 147, 153	8
COMPONENT	-	-	-	142, 143, 147, 153	4
ENDPOINTSTUDY_KJ	-	124, 125, 126, 127	-	-	4
CHEM_ASSESS	-	-	NA	-	0

–, no information available; NA, not available

The interconnection of sheets, as exemplified in Table 13 in the original OpenFoodTox database, was established through a series of specific matching criteria listed below:

- sheet STUDY, variable TOX_ID matched with the rows in the sheet ENDPOINTSTUDY_KJ
- sheet COMPONENT variable SUB_COM_ID matched with the rows in the sheet STUDY
- the sheet CHEM_ASSESS variable HAZARD_ID matched with the rows in the sheet STUDY
- sheets OPINION and QUESTION variable OP_ID matched with the rows in the sheet STUDY
- sheet COM_SYNONYM variable SUB_COM_ID matched with the rows in the sheet COMPONENT

6 IRIS-EPA

It is an established electronic database that contains information on human health effects that may result from exposure to various chemicals in the environment. The IRIS-EPA Assessment Database represents a comprehensive compilation of scientific information on the possible health effects of exposure to various environmental contaminants. The database was downloaded on June 7, 2023. In Annex C, the IRIS-EPA database filtered for BMD is provided. Developed and maintained by the EPA, this database is a critical tool to assess the risks associated with chemical substances in the environment. The IRIS-EPA Assessment Database is publicly available online. The database allows you to search for specific chemical substances or browse a list of assessed substances.

The IRIS-EPA assessments are scientific reports that provide information on a chemical's hazards and, when supported by available data, quantitative toxicity values for cancer and noncancer health effects. IRIS-EPA assessments are not regulations, but they provide a critical part of the scientific foundation for decisions to protect public health across EPA's programs and regions under an array of environmental laws. The IRIS assessments, prepared

and maintained by the EPA, are collected in an electronic database containing information on human health effects that may result from exposure to various environmental chemicals. The heart of the IRIS system is its collection of computer files covering individual chemicals (Samet, 2014). These chemical files contain descriptive and quantitative information in the following categories: Oral reference doses and inhalation reference concentrations (RfDs and RfCs, respectively) for chronic non-carcinogenic health effects. Hazard identification, oral slope factors, and oral and inhalation unit risks for carcinogenic effects (US EPA, 2016).

6.1 Filtering

The code selects columns from the IRIS-EPA sheet corresponding to the mapping information (Table 14) and filters specific data from the intermediate data frames based on the following conditions:

- Data from the sheets 'RfC Toxicity Values' and 'RfD Toxicity Values' are filtered for 'BMD' and 'BMC' in the 'EXPERIMENTAL DOSE TYPE";

- Data from the 'Chemical Details' sheet are filtered for not missing 'CHEMICAL NAME'.

In summary, the code deals primarily with reading, cleaning, and transforming data from multiple sheets, and includes a post-processing step.

6.2 Recap of the mapping

The mapping and matching process specific to the IRIS-EPA database involved aligning and correlating variables to seamlessly integrate data. For example, in addressing "COM_NAME," the chemical details of the IRIS-EPA sheet had a variable named «chemical name».

Table 14. Recap of the mapping in IRIS-EPA

Variable name in the original source	Sheet name in the source	Variable name output
CHEMICAL NAME / CASRN / ORAL EXPOSURE STATUS / INHAL EXPOSURE STATUS	Chemical Details	COM_NAME / COM_CASNUMBER / ROUTE_EXP
CA ROUTE / STUDY CITATION DESC / PT DESCRIPTION / UR EXTRAPOLATION METHOD / SLOPE FACTOR VALUE / SLOPE FACTORY UNITS / PT STUDY SUBJECT / TUMOR SITE / CHEMICAL NAME	WOE Toxicity Values	ROUTE_EXP / PRINCIPAL_TOX_STUDY / KEY_ADVERSE_RESPONSE / DOSE_RESPONSE_MODEL / DEFINITION_FOR_BMD_ESTIMATION / HBGV_VALUE / HBGV_UNIT / STRAINS / GENDER / TARGET_TISSUE / COM_NAME
ROUTE / WOE DESCRIPTION, WOE NARRATIVE / WOE NARRATIVE / CHEMICAL NAME	WOE Details	ROUTE_EXP / TOXICITY / COM_NAME
PRINCIPAL STUDY / STUDY CITATION / PRINCIPAL CRITICAL DESCRIPTION / EXPERIMENTAL DOSE TYPE / POD VALUE / POD UNITS / RFD UNCERTAINTY FACTOR,	RfD Toxicity Values	PRINCIPAL_TOX_STUDY / KEY_ADVERSE_RESPONSE / REFERENCE_POINT / REFERENCE_POINT_VALUE / REFERENCE_POINT_UNIT / UNCERTAINTY_FACTOR / HBGV_VALUE /

Variable name in the original source	Sheet name in the source	Variable name output
RFC UNCERTAINTY FACTOR / RFD VALUE / RFD UNITS / DURATION TYPE / PRINCIPAL CRITICAL EFFECT SYSTEM / CHEMICAL NAME		HBGV_UNIT / STUDY_TYPE / TARGET_TISSUE / COM_NAME
PRINCIPALSTUDY / STUDY CITATION / PRINCIPAL CRITICAL DESCRIPTION / EXPERIMENTAL DOSE TYPE / POD VALUE / POD UNITS / RFC UNCERTAINTY FACTOR, RFC UNCERTAINTY FACTOR / RFC VALUE / RFC UNITS / DURATION TYPE / PRINCIPAL CRITICAL EFFECT SYSTEM / CHEMICAL NAME	RfC Toxicity Values	PRINCIPAL_TOX_STUDY / KEY_ADVERSE_RESPONSE / REFERENCE_POINT / REFERENCE_POINT_VALUE / REFERENCE_POINT_UNIT / UNCERTAINTY_FACTOR / HBGV_VALUE / HBGV_UNIT / STUDY_TYPE / TARGET_TISSUE / COM_NAME

7 CESAR

The CESAR is a database managed by Health Canada, the country's federal health department. CESAR serves as a comprehensive repository of information on the assessment of existing substances in Canada. The purpose of CESAR is to provide access to the results of assessments conducted under the Canadian Environmental Protection Act, 1999 (CEPA). According to the CEPA, Health Canada evaluates the potential risks posed to human health by existing substances in the Canadian environment. The database includes data on substance characteristics, exposure, environmental fate, and toxicological properties (Canada's approach to chemicals, 2022).

The ongoing risk assessment work conducted within the Chemicals Management Plan (CMP) primarily focuses on approximately 4,300 prioritized substances and involves the annual evaluation of 400 to 500 new substance notifications. The categorization process, completed in 2006 based on the information available at that time, identified key priorities. While the CMP continues to address these priorities and new substance notifications, the Government of Canada recognizes that the ongoing generation of new information could contribute to identifying substances of concern. Therefore, Health Canada (HC) and Environment and Climate Change Canada (ECCC) regularly review the available information through the Identification of Risk Assessment Priorities (IRAP) approach, as outlined in the Approach for Identification of Chemicals and Polymers as Risk Assessment Priorities under Part 5 of the Canadian Environmental Protection Act, 1999 (CEPA).

7.1 Access to the source

We accessed the *Status of substances prioritized under the Chemicals Management Plan (2006-2024) as of October 8, 2022* as an Excel file at https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-substances/status-substances-prioritized-cmp.html and selected substances defined as toxic if they are entering or may enter the environment in a quantity or concentration or under

conditions that «constitute or may constitute a danger in Canada to human life or health», according to CEPA, section 64 c).

As a first endeavour, we accessed the overlap between sources by considering the CAS number as the unique identifier. We recall that this identifier is present in the sheet "COMPONENT" of OpenFoodTox, in the sheet "Chemical Details" of IRIS-EPA, in the web-scraped Jecfa.csv and in the file accessed from the Canadian website (Table 15).

Table 15. Presence of CAS number in unfiltered sources

Characteristic	CESAR, N = 4,451[a]	IRIS-EPA, N = 575[a]	JECFA, N = 6,525[a]	OPENFOODTOX, N = 7,605[a]
CASN_missing	86 (1.9%)	6 (1.0%)	4,235 (65%)	1,239 (16%)

(a): n (%)

Figure 6 describes the overlap between sources in two scenarios: one without filtering CESAR by the definition of toxic substance that meets CEPA 64 c) (n= 4,551) and the other when filtering is applied (n = 192).

