WP7: Information and data governance, ethics, technology, data catalogue and quality support

Task 7.6 & 7.7: ConcePTION Data Characterization Protocol & Algorithms for population-based data sources/collections

Version 1.3

Table of Contents

Document history

	Version	Date	Edits
Caitlin Dodd (UMCU), Rosa Gini (ARS)	**0.1**	**June 14, 2019**	
Miriam Sturkenboom (UMCU)	**0.2**	**July 2, 2019**	**Addition of intro/methods**
Marianne Cunnington (GSK), Caitlin Dodd (UMCU)	**0.3**	**July 9, 2019**	**Incorporation of comments**
Caitlin Dodd (UMCU), Miriam Sturkenboom (UMCU)	**0.4**	**September 2, 2019**	**Incorporation of WP1 comments, removal of WP 2 characterization, reformatting**
Caitlin Dodd (UMCU)	**0.5**	**September 23, 2019**	**Incorporation of WP1 comments**

Rosa Gini (ARS), Miriam Sturkenboom	**1.0**	**October 2, 2019**	**Specified component strategy, final comments**
Caitlin Dodd (UMCU)	**1.1**	**November 5, 2019**	**Incorporate comments from data access providers**
Vjola Hoxhaj	**1.2**	**March 9, 2021**	**Update of CDM tables**
Vjola Hoxhaj	**1.3**	**December 2, 2022**	**Update of CDM tables to CDM version 2.2, update of quality indicators, update of literature**

1. List of abbreviations

ACE	Angiotensin-converting enzyme
ADHD	Attention Deficit and Hyperactivity Disorder
ADVANCE	Accelerated development of vaccine benefit-risk collaboration in Europe
AE	Adverse Event
ARB	Angiotensin II Receptor Blockers
ATC	Anatomical Therapeutic Chemical Classification
BNF	British National Formulary
CDM	Common Data Model

CPRD	Clinical Practice Research Datalink
CR	Central Registry
DAP	Data Access Provider
DDD	Defined Daily Dose
DOA	Description of Action
DP	Demonstration Project
EHR	Electronic Health Record
EMR	Electronic Medical Record
ETL	Extract, Transform, and Load
EUROCAT	https://eu-rd-platform.jrc.ec.europa.eu/eurocat. European network of population-based registries for the epidemiological surveillance of congenital anomalies
EUROlinkCAT	https://www.eurolinkcat.eu. H2020 funded project creating linkage between EUROCAT registries and local data
EUROmediCAT	http://euromedicat.eu. European research consortium started as a four year project (2011-2015). Its aim is to build a European system for the evaluation of safety of medication use in pregnancy in relation to the risk of congenital anomalies
FDA	Food and Drug Administration
GAIA	Global Alignment of Immunization safety Assessment in pregnancy
GP	General practitioners
HDQF	Healthcare Data Quality Framework
HTA	Health Technology Assessment
ICD	International Classification of Diseases
ICPC	International Classification of Primary Care
IRB	Institutional Review Board
LMP	Last Menstrual Period
MS	Multiple Sclerosis
OMOP	Observational Medical Outcomes Partnership
OTC	Over The Counter
PAS	Post-Authorization Study
PC	Primary Care
PERISTAT	Perinatal Health Statistics
REB	Research Ethics Board
RWD/RWE	Real World Data/Real World Evidence
SES	Socioeconomic Status
SLE	Systemic Lupus Erythematosus
TIS	Teratology Information Service(s)
TOPFA	Termination of Pregnancy for Fetal Anomaly
UMLS	Unified Medical Language System
VACCO	Vaccine Ontology
WP	Work Package

2. Authors

Main authors of the protocol
Vjola Hoxhaj, Drs (UMCU)
Caitlin Dodd, PhD (UMCU)
Miriam Sturkenboom, PhD (UMCU)

3. Study team

3.1 Principal Investigator

Vjola Hoxhaj, UMCU for data characterization

Rosa Gini, ARS for Algorithm validation

3.2 Study Team members

Rutger van den Bor, UMCU, The Netherlands
Roel van den Bor, UMCU, The Netherlands
Constanza Andaur Navarro, UMCU, The Netherlands
Miriam Sturkenboom, UMCU, The Netherlands

Kirsten Lum, J&J, Us
Marianne Cunnington, GSK, UK

Rosa Gini, ARS, Italy
Claudia Bartolini, ARS, Italy
Giuseppe Roberto, ARS, Italy
Giulia Hyeraci, ARS, Italy
Ippazio Cosimo Antonazzo, ARS, Italy

Gianluca Trifirò, UniMe, Italy
Valentina Ientile, UniMe, Italy
Janet Sultana, UniMe, Italy

Eline Houben, PHARMO Institute for Drug Outcomes Research, The Netherlands
Karin Swart, PHARMO Institute for Drug Outcomes Research, The Netherlands
Ron Herings, PHARMO Institute for Drug Outcomes Research, The Netherlands

Tania Schink, Liebniz Institute for Prevention Research and Epidemiology (BIPS), Germany
Ulrike Haug, Liebniz Institute for Prevention Research and Epidemiology (BIPS), Germany

Talita Duarte Salles, IDIAPJGol, Spain

Maarit Leinonen, Finnish Institute for Health and Welfare (THL), Finland
Visa Martikainen, Finnish Institute for Health and Welfare (THL), Finland
Mika Gissler, Finnish Institute for Health and Welfare (THL), Finland

Anke Rissman, Malformation Monitoring Centre Saxony-Anhalt, Medical Faculty Otto-von-Guericke University, Magdeburg, Germany

Clara Cavero (FISABIO), Congenital Anomalies population-based registry of the Valencia Region, Rare Diseases Research Unit, Foundation for the Promotion of Health and Biomedical Research of the Valencian Region (FISABIO), Valencia, Spain
Laia Barrachina (FISABIO), Spain

Hedvig Nordeng, University of Oslo (UOSL), Norway
Angela Lupattelli (UOSL), Norway
Anders Huitfelds (UOSL), Norway

Ester Garne Hospital Lillebaelt, Denmark

Katrine Strandberg Larsen, University of Copenhagen, Denmark

Vera Ehrenstein, University of Aarhus, Denmark

Tom MacDonald, University of Dundee, Scotland

Rachel Charlton, University of Bath, UK
Anita McGrogan, University of Bath, UK

Helen Dolk, University of Ulster (ULST), UK & Ireland
Maria Loane, University of Ulster (ULST), UK & Ireland

Hanitra Randrianaivo, Registre des Malformations Congenitales de la Reunion

Nicholas Moore, University of Bordeaux, France
Cecile Droz, University of Bordeaux, France
Nicolas Thurin, University of Bordeaus, France

Jorieke van Kammen-Bergman, University Medical Center Groningen (UMCG), The Netherlands

Eugene van Puijenbroek, Stichting Lareb, The Netherlands
Saskia Vorstenbosch, Stichting Lareb, The Netherlands

Awi Wiesel, Universitatskinderklinik Mainz, Germany

Anna Latos Bielenska, Poznan University of Medical Science, Poland

Ingeborg Barisic, Klinikabolnica Sestre milosrdnice, Klinika za djecje bolesti Zagreb, Croatia
Ljubica Boban, Klinikabolnica Sestre milosrdnice, Klinika za djecje bolesti Zagreb, Croatia

Amanda Neville, University of Ferrara (FERR), (UNIFE), Italy
Annarita Armaroli, University of Ferrara (FERR), Italy
Aurora Puccini, University of Ferrara (FERR), Italy

Anna Pierini, CNR Tuscany (CNR-IFC), Italy
Michele Santoro, CNR Tuscany (CNR-IFC), Italy
Alessio Coi, CNR Tuscany (CNR-IFC), Italy

Miriam Gatt, Malta Congenital Anomalies Registry, Directorate for Health Information and Research, Malta

4. Rationale and background

ConcePTION aims to build an ecosystem that can use Real World Data (RWD) to generate Real World Evidence (RWE) that may be used for clinical and regulatory decision making. RWE is required to address the big information gap of medication safety in pregnancy. Regulators and health care professionals are increasingly appreciating the value of RWE, but hesitancy about quality and reliability persists. Although various networks that have been set up to monitor drug safety do use some type of quality indicators (e.g., Sentinel) there is no standardized framework to assess fitness for purpose of RWD.

There is no generally accepted quantitative measure of data quality, but Juran JM et al, 1999 gives an qualitative definition as "…high-quality data are data that are fit for use in their intended operational, decision-making, planning, and strategic roles"[1]. Very importantly, data quality may be adequate when used for one task, but not for another. Therefore, these quality assessments may be called "fit for purpose".

In order to make best use of RWD for generation of evidence across many data sources in a scalable rapid and reproducible manner, many groups and consortia have turned to the use of common data models (CDMs) (Schneeweiss et al., 2020; Trifiró et al 2014; Gini et al 2016). Common data models vary along two axes: 1) the degree to which content is harmonized and 2) their flexibility for use in the conduct of new studies. Along the first axis, CDMs may be structurally (syntactically) harmonized, meaning that data is transformed into a common structure, but the contents remain unchanged, or semantically harmonized, meaning that data is transformed into a common structure and contents are transformed into common concepts. Along the second axis, common data models may be study specific, designed for a set of studies focused on one therapeutic area or one analysis method, or fully reusable for the application of new study questions and designs.

ConcePTION is designed to be a learning healthcare system (LHS). The Institute of Medicine defines a learning healthcare system as a system in which "science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience." (Grossman et al, 2011). In the ConcePTION LHS, we have agreed upon a study-independent syntactically harmonized common data model and aim to assess the quality and fitness for purpose of data in this CDM in a study-independent way (for quality and completeness) and in study design and research question-specific ways (for fitness for purpose).

As reported by Kahn et al 2016, standards for assessment of the quality of observational data used in networks such as the ConcePTION consortium are lacking. Data quality checks employed by these networks typically include checks of consistency with semantic rules, visualization of temporal trends, and rates of codes, events, or exposures. These checks are typically performed both within and between sites. However, no standard rules or thresholds for defining a data source 'fit for purpose' for a specific study exist (Kahn et al, 2013).

4.1 Use cases

In ConcePTION the following *use cases* are important and thus, fit for purpose assessment should focus on these domains:

1. Assessing medication use and vaccine exposure in the general population by age, sex, and source of data: this requires that age and sex from the population members is known, as well as source of the data records
2. Assessing medication use in childbearing age and in pregnancy: this requires that age and sex of the population members is known, as well as pregnancy status
3. Calculation of incidence rates of events in the general population: follow-up needs to be long enough
4. Calculation of prevalence rates of events during pregnancy and before/after: this requires that the onset and ending of a pregnancy is known, as well as the events that occur before/ during and after pregnancy
5. Assessing severity of specific maternal conditions: this requires that healthcare use, or disease severity markers/measures are available
6. Assessing prenatal and antenatal outcomes in relation to drug exposure for signal generation and signal evaluation: this requires that pregnancy duration is known, follow-up is available, and the relevant outcomes and exposures can be measured as well as confounding factors

4.2 Criteria for quality assessment

We reviewed several large initiatives and guidance documents to review what type of quality measures/requirements, and RWD fit for purpose assessment is needed.

The IMI-GetReal project, in their final report, *Advancing Evidence Generation for New Drugs: IMI GetReal's Recommendations on Real-World Evidence*[1] recommended:

- All stakeholders should collaborate to develop and publish minimum requirements for the integrity and quality of RWD sources used to generate RWE submitted for decision making.
- Regulators, Health Technology Assessment (HTA) bodies, payers, researchers and the pharmaceutical industry should collaborate to characterise RWD sources and understand their strengths and weaknesses.
- Characterise barriers to access of RWD.
- Identify and promote efforts to catalogue RWD sources.

Hereby acknowledging that standards and minimum requirements need to be defined. This was shared by the recommendations from The Heads of Medicines Agency Task Force on Big Data which recommended the following in 2018 [2]:

6.3.2. Data quality

Core recommendation

Characterisation of data quality across multiple data sources is essential to understand the reliability of the derived evidence

(Supported by subgroup recommendations # 12, 13, 14, 29, 33, 35, 37b, 41, 42)

- Characterise and document data quality in a sustainable EU inventory.
- Establish minimum sets of data quality standards. Where possible, quality attributes e.g. compliance to GCP requirements should be integrated to facilitate selection of appropriate data sets for analysis.
- Implement data quality control measures.
- Establish a clear framework for the validation of innovative bioanalytical methods e.g. 'omics.

Ownership of the Action: EMA / HMA

The recent publications by Cave A et al. of the European Medicines Agency in July 2019 recommend that RWE should be[3]:

- Derived from data source of demonstrated good quality
- Valid (internal and external validity)
- Consistent (across countries/data sources)
- Adequate (e.g., precision, adequate range of characteristics of population covered, dose and duration of treatment, length of follow-up)

[1] https://www.imi-getreal.eu/Portals/1/Documents/01%20deliverables/2017-03-29%20-%20WP1%20-%20Advancing%20Evidence%20Generation%20for%20New%20Drugs.pdf

[2] https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf

[3] Cave A, Kurz X, Arlett P. Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. Clin Pharmacol Ther. 2019, Jul;106(1):36-39.

