ENKORE

Endocrine disrupting chemicals and knowledge on health-related effects

HORIZON-HLTH-2023-ENVHLTH-02-03

ENKORE Cluster Deliverable

CLUSTER COMMON DATA MANAGEMENT PLAN

Lead project	EDC-MASLD
Lead beneficiary	UFZ
Author(s)	Matej Orešič (EDC-MASLD) Tuulia Hyötyläinen (EDC-MASLD) Jorke Kamstra (EDC-MASLD) You Song (EDC-MASLD, MERLON) Sergio Gomez Olarte (ENDOMIX) Yvonne Kohl (ENDOMIX) Özlem Altay(HYPIEND) Chiara Baudracco (HYPIEND) Laureano Carpio (HYPIEND) Agata Llobet (HYPIEND) Rita Ortega Vallbona (HYPIEND) Laia Tolosa (HYPIEND) Vikas Kumar (MERLON) Antreas Afantitis (NEMESIS) Tassos Papadiamantis (NEMESIS) Haralampos Tzoupis (NEMESIS)
Dissemination level	PU
Type	DMP
Delivery date	15 January 2025

Funded by the European Union

Document History

Version	Date	Description
V 0.1	November 15, 2024	First Draft – Matej Orešič (ORU/EDC-MASLD)
V 0.2	November 16, 2024	Edits following WG2 meeting
V 0.3	November 27, 2024	Consolidated version after edits and input from all ENKORE cluster projects
V 0.4	November 29, 2024	Next consolidated version after edits and input to V0.3 from all ENKORE cluster projects
V 0.5	December 10, 2024	Final draft version for the portal after edits and input to V0.4 from all ENKORE cluster projects
V 0.6	December 17, 2024	Revised version following comments from RTD and edits from NEMESIS project.
V 0.7	January 12, 2025	Revised version following additional round of comments from RTD
V 1.0	January 15, 2025	Finalised deliverable for submission, approved by RTD

Table of contents

1. Introduction

ENKORE (https://enkore-cluster.eu/) is a cluster of five research projects from the call **HORIZON-HLTH-2023-ENVHLTH-02-03 "Health impacts of endocrine-disrupting chemicals: bridging science-policy gaps by addressing persistent scientific uncertainties"**. The cluster aims to optimise synergies, strengthen collaboration, avoid overlaps and increase the impact of the individual projects.

This deliverable is an outcome of the ENKORE Data Analysis Working Group (WG2), which focuses on developing and applying advanced data analysis techniques, ensuring high standards in data management, and protecting sensitive information within ENKORE and across the five participating projects. Specific objectives of the WG2:

- Exchanging best practice approaches to data analysis techniques to extract meaningful insights on EDC-related health effects;
- Discussing and evaluating data management solutions to maintain high data quality, accuracy, and reliability across all projects;
- Providing instructions for data interoperability and facilitating exchange;
- Developing a common data management plan (DMP) for the cluster activities (this deliverable).

EDC-MASLD is coordinating WG2 activities, with Prof. Matej Orešič as the lead. All ENKORE projects have representatives in the WG2 and are actively engaged in the WG activities (WG meetings, email exchanges). WG2 has been holding monthly meetings *via* Zoom. Additional meetings may be set up if circumstances require, as was the case in preparation of this Deliverable. All WG2 meeting minutes, member list, and other relevant material are available in ENKORE cluster Sharepoint WG2 folder.

The main initial focus topic of WG2 was the development of cluster DMP. Following this, other specific activities will be planned (2025 and beyond), which may involve formation of specific subgroups such as for AOP development, interactions and coordination with IPCHEM, data analysis strategies, etc. Existing documents such as DMP will be revisited for potential updates at every WG2 meeting.

Each of the cluster projects has already developed its own DMP and submitted as project deliverables. The aim of the cluster DMP is not to create a synthesis of the five existing DMPs from individual projects, but to identify commonalities and discrepancies for potential harmonization, thus facilitating open science and collaboration between cluster projects, within the research community and with stakeholders. The cluster projects will try their best to use common data and annotations according to specific community standards as well as of common public repositories for various types of data and knowledge.