Among the 192 records filtered (

Appendix), 49 BMD assessments were manually identified. Of these, 14 CAS numbers were common to one of the other three sources. Among the remaining 35, the manual review revealed that 6 were taken from WHO, 3 from EPA, and 22 from Health Canada. The latter refers to risk assessments focused on boric acid and, therefore includes boric acid, its salts, and its precursors. It considers the total exposure of humans and other living organisms to boric acid, whether it is present in environmental media, food, or products. Health Canada has established a BMDL of 2.9 mg B/kg of body weight per day (bw/d) for the decrease in testicular weight derived from two studies of dog toxicity as a point of departure by Health Canada https://www.gazette.gc.ca/rp-pr/p1/2016/2016-07-23/html/notice-avis-eng.html.

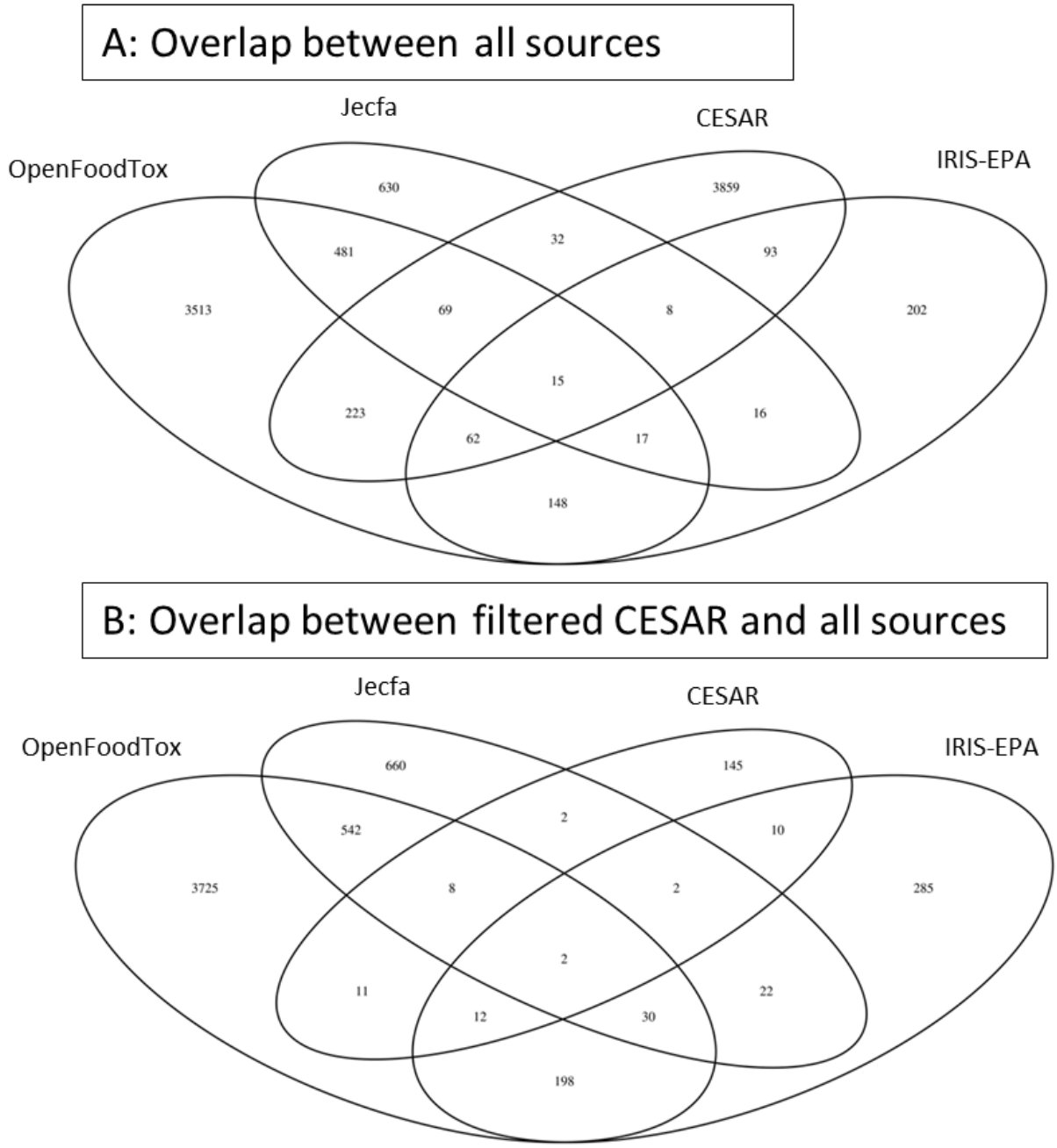

Figure 6. CAS numbers overlap in different sources. Upper panel: Venn diagram without filtering CESAR by CEPA 64 c) definition; Lower panel: Venn diagram with filtering

Given the limited original contribution of Health Canada and the prevalence of BMD evaluations largely included in the other databases, it is decided not to invest further in the review of this inventory.

8 Post processing

Following the completion of data extraction, an additional expert reviewed the database to ensure the accuracy of the inputted information. Although some data were directly uploaded to DistillerSR, others were manually entered. The final stages of the data extraction phase necessitated post-processing efforts, addressing variations in variable levels across different databases. Post-processing was conducted in both DistillerSR and R. Within DistillerSR, thorough checks on fields were performed, and new variables prefixed with "factor_xxx" were created to facilitate future categorization of variables. The variables pertaining to factor levels included levels as detailed in

8.1 Data Source Information

The most relevant variables have been defined to provide toxicological information for chemicals.

The variable 'Year' indicated the publication year of the information source. 'BMD_adopted' specifies whether the JECFA database benchmark dose analysis was adopted.

The database source from which the information was retrieved is reported under the variable 'db_source' and its source of retrieval is under the variable 'information_source_URL'. For this variable, it indicated whether the information was retrieved from OpenFoodTox, IRIS-EPA or JECFA.

Table 3.

8.2 Download from DistillerSR

The download through DistillerSR can be done accessing the project MESE_OC/EFSA/2022/03-SC02_SLR_inventory_BMRs following this path:

1) Reports->Datarama:
    a. Basic option, report settings: report format CSV
    b. Data to Display: Data Extraction
2) Run report-> a CSV will be dowloaded.

In Annex D is available the database as can be downloaded from DistillerSR, the last download was done on 02/02/2023.

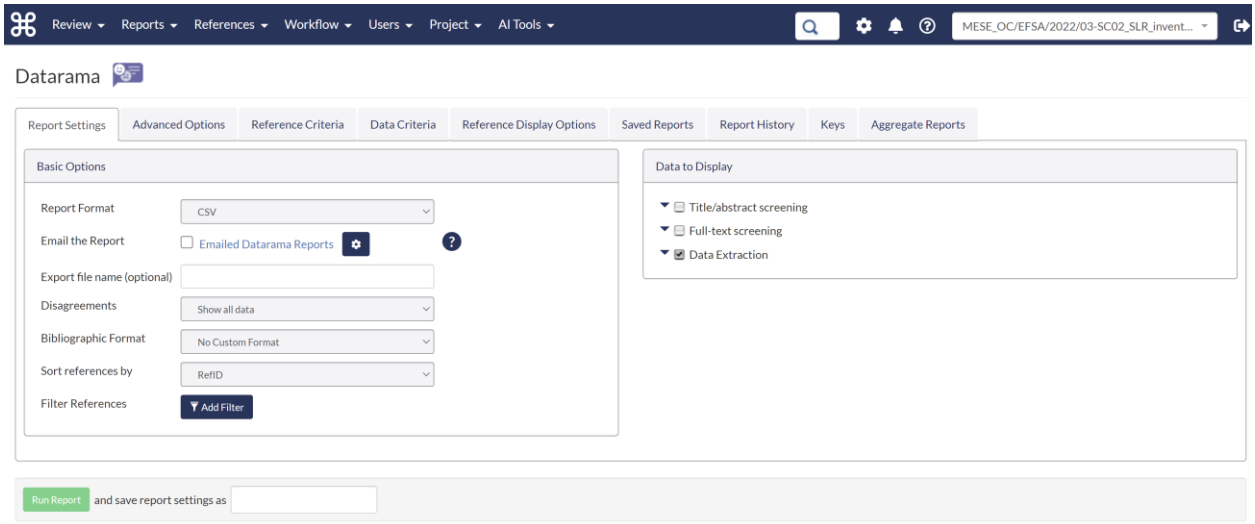

Figure 7. Screenshot from DistillerSR representing the download method

8.3 Data Import and Cleaning in R

The most recent update of the inventory was carried out on 01/02/2024, and, subsequently, it was imported into R for data cleaning and the selection of final variables. The code to import and clean the data is made available. Annex E reports the Inventory cleaned in R.