The FDA guidance for Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies using Electronic Healthcare Data Sets[4] describe under the section 'appropriateness of data sources' that investigators should describe historical accessibility and appropriateness of data.
The description about historical accessibility should include:

- How long the data source has been available to the research community;
- How often this data source has been used for pharmacoepidemiologic safety studies;
- The capability of the selected data source to validate the outcome and other study elements (e.g., exposures, key covariates, inclusion/exclusion criteria) based on the safety question;
- References for any relevant publications, including validation studies of safety outcomes of interest in the proposed study that are captured in the database.

In addition, the FDA states that investigators should demonstrate that each data source contains sufficient clinical granularity to capture the exposures and outcomes of interest in the appropriate setting of care and describe the meta-data well.

The Sentinel Initiative has implemented the FDA guidance and does quality assessment of the Sentinel data sources upon each refresh[5]. Approximately 1,200 data checks are evaluated during each Data Partners data refresh. Each data check is designated a "level 1," "level 2," "level 3," or "level 4" data quality check depending on the complexity of a data characteristic/issue:

o Level 1 data checks review the completeness and content of each variable in each table to ensure that the required variables contain data and conform to the formats specified by the Sentinel Common Data Model (SCDM) specifications (e.g., data types, variable lengths, formats, acceptable values, etc.).

o Level 2 data checks assess the logical relationship and integrity of data values within a variable or between two or more variables within and between tables (e.g., variable ADMITTING_SOURCE in the Encounter table is populated only for inpatient and institutional encounters).

o Level 3 data checks examine data distributions and trends over time, both within a Data Partner's database (by examining output by year and year/month) and across a Data Partner's database (by comparing updated SCDM tables to previous versions of the tables). For example, a level 3 data check would ensure that there are no large, unexpected increases or decreases in records over time.

o Level 4 data checks examine the occurrence and prevalence of nonsensical diagnoses and examine variations in care practices across Data Partners (e.g., the proportion of prostate cancer diagnoses among women). Level 4 checks are designed to provide more targeted data analyses and profiling of the Data Partner data. Level 4 data checks are not necessarily designed to detect and correct errors.

A recent paper by Johnson et al. also advocated for a formal framework to assess data quality in healthcare data, a Healthcare Data Quality Framework (HDQF). The items and domains can be visualized with a heatmap or radar graph. A data quality ontology was described which provides rigorous definitions and

[4] Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data Sets. MAY 2013. https://www.fda.gov/index.php/media/79922/download (accessed August 2019)

[5] Sentinel Data Quality Assurance practices. https://www.sentinelinitiative.org/sites/default/files/data/distributed-database/Sentinel_DataQAPractices_Memo.pdf. Accessed August 2019

can automate the computation of data quality measures[6]. Johnson made a literature overview of different quality dimension/measure definitions that are useful for ConcePTION (see Annex 1).

4.3 Goal

To characterize the data in different data sources that can be accessed for ConcePTION and develop and test algorithms to create variables for ConcePTION use cases. Specific objectives are:

1) To conduct syntactical harmonization of local data sources
2) To describe the completeness, frequencies, and distributions of data
3) To assess the format coherence, structural coherence, and uniqueness of each data source
4) To assess rates of diagnoses, medicinal product use and vaccine exposure as indicators of data quality
5) Perform external validation by benchmarking rates with published literature and between data sources, as well as perform verification within each data source to assess plausibility of data
6) To discuss and assess the relevance and fit-for-purpose of each data source for certain use cases
7) To assess the impact of using different component algorithms to create study variables

7. Methods

5.1 Setting and data sources

All relevant population-based data sources (data sources that capture person-time of follow-up of a defined dynamic or fixed population during which medicine use and/or events can be observed/linked) which aim to participate in one or more of the demonstration projects in ConcePTION and are willing to participate in the ConcePTION database characterization.

The types of population-based data sources to be included in this protocol for characterisation are:

1. **Healthcare claims data sources** – created for operational health care purposes and billing of costs on defined population that is followed over time (for example medicinal product dispensing claims)
2. **General practice databases** – electronic medical records provided by General Practitioners (GPs) on defined population that is followed-up prospectively
3. **Birth cohorts** - recruit pregnant women during pregnancy or at birth, irrespective of exposure, and follow-up prospectively
4. **Medicinal product exposure pregnancy registries:** recruit women who are exposed to specific medicinal product(s) and are followed up prospectively
5. **Demographic/population databases** – includes the population register, residents register, date of birth/death
6. **Linkable Registries:** relevant outcome/exposure data collected for specific purpose when they can be linked to an underlying population file that defines follow-up time
    - **Medical Birth Registries**
    - **Specific Disease or outcomes surveillance registries** e.g., EUROCAT, cancer registries, infectious disease surveillance, death

[6] Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. A Data Quality Ontology for the Secondary Use of EHR Data. *AMIA Annu Symp Proc*. 2015;2015:1937–1946. Published 2015 Nov 5.

- **Child surveillance databases** – record growth and development as measured by community child health teams/public health nurses
- **Educational databases** - created for operational education administration purposed for example school results, special educational needs and attendance
- **Registry of disability** - created for insurance purposes and service delivery
- **Immunization registries**
- **Medical encounter databases:** hospital-based encounters, laboratory measurements, imaging

All data sources that collect case-based data on pregnancy or pregnancy outcomes which cannot be linked to an underlying population will be considered in a separate ConcePTION data characterization protocol (e.g., spontaneous reports, Teratology Information Services (TIS) reports, EUROCAT surveillance that cannot be linked at the individual level to a demographic register).

5.2 Source and study population

For the data characterization and algorithm evaluation the source population comprises of all persons registered in any of the data sources at any time during the study period. From the source population, we will extract a ConcePTION relevant study population of all persons 0-55 years of age.

The study population will be followed from the moment of registration in the data source or start of the study period (01-01-1995) until death, reaching the age of end of follow-up (56) for data characterization. Data for the study population will be extracted from available data sources and converted into a ConcePTION Common Data Model (CDM) and then characterized. See flow in figure 1.

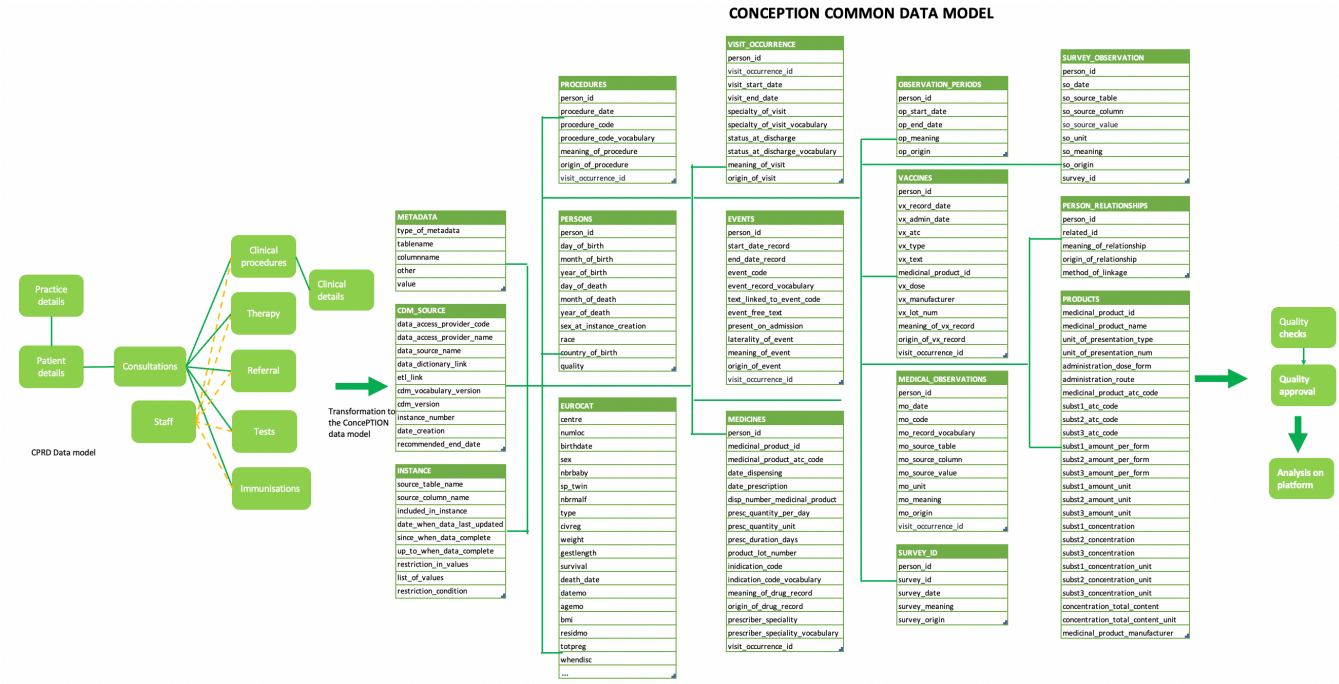

Figure 1: Example using the CPRD (Clinical Practice Research Datalink) data model of how a common data structure can be converted into the ConcePTION CDM. Data characterization will take place with

the second arrow on the data in the CDM. For the purposes of the current protocol, 'Analysis files' are those used in the data characterization exercise. Analysis files for Demonstration Projects (DPs) will be developed at a later stage.

5.3 ConcePTION Common Data Model (CCDM)

Data access providers (DAPs) will be asked to extract all available data of relevance for the ConcepTION study population and convert these data into the ConcePTION Common Data Model (CCDM) using their preferred software for syntactic harmonization. See Figure 2.

The ConcePTION CDM for population-based databases comprises of the following tables:

METADATA TABLES

The metadata tables contain data in a machine-readable format which allows for processing of the data in the CDM.

1. **METADATA**

The METADATA table contains indicators which can act as machine readable guides for code written against the CDM. For instance, whether data in the MEDICINES table represents prescription or dispensing.

2. **CDM_SOURCE**

Contains high-level metadata describing the source data for the current instance such as the name of the source, data access provider, and date of last update.

3. **INSTANCE**

The INSTANCE table contains data on the specific instance of the ConcePTION CDM, such as tables and columns from source data which have been included.

4. **PRODUCTS**

Listing of national product codes for medicinal products. Contains a medicinal product ID, linked to the MEDICINES and VACCINES table. The PRODUCTS table contains detailed data on products at the package level.

CURATED TABLES

Curated tables differ from the other tables of the CDM in that data access providers are asked to create these tables using rule-based algorithms. These tables therefore represent a syntactic and semantic harmonization.

1. **PERSONS**

One row of data per subject present in the data and meeting inclusion criteria for the CDM instance at any point during the study period. Data on each subject includes sex as specified at the date of the instance creation, day of birth, month of birth, year of birth, day of death, month of death, year of death (these may be derived using DAP-specific rules), race, country of birth, and the quality of the record.

2. **OBSERVATION_PERIODS**

One row per period during which a subject is present in the database. One subject can have one or more observation periods. This may be based upon registration in a geographical area, registration in a GP practice, presence in a registry, etc.

3. **PERSON_RELATIONSHIPS**

Contains one row of data for each relationship between two persons identifiable in the database. This relationship may be parent-child, sibling, or shared household status.

ROUTINE HEALTH CARE DATA TABLES

Routine health care data tables capture data observed during routine health care in hospitals, GP offices, pharmacies, outpatient clinics, etc.

1. **VISIT_OCCURRENCE**

Contains an identifier of a visit(visit_occurrence_id) to allow for linkage of diagnoses, procedures, medicinal products, procedures etc in the same visit if this information is available in a database.

2. **EVENTS**

Contains data on events indicated by a diagnosis code or free text. It contains one row per diagnosed event.

3. **MEDICINES**

One record per prescription or dispensing. Contains data required to estimate duration of exposure. Linkage to PRODUCTS table through the medicinal products id to access data on medicinal products at the package level.

4. **PROCEDURES**

Contains data on procedures ordered or completed. For those procedures with an associated result, results and units are recorded in the MEDICAL_OBSERVATIONS table. It contains one row per procedure.

5. **VACCINES**

Contains data on vaccinations with one row per vaccine. Data on dose number for childhood vaccines and manufacturer are accommodated by this table.

6. **MEDICAL_OBSERVATIONS**

Contains observations recorded during routine health care. Can be a result from a laboratory test, or physical measurement, a pathology report, even socio-economic status, smoking etc.

SURVEILLANCE TABLES

Surveillance tables contain data collected for purposes beyond routine health care either for surveillance of specific events or for recording of detailed information related to a unit of observation such as a pregnancy or chronic illness.

1. **EUROCAT**

Contains the EUROCAT or EUROmediCAT (a subset of the EUROCAT) table for those data access providers which have access to this standard table.

2. **SURVEY_ID**

Contains metadata on observations contained in the SURVEY_OBSERVATIONS table and allows for linkage between mothers and infants captured in a medical birth registry.

3. **SURVEY_OBSERVATIONS**

Contains one row per observation in any survey or registry data table – such as a medical birth registry, well child program database, cancer registry, etc.