This cluster DMP is considered a living document that will evolve as the cluster and WG2 work progress.

2. Brief summaries of DMPs from ENKORE projects

This section summarises the types of data being generated in each of the ENKORE projects, data formats and FAIR data.

EDC-MASLD

EDC-MASLD investigates the impact of environmental exposure to endocrine-disrupting chemicals (EDCs) on the internal exposome (metabolome, gut microbiome, epigenome, proteome, immunome) and the degree of liver damage in MASLD – the condition of excessive accumulation of liver fat unrelated to alcohol intake, ranging from simple steatosis to metabolic dysfunction-associated steatohepatitis. EDC-MASLD is particularly focused on interactions between EDC exposure, sex, genotype, diet, socioeconomic and lifestyle factors, via the data and bio-samples available in the unique European MASLD Registry, comprising over 9,000 patients with histologically characterised MASLD.

Project web site: https://edc-masld.eu

EDC-MASLD will generate the following data types and formats:

- **Human data** (chemical exposome including EDCs, metabolomics, lipidomics) from the European MASLD Registry. In a subset, also steroidomics, proteomics (inflammatory markers), RNA-Seq and miRNA (from extracellular vesicles), stool shotgun metagenomics. Validation data from other prospective cohorts, with surrogate MASLD measures in the previous/ongoing EU projects, such as EXPANSE, EPIC and INITIALISE.
- **Survey data:** A web-based survey will be developed and implemented among representative population samples in participating EU countries and the UK with the help of a survey company (*e.g.*, Ipsos) to measure public awareness, perceptions, attitudes and sense of urgency to address the public health and environmental risks associated with EDCs, including MASLD.
- ***In vitro* data** from (1) screening of early key events in the EDC-MASLD (adverse outcome pathways/AOPs) with HepaRG and ZF-L 3D models, and (2) studies of MASLD progression in 3D cultures, liver organoids, and female and male primary hepatocytes. The generated data will include multi-omics and adverse phenotypic outcomes.
- ***In vivo* data** will be collected from experiments with transgenic **zebrafish** (including imaging, histology, and molecular endpoints) as part of toxicity screening as well as studies of MASLD progression and transgenerational effects. Metabolomics, lipidomics, RNA-Seq and ATAC-Seq will be performed to decipher metabolic status of the zebrafish. In the **murine model of MASLD** (B6TAC fed GAN diet), chronic exposure studies will be performed for selected EDCs and their mixtures. Data will include multi-omics and comprehensive mouse phenotyping.

The overall purpose of this data collection is to assemble elements required for the evaluation of EDCs and their mixtures in the context of MASLD progression, a primary objective of the project. The data will contribute to the identification of metabolic pathways and related mechanisms underlying the impacts of EDCs and their mixtures, which will inform the construction of systems biological models and AOPs.

Data formats for omics data will follow community standards, which are also used in public repositories, such as Gene Expression Omnibus (GEO) for transcriptomics data, PRIDE for proteomics and Metabolomics Workbench for metabolomics data. Metabolic sub-networks modulated by EDCs will be made available in BioModels EMBL-EBI database. Open software policy (GNU-GPL like) will be used according to the licencing of the libraries used and will be deposit on GIT repositories. GitHub directory of the EDC-MASLD project will be created, which will also link to GitHub pages of participating research groups. The AOP Knowledgebase (AOP-KB), including the AOP repository AOPwiki, will be a central database for AOP-related knowledge mining and EDC-MASLD AOP submission.

Summarised **biomonitoring data** generated such as from the European MASLD Registry will be provided for the IPCHEM database in accordance with HBM4EU requirements and to PARC.

ENDOMIX

The ENDOMIX project addresses the urgent need to understand the true impact of EDCs on human health to inform regulators and advise citizens. ENDOMIX tackles this challenge by investigating associations and causality between EDCs and adverse health outcomes, focusing on exposure to multiple EDCs during life course including windows of susceptibility and making use of already existing robust data from multiple European cohorts. The knowledge generated will be disseminated to the scientific community, provide a

Funded by the European Union

thorough new evidence base for policy making and will reach citizens to raise awareness about the risks of EDC exposure.