8.3.1 Harmonization of endpoints

The harmonization of endpoints in the database involved the standardization and alignment of diverse data points or parameters. Harmonisation was intended to ensure consistency and comparability between different studies or sources.

Since the inventory was built from data from different sources, harmonization addressed issues related to terminology and nomenclature, ensuring that terms used to describe endpoints were consistent across studies. This helped prevent confusion and misinterpretation when comparing data from different sources.

The variable 'endpoint' was divided into three variables into DistillerSR to facilitate the creation of harmonized levels. The first variable, "endpoint", corresponded to the full description of the key endpoint for which BMD analysis was conducted. This level was harmonized manually across the different databases to obtain a standardized format. Therefore, the variable 'factor_endpoint' was subdivided into change, decrease, incidence, increase, induction, and prevalence to indicate different aspects of the key endpoint. It allowed a detailed and nuanced representation of the variations in the key endpoint's behaviour. This task was performed manually.

The third level "Factor_Endpoint2" indicated more specifically the various responses and enabled a more granular analysis of the specific factors. The levels available for this variable are the following: Liver haemangiosarcoma, Tail muscular atrophy, Anorectic effects, Hepatocellular carcinoma, Liver weight, Kidney weight, Body weight, Erythrocytes, leucocytes and platelets, Serum plasma globulin, Other.

8.3.2 Toxicity

Usually, there are one or two organs that show the major toxic effect. These are referred to as target organs of toxicity of the particular substance. According to the framework derived from TOXRIC (Wu et al., 2022) we have created a database in a longer format according to the toxicity.

9 Inventory

The results shown in the tables in the Appendix C are from the inventory in wide format, whereas the results in Table A.5.1 are from the long format.

9.1 Sources characteristics

Table A.1.1 provides a breakdown of the characteristics by, showing the number and percentage of events for each year in the overall Inventory (N = 593), as well as specific components: IRIS-EPA (N = 108), JECFA (N = 28), and OpenFoodTox (N = 457).

9.2 Population characteristics

Table A.2.1 reports information on the study population, Table A.2.1 reports the strains of easch species according to the database.

Rats are the most common species (65%), and humans are represented to a lesser extent overall (12%) and in the JECFA (25%) component. While rats and humans dominate the dataset, there is notable variability in the number of cases for other species. Fischer 344 rats are prominently featured across all the sources (40%). Despite variability across datasets, the proportion of rat cases remains consistently high, indicating the widespread use of rats in different studies and contexts. In the IRIS-EPA dataset, the B6C3F1 strain of mice stands out with 281 cases, representing 86% of the total. This strain is also notable in the OpenFoodTox dataset, accounting for 84% of the total mouse cases. This suggests a preference for the B6C3F1 strain in the toxicological assessments. The presence of multiple species underscores the importance of comparative studies in toxicology and risk assessment, allowing researchers to assess the impact of substances across different organisms. Overall, the strain data reveal preferences for certain strains, particularly among rodents, in toxicological studies across different datasets. These preferences may be influenced by availability, historical usage, and specific research requirements.

Across all datasets, there is a notable representation of both male and female test subjects, with males comprising 40% and females 23% of the overall dataset. This suggests a balanced approach to sex representation in toxicological studies, which is essential for understanding potential gender-based differences in response to exposures.

9.3 Dose response models

Table A.3.1 reports the descriptive statistics of the statistical characteristics.

The distribution of BMR values across different datasets reveals variations in the selection criteria for defining response levels. For the 10% BMR value, the IRIS-EPA (93%) and OpenFoodTox (95%) components had the highest representation in quantal endpoints. The value of 1 SD BMR was prevalent in IRIS-EPA (71%) continuous endpoints. In the category of 5% BMR, OpenFoodTox showed the highest prevalence (68%). These discrepancies underscore the lack of standardization in selecting BMR values across studies, which may introduce challenges when comparing and synthesizing results. Standardizing BMR selection criteria could enhance consistency and comparability in risk assessment methodologies across diverse regulatory bodies and improve the reliability of toxicological evaluations.

The analysis of BMR value selection for BMD estimation across different datasets indicates disparities in the preferred BMD metrics. Extra risk was the primary adopted definition (465 occurrences). Standard deviation was the predominant definition for continuous endpoints in IRIS-EPA (71%), whereas relative definition for continuous endpoints in OpenFoodTox (61%). The distribution of BMD estimation criteria varies across datasets, with some datasets showing a preference for relative definitions or a combination of different BMR values. These findings underscore the diversity in risk assessment methodologies employed across datasets and

highlight the importance of understanding dataset-specific factors when interpreting BMD estimates.

The multistage was the dose response model most represented in OpenFoodTox (33%) for quantal data, while the log-logistic model was used in 20 IRIS-EPA assessments. A further investigation into the other category suggests that the linear model was prevalent in IRIS-EPA (18 occurrences, 95%) for the continuous data. Model averaging was common in OpenFoodTox (77%). The presence of unknown or unreported factors and model parameters underscores the challenges associated with interpreting and synthesizing toxicological data from heterogeneous sources, emphasizing the importance of standardized reporting practices and data transparency.

Table A.4.1. reports the endpoints available in the three sources in the Inventory. The distribution of endpoints across datasets reveals variations in the types of effects assessed, emphasizing the diversity in toxicological investigations. Most of the endpoints in the IRIS-EPA (66%) and OpenFoodTox (32%) were related to 'incidence', while 'decrease' was observed in 23% of the IRIS-EPA and 12% of the OpenFoodTox endpoints. Induction was a notable endpoint in OpenFoodTox (34%). Induction refers to the process by which a toxic substance or agent triggers a specific biological response or effect in an organism that is measurable and relevant to assessing its toxicity. Liver haemangiosarcoma was the most reported endpoint (29%) for the OpenFoodTox. Hepatocellular carcinoma was the most reported endpoint (11%) for JECFA component. Hepatic effect (14%) and pain sensitivity (14%) were the two most reported endpoints for the IRIS-EPA database.

The predominance of liver haemangiosarcoma and hepatocellular carcinoma endpoints underscores the emphasis on hepatic health assessment, likely due to the liver's crucial role in detoxification processes. Additionally, the inclusion of endpoints such as neurobehavioral effect and neurodevelopmental effect across studies reflects a comprehensive evaluation of neurological outcomes, which is essential given the potential neurotoxic effects of various substances.

Additionally, a significant number of endpoints are categorized as unknown, comprising 276 (60.4%) cases in OpenFoodTox, underscoring the need for standardized reporting practices and improved transparency in data documentation to facilitate accurate interpretation and comparison of toxicological findings across datasets.

9.4 Toxicity and target organs

Table A.5.1 reports the target organ and the toxicity level in the inventory. The distribution of toxicity and target organ levels varies across datasets, reflecting the diverse nature of toxicological studies and the range of health effects evaluated. Overall, the liver was the most frequently targeted organ (50%), particularly in OpenFoodTox (60%), while in IRIS-EPA it was 16%. Liver toxicity emerges as a predominant focus across all datasets, indicating its significance in toxicological research. This underscores the importance of evaluating hepatic effects when assessing the safety of substances. The nervous system was targeted in 5.8% of the total cases, with higher occurrences in IRIS-EPA (24%) and lower in JECFA (14%). The

prevalence of neurotoxicity varies notably across datasets, suggesting differences in the prioritization of neurological health effects. This underscores the need for targeted assessments of neurotoxic outcomes in toxicology studies. The gastrointestinal system was affected in 5.7% of the total cases, with a similar distribution across components.

Carcinogenicity was the most prevalent type of toxicity (22%), with similar occurrences in the three sources. Genotoxicity (17%) and mutagenicity (16%) were less represented, with higher proportions in OpenFoodTox. Consistent assessment of carcinogenicity and genotoxicity across datasets indicates a shared emphasis on evaluating potential cancer risks associated with chemical exposures, reinforcing the importance of prioritizing studies on these endpoints in toxicological research. Hepatotoxicity was observed in 15% of the overall cases, with a relatively higher prevalence in OpenFoodTox (17%). Neurotoxicity, developmental and reproductive toxicity and nephrotoxicity accounted for approximately 4%. Other toxicities had fewer occurrences.

9.5 BMR conversion

In case the relative definition is not used, the conversion was provided under m(0) and sigma for recalculation as BMR*sigma/m(0). Figure 8 shows the histogram of the transformed BMR values.