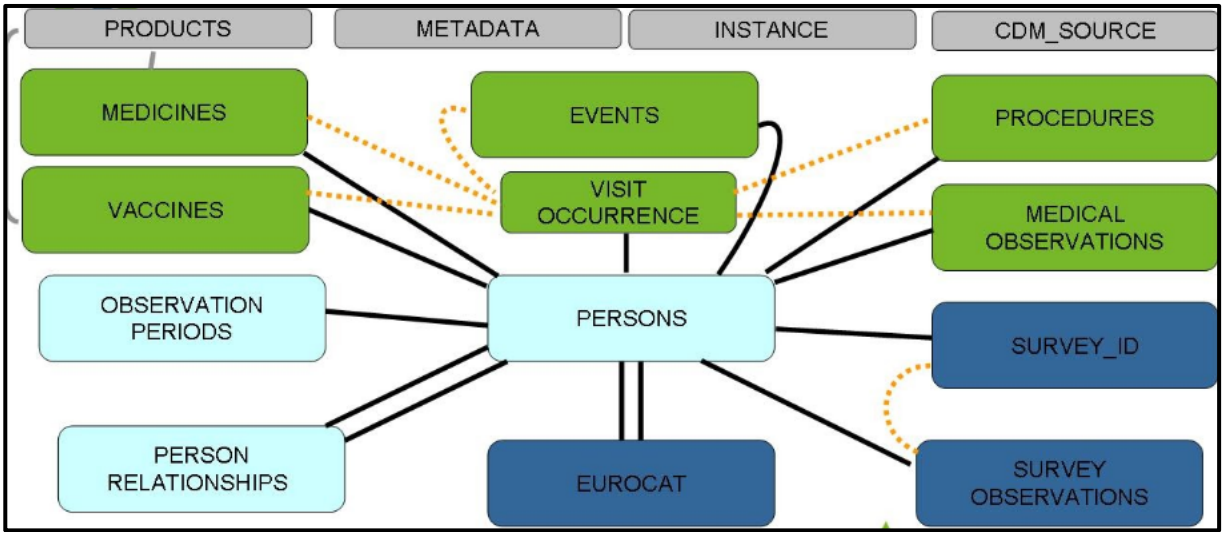

Figure 2. Schematic representation of the ConcePTION CDMv2.2

DAPs will be requested to extract-transform-load their local data (ETL) into the CCDM. The script to carry out to ETL will be uploaded in the ConcePTION catalogue for transparency.

5.4 Data extraction and transformation specification

Data sources will be requested to extract and fill the following type of tables from all the data sources they have available. The CDM specifications will be made available as an Excel workbook with complete definitions for each CDM variable, and versioning control. The following tables provided below provide a high-level description of each CDM table.

METADATA: This is a *mandatory* table. In order to have automated procedures to look at the CDM, DAPs are asked to fill this table indicating presence or absence of each CDM table and non-mandatory column in the instance. DAPs should also indicate values for those tables and columns with finite allowable values in the instance of the CDM.

METADATA table

METADATA	**Metadata**	This table contains some general information about how the origin data fit the CDM: for instance, they are used to describe which tables of the standard CDM are populated in this instance; and what coding systems are used for the various data domains. This information is used by the scripts for quality check (e.g., check that all the target tables that are expected to be findable can indeed be found; and that the coding systems that are observed in the loaded data are indeed those listed here)
Variable	**Mandatory**	**Description**
type_of_metadata	Yes	There are different types of metadata that are recorded, they may be associated with a table or a column, or other.
tablename	Yes	Name of the table whose metadata is recorded.
columnname	Yes	Name of the column whose metadata is recorded.
other	Yes	Other characteristic of the metadata.
value	Yes	Value of the metadata.

CDM_SOURCE: This is a *mandatory* table. DAPs are asked to fill this table describing minimally the data access provider code and name, data source name, CDM version, instance number, date of instance creation, and recommended end date.

CDM_SOURCE table

CDM_SOURCE	**Metadata**	In this table, a high-level, machine-readable description of the instance of the CDM is contained. The scripts of the studies that are deemed to run on this instance will use this information to tailor some choices to the specific DAP and data source.
Variable	**Mandatory**	**Description**
data_access_provider_code	Yes	Code of this DAP organization in the ConcePTION coding system.
data_access_provider_name	Yes	Name of the DAP organization.
data_source_name	Yes	Name of the DAP data source whose subset populates this instance of the CDM (if any).
data_dictionary_link	No	Link to a source where the data dictionary of the original data source can be found.
etl_link	No	Link to a source where the current version of the ETL document of this data source can be found.
cdm_vocabulary_version	No	Version of the ConcePTION CDM this instance conforms to.
cdm_version	Yes	Version of the ConcePTION CDM vocabulary this instance conforms to.
instance_number	Yes	Sequential number of the instances of the CDM that the DAP data_access_provider_code has created on date_creation date from the data source data_source_name it has access to
date_creation	Yes	Date when this CDM instance is populated.
recommended_end_date	Yes	Recommended end date for studies using this instance.

INSTANCE: This is a *non-mandatory* table which DAPs may choose to fill if they would like to provide machine-readable data describing underlying source data on a table-by-table or column-by-column basis.

INSTANCE table

INSTANCE	**Metadata**	This table displays the list of the tables and columns of the origin data dictionary that are mapped to the instance of the CDM, together with date of last update (both in terms of when the data was accessed by the DAPs, and when the data was actually recorded and can be considered complete). This is to be used, together with a machine-readable version of the ETL, to match the inclusion of the study population and the creation of the study variables to the actual data loaded in the CDM instance. The list is restricted to tables and columns of the origin data dictionary that are included in the current ETL document.
Variable	**Mandatory**	**Description**

source_table_name	Yes	Table of the local dictionary that is used in ETL.
source_column_name	Yes	Column of the local dictionary that is used in ETL.
included_in_instance	Yes	Specify whether this column of this table was used to populate this specific instance of the CDM.
date_when_data_last_updated	only if included_in_instance='yes'	Date when the DAP last received this column.
since_when_data_complete	only if included_in_instance='yes'	Date since when the DAP considers this column to contain complete data.
up_to_when_data_complete	only if included_in_instance='yes'	Date up to when the DAP considers this column to contain complete data.
restriction_in_values	only if included_in_instance='yes'	Whether the data uploaded in the CDM were selected based on values of this column.
list_of_values	No	List of values of this column that are included in this instance of the CDM.
restriction_condition	No	Condition involving this column that restricted the data uploaded to the instance of the CDM.

PRODUCTS: This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. In this table, DAPs provide product-level information for medicinal products. This is particularly relevant for data sources with dispensing rather than prescription data as the data contained in the PRODUCTS table allows for calculation of exposure periods based upon dispensed units (contained in the MEDICINES table) and box size (contained in the PRODUCTS table).

PRODUCTS table

PRODUCTS	**Metadata**	This table collects the information associated to each marketed product that may have been prescribed, dispensed, or administered to a patient. It contains one row per product
Variable	**Mandatory**	**Description**
medicinal_product_id	No	Primary key. The medicinal_product_id should be a unique identifier of a specific medicinal product.
medicinal_product_name	Yes	Any substance or combination of substances, which may be administered for treating or preventing disease, with the view to making a medical diagnosis or to restore, correct or modify physiological functions.
unit_of_presentation_type	No	Qualitative term describing the discrete countable entity in which a pharmaceutical product or manufactured item is presented, in cases where strength or quantity is expressed referring to one instance of this countable entity.
unit_of_presentation_num	No	Number of unit presentation type within a medicinal product.

administration_dose_form	No	Pharmaceutical dose form for administration to the patient.
administration_route	No	Route of administration of the pharmaceutical product.
medicinal_product_atc_code	Yes	Unique standardized identification code from the ATC classification system from WHO associated to the medicinal product.
subst1_atc_code	No	Unique standardized identification code from the ATC classification system from WHO associated to the active principle.
subst2_atc_code	No	Unique standardized identification code from the ATC classification system from WHO associated to the active principle.
subst3_atc_code	No	Unique standardized identification code from the ATC classification system from WHO associated to the active principle.
subst1_amount_per_form	No	Quantity of the first active principle contained in the medicinal product
subst2_amount_per_form	No	Quantity of the second active principle contained in the medicinal product
subst3_amount_per_form	No	Quantity of the third active principle contained in the medicinal product
subst1_amount_unit	No	Unit of measure of the quantity of the first active principle contained in the medicinal product
subst2_amount_unit	No	Unit of measure of the quantity of the second active principle contained in the medicinal product
subst3_amount_unit	No	Unit of measure of the quantity of the third active principle contained in the medicinal product
subst1_concentration	No	Strength or quantity contained into a single unit of presentation or dose form.
subst2_concentration	No	Strength or quantity contained into a single unit of presentation or dose form.
subst3_concentration	No	Strength or quantity contained into a single unit of presentation or dose form.
subst1_concentration_unit	No	Unit of measure of the strength or quantity by which a particular type of unit of presentation or dose form is described.
subst2_concentration_unit	No	Unit of measure of the strength or quantity by which a particular type of unit of presentation or dose form is described.
subst3_concentration_unit	No	Unit of measure of the strength or quantity by which a particular type of unit of presentation or dose form is described.
concentration_total_content	No	Total content of a single unit such as particular type of pharmaceutical unit of presentation or dose form.
concentration_total_content_unit	No	Unit of measure of the concentration total content.
medicinal_product_manufacturer	No	Name of the manufactured of the pharmaceutical product.

PERSONS: This is a *mandatory* table. All fields need to be filled for the study population. DAPs are asked to decide upon a local algorithm to determine dates of birth (day, month and year of birth) and death (day, month and year of death) as well as sex of the person.

PERSONS table

PERSONS	**Curated tables**	This table records persons that are to enter analysis of this instance of the CDM
Variable	**Mandatory**	**Description**
person_id	Yes	
day_of_birth	No	
month_of_birth	No	
year_of_birth	Yes	
day_of_death	No	
month_of_death	No	
year_of_death	Yes	
sex_at_instance_creation	Yes	Sex of the person in the moment when in the instance of the CDM is created.
race	No	
country_of_birth	No	
quality	No	A judgement on the quality of the variables recorded in this table.

OBSERVATION_PERIODS: This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. All fields to be filled for each person in the study population and their periods of follow-up as well as the provenance (source) of the data on follow-up. One person may have multiple observation periods, in one or more data sources if they can be linked. For example, if you have a national population-based database you may be able to follow all subjects from birth to death, also you may be able to link to a vaccination register only from first of January 2010. In that instance you enter two records for the person_id, one with provenance demographic_register, one with provenance vaccine_register.

OBSERVATION_PERIODS table

OBSERVATION_PERIODS	**Curated tables**	Periods during which data is collected in the database for this person. This table is a starting point to define the study population of all studies based on this instance
Variable	**Mandatory**	**Description**
person_id	Yes	
op_start_date	Yes	
op_end_date	Yes	
op_origin	Yes	Represents what mechanism originated the record.
op_meaning	Yes	Represents the semantic of the record.

PERSON_RELATIONSHIPS: This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. If mother-child linkage is available in a data source, we ask data access providers to fill the PERSON_RELATIONSHIPS table for this linkage.

PERSON_RELATIONSHPS table

PERSON_RELATIONSHIPS	**Curated table**	For any person, this table collects the pairing with the identifier of mother or of other relationships that may be available
Variable	**Mandatory**	**Description**
person_id	Yes	
related_id	Yes	
origin_of_relationship	Yes	Where the information about the relationship comes from.
meaning_of_relationship	Yes	Which type of relationship there is between the mother and the person.
method_of_linkage	Yes	How the linkage was performed.

VISIT_OCCURRENCE: This is a *non-mandatory* table which DAPs may choose to fill if they would like to provide linkage among observations occurring within the same healthcare visit. Contains an identifier of a visit to allow for linkage of diagnoses, procedures, dispensings etc. in the same visit if this information is available in a data source.

VISIT_OCCURRENCE table

VISIT_OCCURRENCE	**Routine healthcare data**	This table contains a summary description of the visits during which records of EVENTS, PROCEDURES, but possibly also MEDICAL_OBSERVATIONS, VACCINES or MEDICINES were recorded. This serves both to collect visit-level information, and to enable grouping sets of records that were recorded concurrently
Variable	**Mandatory**	**Description**
person_id	Yes	
visit_occurrence_id	Yes	Visit identifier.
visit_start_date	Yes	Date when the visit starts, or, if it is just a one-day visit, date of the visit.
visit_end_date	No	Date when the visit ends (only for visits that may last more than one day, such as a hospital admission).
speciality_of_visit	No	Specialty of the visit, or if this is a hospital admission, specialty of the discharge ward.
speciality_of_visit_vocabulary	No	Coding system of the specialty.
status_at_discharge	No	Outcome of the visit.
status_at_discharge_vocabulary	No	Vocabulary of outcome of the visit.
meaning_of_visit	Yes	
origin_of_visit	Yes	

EVENTS: This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. We would like to ask data access providers to extract diagnosis codes for the following events (diagnoses, see box 1), when there is any occurrence during the study period. This selection of events is based on the proposed demonstration studies in WP1, but they may suffer modification in a later stage.

Box 1. List of diagnosis events to be extracted

Attention Deficit Hyperactivity Disorder (ADHD)
Autism spectrum disorder
Bacterial Infections
Bipolar disorder
Breast Cancer
Depression/ anxiety
Digestive disorders
Epilepsy
Foetal growth restriction
Gestational Diabetes
Hearing Impairment
Hyperemesis gravidarum
Induced terminations of pregnancy -elective
Low birth weight
Major congenital anomalies
Maternal death
Microcephaly
Migraine
Multiple gestation
Multiple sclerosis

Neonatal death Pain Pre-eclampsia Preterm birth

Events for initial data characterization and algorithm development will be defined using a standard template to arrive at the codes that need to be extracted (See annex 3). Codes will be provided to the DAPs by the ConcePTION Definitions Task Force.

