

## IMI2 821520 - ConcePTION

### ConcePTION

#### WP7 – Information and data governance, ethics, technology and data catalogue and quality support

## D7.5 Report on existing common data models and proposals for ConcePTION

<b>Lead contributor</b>	Caitlin Dodd (UMCU)
	Rosa Gini (ARS)
<b>Other contributors</b>	Miriam Sturkenboom (UMCU) Vjola Hoxhaj, Marieke Hollestelle (UMCU)
	Nicolas Thurin (BPE)
	Claudia Bartolini, Olga Paoletti, Giuseppe Roberto (ARS)
	Marianne Cunningham, Betsy Georgiou (GSK)
	Hedvig Nordeng (UOSL)
	Maria Loane (UUIster)
	Maarit Leinonen (NIHW)
	Vera Ehrenstein (AUH)
	Christine Damase-Michel, Anna-Belle Beau (CHUT)
	JEH van Kammen (UMCG)
	Karin Swart, Eline Houben (PHARMO)
	Talita Duarte (IDIAP)
	Ulrike Haug, Tania Schink (BIPS)
	Clara Caverio (FISABIO)
	Amanda Neville (UNIFE)
	Anna Pierini (CNR-IFC)

Janet Sultana (UNIME)
Miriam Gatt
Sue Jordan, Daniel Thayer, Ieuan Scnalon (USWAN)
Eugene van Puijenbroek (LAREB)
Luke Richardson (UMAN)

<b>Due date</b>	31-03-2020
<b>Delivery date</b>	04-08-2020
<b>Deliverable type</b>	Report
<b>Dissemination level</b>	PU

## Document History

Version	Date	Description	Non-contributor reviewers (if applicable)
V0.1	2 March 2020	First draft section 1 (introduction) and section 4 (interviews) (Rosa Gini, ARS)	
V0.2	20 March 2020	Updated CDM section with first results	
V0.3	08 April 2020	Updated overview of existing models and CDM v1.0 (Caitlin Dodd, UMCU)	
V0.4	21 April 2020	First draft section 7 (CDM work for WP2) (Eugene van Puijenbroek, LAREB)	
V0.5	12 May 2020	Inclusion of CDM v2.0 and 2.01 with list of vocabularies (Rosa Gini, ARS and Caitlin Dodd, UMCU)	
V0.7	28 May 2020	ETL section	
V0.9	30 June 2020	Wrapping up and sending for review (Miriam Sturkenboom)	Several persons in the consortium
V1.0	23 July 2020	Incorporate comments of partners (Rosa Gini) & lay-out (Miriam Sturkenboom)	

## Summary

In this deliverable we describe the creation of the ConcePTION common data model (CDM) for secondary re-use of electronic health care and surveillance data which will be used for demonstration studies in WP1. We also describe current approaches and status for WP2.

ConcePTION partners have conducted many multi-site studies across Europe and globally, using the principle of a common protocol, and CDM for primary data collection and secondary use of health data (1-4). Strategies deployed differed. They included: strategies where everybody conducts the study locally using the common protocol (e.g. for EUROMediCAT, IMI-PROTECT), strategies where data are pooled in a central database and then analysed (e.g. in Nordic countries), or strategies where data are transformed in a common data model locally and then analysed in a distributed manner (e.g. in ADVANCE). Based on the learning from these projects we know that the following factors are key to success:

- 1) A common protocol, CDM and a distributed analysis are an efficient way of conducting multi-database studies, while complying with General Data Protection Regulation (GDPR) rules. The analysis goes to the data rather than the data to the analysis. Using a common script helps as not all sites have the analysts to code all the steps and even then, it is difficult to make it exactly the same.
- 2) The CDM structure (tables and variable names) should remain stable in order to re-use modules of the analytics.
- 3) Common analytics need to be written in a coding language that is widely understood and used, and must be open source so that every partner can (re-)use it without buying a license.
- 4) The provenance of the data needs to be recorded in the CDM and utilized in the analysis because of the large variety of data sources in Europe.
- 5) Full transparency is needed in mapping from local data to CDM, therefore the local Extraction transformation & Load (ETL) design and script should be made available as well as the local data dictionary
- 6) Semantic mapping and creation of study variables should be transparent and adapted to the data sources and to the study. Flexibility is needed to run multiple algorithms to create study variables from a central coding perspective, the local Data Access Provider (DAP) is important to help interpretation.
- 7) The 'instance' of a common data model (the specific version of data in the common data model) does not need to include all data from the data source, but at least the data required for the study protocol. This is especially relevant in countries where data access is limited to study specific datasets. Because the CDM is constant in structure (see point 2), the ETL design and script can be re-used quickly for other studies.
- 8) Quality checks (completeness, logics and benchmarking) are required to assess whether the ETL was successful and data are fit for purpose.
- 9) Pooling of results of local analyses should be done on a shared platform where people can work together, remotely (see D7.2).