Project web site: https://endomix.eu/

ENDOMIX will generate the following data types and formats:

- **Human data:** ENDOMIX will re-use demographic, risk factor covariates, health outcomes, and epigenetic data from European cohorts to assess associations between prenatal and early-life exposure to EDC and child health through inflammatory and metabolic changes. Cohort data is stored in Statistica software format (.sta) but can be exported in simplified formats (.csv). Omics and epigenetic data are also available for specific windows of susceptibility (.RData, .csv, .dat, .txt, .idat). Additionally, we will generate new proteomic data using the Olink platform (.csv, .xlsx) and LC-MS spectra of EDC chemical standards for chemical annotation in existing metabolomic data (.mzML, and .tsv). Microbiome data comprise raw sequencing files from 16S rRNA amplicon sequencing analysis (.fastq).
- ***In vitro* data:** In the context of EDC mixture exposure, immune cells alone and in coculture with barrier and organ systems will be phenotyped. The output includes data derived from flow and mass cytometry (.fcs), immunofluorescence (.jpg, tif), cytometry bead arrays (.vg, fcs), and gene expression analysis (.csv, .txt), phase-contrast images (.tif/.png), and concentration-response curves (.pdf). Models of physiological barriers will generate numeric data, such as impedance values and fluorescence, luminescence, and optical density units in ELISA assay (.xlsx).
- ***In vivo* data:** ENDOMIX contemplates evaluating how exposure to EDC mixtures can impact neurodevelopment in transgenic zebrafish, the life cycle of C. elegans, and the onset of immune-mediated diseases (e.g., rheumatoid arthritis and allergies) in transgenerational murine models. Therefore, organs and tissues targeted by EDCs will be prepared for ex vivo phenotyping by histopathology (.tif), immunoassays (.csv, .facs, .xlsx), and omics approaches (.csv, .fastq).

HYPIEND

HYPIEND is one of the first projects to study EDCs effects on the hypothalamus-pituitary axis using a multidisciplinary approach, including preclinical models and two European-wide clinical studies. The findings will be used to delineate interventional strategies for minimizing EDC exposure and consequences on the neuroendocrine system in pregnant and breastfeeding women, infants and pre-pubertal children.

Project web site: https://hypiend.eu/

HYPIEND will generate the following data types and formats:

- ***In vitro* data** collected from cell-based assays and organoid models derived from pluripotent stem cells to assess EDC effects.
- ***In vivo* data** from experiments using zebrafish models for developmental and functional analyses and gene expression studies. Additionally, animal studies will be used to examine placental and blood-brain barrier models, along with the multi-generational effects of EDCs.
- **Survey data** gathered from human interventional studies for two vulnerable groups: pregnant women, including their infants, and prepubertal children. This data will be obtained through questionnaires that will assess various factors, including anthropometric measurements, physical activity, diet, quality of life, socio-demographic data, and medical history (both before and after birth). Additionally, questionnaires will evaluate participants' knowledge of EDCs, their perceived exposure, the Health Action Process Approach (HAPA) constructs, behavioural and psychological patterns, neurodevelopment, and lifestyle habits related to EDC exposure.

Funded by the European Union

- **Human data** obtained from these interventional studies, focusing on blood and urine EDC analyses and gut microbiome assessments.
  - For the perinatal study (mothers and infants): **Urine** (metabolomics and iodine concentration). **Plasma** (metabolomics and immunoassays). **Peripheral blood** (Gene expression (Q-PCR) and epigenetic). **Breast milk** (metabolomics). **Faeces** (shotgun sequencing).
  - For the prepubertal study (children): **Urine** (metabolomics and iodine concentration). **Faeces** (shotgun sequencing). **Blood (DBS)** (DNA methylation analysis and immunoassays).
- **Computational data** generated using computational models and machine-learning methods.

The primary goal of this data collection is to enhance our understanding of the impact of EDCs on the hypothalamus-pituitary (HP) axis. These data will create a comprehensive knowledge base linking EDC exposure and adverse health outcomes.