The extracted data should be formatted using the following structure:

EVENTS table

EVENTS	**Routine healthcare data**	This table collects diagnoses, symptoms and signs ('events') observed during routine healthcare, such as a hospital admission, a primary care or specialist visit, or other.
Variable	**Mandatory**	**Description**
person_id	Yes	A foreign key to the person in "person" table who experimented the event
start_date_record	Yes	Start date of the visit that led to the recording of the event code of free text
end_date_record	No	End date of the visit that led to the recording of the event code of free text
event_code	Yes, unless 'event_free_text' is filled in	Code characterizing the event according to the vocabulary defined in event_record_vocabulary
event_record_vocabulary	Yes	Vocabulary to which the event_code belongs to; or, if the record contains event_free_text, this column contains the indication 'free text'
text_linked_to_event_code	No	If in the original record the code is modified by a text, include this text here
event_free_text	No	Use this cell if in the record there is no code, just a text
present_on_admission	No	Indicates the presence of the event at the start of the visit or hospital admission
laterality_of_event	No	Laterality of event.
meaning_of_event	Yes	This is a ConcePTION classification of the nature of the original record associated with this event
origin_of_event	Yes	This is a ConcePTION classification of the purpose why the record was recorded
visit_occurrence_id	No	A foreign key linking this record to the VISIT_OCCURRENCE table

MEDICINES: This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. We ask that all medicines in the classes listed below (see box 2) with a date of dispensing or prescription within the study period, to be extracted, thus, to be able to characterize the fitness of purpose of the data and be prepared for future studies. Please provide Anatomical Therapeutic Chemical Classification (ATC) codes as much as possible.

Box 1. List of medicines to be extracted

Agents acting on the renin-angiotensin system (C09)
Analgesics (N02)
Antibacterials for systemic use (J01)
Antidepressants (N06A)
Antiemetics and antinauseants (A04A)
Antiepileptics (N03A)
Antihypertensives (C02)
Antineoplastic agents (L01)
Anti-Parkinson drugs (N04)
Antipsychotics (N05A)
Antivirals for systemic use (J05)
Betablockers (C07)
Calcium blockers (C08)
Corticosteroids for systemic use (H02)
Diuretics (C03)
Drugs for obstructive airway diseases (R03)
Drugs used in Diabetes (A10)
Endocrine therapy (L02)
Immunostimulants (L03)
Immunosuppressants (L04)
Muscle relaxants (M03)
Other nervous system drugs (N07)

MEDICINES table

MEDICINES	**Routine healthcare data**	This table collects data on medicinal products prescriptions, dispensings or administrations occurred during routine healthcare.
Variable	**Mandatory**	**Description**
person_id	Yes	A foreign key to the person in PERSONS table
medicinal_product_id	No	Foreign key to the PRODUCTS table. The medicinal_product_id should be a unique identifier of a specific medicinal product.
medicinal_product_atc_code	Yes	ATC classification system code attributed to the medicinal product.
date_dispensing	Yes, unless 'date_prescription' is populated	Date when the medicinal product that led to the recording was dispensed or administrated to the patient
date_prescription	Yes, unless 'date_dispensing' is populated	Date when the medicinal product that led to the recording was prescribed
disp_number_medicinal_product	No	Number of dispensed units of medicinal_product_id.
presc_quantity_per_day	No	Prescribed quantity of medicinal product to be taken daily.
presc_quantity_unit	No	Unit of measure of the prescribed daily quantity.
presc_duration_days	No	Number of days of medication as prescribed.
product_lot_number	No	An identifier assigned to a particular quantity or lot of medicinal products from the manufacturer.

inidication_code	No	Single identifier of a condition/indication for which the medicinal product was prescribed/dispensed.
indication_code_vocabulary	No	Coding system referring to indication code.
meaning_of_drug_record	Yes	Nature of the original record having originated the drug record.
origin_of_drug_record	Yes	Name of the source table that originated the record.
prescriber_speciality	No	Profile of the healthcare professional who has prescribed the medicinal product.
prescriber_speciality_vocabulary	No	Coding system of the speciality.
visit_occurrence_id		Identifier of the prescription. A foreign key linking this record to the VISIT_OCCURRENCE table, indicating the visit where the drug was prescribed or dispensed.

PROCEDURES: This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. A procedure is a course of action intended to achieve a result in the delivery of care. We would like to ask DAPS to extract the procedures listed below:

PROCEDURES table

PROCEDURES	**Routine healthcare**	This table collects procedures administered during routine healthcare. Can be a surgery, or a diagnostic procedure, a rehabilitation procedure, a therapeutical procedure.
Variable	**Mandatory**	**Description**
person_id	Yes	
procedure_date	Yes	
procedure_code	Yes	
procedure_code_vocabulary	Yes	
visit_occurrence_id	No	A foreign key linking this record to the VISIT_OCCURRENCE table.
meaning_of_procedure	Yes	
origin_of_procedure	Yes	

VACCINES: This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. We ask that all vaccines (ATC code beginning J07) with a date of dispensing or administration within the study period be extracted. See **Annex 4** for a listing of requested drug classes.

VACCINES table

VACCINES	**Routine healthcare data**	This table collects dispensations or administrations of vaccines.
Variable	**Mandatory**	**Description**
person_id	Yes	
vx_record_date	Yes, if vx_admin_date is missing	
vx_admin_date	Yes, if vx_record_date is missing	
vx_atc	Yes, if vx_type is missing	
vx_type	Yes, if vx_atc is missing	
vx_text	No	
medicinal_product_id	No	Foreign key to the PRODUCTS table.

origin_of_vx_record	Yes	Name of the source table that originated the record
meaning_of_vx_record	Yes	
vx_dose	No	Dose, particularly for childhood vaccines (1, 2, 3, Booster, etc.)
vx_manufacturer	No	Name of vaccine manufacturer
vx_lot_num	No	
visit_occurrence_id	No	External key to VISIT_OCCURRENCE

MEDICAL_OBSERVATIONS: This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. Based on the demonstration studies' needs, we will initially focus on the type of measurements described in box 3. An exact list of definitions will be provided.

We would like to ask data access providers to extract the following observations (incl. Measurements), if available (see box 3):

Box 3. List of observations and measurements to be extracted

Alcohol use
Apgar score
BMI and/or its components
Breastfeeding duration
Breastfeeding exclusivity
Breastfeeding status
Educational level
Folic acid use
Gestational age at birth
Last menstrual period
Mode of delivery
Smoking status
Socio-economic status and/or proxies of SES

MEDICAL_OBSERVATIONS table

MEDICAL_OBSERVATIONS	**Routine healthcare data**	This table collects observations recorded during routine healthcare. Can be a result from a laboratory test, or a physical measurement, but also level of education, or sex, or a pathology report
Variable	**Mandatory**	**Description**
person_id	Yes	A foreign key to the person in PERSONS table.
mo_date	Yes	
mo_code	No	
mo_record_vocabulary	No	
mo_source_table	No	
mo_source_column	No	
mo_source_value	Yes	

mo_unit	No	
mo_meaning	Yes	
mo_origin	Yes	Name of the source table that originated the record.
visit_occurrence_id	No	

EUROCAT: This is a *mandatory* table for EUROCAT DAPs. It is a copy of the locally held EUROCAT table, with identifiers removed or recoded if necessary.
See **Annex 2** for the complete EUROCAT CDM.

SURVEY_ID: This is a *non-mandatory* table. This table should be filled by those DAPs choosing to fill the corresponding SURVEY_OBSERVATIONS table.

SURVEY_ID table

SURVEY_ID	**Surveillance**	This table contains a summary description of the survey during which records of SURVEY_OBSERVATIONS were recorded. This serves both to collect survey-level information, and to enable grouping sets of records that were recorded concurrently
Variable	**Mandatory**	**Description**
person_id	Yes	Person whose information is collected in this survey.
survey_id	Yes	Identifier of the survey.
survey_date	Yes	Date when the survey is recorded.
survey_meaning	Yes	The meaning of this survey for this person.
survey_origin	Yes	Name of the source table that originated the record.

SURVEY_OBSERVATIONS: This is a *non-mandatory* table. This table should be filled by those DAPs with access to surveillance data which may help to define study outcomes.

SURVEY_OBSERVATIONS table

SURVEY_OBSERVATIONS	**Surveillance**	List of observations in a survey
Variable	**Mandatory**	**Description**
person_id	Yes	
so_date	Yes	
so_source_table	Yes	
so_source_column	Yes	
so_source_value	Yes	
so_unit	No	
so_meaning	Yes	
so_origin	Yes	Name of the source table that originated the record.
Survey_id	Yes	

5.4 Analysis and indicators

Each data access provider (DAP) will be responsible for the extraction, transformation, and loading of their original data to the ConcePTION CDM. Standardized scripts will be written by the group of statisticians and data engineers in R for data characterization, to run against data in the ConcePTION common data model. R scripts and instructions will be sent to participating DAPs using a task management system.

The DAP is responsible for converting data into the CDM using their preferred software and subsequently running the provided R script against the CDM-converted data. The results of the R-script will be submitted to a computing platform that can be accessed remotely by DAPs and ConcePTION partners and participating DAPs using authentication. Access to each DAP's results on the platform will be limited to the data access provider, WP1 statisticians, and WP7 statisticians.

Data quality will be assessed according to a clear framework based on the ADVANCE database characterization process[7], the United States FDA Sentinel System data quality indicators[8,10], the Observational Health Data Sciences and Informatics (OHDSI) data quality dashboard (in development)[9], and EUROCAT indicators for population-based healthcare data sources. The data quality and characterization checks described below will take place in collaboration with partners. All data will remain local and only summary measures described below will be inspected in collaboration with WP7 partners and the task force for data transformation. This process will proceed iteratively in collaboration with each DAP until consensus on fitness for purpose has been reached between WP7 and the DAP, the result of this consensus process and some core results will be made available on the catalogue in a private area for inspection by investigators and DAPs. For all indicators and characterization output resulting in a cell count less than 5, counts will not be reported and will be replaced with "<5" programmatically.

All data sources with data in the EUROCAT CDM will be characterized according to the EUROCAT data quality indicators (https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/DQI-List-of-Data-Quality-Indicators-since-2012.pdf). See Annex 4.

Level 1 data checks review the completeness and content of each variable in each table of the D2 CDM to ensure format, structural and relational coherence of the data (e.g., date format, data types, variable lengths, formats, acceptable values, etc.). Also, the plausibility of the data is assessed (e.g., distribution of date variables over time, distribution of continuous variables etc.), as well as the uniqueness of records (through a duplicated records check).

> This is a check conducted in collaboration with DAPs to verify that the extract, transform, and load (ETL) procedure to convert from source data to the *ConcePTION CDM* has been completed as expected. All D2 CDM tables will be undergo a verification of formatting procedure to ensure the loading of each table is correct. Formats for all values will be assessed and compared to a list of acceptable formats. Frequency tables of variables with finite allowable values will be created to identify unacceptable values. Distribution of continuous variables and date variables will also be constructed. All tables will be checked for duplicated data.

Level 1b data checks map all extracted and loaded data in the common data model at the value level for each D2 CDM table. The precision and coherence of data can also be assessed from the Level 1b output.

> Level 1b gives the possibility to the data access providers and the principal investigator to assess question specific determinants by reviewing all values and variables present in the database. Through the level 1b output the semantic coherence (e.g., the use of ATC codes in the MEDICINES

[7] http://www.advance-vaccines.eu/app/archivos/publicacion/78/ADVANCE_D5.4_finalfingerprintreportv1.020190620.pdf
[8] https://www.sentinelinitiative.org/sentinel/data-quality-review-and-characterization
[9] https://github.com/OHDSI/DataQualityDashboard/tree/master/docs

and PRODUCTS table is present at the same level, smoking expressed as a categorical yes/no variable and as number of cigarettes per day in the same database etc.) and precision of data (e.g., the duration of treatment can be expressed in days or weeks) can be examined.

Level 2 data checks assess the logical relationship and integrity of data values within a variable or between two or more variables within and between tables. Level 2 verifies the temporal plausibility of data by looking at records occurring outside of recorded person time. Other checks include proportion of observations related to a person id that is not in the PERSONS table, proportion of parents younger than 12 years old etc.

In this check, we will assess the proportion of observations occurring before a recorded birth date, observations occurring after a recorded death date and, observations occurring outside of observation periods. Also, assessment of records related to a visit occurrence id that occur before or after a visit start or end date will be performed. The uniqueness of the data will be verified by comparing person id linked to a visit id between all D2 CDM tables and the VISIT OCCURRENCE table.

Following completion of level 1, 1b and, 2 checks, WP7 will review results with DAPs and assess any detected errors. Only after these errors have been resolved to the satisfaction of the DAP will characterization proceed with level 3 checks.

Level 3 data checks examine data distributions and trends over time, both within a DAP's database (by examining output by year) and across a Data Partner's databases (by comparing updated CDM tables to previous versions of the tables). Logical constraints of data will be verified, for example rates of sex-specific diseases (i.e., higher rates of breast cancer in males compared top females) or age-specific medication use (i.e., high rates of narcotic medication in children) etc. Also, the coverage of data will be assessed by looking at the population distribution and comparing it to the source data.