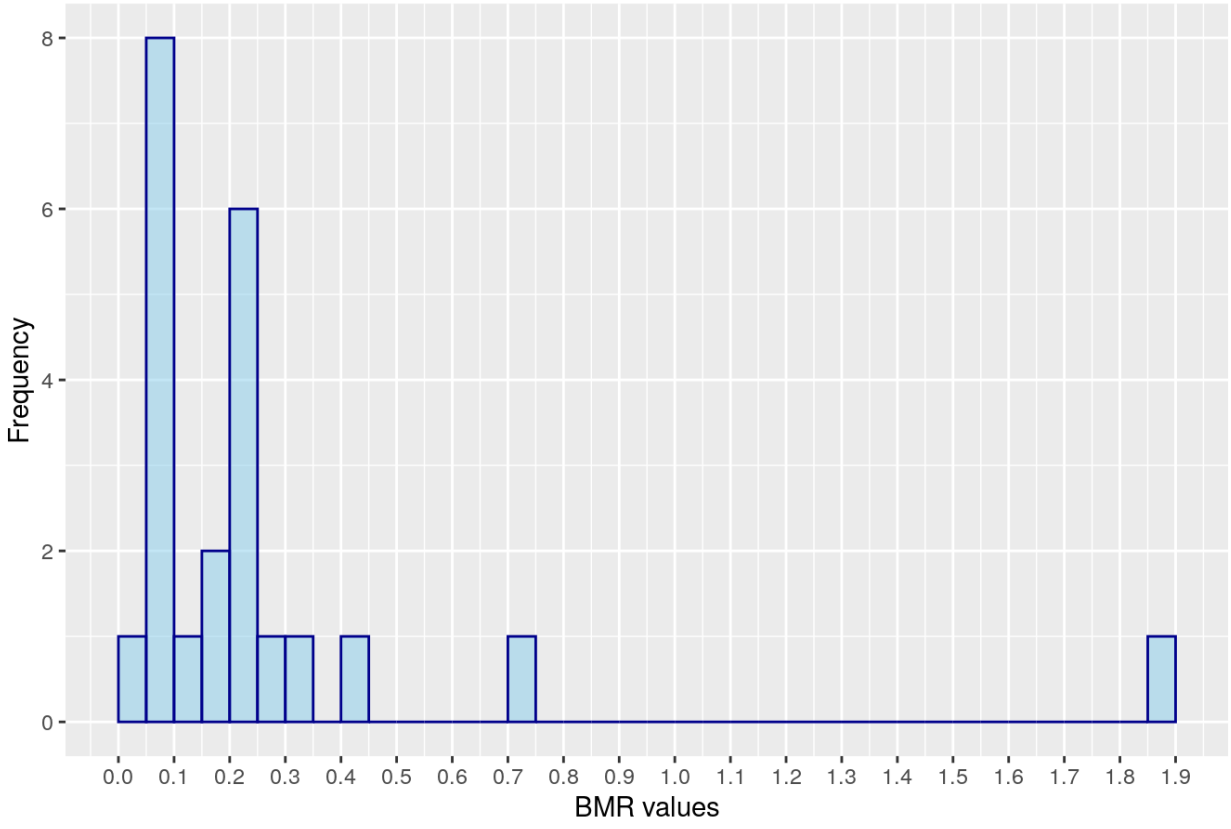

Figure 8. Histogram of BMR values converted

10 Conclusions

This study aimed to define BMR for specific endpoints in toxicological studies and use the BMD modelling framework to establish HBGV or RP. Through a comprehensive search, an inventory of BMR was created for various endpoints, types of studies, substances evaluated, and other aspects of the assessment characterization.

The search followed two main steps to gather information. First, primary sources of information were used, including published databases and reports. These primary sources served as valuable data repositories for the study. Second, secondary sources of information were explored, including risk assessment reports and articles from various risk assessment agencies. A third source of information was consulted to fill in the missing information related to the chemical substance itself.

The findings highlight the diversity in risk assessment methodologies employed across sources. This diversity may reflect differences in regulatory requirements, scientific practices, or research objectives, underscoring the need for flexibility and adaptability in risk assessment approaches. Variations observed in endpoint classification, toxicity profiles, and species sensitivity highlight the importance of considering dataset-specific factors when interpreting toxicological findings and extrapolating results to human health risks. Given the variability observed, transparency in reporting statistical models and BMD estimation criteria becomes crucial for ensuring reproducibility and facilitating comparisons across studies. Clear documentation of methods and rationale behind model selection can enhance the credibility and utility of risk assessment findings.

The inclusion of various species and endpoints in toxicological studies contributes to a robust understanding of potential health risks associated with exposure to different substances, thereby informing regulatory decisions and public health policies.

The findings underscore the need for standardized reporting practices, data transparency, and collaborative efforts in toxicological research to enhance the reliability and utility of toxicological data for regulatory decision-making and public health protection.

References

Bhalla TC, Monika, Sheetal, & Savitri, 2019. Chapter 12—International Laws and Food-Borne Illness. In R. L. Singh & S. Mondal (A c. Di), Food Safety and Human Health (pp. 319–371). Academic Press. https://doi.org/10.1016/B978-0-12-816333-7.00012-6

Canada's approach on chemicals substances, 2022. https://www.canada.ca/en/health-canada/services/chemical-substances/canada-approach-chemicals.html

Crump K, 1984. A new method for determining allowable daily intakes*1. Fundamental and applied toxicology, 4(5).

Crump KS, 1995. Calculation of Benchmark Doses from Continuous Data. Risk Analysis, 15(1), 79–89. https://doi.org/10.1111/j.1539-6924.1995.tb00095.x

Dekkers S, de Heer C and Rennen MAJ, 2001. Critical effect sizes in toxicological risk assessment: A comprehensive and critical evaluation. Environmental Toxicology and Pharmacology, 10(1–2), 33–52. https://doi.org/10.1016/s1382-6689(01)00068-0

Dorne. C. M., Richardson, J., Livaniou, A., Carnesecchi, E., Ceriani, L., Baldin, R., Kovarich, S., Pavan, M., Saouter, E., Biganzoli, F., Pasinato, L., Zare Jeddi, M., Robinson, T. P., Kass, G. E. N., Liem, A. K. D., Toropov, A. A., Toropova, A. P., Yang, C., Tarkhov, A., … Bassan, A. (2021). EFSA's OpenFoodTox: An open source toxicological database on chemicals in food and feed and its future developments. Environment International, 146, 106293. https://doi.org/10.1016/j.envint.2020.106293

Samet J, 2014. Preface. In Review of EPA's Integrated Risk Information System (IRIS) Process. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK230071/

More SJ, Bampidis V, Benford D, Bragard C, Halldorsson TI, Hernández-Jerez AF, Bennekou S H, Koutsoumanis K, Lambré C, Machera K, Mennes W, Mullins E, Nielsen SS, Schrenk D, Turck D, Younes M, Aerts M, Edler L, Sand S, … Schlatter J, 2022. Guidance on the use of the benchmark dose approach in risk assessment. EFSA Journal, 20(10), e07584. https://doi.org/10.2903/j.efsa.2022.7584

Munns WR, Kroes R, Veith G, Suter Ii GW, Damstra T and Waters MD, 2003. Approaches for Integrated Risk Assessment. Human and Ecological Risk Assessment: An International Journal, 9(1), 267–272. https://doi.org/10.1080/713609863

R Core Team, 2023. R: A Language and Environment for Statistical Computing [Software]. R Foundation for Statistical Computing. https://www.r-project.org/

Fernández R, 2021. Risk analysis of food safety hazards. FoodDrinkEurope. https://www.fooddrinkeurope.eu/risk-analysis-of-food-safety-hazards/

Slob W, 2017. A general theory of effect size, and its consequences for defining the benchmark response (BMR) for continuous endpoints. Critical Reviews in Toxicology, 47(4), 342–351. https://doi.org/10.1080/10408444.2016.1241756

Systematic Review and Literature Review Software by DistillerSR, 2024. DistillerSR. https://www.distillersr.com/

US EPA, 2016. Chemicals and Toxics Topics [Collections and Lists]. https://www.epa.gov/environmental-topics/chemicals-and-toxics-topics

US EPA, 2022. Benchmark Dose Software. Version 3.3. User Guidance. https://ordspub.epa.gov/ords/eims/eimscomm.getfile?p_download_id=545595

WHO, 2020. Chapter5-dose-response.pdf. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.who.int/docs/default-source/food-safety/publications/chapter5-dose-response.pdf?sfvrsn=32edc2c6_5

WHO, 2023. WHO|JECFA Database. https://apps.who.int/food-additives-contaminants-jecfa-database/