Data formats for omics data will adhere to community standards formats, leveraging public repositories to ensure accessibility and reproducibility. Specifically, transcriptomics, metagenomics, and epigenomics/ethics data will be deposited in the Gene Expression Omnibus (GEO), while databases, metadata, transcriptomics, and proteomics will utilize Zenodo. Genomic data will be stored in BioSamples, and metagenomics data will be submitted to the European Nucleotide Archive (ENA). AOP-wiki will serve as a central resource to contextualize observed Adverse Outcome Pathways.

MERLON

The MERLON project aims to improve our knowledge on how EDC exposures at critical life stages affect sex development and reproductive health, with a key goal to improve the current test battery for EDC identification and to promote the transition from conventional animal testing to non-animal based new approach methodologies (NAMs). It brings together world-leading experts in endocrinology, chemical safety assessment, developmental and molecular biology, epidemiology, toxicogenomics, toxicokinetics-toxicodynamics (TKTD) modelling, regulatory toxicology, and psychology to investigate EDC-mediated effects on sexual development, providing human data on the role of EDC exposure during fetal development and changes in mini-puberty, connecting to puberty, reproductive function, and gender incongruence.

Project web site: https://merlon.dtu.dk/

In MERLON, new data related to ED relevant mechanisms- and modes-of-action, as well as biomarkers for adverse outcomes will be generated through in vivo, in vitro, and in silico studies. This will include transcriptomics, proteomics, metabolomics, and advanced multi-omics integration. The data aims to bridge scientific research with regulatory frameworks for EDC evaluation.

Human data collection

MERLON leverages human cohort data reuse, including:

- *Cohort Studies*: Data from the gender incongruence cohorts as well as data on chemical exposure, hormone profiles, and health outcomes from the COPANA study.
- COPANA NeuroTop: All participants in COPANA will be invited to take part in a follow up study with a focus on the neurobehavioral development of the children.
- *Experimental Data*: Derived from ex vivo tissue cultures for mechanistic insights.

In vitro data

Experiments will employ human tissue models, primary cell cultures, and innovative systems like liver organoids and 3D cultures to study EDC-induced effects and disease progression. MERLON will produce a high volume and several types of transcriptomics data, both from single-cell and bulk-RNA sequencing of tissues and cells collected from other experiments within the consortium. For model organisms, count matrices (.mtx) and fastq files will be submitted to public repositories (*e.g.*, GEO, ArrayExpress). For humans, count matrices (.mtx) will be submitted to public repositories (*e.g.*, GEO, ArrayExpress) and fastq files will be submitted to the EGA repository Transcriptomics and multi-omics data will underpin pathway analyses.

In vivo data

Animal models will include:

- *Rodents*: For assessing endocrine and reproductive disruptions.
- *Zebrafish*: To study multi-generational effects and toxicity through advanced molecular profiling.

Data formats and standards

Generated data will align with FAIR principles, ensuring it is findable, accessible, interoperable, and reusable. Standardized formats such as CSV, fastq, and mtx will facilitate integration and analysis. Public repositories like GEO, TOXsIgN, and AOPwiki will host data for broader access post-publication.

NEMESIS

NEMESIS addresses the adverse metabolic effects of EDCs through a multidisciplinary approach and responds to unmet regulatory needs regarding EDCs. NEMESIS aims to elucidate the mechanisms and dose-dependency of metabolic disruption by EDCs and their mixtures in liver and pancreas as well as their effects on gut microbiota through in silico, in vitro, in vivo, epidemiological and systems biology approaches. The consortium will support risk assessment by improving regulatory testing guidelines by incorporating metabolic endpoints and developing AOPs and IATAs. Engagement with citizens and stakeholders ensures effective risk communication and maximizes the impact of NEMESIS on policy development.