A level 3 data check would ensure that there are no large, unexpected increases or decreases in records over time which do not have an appropriate explanation (such as changes in the number of subjects included in the database or known changes in treatment recommendations). In this check, we will calculate incidence of events, drug use and vaccine exposure by database and calendar year. By comparing these types of summary measures across databases, over time, and against external statistics and literature sources, anomalies and errors which can be corrected in partnership with DAPs will appear. Also, the output of level 3 will give the possibility to the PI to determine the study specific database relevance.

<table>
<tr><th colspan="3">Level 1: Data Structure Characterization & Completeness*
*All frequency tables constructed overall and by calendar year
from latest of 1995 or year of DB initiation until the present</th></tr>
<tr><td colspan="3">● Compliance with formatting requirements in each CDM table (by table).
● Completeness of the data in each CDM table (by table).
● Frequency tables for all categorical variables in each CDM table (by table).
● Distributions for continuous variable and dates in each CDM table (by table).
● Proportion of incomplete dates in each CDM table (by table).
● Uniqueness of data in each CDM table (by table).</td></tr>
<tr><td>Indicator</td><td>Precedent</td><td>First Characterization Round</td></tr>
<tr><td>Number of subjects</td><td>Sentinel Level 1 quality check</td><td>Yes</td></tr>
<tr><td>Number of subjects with missing sex data</td><td>Sentinel Level 1 quality check</td><td>Yes</td></tr>
</table>

Percentage of subjects with missing sex data • Percentage with 'unknown' and 'other'	Sentinel Level 1 quality check	Yes
Number of subjects with missing date of birth	Sentinel Level 1 quality check	Yes
Percentage of subjects with missing date of death	Sentinel Level 1 quality check	Yes

Level 2: Data Relational Logic Characterization*
***All characterizations conducted by calendar year from latest of 1995 or year of DB initiation until the present**

- Proportion of records with dates before birth date for each person in each CDM table (by table).
- Proportion of records with dates after death date for each person in each CDM table (by table).
- Proportion of records with dates outside of observation dates for each person in each CDM table (by table).
- Proportion of records with person id not in PERSONS table.
- Proportion of records linked to a visit id with date prior to visit start date.
- Proportion of records linked to a visit id with date after to visit end date.
- Proportion of records linked to a visit and with a person id different from the person id in VISIT OCCURRENCE
- Proportion of parents younger than 12 years old.

Indicator	Precedent	First Characterization Round
Percentage of subjects with data outside of observation periods	OHDSI data quality dashboard, Sentinel Level 2 quality check	Yes
Percentage of subjects with data before their recorded date of birth	OHDSI data quality dashboard, Sentinel Level 2 quality check	Yes
Percentage of subjects with data after their last recorded date of death	OHDSI data quality dashboard, Sentinel Level 2 quality check	Yes

Level 1b: Data Characterization at value level*
All characterizations conducted in the D2 CDM tables

- Frequency analysis for each variable value in each CDM table (by table).
- Frequency analysis for each combination of variable values in each CDM table (by table).

Level 3: Data Content Characterization*
***All characterizations conducted by calendar year from latest of 1995 or year of DB initiation until the present**

- Quality indicators will be assessed at value and column level for each CDM table.

Domain	Indicator	Definition	First Characterization Round
Study and source population	Number of subjects by age band and sex in the source population[a]	Number of subjects by sex and 5-year age bands (i.e., 1-4, 5-9, 10-11 etc.) in the source population[a]	Yes
Study and source population	Number of subjects by age band at original start observation[c] over time in the source population[a]	Number of subjects by 5-year age bands (i.e., 0, 1-4, 5-9, 10-11 etc.) calculated at original start follow-up[c] by year of observation in the source population[a]	Yes

Study and source population	Number of subjects by age band at original start observation over time in the study population[b]	Number of subjects by 5-year age bands (i.e., 0, 1-4, 5-9, 10-11 etc.) calculated at original start follow-up[c] by year of observation in the study population[b]	Yes
Study and source population	Person-years of follow-up by age band at original start observation[c] over time in the source population[a]	Person-years of follow-up by 5-year age bands (i.e., 0, 1-4, 5-9, 10-11 etc.) calculated at original start follow-up[c] by year of observation in the source population[a]	Yes
Study and source population	Person-years of follow-up by age band at original start observation[c] over time in the study population[b]	Person-years of follow-up by 5-year age bands (i.e., 0, 1-4, 5-9, 10-11 etc.) calculated at original start follow-up[c] by year of observation in the study population[b]	Yes
Study and source population	Average number of observation periods per subject by age band at original start observation[c] over time in the source population[a]	Number of observation periods per subject (mean) by 5-year age bands (i.e., 0, 1-4, 5-9, 10-11 etc.) calculated at original start follow-up[c] by year of observation in the source population[a]	Yes
Study and source population	Average number of observation periods per subject by age band at original start observation[c] over time in the study population[b]	Number of observation periods per subject (mean) by 5-year age bands (i.e., 0, 1-4, 5-9, 10-11 etc.) calculated at original start follow-up[c] by year of observation in the study population[b] (should always be equal to one)	Yes
Study and source population	Number of subjects by age band at start observation[d] over time in the study population[b]	Number of subjects by 10-year age bands (i.e., 0, 1-9, 10-19, 20-29 etc.) calculated at start follow-up[d] by year of observation in the study population[b]	Yes
Study and source population	Number of subjects by sex, age band at start observation[d] over time in the study population[b]	Number of subjects by sex, 10-year age bands (i.e., 0, 1-9, 10-19, 20-29 etc.) calculated at start follow-up[d] by year of observation in the study population[b]	Yes
Study and source population	Average person-years of follow-up by sex, age band at start observation[d] over time in the study population[a]	Person-years of follow-up (mean, median) by sex, 10-year age bands (i.e., 0, 1-9, 10-19, 20-29 etc.) calculated at start follow-up[d] by year of observation in the study population[a]	Yes
Study and source population	Proportion of subjects by sex and age band over time in the study population[b]	Proportion of subjects by sex, 10-year age bands (i.e., 0, 1-9, 10-19, 20-29 etc.) by year of observation in the study population[b]	Yes
Study and source population	Person-years of follow-up by age band over time in the study population[b]	Person-years of follow-up by 10-year age bands (i.e., 0, 1-9, 10-19, 20-29 etc.) by year of observation in the study population[b]	Yes
Study and source population	Person-years of follow-up by sex and age band over time in the study population[b]	Person-years of follow-up by sex and 10-year age bands (i.e., 0, 1-9, 10-19, 20-29 etc.) by year of observation in the study population[b]	Yes
Study and source population	Proportion of subjects by sex and age band over time (month and year) in the study population[b]	Proportion of subjects by sex, 10-year age bands (i.e., 0, 1-9, 10-19, 20-29 etc.) by month and year of observation in the study population[b]	Yes

Study and source population	Person-years of follow-up by sex and age band over time (month and year) in the study population[b]	Person-years of follow-up by sex and 10-year age bands (i.e., 0, 1-9, 10-19, 20-29 etc.) by month and year of observation in the study population[b]	Yes
Study and source population	Distribution of time difference between date of birth and cohort entry date	Proportion of records categorised in week differences between date of birth and cohort entry date for each subject	Yes
Study and source population	Person-years of follow-up by year of birth in the study population[b]	Person-years of follow-up by year of birth in the study population[b], for subjects present in the database within 8 weeks of their date of birth	Yes
Dates	Distribution of start and end dates of observation in the study population[b]	Distribution of original start and end date of observation[c] and, after application of time exclusion criteria by month and year of observation in the study population[b]	Yes
Dates	Distribution of birthdates over time in the study population[b]	Distribution of birthdates by day, month, and year of birth in the study population[b]	Yes
Medicines	Completeness of medicines records overall and by source of record in the medicines study population[e]	Proportion of records with missing information on indication code, prescriber speciality, dispensed or prescribed quantity and unit overall and by source of data in the medicines study population[e]	Yes
Medicines	Number of prescriptions or dispensations by ATC level 1 and source of data over time in the medicines study population[e]	Number of prescriptions/dispensations as defined by a record of a prescription or a dispensation in the medicines study poulation[e] by ATC code level 1 and source of data over time	Yes
Medicines	Number of prescriptions or dispensations by sex, ATC level 1, 2, 3, and source of data over time in the medicines study population[e]	Number of prescriptions/dispensations (mean, median, total) as defined by a record of a prescription or a dispensation in the medicines study poulation[e] by sex, ATC code level 1, 2 ,3 and source of data over time	
Medicines	Number of users by sex, ATC level 1, 2, 3, 7 and source of data over time in the medicines study population[e]	Number of subjects having at least one record of a prescription or dispensation in the medicines study poulation[e] by sex, ATC code level 1, 2 ,3, 7 and source of data over time	
Medicines	Incidence of prescriptions or dispensations, of the following classes of medicinal products: Agents acting on the renin-angiotensin system (C09) Analgesics (N02) Antibacterials for systemic use (J01) Antidepressants (N06A) Antiemetics and antinauseants (A04A) Antiepileptics (N03A) Antihypertensives (C02) Antineoplastic agents (L01) Anti-Parkinson drugs (N04)	Incidence of exposures as defined by at least one record of a dispensation or prescription by year of age denominator is person-time	

	Antipsychotics (N05A) Antivirals for systemic use (J05) Betablockers (C07) Calcium blockers (C08) Corticosteroids for systemic use (H02) Diuretics (C03) Drugs for obstructive airway diseases (R03) Drugs used in Diabetes (A10) Endocrine therapy (L02) Immunostimulants (L03) Immunosuppressants (L04) Muscle relaxants (M03) Other nervous system drugs (N07)		
	Person years of follow-up in female subjects of childbearing age	Person years of follow-up for women between the ages of 12-55 (inclusive), median	Yes
	Duration of follow-up in female subjects of childbearing age from start of follow-up by age and type of source	Duration of follow-up from start, median	Yes
	Person years of follow-up in children	Person years of follow-up between birth and age 18 (inclusive) median	Yes
	Number of pregnancy-related codes	Count of all pregnancy, end of pregnancy, antenatal care, or delivery-related codes	Yes
	Rate of pregnancy-related codes	Incidence rate of all pregnancy, end of pregnancy, antenatal care, or delivery-related codes for women of childbearing age	Yes
	Number of labor and delivery records	Total number of births (alive or dead)	No
	Number of children present from birth	Number of children with an observation period including their date of birth	No
	Completeness of follow-up surrounding pregnancy	Percentage of subjects with at least one pregnancy-related code who also have at least one year of follow-up	Yes
	Follow-up time from birth	Time from recorded date of birth until exit from the data source (% with 6 months, 1 year, 2 years, 3 years, 5 years, 6 years or more) by year of birth	Yes
	Incidence of prescriptions, dispensations, or exposure to the following classes of medicinal products in women of child-bearing age: ACE Inhibitors/Angiotensin II Receptor Blockers (ARB) (C09) Analgesics (N02) Antiasthmatics (R03A) Antibacterials (J01) Antidepressants (N06A) Antiemetics (A04A)	Incidence of exposures as defined by at least one record of a dispensation or prescription by year of age denominator is person-time	Yes

	Antiepileptics (N03A) Antihypertensives (C02) Antineoplastic agents (L07) Anti-Parkinson drugs (N04) Antipsychotics (N05A) Antivirals (J05) Betablockers (C07) Calcium blockers (C08) Corticosteroids for systemic use (H02) Diuretics (C03) Drugs used in Diabetes (A10) Endocrine therapy (L02) Immunostimulants (L03) Immunosuppressants (L04) Muscle relaxants (M03) Other nervous system drugs (N07) Vaccines (J07)		
	Users per calendar year of the drugs listed above in women of childbearing age	Use of medicinal products as defined by at least one record of a dispensation or prescription, denominator is persontime	Yes
	Incidence of prescriptions, dispensations, or exposure to the following classes of medicinal products in women of child-bearing age by year of age: ACE Inhibitors/Angiotensin II Receptor Blockers (ARB) (C09) Analgesics (N02) Antiasthmatics (R03A) Antibacterials (J01) Antidepressants (N06A) Antiemetics (A04A) Antiepileptics (N03A) Antihypertensives (C02) Antineoplastic agents (L07) Anti-Parkinson drugs (N04) Antipsychotics (N05A) Antivirals (J05) Betablockers (C07) Calcium blockers (C08) Corticosteroids for systemic use (H02) Diuretics (C03) Drugs used in Diabetes (A10) Endocrine therapy (L02) Immunostimulants (L03) Immunosuppressants (L04) Muscle relaxants (M03) Other nervous system drugs (N07) Vaccines (J07)	Incidence of exposures as defined by at least one record of a dispensation or prescription by year of age denominator is person-time	Yes
	Users per calendar year of the drugs listed above in women of child-bearing age by year of age	Use of medicinal products as defined by at least one record of a dispensation or prescription, denominator is person-time	Yes

	Incidence of prescriptions, dispensations, or exposure to the following classes of medicinal products in women who have at least one pregnancy-related code in the calendar year: ACE Inhibitors/Angiotensin II Receptor Blockers (ARB) (C09) Analgesics (N02) Antiasthmatics (R03A) Antibacterials (J01) Antidepressants (N06A) Antiemetics (A04A) Antiepileptics (N03A) Antihypertensives (C02) Antineoplastic agents (L07) Anti-Parkinson drugs (N04) Antipsychotics (N05A) Antivirals (J05) Betablockers (C07) Calcium blockers (C08) Corticosteroids for systemic use (H02) Diuretics (C03) Drugs used in Diabetes (A10) Endocrine therapy (L02) Immunostimulants (L03) Immunosuppressants (L04) Muscle relaxants (M03) Other nervous system drugs (N07) Vaccines (J07)	Incidence of exposures as defined by a record of a dispensation or prescription in women with pregnancy event in the same calendar year	No
	Users per calendar year of the drugs listed above in women who have at least one pregnancy-related code in the calendar year.	Use of medicinal products as defined by at least one record of a dispensation or prescription, denominator is person-time	Yes
	Percent of events with missing diagnosis codes	Proportion of all records in the Event table without a corresponding code for diagnosis, by provenance	Yes
	Distribution (counts over time) of component algorithms identifying the following events: Breast Cancer Depression/ anxiety Epilepsy Fetal growth restriction Gestational Diabetes Induced terminations of pregnancy - elective Maternal death Migraine Multiple gestation Multiple sclerosis	Counts of occurrence of component per event (codes to be proposed by the Definitions Task Force with mapping by Codemapper[10]) by provenance	Yes

[10] Becker, Benedikt FH, et al. "CodeMapper: semiautomatic coding of case definitions. A contribution from the ADVANCE project." Pharmacoepidemiology and drug safety 26.8 (2017): 998-1005.