Wu L, Yan B, Han J, Li R, Xiao J, He S and Bo X, 2022. TOXRIC: A comprehensive database of toxicological data and benchmarks. Nucleic Acids Research, 51(D1), D1432–D1445. https://doi.org/10.1093/nar/gkac1074

Abbreviations

ADIs	Acceptable daily intakes
BMC	Benchmark concentration
BMCL	Benchmark concentration lower limit
BMD	Benchmark dose
BMDL	Benchmark dose lower limit
BMR	Benchmark response
CAC	Codex alimentarius commission
CAS	Chemical abstracts service
CEPA	Canadian environmental protection act
CESAR	Canada's existing substances assessment repository
CMP	Chemicals management plan
CSF	Cancer slope factor
DSL	Canada's domestic substances list
ECCC	Environment and Climate Change Canada
EFSA	European food safety authority
EPA	Environmental protection agency
FAO	Food and Agriculture Organization
FAS	Food additive status
HBGV	Health based guidance value
HC	Health canada
IRIS-EPA	Integrated risk information system-united states environmental protection agency
IRAP	Identification of Risk Assessment Priorities
JECFA	Joint FAO/WHO Expert Committee on Food Additives
LOAEL	Lowest observed adverse effect level
NITE	National Institute of Technology and Evaluation
NOAEL	No observed adverse effect level
Rfc	Inhalation reference concentration
RfD	Oral reference dose
RP	Reference point
SD	Standard deviation
TDIs	Tolerable daily intakes
TRS	Technical report series

WHO	The world health organization

Appendix A – List of variables for the Inventory

For each variable, is reported the corresponding in EFSA catalogue and, when available, the levels in the catalogue.

Variable name	EFSA catalogue	EFSA levels
Year	No	No
BMD adopted	No	No
db_source	No	No
information_source_URL	No	No
com_name	PARAMCLS	The terms in the catalogue are numerous. Check which ones are of interest
substance_CASRN_number	No	No
molecular_formula	No	No
substance_description	No	No
assessment_type	No	No
exposure_routes	ROUTEXP	OECD1842PL DERMAL OECD1983PL IMPLANTATION OECD2400PL SUBCUTANEOUS OECD0PL OTHER OECD1990PL INFUSION OECD1991PL INHALATION: UNSPECIFIED OECD1992PL INHALATION: AEROSOL OECD1993PL INHALATION: DUST OECD1994PL INHALATION: GAS OECD1995PL INHALATION: VAPOUR OECD2010PL INTRAMUSCULAR OECD2011PL INTRAPERITONEAL OECD2012PL INTRATRACHEAL OECD2013PL INTRAVENEOUS OECD2228PL ORAL: CAPSULE OECD2229PL ORAL: DRINKING WATER OECD2230PL ORAL: FEED OECD2231PL ORAL: GAVAGE OECD2234PL ORAL: UNSPECIFIED OECDX001PL INTRAVAGINAL OECDX002PL VECTOR BITE
principal_toxicological_study	No	No
principal_toxicological_study_DOI/URL	No	No
Endpoint	No	No
Factor_Endpoint	No	No
Factor_Endpoint2	No	No
key_adverse_response	No	No
dose_response_model	No	No

selected_BMR_value	No	No
definition_for_bmd_estimation	No	No
reference_point	ENDPOINT_HGV	BMDL01[CHD003EP] BMDL005[CHD018EP] BMDL10[CHD019EP] BMDL[CHD009EP] BMD05 [CHD080EP]
reference_point_value	No	No
reference_point_unit	UNIT	mg/kg bw/day [G211A]
uncertainty_factor	No	No
health_based_guidance_values	ENDPOINT_HGV	ADI[CHD030EP] RfD[CHD046EP] TDI[CHD046EP]
HBGV_value	No	No
HBGV_unit	UNIT	mg/kg bw/day [G211A]
study_duration_type	TEST_TYPE	chronic/long term toxicity [CHD004TT] short term dietary toxicity [CHD010TT] subacute [CHD012TT] subchronic [CHD013TT] acute toxicity [CHD016TT] short-term toxicity [CHD017TT] long-term toxicity [CHD018TT]
study_species	No	No
study_strains	No	No
sex_of_test_subjects	Male [G1]	Male [G1]
study_experimental_model	TEST_TYPE	in vitro [CHD061TT]
type_of_substance_toxicity	TOXICITY	TOX01A SYSTEMIC TOX02A HEPATOTOXICITY TOX17A NoT REPORTED TOX03A NEPHROTOXICITY TOX04A NEUROTOXICITY TOX05A REPRODUCTIVE TOX06A TERATOGENIC TOX07A DEVELOPMENTAL TOX08A PULMONARY AND CARDIAC TOX09A IMMU NoTOXICITY TOX10A HEMOPOIETIC TOX11A IRRITATION TOX12A ENDOCRINE TOX13A MUSCLO-SKELETAL TOX14A NoNE TOX18A Digestive TOX19A Other toxicities
target_organ	TARGET TISSUE	Cardiovascular / Haematological [TT001A] Digestive [TT002A] Glandular [TT003A] Neurologic [TT004A] Respiratory [TT005A] Urogenital [TT006A] Other [TT007A]

mechanism_of_toxicity	No	No
single_grouped_chem_assess	No	No
specify_group_chem_assessment	No	No
type_of_data	No	No
statistical_paradigm	No	No
m(0)	No	No
Sigma	No	No
transformed_BMR_value	No	No

Appendix B – Cesar

The Table reports the 49 BMD assessments filtered among the 192 available under section 63c in CESAR. Of these, 14 CAS numbers were common to one of the other three sources.

Chemical Name	CASN	Prioritization Mechanism	Initiative	Assessment Document	Status of Publication	Section	Sources from which the BMD is derived							Overlap of sources		
							JECFA	US EPA	NTP	Health Canada	WHO	EFSA	ATSDR	OFT	IRIS-EPA	JECFA
1-Butanol	71-36-3	ITC	CMP3	Alcohols	DSAP	64c	Yes	NA	NA	NA	NA	NA	NA	2	1	0
Benzenemethanol	100-51-6	ITC	CMP3	Alcohols	DSAP	64c	Yes	NA	NA	NA	NA	NA	NA	1	0	6
Methanol	67-56-1	ITC	CMP3	Alcohols	DSAP	64c	Yes	NA	NA	NA	NA	NA	NA	2	1	2
Propane, 1-bromo-	106-94-5	ITC	CMP3	Alkyl Halides Group	DSAP	64c	NA	Yes	NA	NA	NA	NA	NA	0	0	0
9,10-Anthracenedione, 1-hydroxy-4-[(4-methylphenyl)amino]-	81-48-1	ITC	CMP3	Anthraquinones Group	FSAP	64c	NA	NA	Yes	NA	NA	NA	NA	0	0	0

2-Naphthalenol, 1-[[4-(phenylazo)phenyl]azo]-	85-86-9	ITC	CMP2	Aromatic Azo and Benzidine-based Substance Grouping: Certain Azo Solvent Dyes	FSAP	64c	NA	NA	Yes	NA	NA	NA	NA	1	0	0
2-Naphthalenol, 1-[(4-methyl-2-nitrophenyl)azo]-	2425-85-6	ITC	CMP1 - Batch 3	Aromatic Azo and Benzidine-based substance grouping: Pigments	FSAP	64c	NA	NA	Yes	NA	NA	NA	NA	0	0	0
Ammonium boron oxide ((NH4)B5O8)	12007-89-5	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Borate(1-), tetrafluoro-, ammonium	13826-83-0	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0

Borate(1-), tetrafluoro-, cadmium (2:1)	14486-19-2	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Borate(1-), tetrafluoro-, lead(2+) (2:1)	13814-96-5	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Borax (B4Na2O7.10H2O)	1303-96-4	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boric acid	11113-50-1	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	2	0	0
Boric acid (H3B3O6)	13460-51-0	RICL	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0

Boric acid (H3BO3)	10043-35-3	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	1
Boric acid (H3BO3), ammonium salt (1:?)	27522-09-4	RICL	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boric acid (H3BO3), compd. with 1-amino-2-propanol (1:1)	68003-13-4	RICL	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boric acid (H3BO3), sodium salt	13840-56-7	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boric acid (H3BO3), triammonium salt	22694-75-3	RICL	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boric acid (HBO2), lithium salt	13453-69-5	ITC	CMP2	Boric Acid, its salts and precursors	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0

				substance grouping												
Boric acid, 2-aminoethyl ester	68130-12-1	RICL	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boric acid, ammonium salt	11128-98-6	RICL	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boric acid, zinc salt	1332-07-6	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boron lithium oxide (B4Li2O7)	12007-60-2	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boron oxide (B2O3)	1303-86-2	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0