Project web site: https://www.nemesis-project.eu/

NEMESIS will generate the following data types and formats:

- **Human data** (chemical exposome including EDCs, EDC predictive biomarkers, metabolomics, transcriptomics, toxicogenomics) from four mother-child cohorts (ENVIRONAGE, INMA, KuBiCo, and NorthPop) and two adult cohorts (NSHDS and COT).
- **Survey data:** A survey will be developed in English by UNL and translated into different EU languages. The aim of the survey is to collect baseline data on knowledge, literacy, risk perceptions on EDCs exposure, and the acceptability of policy measures. Focus groups and interviews with populations at increased EDC exposure risk in NEMESIS partner countries will be employed to complement these data.
- ***In vitro* data** from (1) hepatoxicity, liver development, hepatic metabolism, nuclear receptor interactome, and mitochondrial function studies in 3D liver cultures using various cells, e.g., continuous cell lines, primary human hepatocytes (PHH), non-parenchymal cells, hiPSCs, cholangiocytes, endothelial cells, hepatic stellate cells, and macrophages; (2) dose-dependent effects on primary human pancreatic islet and mature human pancreatic β cells; (3) EDC effects on gut microbiota.

Funded by the European Union

- ***In vivo* data** (including imaging, histology, and molecular endpoints) regarding development, growth, inflammation, and metabolic functions will be acquired from zebrafish. Selected EDC metabolic disturbances will also be studied using C57BL/6 mice newborn pups fed with Western-type high fat diet to study potential effects leading to NAFLD. RNA-sequencing will be used to characterise transcriptome of selected tissues. Plasma and liver tissue will be analysed with metabolomics.
- **Computational data** will be developed from the combination of existing and newly generated in silico, toxicogenomics, and biological annotation data. These will be used to develop a computational framework utilising AOP modelling, network analysis, and machine learning to generate mechanistic insights into the metabolic disruption caused by EDC exposure.

The overall purpose of this data collection will be to assemble elements required for the evaluation of EDCs and their mixtures. These will be used to improve assessment of metabolic endpoints in testing guidelines and adopt alternative models to animal testing. Adverse Outcome Pathways (AOP)s and Integrated Approaches to Testing and Assessment (IATA).

Data formats for omics data will follow community standards, which are also used in public repositories, such as Gene Expression Omnibus for transcriptomics data and Metabolomics Workbench for metabolomics data. NEMESIS data will generally be made available through a dedicated database based on the NovaMechanics's Pharos database solution to accommodate dose-response data and omics profiles of ED-related exposures. The database is fully connected to chemical-related databases, e.g., PubChem, UNICHEM, Chemspider, and CIR through APIs. If needed, other specialised databases will be examined. Computational tools and models will be made available through the NEMESIS Cloud Platform leveraging the NovaMechanics's Enalos Cloud Platform. AOPwiki will be a central resource interrogated to put the observed Adverse Outcome Pathways into context. Data and models along with respective technical, scientific, bibliographical, and provenance metadata will be FAIRified based on the FAIR data principles and their interpretations from the GoFAIR Foundation.

3. Types of data being generated within ENKORE

Human data – omics

- Plasma/serum chemical exposome including EDCs (EDC-MASLD, ENDOMIX, HYPIEND, MERLON)
- Plasma/serum metabolomics (EDC-MASLD, ENDOMIX, HYPIEND, NEMESIS)
- Plasma/serum lipidomics (EDC-MASLD, HYPIEND)
- Plasma/serum steroidomics (EDC-MASLD, HYPIEND)
- Plasma/serum proteomics (EDC-MASLD, ENDOMIX)
- Plasma/serum thyroid hormones, kisspeptin, cytokines (HYPIEND)
- Blood epigenetics (HYPIEND)
- *Ex vivo* tissue culture (MERLON)
- Urine EDCs, iodine (HYPIEND)
- Breast milk EDCs (HYPIEND)
- Toxicogenomics (NEMESIS)
- Plasma 4β-hydroxycholesterol-based integrative exposure indices (NEMESIS)
- Tissue RNA-Seq/transcriptomics (EDC-MASLD, NEMESIS)
- Tissue miRNA (EDC-MASLD, HYPIEND)
- Tissue metabolomics (NEMESIS)
- Stool 16srRNA or shotgun metagenomics (EDC-MASLD, ENDOMIX, HYPIEND, NEMESIS)