	Pain Pre-eclampsia Rheumatoid arthritis Spontaneous abortions Stillbirth Systemic Lupus Erythematosus (SLE) Termination of Pregnancy for Fetal Anomaly (TOPFA) Attention Deficit Hyperactivity Disorder (ADHD) Autism spectrum disorder Low birth weight Major congenital anomalies Microcephaly Neonatal death Preterm birth		
	Incidence rates of first occurrences of all events listed above	Numerators are incident cases of the event (no diagnosis code in prior year), denominator person-time, both in children and women of childbearing age	Yes
	Incidence rates of first occurrences of the events listed above by year of age	Numerator are incident cases of the event (no diagnosis code in prior year), denominator person-time, both in children and women of childbearing age	Yes
	Healthcare seeking behaviour, Healthcare contacts	Incidence of records in diagnosis or measurement table with a unique date, denominator person-time, both in children and women of childbearing age	Yes
	Availability of folic acid use in women of childbearing age	Percentage of women of childbearing age with at least one record of folic acid use.	No
	Availability of folic acid use in women who experience a pregnancy event	Percentage of subjects with at least one pregnancy-related code who also have at least one record of folic acid use.	No
	Availability of smoking status in women of childbearing age	Percentage of women of childbearing age with at least one recorded smoking status.	Yes
	Availability of smoking status in women who experience a pregnancy event	Percentage of subjects with at least one pregnancy-related code who also have at least one recorded smoking status.	Yes
	Availability of alcohol use status in women of childbearing age	Percentage of women of childbearing age with at least one recorded alcohol use status.	Yes
	Availability of alcohol use status in women who experience a pregnancy event	Percentage of subjects with at least one pregnancy-related code who also have at least one recorded alcohol use status.	Yes
	Availability of education level in women of childbearing age	Percentage of women of childbearing age with at least one recorded education level.	Yes
	Availability of education level in women who experience a pregnancy event	Percentage of subjects with at least one pregnancy-related code who also have at least one recorded education level.	Yes
	Availability of breastfeeding status in women who experience a pregnancy event	Percentage of subjects with at least one pregnancy-related code who also have	Yes

		at least one recorded breastfeeding status.	
	Availability of breastfeeding exclusivity in women who experience a pregnancy event	Percentage of subjects with at least one pregnancy-related code who also have at least one recorded breastfeeding exclusivity status.	No
	Availability of Last Menstrual Period (LMP) in women who experience a pregnancy event	Percentage of subjects with at least one pregnancy-related code who also have at least one recorded Last Menstrual Period date	Yes
	Availability of BMI in women of childbearing age	Percentage of women of childbearing age with at least one recorded BMI, or weight and height measured on the same day.	Yes
	Availability of BMI in women who experience a pregnancy event	Percentage of subjects with at least one pregnancy-related code who also have at least one recorded BMI, or weight and height measured on the same day..	Yes
	Availability of BMI in children	Percentage of children with at least one record of BMI, or weight and height measured on the same day.	Yes
	Availability of SES in women of childbearing age	Percentage of women of childbearing age with at least one recorded SES.	Yes
	Availability of SES in women who experience a pregnancy event	Percentage of subjects with at least one pregnancy-related code who also have at least one recorded SES.	Yes
	Availability of SES in children	Percentage of children with at least one record of SES.	Yes

[a] source population- all subjects present in the PERSONS D2 CDM table as provided by the DAP

[b] study population- source population after application of time exclusion criteria (i.e., start of follow up set to 365 days before start of study if earlier and end of follow up set to whichever comes first between end of study date, date of death, date of creation of instance or recommend end date retrieved from the D2 CDM SOURCE table or date of reaching 55 years old)

[c] original start of follow up- the original date of start of follow up as retrieved by the D2 OBSERVATION PERIODS table

[d] start of follow up- the original date of start of follow up after application of time exclusion criteria (i.e., start of follow up set to 365 days before start of study if earlier)

[e] medicines study population- the study population after merging with the D2 MEDICINES table and exclusion of empty records

Benchmarking

Level 3 data checks can also be understood as one stage of "benchmarking", which is typically defined as the evaluation of a process or product against a standard. In the context of this protocol, benchmarking is the process of comparing data in ConcePTION CDM format held by a DAP to external independent data, published studies, other data sources participating in the benchmarking process, and to itself over time.

Comparing data held by DAPs to external data sources: these external data sources may include WHO Global Health Observatory[11] data on metrics such as neonatal, under-five, maternal, and adult mortality rates; adolescent and overall birth rates, stillbirth rates, and population-based incidence of depression.

[11] https://www.who.int/gho/en/

We may also compare age and sex distributions in the populations held by DAPs to publicly available census data held by Eurostat[12] to ascertain representativeness.

WITHIN
Comparing data to itself over time: all metrics listed in the tables above will be calculated by year and plotted vs. calendar time for visual assessment. These comparisons will elucidate population shifts; differences in prescription, diagnostic, and coding patterns over time; availability of covariates over time, and may reveal errors in the recording of data if these differences cannot be explained by known changes in the underlying population or database practices.

INTERNAL
Comparing data to other data sources participating in characterization: all metrics characterizing diagnoses and prescriptions or dispensations will be calculated by year and compared across DAPs. These comparisons will elucidate differences in prescription, diagnostic, and coding patterns across DAPs and may reveal errors in the recording of medicinal products and events if these differences cannot be explained by exogenous factors such as known policy differences.

EXTERNAL
Comparing estimates to those from published data: in the process of populating the ConcePTION catalogue and defining events of interest according to the process outlined in appendix 4, published studies related to events of interest (e.g., in EuroPeristat) as well as published studies conducted within each data source will be collated. For each event included in the data source characterization exercise, published rates will be reported along with rates calculated in the characterization exercise. Descriptions of the populations and time periods under study in each publication will be reported to allow interpretation of differences.

COMPONENT STRATEGY
According to the input of the Definitions Task Force, we will identify the events from Level 3 (Gestational age, mother-child linkage, Breast Cancer, Depression/anxiety,…) according to different algorithms, and will benchmark within and across databases the impact of changing algorithm. This will be used to obtain evidence on the best possible algorithm per database, and to quantify validity, according to the *component strategy* (ref Gini R, Dodd CN, Bollaerts K, Bartolini C, Roberto G, Huerta-Alvarez C, et al. Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project. Vaccine. In press. 10.1016/j.vaccine.2019.07.045)

8. Data management

ConcePTION will work according to a distributed (I.e., federated) network approach, with a common protocol for data characterization, a common data model and common analytics. That means that the data remain local, and the analytics is sent to DAPs and not the other way around.
Requests for data characterization, instructions and scripts will be sent and managed through the ConcePTION Task management system, via which a note will arrive in the e-mail that there is a task to be done. Task management will be conducted by UMC Utrecht and ARS.

[12] https://ec.europa.eu/eurostat/data/database

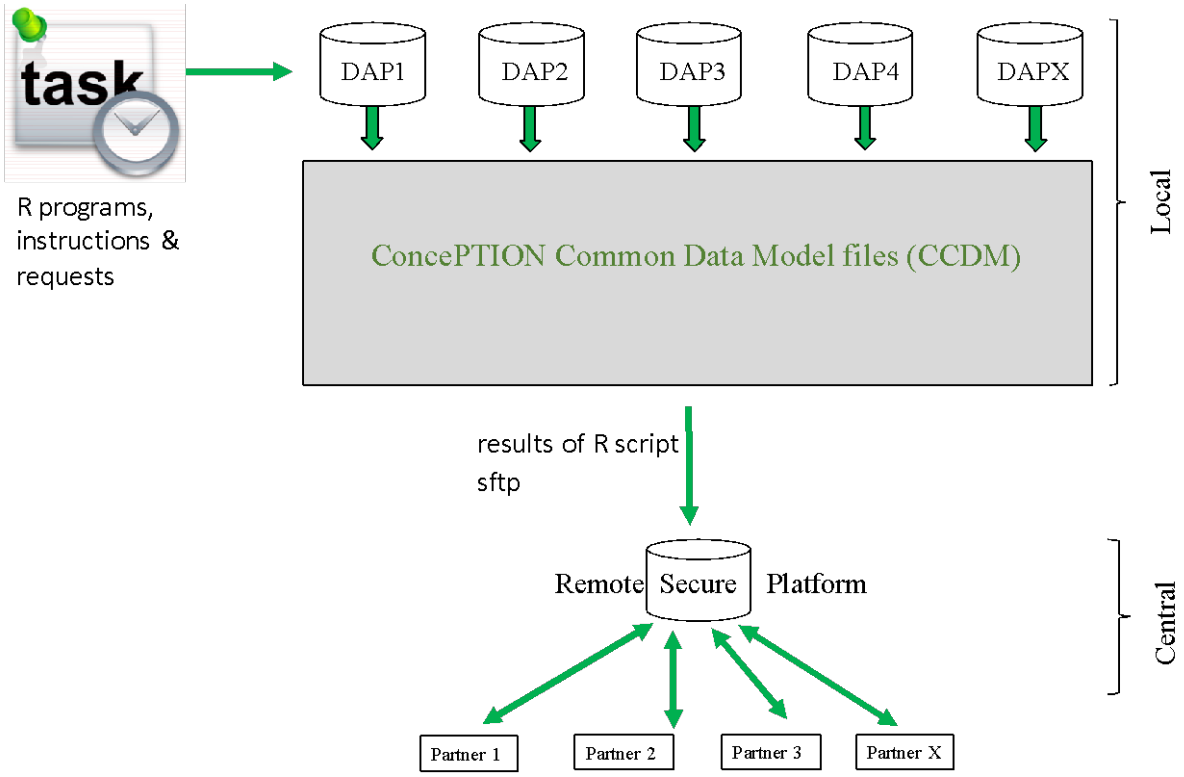

Figure 3: Data flow in ConcePTION

The DAP will perform the following tasks

1. The DAP will extract and transform the data locally into the required format
2. The DAP will run the R script which is coded and verified by the group of statisticians and distributed to the DAP by the coordinating center
3. The coordinating center will provide a detailed statistical analysis plan for the DAP, with full transparency on the data that will be shared and the analytical steps
4. The R-script will output graphics and summary tables for easy review by the DAP prior to submitting results
5. Once happy the DAP will submit the results of the R-script (anonymized) to the platform for further analysis and pooling
6. The DAP and Study team will review the quality indicators and discuss with the DAP about modalities to improve the data
7. If needed, steps 1-6 can be repeated until the DAP is satisfied
8. The DAP sends a results release form
9. Selected graphics/results will be put on the catalogue for inspection by the consortium members, the DAP will choose the level of sharing

The stepwise and potentially iterative process is graphically displayed in figure 4.