Boron silver oxide (B4Ag2O7)	12271-95-3	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Boron sodium oxide (B4Na2O7)	1330-43-4	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	1	0	1
Boron zinc oxide (B6Zn2O11)	12767-90-7	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Cobalt, borate neodecanoate complexes	68457-13-6	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Perboric acid (HBO(O2)), sodium salt	04/04/7632	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
Perboric acid (HBO(O2)), sodium salt, monohydrate	10332-33-9	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0

Perboric acid, sodium salt	11138-47-9	ITC	CMP2	Boric Acid, its salts and precursors substance grouping	DASP-RDAE	64a, c	NA	NA	NA	Yes	NA	NA	NA	0	0	0
1,3,5-Triazine-2,4,6-triamine	108-78-1	ITC	CMP2	Certain organic flame retardants substance grouping - Melamine	DSAP	64c	NA	NA	NA	NA	Yes	Yes	NA	2	0	0
2-Propanol, 1,3-dichloro-, phosphate (3:1)	13674-87-8	ITC	CMP2	Certain organic flame retardants substance grouping - TCPP and TDCPP	DSAP	64c	NA	Yes	NA	NA	NA	NA	NA	0	0	0
2-Propanol, 1-chloro-, phosphate (3:1)	13674-84-5	ITC	CMP2	Certain organic flame retardants substance grouping - TCPP and TDCPP	DSAP	64c	NA	Yes	NA	NA	NA	NA	NA	0	0	0
Distillates (coal tar), upper	65996-91-0	ITC	CMP1 - Petroleum	Coal tars and their distillates	FSAP	64a, c	NA	NA	NA	Yes	Yes	NA	NA	0	0	0
Pitch, coal tar, high-temp.	65996-93-2	ITC	CMP1 - Petroleum	Coal tars and their distillates	FSAP	64a, c	NA	NA	NA	Yes	Yes	NA	NA	0	0	0

Tar oils, coal	65996-82-9	ITC	CMP1 - Petroleum	Coal tars and their distillates	FSAP	64a, c	NA	NA	NA	Yes	Yes	NA	NA	0	0	0
Tar, coal	8007-45-2	ITC	CMP1 - Petroleum	Coal tars and their distillates	FSAP	64a, c	NA	NA	NA	Yes	Yes	NA	NA	0	0	0
Tar, coal, high-temp.	65996-89-6	ITC	CMP1 - Petroleum	Coal tars and their distillates	FSAP	64a, c	NA	NA	NA	Yes	Yes	NA	NA	0	0	0
Tar, coal, low-temp.	65996-90-9	ITC	CMP1 - Petroleum	Coal tars and their distillates	FSAP	64a, c	NA	NA	NA	Yes	Yes	NA	NA	0	0	0
Phenol, isopropylated, phosphate (3:1)	68937-41-7	ITC	CMP3	Flame Retardants Group	DSAP	64a, c	NA	NA	NA	NA	NA	NA	Yes	0	0	0
Phosphoric acid, triethyl ester	78-40-0	ITC	CMP3	Flame Retardants Group	DSAP	64c	NA	NA	NA	NA	NA	NA	Yes	0	0	0
2-Furanmethanol	98-00-0	ITC	CMP3	Furan compounds	DSAP	64c	NA	NA	Yes	NA	NA	NA	NA	3	0	0
Furan, tetrahydro-	109-99-9	ITC	CMP3	Furan compounds	DSAP	64c	NA	NA	Yes	NA	NA	NA	NA	0	1	0

Carbamic acid, ethyl ester	51-79-6	ITC	CMP2	Internationally Classified Substance Grouping	FSAP	64c	NA	NA	Yes	NA	Yes	NA	NA	1	1	0
Methanone, diphenyl-	119-61-9	ITC	CMP3	Methanone, diphenyl- (Benzophenone)	FSAP	64c	NA	NA	NA	NA	NA	Yes	NA	2	0	0
Quinoline	91-22-5	ITC	CMP1	Quinoline	FSAP	64a, c	NA	NA	NA	Yes	Yes	NA	NA	0	1	0
Methanaminium, N-[4-[[4-(dimethylamino)phenyl]phenylmethylene]-2,5-cyclohexadien-1-ylidene]-N-methyl-, chloride	569-64-2	ITC	CMP3	Triarylmethanes Group	FSAP	64a, c	Yes	NA	NA	NA	NA	NA	NA	1	0	0

ITC, Identified at the time of Categorization; RICL, Revised In Commerce List; DSAP, Draft screening assessment published; FSAP, Final screening assessment published; DASP-RDAE, Draft screening assessment published; Revised draft assessment expected October 2022 - March 2024; CMP, Chemicals Management Plan; USEPA. United States Environmental Protection Agency; WHO, World Health Organization; EFSA, European Food Safety Authority; ATSDR, Agency for Toxic Substances and Disease Registry; CASN, Chemical Abstracts Service Number

Appendix C – Descriptive Statistics of the Inventory

A.1.1. Sources characteristics

Year	Overall, N = 593[a]	IRIS-EPA, N = 108[a]	JECFA, N = 28[a]	OpenFoodTox, N = 457[a]
1995	4 (0.7%)	4 (3.7%)	0 (0%)	0 (0%)
1997	1 (0.2%)	1 (0.9%)	0 (0%)	0 (0%)
1998	4 (0.7%)	4 (3.7%)	0 (0%)	0 (0%)
2000	2 (0.3%)	2 (1.9%)	0 (0%)	0 (0%)
2001	4 (0.7%)	4 (3.7%)	0 (0%)	0 (0%)
2002	4 (0.7%)	4 (3.7%)	0 (0%)	0 (0%)
2003	5 (0.8%)	5 (4.6%)	0 (0%)	0 (0%)
2004	8 (1.3%)	8 (7.4%)	0 (0%)	0 (0%)
2005	3 (0.5%)	3 (2.8%)	0 (0%)	0 (0%)
2006	4 (0.7%)	1 (0.9%)	2 (7.1%)	1 (0.2%)
2007	24 (4.0%)	2 (1.9%)	2 (7.1%)	20 (4.4%)
2008	21 (3.5%)	3 (2.8%)	1 (3.6%)	17 (3.7%)
2009	38 (6.4%)	14 (13%)	0 (0%)	24 (5.3%)
2010	29 (4.9%)	14 (13%)	0 (0%)	15 (3.3%)
2011	103 (17%)	6 (5.6%)	13 (46%)	84 (18%)
2012	43 (7.3%)	4 (3.7%)	1 (3.6%)	38 (8.3%)
2013	14 (2.4%)	5 (4.6%)	2 (7.1%)	7 (1.5%)
2014	8 (1.3%)	1 (0.9%)	0 (0%)	7 (1.5%)
2015	31 (5.2%)	0 (0%)	2 (7.1%)	29 (6.3%)
2016	40 (6.7%)	15 (14%)	4 (14%)	21 (4.6%)
2017	138 (23%)	5 (4.6%)	0 (0%)	133 (29%)
2018	23 (3.9%)	3 (2.8%)	0 (0%)	20 (4.4%)
2019	24 (4.0%)	0 (0%)	0 (0%)	24 (5.3%)
2020	17 (2.9%)	0 (0%)	0 (0%)	17 (3.7%)
2021	1 (0.2%)	0 (0%)	1 (3.6%)	0 (0%)

(a): n (%)

A.2.1. Population characteristics

Table A.2.1. Characteristics of the population

Characteristic	Overall, N = 593[(a)]	IRIS-EPA, N = 108[(a)]	JECFA, N = 28[(a)]	OpenFoodTox, N = 457[(a)]
Factor_Study species				
American mink	29 (4.9%)	0 (0%)	0 (0%)	29 (6.3%)
Cat	1 (0.2%)	0 (0%)	0 (0%)	1 (0.2%)
Cattle	1 (0.2%)	0 (0%)	0 (0%)	1 (0.2%)
Dog	6 (1.0%)	2 (1.9%)	1 (3.6%)	3 (0.7%)
Hamster	1 (0.2%)	1 (0.9%)	0 (0%)	0 (0%)
Human	70 (12%)	4 (3.7%)	7 (25%)	59 (13%)
Mallard duck	2 (0.3%)	0 (0%)	0 (0%)	2 (0.4%)
Monkey	1 (0.2%)	1 (0.9%)	0 (0%)	0 (0%)
Mouse	80 (13%)	28 (26%)	7 (25%)	45 (9.8%)
Other	7 (1.2%)	1 (0.9%)	1 (3.6%)	5 (1.1%)
Pig	9 (1.5%)	0 (0%)	2 (7.1%)	7 (1.5%)
Rat	386 (65%)	71 (66%)	10 (36%)	305 (67%)
Factor_Study_strains				
B6C3F1	55 (16%)	24 (26%)	4 (36%)	27 (11%)
Fischer 344	138 (40%)	21 (23%)	5 (45%)	112 (46%)
Other	17 (4.9%)	8 (8.7%)	2 (18%)	7 (2.9%)
Sprague-Dawley	60 (17%)	20 (22%)	0 (0%)	40 (17%)
Wistar	53 (15%)	19 (21%)	0 (0%)	34 (14%)
Wistar Han	21 (6.1%)	0 (0%)	0 (0%)	21 (8.7%)
Unknown	249	16	17	216
Factor_Sex of test subjects				
Female	122 (23%)	25 (24%)	3 (17%)	94 (23%)
Male	209 (40%)	57 (55%)	2 (11%)	150 (37%)
Male/Female	197 (37%)	22 (21%)	13 (72%)	162 (40%)
Unknown	65	4	10	51