Human data – health outcomes and clinical data (main focus)

- Physical measures (weight, height, etc) and demographic characteristics (EDC-MASLD, HYPIEND)
- MASLD histology, liver enzymes, HOMA-IR (EDC-MASLD)
- Birth outcomes (birth weight, gestational age, …) (HYPIEND)
- Neurodevelopmental data (HYPIEND, MERLON)
- Pubertal development data (HYPIEND)

Survey data

- Web-based survey, with structured questionnaire (EDC-MASLD, ENDOMIX, HYPIEND, NEMESIS)
- Focus groups and interviews (NEMESIS)

In vitro model data

- 2D Cell lines HepaRG and ZF liver cells – hepatocytes → multi-omics, toxicity outcomes (EDC-MASLD, HYPIEND, NEMESIS)
- Human primary trophoblast cells → multi-omics, toxicity outcomes, metabolic parameters (NEMESIS)
- Cell lines – hepatocytes, trophoblast, endothelial cells, intestinal cells → multi-omics, toxicity outcomes, immunophenotyping (ENDOMIX)
- Cell lines – reporter cell lines, neuronal and hepatic models → toxicity outcomes (HYPIEND)
- 3D hepatocyte co-culture models → multi-omics, toxicity outcomes (EDC-MASLD, NEMESIS)
- Liver organoids and primary hepatocytes; placental spheroids and organoids, lung spheroids, intestine spheroids and organoids → multi-omics, toxicity outcomes, gene expression, ELISA (ENDOMIX)
- Hypothalamus organoids → toxicity outcomes (HYPIEND)
- Placental explants, spheroids and organoids → multi-omics, toxicity outcomes, gene expression, metabolic outcomes (NEMESIS)
- Barrier models of placenta, lung, gut, and brain → gene expression, toxicity outcomes, immunophenotyping (ENDOMIX)
- Human primary immune cells for toxicology (EDC mixtures) assays → gene expression, immunophenotyping (ENDOMIX)
- Human tissue models (MERLON)
- Primary cell cultures (MERLON)
- Liver organoids and 3D cultures for disease progression (MERLON)
- *In vitro* experiments (human and rat)

In vivo data

- Zebrafish including transgenic zebrafish
  - multi-omics, imaging, toxicity screening (EDC-MASLD, HYPIEND, NEMESIS)
  - studies of MASLD progression and transgenerational effects (EDC-MASLD)
  - studies of the impact of immunotoxicants on the innate immune system (ENDOMIX)
- *Caenorhabditis elegans*
  - testing the immunotoxic effect of EDCs on the whole life cycle of the organism (ENDOMIX)
- murine models
  - murine model of MASLD (B6TAC fed GAN diet) for chronic exposure studies, comprehensive phenotyping, multi-omics (EDC-MASLD)
  - wild-type strains (C57BL/6 female and BALB/c male) for prenatal exposure studies on reproductive, immunological, and metabolic outcomes (ENDOMIX, HYPIEND) in the offspring.

  - C57BL/6 mice fed with Western type high-fat diet for exposure and multi-omics studies (NEMESIS).
  - Rat and mouse in vivo data *e.g.* histopathology data and images, transcriptomic analysis of rat organs and analysis of changes in neuroendocrine activity, affecting key reproductive neuronal cell type (MERLON).

Computational data

- Genome-scale metabolic models of MASLD (EDC-MASLD)
- Quantitative structure-activity relationship (QSAR) models predictions (HYPIEND)
- Multiple ligand simultaneous docking (MLSD) 3D poses (HYPIEND)
- Physiological-Based Kinetic (PBK) models (MERLON)
- Computational framework utilising AOP modelling, network analysis, and machine learning (EDC-MASLD, ENDOMIX, HYPIEND, MERLON, NEMESIS)

Communication and dissemination data

- Web site analytics (all cluster projects).
- Newsletter subscriptions (all cluster projects).
- Social media statistics (all cluster projects).
- Event statistics and subscriptions (all cluster projects).