In the ConcePTION project that started in 2019, a large ecosystem is built for medicines safety in pregnancy. As part of the ConcePTION project, ARS, UMCU, BPE, USWAN and

GSK have spent one year to further improve the evidence generation pipeline for different types of secondary use data sources. The following improvements were made based on the pipeline that was drafted in project proposal:

- We added tables to the CDM to be able to incorporate survey data, EUROCAT tables, mother-child linkage, provenance of records, and meta-data.
- We made the process more transparent and flexible: when populating the CDM, DAPs now only need to do the syntactic mapping, the semantic mapping is done after quality checks and with a centralized script based on the needs of the study team.
- We have improved the workflow for mapping data to the CDM which allows for quick running of multiple studies. Each DAP will :
  1. Supply: meta-data information, a data dictionary, and information on local data from a in-depth interview.
  2. DAPs design their ETL process in a standard template
  3. DAPs develop an ETL program based on the ETL design
  4. DAPs run standard R scripts to do level 1, 2, 3 quality checks (see section 3)
  5. DAPs review output and adapt ETL if needed

The first three steps only need to be taken once unless changes to data collection practise, recording or variable coding occur. Steps 4-5 need to be repeated for each new instance of the CDM (e.g. with update or new study). After the 5<sup>th</sup> step the DAP is ready to execute the study specific scripts. This is a generic process that will also be used for WP2 data in year 2.

- We have consistently adopted R as a central programming language against the CDM (R is open source and can be used by all DAPs), and streamlined the programming pipeline. Not only quality checks (step 4 above) are programmed and ready to be re-used, but also the structure of study scripts has been modularized and functions have been developed in order to be re-used to quickly compose new scripts.
- We operate an IT infrastructure that allows for efficient and transparent work across multiple network studies (See D7.2).

In this deliverable we describe the steps that were taken to create and improve the CDM and the methods for designing the ETL and study variables. A small description is provided on the development of the CDM for spontaneous reports and prospective monitoring.

# Contents

<b>Document History .....</b>	<b>2</b>
<b>Summary .....</b>	<b>3</b>
<b>1. Introduction: model &amp; terminology .....</b>	<b>7</b>
<b>2. Background: overview of existing common data models for secondary use of health care data .....</b>	<b>8</b>
<b>2.1 Types of Common Data Models .....</b>	<b>8</b>
<b>2.2 Existing Common Data Models for secondary re-use of data collected for other purposes.....</b>	<b>8</b>
2.2.1 Common Data Models for Spontaneous reporting systems databases .....	9
2.2.2 Vaccine Safety Datalink.....	10
2.2.3 United States FDA Sentinel .....	11
2.2.4 OMOP .....	11
2.2.5 PEDSnet .....	13
2.2.6 Other common data models.....	14
<b>3. ConcePTION CDM for health care data sources v1.0 .....</b>	<b>15</b>
<b>4. Processes to create ConcePTION CDMv2.01 .....</b>	<b>16</b>
<b>4.1 Indepth interviews with Data Access Providers .....</b>	<b>16</b>
4.1.1 Request to DAPs their data dictionary .....	16
4.1.2 Conduct one-to-one interviews. ....	16
<b>4.2 Finalize the CDM .....</b>	<b>18</b>
4.2.1 CDM recommendations following each interview.....	18
4.2.2 Comparison with existing common data models .....	19
4.2.3 Search for standard for medical birth registers .....	20
4.2.4 Update to ConcePTION CDM v2.0 .....	21
4.2.5 Update to ConcePTION CDM v2.01 .....	22
4.2.6 ConcePTION_CDM v2.01: tables.....	22
4.2.7 ConcePTION CDM v2.01: vocabulary .....	26
<b>5. Extract, Transform, and Load template and process for specification .....</b>	<b>28</b>
<b>5.1 ETL template .....</b>	<b>28</b>
<b>5.2 ETL process specification for the ConcePTION DAPs .....</b>	<b>30</b>
<b>5.3 ETL documents .....</b>	<b>30</b>
<b>6. Semantics superimposed on the ConcePTION CDM .....</b>	<b>31</b>
<b>6.1 Template of a Statistical Analysis Plan .....</b>	<b>31</b>
<b>6.2 Design study variables .....</b>	<b>31</b>
6.2.1 Study variables based on <i>Surveillance</i> or <i>Curated</i> tables .....	31
6.2.2 Study variables based on <i>Routine healthcare data</i> tables .....	33
<b>6.3 Modular programming of study variables .....</b>	<b>34</b>
<b>7. CDM work for WP2.....</b>	<b>34</b>
<b>7.1 CDM development.....</b>	<b>34</b>
<b>7.2 Multiple CDMs .....</b>	<b>34</b>
<b>7.3 CDE and CDM .....</b>	<b>34</b>
<b>7.4 Pilot study 1 - Conversion of data from different type of data sources .....</b>	<b>35</b>
<b>7.5 Pilot study 2 - Possibility to convert from SRS data.....</b>	<b>35</b>
<b>8. References .....</b>	<b>35</b>
<b>Appendix 1. Instructions for task “provide data dictionary” .....</b>	<b>37</b>
<b>Appendix 2. Details of the ConcePTION CDM v1.0 .....</b>	<b>39</b>