(a): n (%)

Tables A.2.2. Strains of each species according to the database

Characteristic	OpenFoodTox										
Factor_Study_strains	Dog, N = 3[a]		Human, N = 59[a]		Mouse, N = 45[a]	Other, N = 5[a]	Rat, N = 305[a]	Pig, N = 7[a]	American mink, N = 29[a]	Cat, N = 1[a]	Cattle, N = 1[a]
B6C3F1					27 (84%)	0 (0%)	0 (0%)				
Fischer 344					0 (0%)	0 (0%)	112 (54%)				
Other					5 (16%)	1 (100%)	1 (0.5%)				
Sprague-Dawley					0 (0%)	0 (0%)	40 (19%)				
Wistar					0 (0%)	0 (0%)	34 (16%)				
Unknown	3		59		13	4	97	7	29	1	1
Wistar Han					0 (0%)	0 (0%)	21 (10%)				
	JECFA										
Factor_Study_strains	Dog, N = 1[a]		Human, N = 7[a]		Mouse, N = 7[a]	Other, N = 1[a]	Rat, N = 10[a]	Pig, N = 2[a]			
B6C3F1					3 (60%)		1 (17%)				
Fischer 344					1 (20%)		4 (67%)				
Other					1 (20%)		1 (17%)				
Sprague-Dawley											
Wistar											
Unknown	1		7		2	1	4	2			
Wistar Han											
	IRIS-EPA										
Factor_Study_strains	Dog, N = 2 [a]	Hamster, N = 1[a]	Human, N = 4[a]	Monkey, N = 1[a]	Mouse, N = 28[a]	Other, N = 1[a]	Rat, N = 71[a]				
B6C3F1					24 (86%)		0 (0%)				
Fischer 344					0 (0%)		21 (33%)				
Other					4 (14%)		4 (6.3%)				
Sprague-Dawley					0 (0%)		20 (31%)				
Wistar					0 (0%)		19 (30%)				

Unknown	2	1	4	1	0	1	7				
Wistar Han											

(a): n (%)

A.3.1. Dose-response models

Characteristic	IRIS-EPA		JECFA		OpenFoodTox	
	Continuous, N = 48[a]	Quantal, N = 60[a]	Continuous, N = 5[a]	Quantal, N = 23[a]	Continuous, N = 119[a]	Quantal, N = 338[a]
Factor_Selected BMR value						
1 SD	34 (71%)	0 (0%)	2 (40%)	0 (0%)	1 (0.8%)	0 (0%)
10%	7 (15%)	56 (93%)	1 (20%)	20 (87%)	16 (13%)	322 (95%)
200%	1 (2.1%)	0 (0%)				
5%	6 (13%)	3 (5.0%)	2 (40%)	0 (0%)	81 (68%)	3 (0.9%)
50%	0 (0%)	1 (1.7%)				
0.5%			0 (0%)	3 (13%)		
1%					20 (17%)	13 (3.8%)
3%					1 (0.8%)	0 (0%)
Factor_Definition for BMD estimation						
Extra risk	4 (8.3%)	56 (93%)	1 (20%)	22 (96%)	42 (35%)	336 (99%)
Relative definition	10 (21%)	1 (1.7%)	2 (40%)	1 (4.3%)	73 (61%)	1 (0.3%)
Standard deviation	34 (71%)	0 (0%)	2 (40%)	0 (0%)	1 (0.8%)	0 (0%)
Summed risk	0 (0%)	3 (5.0%)				
Unknown[b]					4 (3 %)	1 (0.3%)
Factor_Dose response model						
Dichotomous Hill	0 (0%)	1 (1.7%)				
Exponential	1 (2.1%)	0 (0%)			5 (5.0%)	23 (7.4%)
Gamma	0 (0%)	4 (6.7%)			0 (0%)	5 (1.6%)
Hill	9 (19%)	0 (0%)			14 (14%)	2 (0.6%)
Log-logistic	0 (0%)	12 (20%)	0 (0%)	7 (33%)	4 (4.0%)	36 (12%)
Log-Probit	0 (0%)	4 (6.7%)			4 (4.0%)	3 (1.0%)
Logistic	0 (0%)	1 (1.7%)			4 (4.0%)	22 (7.1%)

Multistage	1 (2.1%)	19 (32%)	0 (0%)	4 (19%)	0 (0%)	103 (33%)
Other	19 (40%)	12 (20%)	4 (100%)	8 (38%)	65 (64%)	113 (36%)
Polynomial	12 (25%)	4 (6.7%)			2 (2.0%)	0 (0%)
Power	6 (13%)	0 (0%)				
Weibull	0 (0%)	3 (5.0%)	0 (0%)	1 (4.8%)	0 (0%)	3 (1.0%)
Probit			0 (0%)	1 (4.8%)		
Unknown			1	2	18	28
Log-Normal					3 (3.0%)	0 (0%)
Factor_Dose response model_comment						
Linear	18 (95%)	2 (17%)			4 (6.2%)	0 (0%)
Linear; polynomial; power; hill	1 (5.3%)	0 (0%)				
Model average	0 (0%)	2 (17%)			1 (1.5%)	87 (77%)
Multistage-weibull	0 (0%)	7 (58%)				
Weibull and linear	0 (0%)	1 (8.3%)				
Unknown	29	48	1	15	54	225
Exponential and hill			2 (50%)	1 (13%)	49 (75%)	3 (2.7%)
Gamma multi-hit; multistage; quantal-linear			0 (0%)	1 (13%)		
Gamma, logistic, log-logistic, multistage, probit, log-probit, weibull, quantal linear			0 (0%)	1 (13%)		
Hill and bilinear			2 (50%)	0 (0%)		
Model average			0 (0%)	1 (13%)		
Multistage			0 (0%)	1 (13%)	0 (0%)	1 (0.9%)
Probit and quantal-linear			0 (0%)	3 (38%)		
Additive					4 (6.2%)	4 (3.5%)
Gamma-multihit, log-logistic, multistage, probit, quantal linear, quantal quadratic, weibull					0 (0%)	1 (0.9%)
Linear extrapolation					0 (0%)	4 (3.5%)

Log-probit and multistage				0 (0%)	4 (3.5%)
Logarithmic				2 (3.1%)	0 (0%)
Multistage-cancer				0 (0%)	1 (0.9%)
Null, gamma, logistic, log-logistic, probit, log-probit, multi-stage, weibull, full				0 (0%)	1 (0.9%)
One-stage				0 (0%)	2 (1.8%)
Piecewise linear				4 (6.2%)	0 (0%)
Quantal linear				0 (0%)	1 (0.9%)
Quantal-linear				0 (0%)	1 (0.9%)
Quantal-quadratic				0 (0%)	3 (2.7%)
Square root				1 (1.5%)	0 (0%)

(a): n (%)
(b): The assessments were based on relative risk definition.