4. Data formats within ENKORE

Exposomics, metabolomics, lipidomics will range from raw files produced by machines (*e.g.*, by MS) to structured and standardized formats (*e.g.*, mzML for MS data). Metabolomics as well as chemical exposure data will be annotated according to the Metabolomics Standards Initiative (MSI).

Proteomics data using targeted inflammatory panel will be provided in Excel format. Proteomics data will be annotated by using community standards as set by the HUPO Proteomics Standards Initiative.

Transcriptomics and epigenomics data will range from fastq files (.fq) produced by the Illumina sequencing platform to normalized and filtered files for the gene expression datasets (.xlsx; .csv; R data frames) as well as matching gene set resulting from Gene Set Enrichment Analysis (GSEA) search. For study publication, expression datasets will be published on Gene Expression Omnibus (GEO-ncbi). Functional genomics data will be annotated with MIAME and MINSEQE-compliant metadata.

Metagenomics data will range from raw FASTQ files (.fq) generated by high-throughput sequencing platforms to processed files, such as annotated feature tables (.csv, .biom) and functional profiles.

Genome-scale metabolic models computed within the project will be exported in standardized file format such as Systems Biology Mark-up Language (SBML) and submitted to the BioModels EMBL-EBI database.

Physiological-Based Kinetic (PBK) models developed in the MERLON project will adhere to OECD standards and will be harmonized using PBK ontology in collaboration with the EU PARC project. Open-source code (in SBML and R format) will be shared via GitHub repository of lead partner (IISPV).

In vitro* data** will consist of image analysis (.tif) of hepatocytes, and zebrafish liver tissue, and a variety of read-outs from different equipment and gene expression data. Data handling will be performed in MS Excel (xlsx). Similar to *in vitro* data, ***in vivo read-outs will be based on image analysis (.tif), omics (as listed above), immunophenotyping (.csv, .facs) and Excel data handling.

AOP data will be developed and submitted to the AOPWiki according to the OECD Handbook AOP development and assessment. The AOP data will be publicly accessible through direct downloading in an XML format.

Public data available in online repositories including, but not limited to, Tox21, ToxCast, PubChem Bioassay will be collected to feed machine learning models required to predict Molecular Initiating Events (MIEs), and metabolic network published in 2018 for human metabolic modelling (Recon3D). Human exposure data will be gathered from IPCHEM. CTD (Comparative Toxicogenomics Database) will be used as resource on current knowledge on chemical-gene-disease associations. The mass spectrometry proteomics data will be deposited to the ProteomeXchange Consortium via the PRIDE. Finally, AOPwiki will be a central resource interrogated to put the observed Adverse Outcome Pathways into context.

Summarised **biomonitoring data** will be provided for the IPCHEM database in accordance with HBM4EU requirements and to PARC.

Dissemination and communication data is described in the Deliverable '*Cluster dissemination and communication strategy*'.

5. FAIR data practices

5.1. Making data findable, including provisions for metadata

Data will be made findable by using Metadata stored as much as possible in dedicated and computer-readable formats (*e.g.*, ISA-tab or dedicated XML formats).

Clinical and different types of experimental legacy data will be integrated based on existing common data models such as the OMOP CDM (https://www.ohdsi.org/data-standardization/the-common-data-model/).

Cohort metadata from ENDOMIX is partiality available in the European Child Cohort Network (https://data-catalogue.molgeniscloud.org/catalogue/catalogue/#/networks-catalogue), the EU Child Network Variable Catalogue (https://lifecycle-project.eu/for-scientists/variable-catalogue/), and the Dementia Platform UK (https://www.dementiasplatform.uk/). These catalogues include the list of variables that have been harmonized from previous collaborative projects.

Omics data will be annotated using community standards, such as Metabolomics Standards Initiative (MSI), HUPO Proteomics Standards Initiative, and (for transcriptomics) MIAME and MINSEQE. As the cluster projects and WG2 work progress, we will work towards the adoption of regulatory accepted guidelines such as OECD Omics reporting guidance document.

Datasets and related metadata may undergo curation and modifications. To ensure traceability of these changes a versioning system (*e.g.*, GIT) will be used. DOIs will be associated to datasets and workflows.