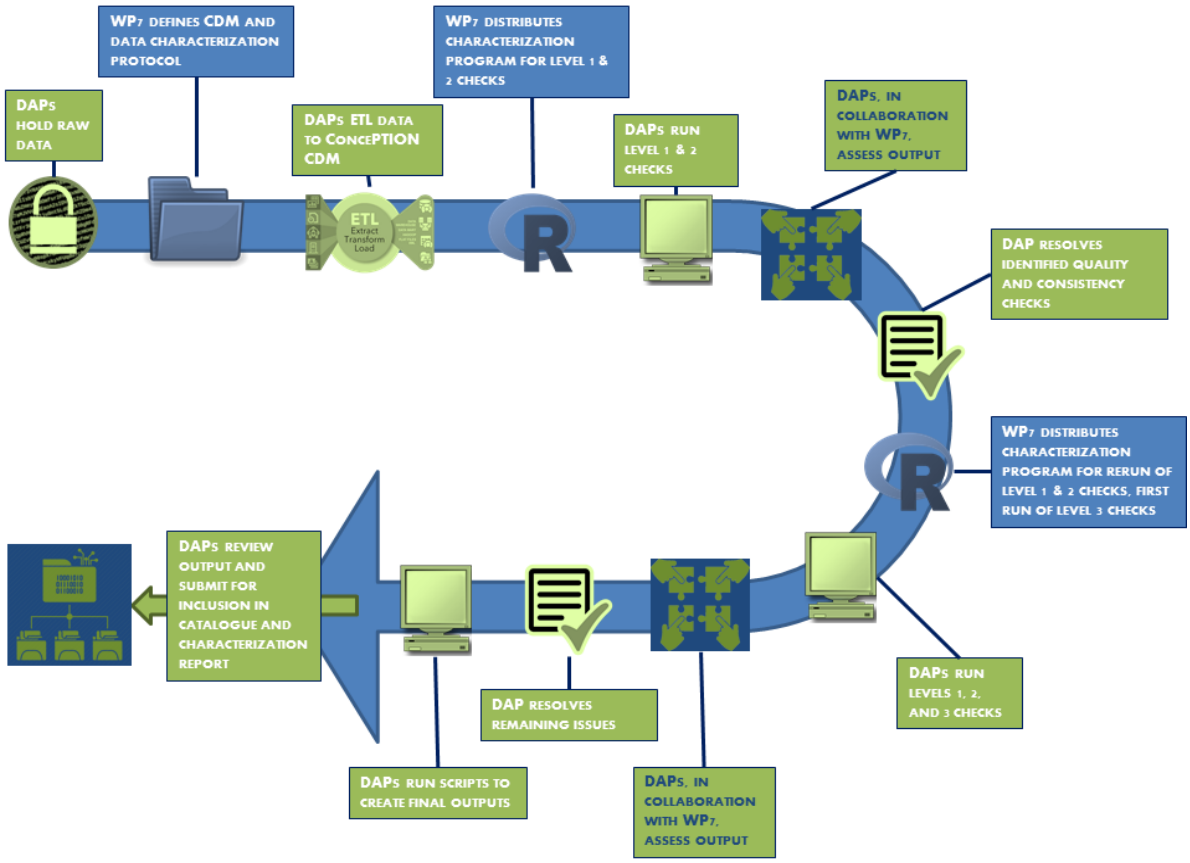

Figure 4. ConcePTION data characterization process

9. Quality Control

7.1 Record Retention

Documents that individually and collectively permit evaluation of the conduct the quality of the data produced will be retained for a period of 5 years in accordance with Good Participatory Practice (GPP) guidelines. These documents should be retained for a longer period, however, if required by the applicable regulatory requirements or by an agreement between study partners. It is the responsibility of the principal investigator to inform the other investigators/institutions as to when these documents no longer need to be retained.

Study records or documents may also include the analysis files, syntaxes (usually stored at the site of the database), and questionnaires.

7.2 Limitations of the Methods

This protocol addresses data characterization and a structured approach to assess the quality of different data sources, prior to addressing certain hypothesis evaluation studies. Data quality and fitness for purpose are imprecise concepts, and depend on the use cases. We have defined the use cases and will assess with the DAPs and study teams what will be appropriate indicators to assess fitness for purpose. There are no pre-defined thresholds nor benchmarks.

7.3 Advisory Committee

The ConcePTION Strategic Advisory Board.

10. PROTECTION OF HUMAN SUBJECTS

8.1 Regulatory and Ethical Compliance

This study is non-interventional, based on secondary use of data. Therefore, the reporting of suspected adverse reactions in the form of individual case safety reports (ICSRs) is not required. Reports of adverse events/reactions should be summarized as part of any interim analysis and in the final study report unless the protocol provides differently.
This data characterization is not considered as a PASS because the aim is not of "identifying, characterizing or quantifying a safety hazard, confirming the safety profile of the medicinal product, nor of measuring the effectiveness of risk management measures."

While the data characterization is being conducted, the marketing authority holder (MAH) shall monitor the results generated and consider its implications for the risk-benefit balance of the medicinal product concerned. Any new information which might influence the evaluation of this risk-benefit balance shall be communicated to the competent authorities of Member States in which the medicinal product has been authorized. The channel for communicating this information is the notification of an Emerging Safety Issue.

This study is compliant with the provisions of the ENCePP Code of Conduct, Revision 4.

8.2 Informed Consent

Data bases with an Institutional Review Board (IRB) approval indicating that informed consent is waived and the rational for this decision will be maintained.

8.3 Responsibilities of the Investigator and IRB/IEC/REB

The protocol and waiver of informed consent must be reviewed and approved by a properly constituted institutional review board/independent ethics committee/research ethics board (IRB/IEC/REB) before study start. A signed and dated statement that the protocol has been approved by the IRB/IEC/REB and waiver of informed consent must be given to the principal investigator before study initiation.

11. MANAGEMENT AND REPORTING OF ADVERSE EVENTS/ ADVERSE REACTIONS

N/A

9.1 PLANS FOR DISSEMINATING AND COMMUNICATING RESULTS

Registration in Public Database(s)
Principal investigator assures that the key design elements of this protocol will be posted in the EU Post-Authorisation Studies (PAS) database in compliance with current regulations.
Principal investigator also assures that key results of this study will be posted in the EU PAS database within the required time-frame from completion of the data collection where applicable and in compliance with current regulations.

Publications
Further to legislated data disclosure, the results of this study may be published as scientific papers in peer-reviewed journals. Preparation of such manuscripts will be prepared independently by the investigators and in accordance with the current guidelines of STrengthening the Reporting of OBservational studies in Epidemiology (STROBE). The ConcePTION Steering Committee will be entitled to

view the results and interpretations included in the manuscript and provide comments prior to submission of the manuscript for publication.

9.2 Inclusion of results in the ConcePTION catalogue

Results of the quality indicators may be displayed on the ConcePTION catalogue dashboard after approval of the DAP. For all indicators and characterization output resulting in a cell count less than 5, counts will not be reported and will be replaced with "<5" programmatically.

9.3 Use of data in demonstration projects

The data extracted for the data characterization & validation of algorithms may be re-used in the demonstration projects, for which separate protocols will be written.

Annex 1: Quality definitions from Johnson et al.

Concept	Definition
RepresentationIntegrity	Aspects of the Representation that reassure that data was not corrupted or subject to data entry errors.
RelativeCorrectness	Assesses the quality of a Representation by comparing it to its counterpart in another Dataset which is a “relative standard”, computed as PPV.
RepresentationCorrectness	A correct Representation has high accuracy and is complete.
Reliability	The data is correct and suitable for the Task.
RepresentationConsistency	The data is a valid value and format for its Data Value Type and all of the Representations for the same information have the same values.
DomainConsistency	Concepts in the Domain are represented in the data and the data satisfies syntactic and semantic rules. Constraints for the Domain are satisfied.
CodingConsistency	Representations that are of coded text data type must be correctly mapped to an enumerated list or a terminology.
DomainMetadata	Meta-data exists to describe the Domain and it is logically consistent.
RepresentationComplete	Domain independent extent to which data is not missing.
DomainComplete	The extent to which information is present or absent as expected.
RelativeCompleteness	The extent to which a truth about the world is represented in the data. This is computed as sensitivity relative to another Dataset.
Sufficiency	The data has sufficient Representations along a given dimension (i.e. time, patient, encounter) to perform the Task.
DomainCoverage	The data can represent the values and concepts required by the Domain.
TaskCoverage	The data contains all of the information required by the Task.
Flexibility	The extent to which the data is sufficient to be used by many Tasks.
Relevance	The data is sufficient for the Task and conforms to the Domain.
RepresentationCurrent	Calculation for time difference between when an observation was made and when it was entered into the system.
DatasetCurrent	Time difference between when a Dataset was updated and when it was made available. For example, periodic updates to a repository.
TaskCurrency	The Data is sufficiently up-to-date for the requirements of the Task.

Annex 2: EUROCAT CDM

EUROCAT table (see https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/Full_Guide_1_4_version_28_DEC2018.pdf.)

MPersonID	Character	Patient identification code for mother, if also in EHR databases
CPersonID	Character	Patient identification code for child if in EHR database
Centre	Character	CentreNumber
NUMLOC	Character	localID
BIRTHDATE	Character	Date of Birth ddmmyyyy
SEX	Numeric	Chromosomal sex 1 = Male 2 = Female 3 = Indeterminate 9 = Not known
NBRBABY	Number	Number of babies/fetuses delivered 1 = Singleton 2 = Twins 3 = Triplets 4 = Quadruplets 5 = Quintuplets 6 = Sextuplets or more 7 = Multiple birth, number of babies not known 8 = Singleton at time of delivery/termination, but known to have been a multiple pregnancy at an earlier stage in pregnancy 9 = Not known
SP_TWIN	string	Specify twin type of birth, like or unlike, zygosity (Free text)
NBRMALF	number	Number of malformed in multiple set 1 = One 2 = Two 3 = Three 4 = Four 5 = Five 6 = Six or more 9 = Not known
Type	Character	Type of Birth 1=livebirth 2-stillbirth 3=spontaneous abortion 4=TOPFA 9-not known
CIVREG	Character	Civil registration status 1=livebirth 2-stillbirth 3=no civil registration 9-not known
weight	number	Birthweight in grams 9999=not known
gestlength	number	Length of gestation in completed weeks 99=not known

survival	number	Survival beyond one week of age 1=yes 2=no 3=alive at discharge <1 week 9=not known
Death_date	Character	Date of death Ddmmyy 99=died unknown date 222222= known to be alive at 1 year 333333=not known if alive or dead at 1 year
datemo	Character	Date of birth of mother Ddmmyy 99 = Not known day or month 44 = Not known year
Agemo	Character	Age of the mother at delivery in years 99=not known
BMI	Numeric	2 digits 97=exact BMI NK but <30 98=exact BMI NK but>=30 99=not known
RESIDMO	Character	Mother's residence code Local code
TOTPREG	Character	Total number of previous pregnancies 00 = None 01 = One 02 = Two 03 = Three etc 20 = Twenty or more 99 = Not known
WHENDISC	Character	When was anomaly discovered 1 = At birth 2 = Less than 1 week 3 = 1-4 weeks 4 = 1-12 months 5 = Over 12 months 6 = Prenatal diagnosis in live fetus 7 = At abortion (spontaneous) 9 = Not known 10 = Postnatal diagnosis, age not known
CONDISC	Character	Condition at discovery 1 = Alive 2 = Dead 9 = Not known
AGEDISC	Character	If prenatally diagnosed gestational age at discovery in completed weeks
FIRSTPRE	Character	First positive prenatal test 1 = Ultrasound at gestational age (GA) < 14 weeks 2 = Ultrasound at GA 14-21 weeks 3 = Ultrasound at GA ≥ 22 weeks 4 = Ultrasound GA not known 5 = Serum/ combined screening

		6 = CVS or amniocentesis 7 = Other test positive 8 = Test(s) performed, result negative 9 = Not known 10 = No test performed 11 =Fetal karyotype on maternal blood
SP_FIRSTPRE	Character	Free text "other'' first prenatal test
KARYO	Character	Karyotype of infant/fetus 1 = Performed, result known 2 = Performed, results not known 3 = Not performed 4 = Probe test performed 8 = Failed 9 = Not known
SP_KARYO	Character	Free text of specific karyotype
GENTEST	Character	Genetic test 1 = specific genetic test positive 2 = specific genetic test negative 3 = Specific genetic test not performed 9 = Not Known if genetic test is performed or result not known
SP_GENTEST	Character	Free text of specific type of genetic test
PM	Character	Post mortem examination 1 = Performed, results known 2 = Performed, results not known 3 = Not performed 4 = Macerated fetus 9 = Not known
SURGERY	Character	First surgical procedure for malformation 1 = Performed (or expected) in the first year of life 2 = Performed (or expected) after the first year of life 3 = Prenatal surgery 4 = No surgery required 5= Too severe for surgery 6 = Died before surgery 9 = Not known
SYNDROME	Character	Syndrome or association ICD 10 First 4 digits are ICD10 5th digit = BPA supplement or leave blank
SP-SYNDROME	Character	Written description of the specific syndrome
MALFO1	Character	Malformation ICD 10 First 4 digits are ICD10 5th digit = BPA supplement or leave blank
SP_MALFO1	Character	Specify malformation free text
MALFO2	Character	Malformation ICD 10 First 4 digits are ICD10 5th digit = BPA supplement or leave blank
SP_MALFO2	Character	Specify malformation free text

MALFO3	Character	Malformation ICD 10 First 4 digits are ICD10 5th digit = BPA supplement or leave blank
SP_MALFO3	Character	Specify malformation free text
MALFO4	Character	Malformation ICD 10 First 4 digits are ICD10 5th digit = BPA supplement or leave blank
SP_MALFO4	Character	Specify malformation free text
MALFO5	Character	Malformation ICD 10 First 4 digits are ICD10 5th digit = BPA supplement or leave blank
SP_MALFO5	Character	Specify malformation free text
MALFO6	Character	Malformation ICD 10 First 4 digits are ICD10 5th digit = BPA supplement or leave blank
SP_MALFO6	Character	Specify malformation free text
MALFO7	Character	Malformation ICD 10 First 4 digits are ICD10 5th digit = BPA supplement or leave blank
SP_MALFO7	Character	Specify malformation free text
MALFO8	Character	Malformation ICD 10 First 4 digits are ICD10 5th digit = BPA supplement or leave blank
SP_MALFO8	Character	Specify malformation free text
PRESYN	Character	Prenatal diagnosis for syndrome 1 = Yes, this anomaly was diagnosed prenatally 2 = No, this anomaly was diagnosed postnatally 3 = This anomaly partially prenatally diagnosed 9 =Not known
PREMAL1	Character	Prenatal diagnosis for malformation as PRESYN
PREMAL2	Character	Prenatal diagnosis for malformation as PRESYN
PREMAL3	Character	Prenatal diagnosis for malformation as PRESYN
PREMAL4	Character	Prenatal diagnosis for malformation as PRESYN
PREMAL5	Character	Prenatal diagnosis for malformation as PRESYN
PREMAL6	Character	Prenatal diagnosis for malformation as PRESYN
PREMAL7	Character	Prenatal diagnosis for malformation as PRESYN
PREMAL8	Character	Prenatal diagnosis for malformation as PRESYN
OMIM	Character	Type of mendelian inheritance