A.4.1. Endpoints

Characteristic	IRIS-EPA, N = 108[a]	JECFA, N = 28[a]	OpenFoodTox, N = 457[a]
Factor_Endpoint			
Change	0 (0%)	3 (11%)	42 (9.2%)
Decrease	25 (23%)	1 (3.6%)	56 (12%)
Incidence	71 (66%)	23 (82%)	148 (32%)
Increase	12 (11%)	0 (0%)	38 (8.3%)
Induction	0 (0%)	1 (3.6%)	156 (34%)
Prevalence	0 (0%)	0 (0%)	17 (3.7%)
Factor_Endpoint2			
Anorectic effects	0 (0%)	0 (0%)	22 (4.8%)
Body weight	8 (7.4%)	1 (3.6%)	16 (3.5%)
Erythrocytes, leucocytes and platelet	0 (0%)	0 (0%)	27 (5.9%)
Hepatocellular carcinoma	6 (5.6%)	3 (11%)	25 (5.5%)
Kidney weight	0 (0%)	1 (3.6%)	3 (0.7%)
Liver haemangiosarcoma	2 (1.9%)	1 (3.6%)	134 (29%)
Liver weight	4 (3.7%)	0 (0%)	17 (3.7%)
Other	88 (81%)	22 (79%)	181 (40%)
Tail muscular atrophy	0 (0%)	0 (0%)	32 (7.0%)
Factor_Endpoint2_comment			
Bladder cancer	0 (0%)	1 (4.5%)	4 (2.2%)
Bronchial adenocarcinoma	0 (0%)	0 (0%)	4 (2.2%)
Cancer	0 (0%)	0 (0%)	2 (1.1%)
Cardiac effect	0 (0%)	1 (4.5%)	5 (2.8%)
Cholinergic effect	0 (0%)	1 (4.5%)	0 (0%)
Contact dermatitis	0 (0%)	0 (0%)	3 (1.7%)
Gastrointestinal carcinoma	1 (1.1%)	1 (4.5%)	17 (9.4%)
Gastrointestinal effect	1 (1.1%)	1 (4.5%)	8 (4.4%)
Goblet cells	0 (0%)	0 (0%)	4 (2.2%)
Hardarian gland tumors	0 (0%)	1 (4.5%)	1 (0.6%)
Hematologic effect	3 (3.4%)	0 (0%)	9 (5.0%)
Hepatic effect	12 (14%)	2 (9.1%)	7 (3.9%)
Hepatocellular and bronchoalveolar carcinomas/adenomas	1 (1.1%)	0 (0%)	0 (0%)
Immune effect	6 (6.8%)	0 (0%)	10 (5.5%)
Irritation	2 (2.3%)	0 (0%)	0 (0%)
Kidney cancer	0 (0%)	0 (0%)	1 (0.6%)
Kidney effect	6 (6.8%)	2 (9.1%)	27 (15%)
Liver and kidney weight	0 (0%)	0 (0%)	1 (0.6%)

Liver and lung adenomas/carcinomas	1 (1.1%)	0 (0%)	0 (0%)
Lung cancer	0 (0%)	1 (4.5%)	4 (2.2%)
Mammary tumor	0 (0%)	1 (4.5%)	0 (0%)
Mesothelioma	1 (1.1%)	0 (0%)	0 (0%)
Multiple organs adenomas/carcinomas	5 (5.7%)	2 (9.1%)	0 (0%)
Nasal tumors	1 (1.1%)	0 (0%)	0 (0%)
Nervous effect	5 (5.7%)	1 (4.5%)	3 (1.7%)
Neurobehavioral effect	3 (3.4%)	0 (0%)	3 (1.7%)
Neurodevelopmental effect	4 (4.5%)	2 (9.1%)	9 (5.0%)
Ocular effect	0 (0%)	0 (0%)	1 (0.6%)
Olfactory atrophy	3 (3.4%)	0 (0%)	0 (0%)
Oral cavity cancer	0 (0%)	0 (0%)	1 (0.6%)
Osteosarcoma	0 (0%)	0 (0%)	2 (1.1%)
Pain sensitivity	12 (14%)	0 (0%)	0 (0%)
Prostatitis	1 (1.1%)	0 (0%)	0 (0%)
Reproductive effect	4 (4.5%)	0 (0%)	10 (5.5%)
Reproductive/endocrine tumors	2 (2.3%)	0 (0%)	0 (0%)
Respiratory effect	11 (13%)	0 (0%)	0 (0%)
Respiratory/immune effect	1 (1.1%)	0 (0%)	0 (0%)
Respiratory/nervous efffect	1 (1.1%)	0 (0%)	0 (0%)
Serum bile acids	0 (0%)	0 (0%)	4 (2.2%)
Serum cholesterol	0 (0%)	0 (0%)	5 (2.8%)
Skin lesions	0 (0%)	1 (4.5%)	8 (4.4%)
Splenic cancer	0 (0%)	0 (0%)	3 (1.7%)
Tail muscular atrophy	0 (0%)	1 (4.5%)	0 (0%)
Thyroid adenoma	0 (0%)	1 (4.5%)	0 (0%)
Thyroid effect	1 (1.1%)	2 (9.1%)	3 (1.7%)
Total proteins	0 (0%)	0 (0%)	22 (12%)
Unknown	20	6	276

(a): n (%)

A.5.1. Toxicity and target organs

Characteristic	Overall, N = 1,296[a]	IRIS-EPA, N = 159[a]	JECFA, N = 146[a]	OpenFoodTox, N = 991[a]
Target_organ_level				
Liver	723 (50%)	27 (16%)	17 (11%)	679 (60%)
Other	106 (7.4%)	16 (9.7%)	12 (8.1%)	78 (6.9%)
Nervous	84 (5.8%)	39 (24%)	21 (14%)	24 (2.1%)
Gastrointestinal	82 (5.7%)	9 (5.5%)	7 (4.7%)	66 (5.9%)
Kidney	70 (4.9%)	7 (4.2%)	12 (8.1%)	51 (4.5%)
Skin	58 (4.0%)	0 (0%)	16 (11%)	42 (3.7%)
Lung	56 (3.9%)	23 (14%)	16 (11%)	17 (1.5%)
Musclo-skeletal	42 (2.9%)	2 (1.2%)	0 (0%)	40 (3.6%)
Blood	41 (2.9%)	4 (2.4%)	0 (0%)	37 (3.3%)
Immune	39 (2.7%)	12 (7.3%)	0 (0%)	27 (2.4%)
Reproductive	37 (2.6%)	10 (6.1%)	6 (4.1%)	21 (1.9%)
Bladder	28 (1.9%)	0 (0%)	16 (11%)	12 (1.1%)
Heart	27 (1.9%)	0 (0%)	14 (9.5%)	13 (1.2%)
Developmental	19 (1.3%)	12 (7.3%)	0 (0%)	7 (0.6%)
Thyroid	17 (1.2%)	4 (2.4%)	10 (6.8%)	3 (0.3%)
Spleen	5 (0.3%)	0 (0%)	0 (0%)	5 (0.4%)
Esophagus	2 (0.1%)	0 (0%)	0 (0%)	2 (0.2%)
Ocular	2 (0.1%)	0 (0%)	1 (0.7%)	1 (<0.1%)
Toxicity_level				
Carcinogenicity	317 (22%)	28 (17%)	29 (20%)	260 (24%)
Genotoxicity	234 (17%)	2 (1.2%)	21 (14%)	211 (19%)
Mutagenicity	221 (16%)	2 (1.2%)	10 (6.8%)	209 (19%)
Hepatotoxicity	215 (15%)	20 (12%)	9 (6.1%)	186 (17%)
Systemic toxicity	111 (7.9%)	1 (0.6%)	1 (0.7%)	109 (9.9%)
Neurotoxicity	68 (4.8%)	39 (24%)	16 (11%)	13 (1.2%)
Developmental and Reproductive	64 (4.5%)	22 (13%)	16 (11%)	26 (2.4%)
Nephrotoxicity	61 (4.3%)	7 (4.3%)	17 (12%)	37 (3.4%)
Immunotoxicity	39 (2.8%)	12 (7.3%)	14 (9.5%)	13 (1.2%)
Respiratory Toxicity	34 (2.4%)	23 (14%)	1 (0.7%)	10 (0.9%)
Irritation and Corrosion	18 (1.3%)	6 (3.7%)	1 (0.7%)	11 (1.0%)
Endocrine Disruption	13 (0.9%)	1 (0.6%)	4 (2.7%)	8 (0.7%)
Cardiotoxicity	10 (0.7%)	0 (0%)	5 (3.4%)	5 (0.5%)
Acute toxicity	4 (0.3%)	0 (0%)	3 (2.0%)	1 (<0.1%)
Clinical Toxicity	1 (<0.1%)	1 (0.6%)	0 (0%)	0 (0%)
Ecotoxicity	1 (<0.1%)	0 (0%)	0 (0%)	1 (<0.1%)

Unknown[b]	27	1	1	25

(a): n (%)
(b): Hemopoietic toxicity is reported in 25 assessments, ocular in 2 assessments

Annex A – DB matching

Annex B – Original OpenFoodTox

Annex C – Original IRIS-EPA

Annex D – R markdown file to import the DistilleR extractions

Annex E – Inventory cleaned in R