5.2. Making data accessible

Data will be made accessible by using public repositories to grant public access to the data

- ***Horizon Results Platform*** for major cluster data deliverables (https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/horizon-results-platform).
- ***Gene Expression Omnibus*** (https://www.ncbi.nlm.nih.gov/geo/) for transcriptomics metagenomics,

Funded by the European Union

and epigenomics data;

- ***PRIDE*** (https://www.ebi.ac.uk/pride/) for proteomics;
- ***Metabolomics Workbench*** (https://www.metabolomicsworkbench.org/) for metabolomics data, with depositions to be linked to GNPS resource (https://gnps.ucsd.edu/) for mining unannotated spectra ('unknown' chemicals), for potential matches;
- ***Zenodo*** (https://zenodo.org/) for databases, metadata, transcriptomics and proteomics;
- ***BioSamples*** (https://www.ebi.ac.uk/biosamples/) for genomic data;
- ***ENA*** *(European Nucleotide Archive) for* metagenomics data;
- ***FlowRepository*** *(http://flowrepository.org/)* for traditional and spectral flow cytometry and mass cytometry data;
- ***IPCHEM - the Information Platform for Chemical Monitoring*** (https://ipchem.jrc.ec.europa.eu/) for summary data on human health impacts of specific EDCs.
- ***Protocols and SOPs*** will be uploaded to and published via online resources as zenodo (https://zenodo.org/) or protocols.io (https://protocols.io).
- ***Metabolic sub-networks*** modulated by EDCs and other models will be made available in BioModels EMBL-EBI database. AOPwiki (https://aopwiki.org/) will be a central resource interrogated to put the observed Adverse Outcome Pathways into context.
- ***Open software policy*** (GNU-GPL like) will be used according to the licencing of the libraries used and will be deposited on GIT repositories.
- Projects will create their own ***GitHub directories***, which will also link to GitHub pages of other ENKORE projects when and where relevant.
- Centralised ***links to deposited data, methods, code, and SOPs*** will be made available in a table deposited in Zenodo.

5.3. Making data interoperable and increase data re-use

To ensure data interoperability within the ENKORE cluster, we will follow key strategies that include the creation of a DMP (this document) and FAIR principles. In addition, we will adopt standardized formats and ontologies by employing widely accepted data formats and domain-specific ontologies (as described in **Section 4**) to facilitate consistent data annotation and understanding between partners and other potential stakeholders.

Long-term preservation of the data will be ensured through deposition in repositories mentioned above for minimum 10 years after the completion of the projects, with data retention policies aligned with EU regulations.

ENKORE will adhere to the principles of transparency, openness, and involvement of all relevant stakeholders in the investigations of health impacts and underlying mechanisms-of-action of EDCs in project-specific settings. All experimental model data generated in the project will thus be made accessible. However, there are other values such as data privacy (and the rules and regulations that make the protection of the data of individuals a legal requirement) that may limit our ability to provide access to all **human** data beyond limited metadata and summary statistics. Therefore, accessing subject/patient-level data within the cluster projects is something that will only be possible after rigorous evaluation by our experts for medical data governance and ethics, and legal aspects; which will be done at the level of individual cluster projects, *e.g.*, by projects' ethics committees. The responsibility for ensuring secure data access, in compliance with GDPR, will thus rest with the data owner project. This includes implementing measures such as data use agreements, anonymization or pseudonymization techniques, and the use of secure data-sharing platforms to safeguard sensitive information.

Funded by the European Union

However, any such restrictions will not limit our ability to share and deposit the relevant chemical biomonitoring data (summary data on human health impacts of specific EDCs) in IPCHEM database through involvement with the European Commission's Joint Research Centre (JRC), as outlined in the call text, as well as as part of collaboration with PARC.

6. Conclusion

The aim of the cluster DMP is to facilitate open science and collaboration between cluster projects, within the research community and with stakeholders, through use of common data and annotations according to specific community standards as well as of common public repositories for various types of data and knowledge.

This is a living document that will be revisited at regular WG2 meetings and updated as the ENKORE cluster and Working Group 2 work progress.

Funded by the European Union