		Full codes can be found on the OMIM website http://www.ncbi.nlm.nih.gov/omim/
ORPHA	Character	Rare disease code https://www.orpha.net/consor/cgi-bin/Disease.php?lng=EN
ASSCONCEPT	Character	Assisted conception 0 = No 1 = Induced ovulation only 2 = Artificial insemination 3 = IVF 4 = GIFT 5 = ICSI 6 = Egg donation 8 = Other 9 = Not known 10 = Assisted conception, type unknown
OCCUPMO	Character	Mother's occupation at time of conception **4 digit** code ISCO 9999 = Not known (do NOT use 9, 99 or 999 for not known) Links for ISCO classifications: http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm EUROCAT Supplement: 9991 = Employed (including self-employed), but occupation unknown 9995 = Housewife 9996 = Student 9997 = Unemployed 9999 = Not known whether employed or not
ILLBEF1	Character	Illness before pregnancy ICD 10 0 = No illness 1 = Yes, but no information available 9 = Not known
ILLBEF2	Character	As above
MATDIAB	Character	Maternal pregestational diabetes 1= Yes, type 1 diabetes (IDDM) 2= Yes, type 2 diabetes (NIDDM) 3 = Yes, type MODY* (all types) 4 = Yes, type not known 5 = No, but impaired glucose intolerance 6 = No pregestational diabetes 9 = Not known
HbA1c	Character	Glycated hemoglobin value in mmol/mol units 999 = Not known 3 digits
ILLDUR1	Character	Illness during pregnancy ICD 10 0 = No 1 = Yes, but no information available

		9 = Not known
ILLDUR2	Character	As for ILLDUR1
FOLIC_G14	Character	Folic acid supplementation 1 = Folic acid taken pre and post-conceptionally 2 = Folic acid taken only post-conceptionally 3 = Folic acid not taken 4 = Folic acid taken, timing unknown 9 = Not known if folic acid taken
FIRSTTRI	Character	First trimester medication 1 = Yes, medication taken in first trimester 2 = No medication taken in first trimester 3 = Undetermined 4 = Medication taken, but timing unknown 9 = Not Known
DRUGS1	Character	Drug used in first trimester ATC code http://www.whocc.no/atcddd/.
SPDRUGS1	Character	Specify drug exposure (free text)
DRUGS2	Character	Drug used in first trimester ATC code http://www.whocc.no/atcddd/.
SPDRUGS2	Character	Specify drug exposure (free text)
DRUGS3	Character	Drug used in first trimester ATC code http://www.whocc.no/atcddd/.
SPDRUGS3	Character	Specify drug exposure (free text)
DRUGS4	Character	Drug used in first trimester ATC code http://www.whocc.no/atcddd/.
SPDRUGS4	Character	Specify drug exposure (free text)
DRUGS5	Character	Drug used in first trimester ATC code http://www.whocc.no/atcddd/.
SPDRUGS5	Character	Specify drug exposure (free text)
EXTRA-DRUGS	Character	Extra drugs if more than five above
CONSANG	Character	Consanguinity 0 = Not related or relationship more distant than second cousin 1 = Relationship of second cousin or closer 9 = Not known
SP_CONSANG	Character	Specific information on consanguinity (free text)
SIBANOM	Character	Sibs with anomalies 1 = Same 2 = Other 3 = Same and other 4 = No 9 = Not known
SP_SIBANOM	Character	Specific type of anomaly (free text)
PREVSIB	Character	Previous malformed siblings notified to EUROCAT 1 = Yes 2 = No 9 = Not known

SIB1	Character	SIB local ID number notified to the Central Registry
SIB2	Character	SIB local ID number notified to the Central Registry
SIB3	Character	SIB local ID number notified to the Central Registry
MOANOM	Character	Mother's family with anomalies 1 = Same 2 = Other 3 = Same and other 4 = No 9 = Not known
SP_MOANOM	Character	Specify type of anomaly (free text)
FAANOM	Character	Father's family with anomalies (as in MOANOM)
SP_FAANOM	Character	Specify type of anomaly (free text)
MATEDU	Character	Maternal education 1 = Elementary and lower secondary 2 = Upper secondary 3 = Tertiary 9 = Not known
SOCM	Character	Socioeconomic status of mother 1 = Upper non-manual 2 = Lower non-manual 3 = Skilled manual 4 = Unskilled manual 5 – Self employed/artisan 6 = Farmer 8 = Other/Student 9 = Not known
SOCF	Character	Socioeconomic status of father 0 = Single mother, no father recorded 1 = Upper non-manual 2 = Lower non-manual 3 = Skilled manual 4 = Unskilled manual 5 – Self-employed/artisan 6 = Farmer 8 = Other/Student 9 = Not known
MIGRANT	Character	Migrant status 1 = Mother migrated from outside EU during pregnancy 2 = Mother migrated from outside EU during adult life (from age 18) 3 = Mother not a migrant as defined in 1 or 2 4 = Other (specify in text) 9 = Not known

https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/Full_Guide_1_4_version_28_DEC2018.pdf

Annex 3: Template for clinical definition

Event:
Outcome/covariate:
Partner acronym:
Version:
Status:

Contributing authors & expertise

author	Expertise

Modifications table:

Name	Action

1. Event definition

(existing learned societies/literature: e.g. Sentinel, OMOP health outcomes of interest, EUROCAT, GAIA definitions, PERISTAT…). List multiple if they exist and provide references

2. Synonyms / lay terms used
(please search/google/used medical textbooks & literature)

3. Laboratory tests done specific for event
(please search/google/used medical textbooks & literature)

4. Diagnostic tests done specific for event
(please search/google/used medical textbooks & literature) (e.g. echo, X-ray)

5. Medicinal products used to treat event
(please search/google/used medical textbooks & literature): these may be proxies to identify/confirm the event

6. Procedures used specific for event treatment
(please search/google/used medical textbooks & literature) (example surgery) these may be proxies to identify/confirm the event

7. Setting (outpatient specialist, in-hospital, GP, emergency room) where condition will be most frequently /reliably diagnosed

8. Diagnosis codes or algorithms used in different papers to extract the events in Europe/USA
ICD-9/CM
ICD-10
ICD-10/CM
READ
ICPC
SNOMED
MEDDRA
Laboratory

9. Experience of participating data sources to extract the events prior to ConcePTION (to be completed by each data source, if no experience please state NA)

Data source	Codes used	Medicinal products used	Procedures	Tests	Chart Validation conducted ?	References studies, occurrence of the condition in the database population and validations (number)
The Norwegian Prescription Database (NorPD)						

linked to the Medical Birth Registry of Norway (MBRN)						
Lillebaelt: Pseudonymised EUROCAT data from Funen						
Uni. Copenhagen The Danish National Prescription Registry, The Danish Medical Birth registry and The national Patient registry						
University Aarhus: Population based linkage with The Danish National Prescription Registry, The Danish Medical Birth registry and The national Patient registry						
University of Dundee, Scottish data						
Uni. Bath CPRD						
Uni Ulster: EUROmediCAT central						
USWAN: SAIL and CARIS						
Uni Ulster ADRC-NI						
Uni. Toulouse, EFEMERIS						
Uni. Toulouse, POMME						
Uni. Toulouse, SNDS						
Uni. Toulouse, SNIIRAM						
Uni. Bordeaux: SNDS						
Uni Groningen NL EUROCAT						
Uni Groningen NL IADB						
LAReb: pregnant						
PHARMO						
Uni Mainz: CT						
BIPS: GePARD						
Uni Poznan:PRCM						
Croatia: Eurocat						
Congenital Anomalies						

population-based registry of the Valencia Region, Foundation for the Promotion of Health and Biomedical Research of the Valencian Region (FISABIO), Valencia, Spain						
The Information System for Research in Primary Care (SIDIAP), Catalonia, Spain						
Uni Ferrara: IMER Congenital Anomalies population-based registry of the Emilia Romagna region						
ARS						
CNR-IFC Tuscany Congenital Anomalies population-based registry of the Tuscany Region						
Uni Messina: Caserta						
Uni Messina: Sicily						
Malta: Eurocat						
Malformation Monitoring Centre Saxony-Anhalt, Medical Faculty Otto-von-Guericke University Magdeburg, Germany						

10. Proposed codes by Codemapper

11. Extracted codes (upon characterization)

12. Validation studies for codes

12. References

Annex 4. EUROmediCAT Data Quality Indicators for EUROCAT table

Source: https://eu-rd-platform.jrc.ec.europa.eu/eurocat/data-collection/data-quality
Reference: Loane, Maria, et al. "Paper 3: EUROCAT data quality indicators for population-based registries of congenital anomalies." Birth Defects Research Part A: Clinical and Molecular Teratology 91.S1 (2011): S23-S30.

Ascertainment
Ratio of Spina bifida to Anencephalus.
Prevalence NTD = Neural Tube Defects
Prevalence selected cardiac malformations
Hypoplastic left heart, Transposition of Great Vessels, Tetralogy of Fallot, Coarctation of aorta or Common arterial truncus.
Prevalence selected postnatal diagnosed malformations
Corpus callosum anomalies, Cataract, Coarctation of aorta, Hirschprung's disease, Unilateral renal agenesis, or Craniosynostosis.
Prevalence non-chromosomal syndromes
Prevalence malformed fetal deaths calculated using total births.
Total number of cases
Total major congenital anomaly prevalence – >200 per 10,000 births expected
Prevalence of subgroup “anencephalus” – compare with the EUROCAT average
Prevalence of subgroup “severe cardiac defects” – compare to the EUROCAT average Prevalence of selected congenital anomalies often diagnosed after the neonatal period – compare to the EUROCAT average
Includes codes for corpus callosum anomalies (Q040), cataract (Q120), coarctation of aorta (Q251), Hirschprung (Q431) and craniosynostosis (Q750).
Prevalence of subgroup “Genetic syndromes and microdeletions” – compare to the EUROCAT average
Prevalence of malformed fetal deaths – compare to the EUROCAT average
Down syndrome: Observed/Expected ratio by maternal age – compare to the EUROCAT average

Accuracy of Diagnosis
% of possible multiple malformations in database excluding chromosomal or syndrome cases using the EUROCAT flow-chart for multiple malformations
% fetal deaths and terminations of pregnancy with post-mortem examination performed
% of chromosomal cases with a karyotype performed
% of non-chromosomal / non-syndrome multiple malformation cases in database with known karyotype Using the EUROCAT flow-chart for multiple malformations
Prevalence of selected Q-BPA extension codes (restricted to registries that use ICD10 coding).
Prevalence of selected Q-chapter unspecified codes (restricted to registries that use ICD10 coding).
% fetal deaths with postmortem examination carried out – compare to the EUROCAT average
% TOPFA (GA ≥ 15 weeks) with post-mortem carried out – compare to the EUROCAT average
% chromosomal cases (except trisomy 13, 18 and 21) with karyotype text – compare to the EUROCAT average
% potential multiple anomalies with known karyotype – compare to the EUROCAT average
Prevalence of selected 4-digit Q-BPA codes – compare to the EUROCAT average
% livebirths with ASD, VSD, hydronephrosis, hypospadia or club foot with known data on surgery – compare to the EUROCAT average

Completeness of information in EUROCAT tables

Completeness of information describes the amount of complete valid data transmitted to Central Registry (eg. “Not known” values, invalid values, or missing/blank fields are counted as incomplete information).

Number of core variables 90% complete - compare to total number of core variables
11 variables: Sex, Nbrbaby, Nbrmalf, Type, Weight, Gestlength, Survival,Whendisc, Agedisc, Agemo, Civreg

Number of non-core variables 80% complete – compare to total number of non-core variables
26 variables: Death_date, Condisc, Karyo, PM, Datemo, Residmo, Totpreg, Occupmo, Assconcept, Illbef1, Illdur1, Drugs1, Consang, Prevsib, Sib1, Sibanom, Moanom, Faanom, Firstpre, Surgery, Folic, Matedu, Socm, Socf, Migrant, Aetiology

% TOPFA with civil registration known – compare to the EUROCAT average
% livebirths with one week survival known – compare to the EUROCAT average Medication exposure recorded using 7 digit ATC codes – yes or no
% of ATC codes with 7 digits and in correct format – compare to the EUROCAT average
% genetic syndromes + microdeletions with syndrome text complete – compare to the EUROCAT average
% malformation1 text complete – compare to the EUROCAT average
Number of unresolved data edits (excluding free text fields) – compare to the EUROCAT average
(If Central Registry changes information in the central database, CR staff will request that the change be made in the local registry data also. If the change is not made and data are re-submitted to CR, this will be flagged as an ‘unresolved data edit’. This is a new feature of the central database.)