<b>Appendix 3. Details of ConcePTION CDM v2.0 .....</b>	<b>43</b>
<b>Appendix 4. Answer sheet for interview with DAP .....</b>	<b>45</b>
<b>Appendix 5. Results from analysis of interviews for the purpose of updating the CDM .....</b>	<b>49</b>
<b>Appendix 6. ETL template v1.0 .....</b>	<b>56</b>
<b>Appendix 7: Template statistical analysis plan based on CDM .....</b>	<b>89</b>
<b>8 References .....</b>	<b>100</b>

## 1. Introduction: model & terminology

ConcePTION aims to create an ecosystem for the rapid and robust generation of evidence on the safety of medications in pregnancy and lactation, using both existing and newly generated real-world data.

To describe and analyse existing data in population-based studies, ConcePTION is using “a structured and stable common data model that is filled with study-specific data”(2). This strategy implies that a Common Data Model (CDM) structure is specified, and has the potential to host all the data that may be relevant for ConcePTION studies. Data Access Providers (DAPs) create a procedure to Extract, Transform and Load (ETL) their data to the CDM; this is known as a data instance. No data is permanently stored in the CDM by DAPs. On the contrary, whenever a study protocol is approved, each DAP extracts or obtains from its partner organizations the data needed for that study, and transforms and loads it to the CDM using the ETL procedure developed beforehand. The central group responsible for the study then creates a R procedure that runs against the CDM and shares it with the DAPs. Each DAP runs the procedure and shares its output with the central group, via a secure Platform (see Deliverable 7.2).

In ConcePTION, a CDM is therefore intended as a tool to enable transparent and reproducible data processing and data analysis. However, it must at the same time ensure that local data characteristics are preserved: local strengths must be exploited to improve the quality of the evidence generated, and local weaknesses must be addressed to protect quality.

Therefore, a) a CDM was chosen to represent, as faithfully as possible, the local data ('syntactic' harmonization), and b) harmonization of semantics has been deferred to the data processing phase, in order to tailor it to the study and to the DAP.

This deliverable aims to provide an overview of the processes and methods by which we arrived at the CDM for secondary re-use of health care data.

## 2. Background: overview of existing common data models for secondary use of health care data

The data held by institutions providing or recording health care varies widely. It can be generated in the course of routine care in hospitals and primary care practitioners' offices either for record keeping purposes or for billing. Additionally, data may be collected and maintained in spontaneous reporting system databases held by pharmaceutical companies or public health institutions for detection of adverse events following exposures to medicines and vaccines. Data may also be stored in single purpose systems such as medical birth, cancer, or perinatal registries, which may be linkable to other data sources.

In recent years, the amount and diversity of data available to researchers has grown and continues to grow exponentially. In order to make best use of this data for research purposes, it must be harmonized. Harmonization of disparate data sources allows for interoperability, the use of common tools across data sources, and reusability of data.

### 2.1 Types of Common Data Models

CDMs vary in terms of scope (protocol-based or protocol independent) and harmonization (syntactic or semantic harmonization). Definitions are provided below:

- *Protocol-based*: Source data extracted, transformed, and loaded to the CDM is limited to that required for a specific protocol or set of protocols. It is the subset of the source data required to answer a predefined set of study questions.
- *Protocol-independent*: Source data extracted, transformed, and loaded to the CDM is minimally a subset of the source data and maximally the entirety of the source data. If a subset, this subset is not limited according to data deemed relevant to a study question or set of study questions. Rather, it is potentially applicable to as yet undefined study questions.
- *Syntactic harmonization*: Syntax is defined as a connected or orderly system, harmonious arrangement of parts or elements (Merriam-Webster Dictionary). Syntactic harmonization is the arranging of data elements into a common structure without altering their content or meaning. Source data is extracted, transformed, and loaded to a CDM harmonized in terms of *structure* across data sources. The content of the tables and columns of the data in the CDM remains in its original format and is therefore allowed to remain heterogeneous amongst data sources.
- *Semantic harmonization*: Semantics is defined as of or relating to meaning in language, or the meaning or relationship of meanings of a sign (data element) or set of signs (Merriam-Webster Dictionary). Semantic harmonization is the derivation of common variables from the combination or restructuring of various data elements. Source data is extracted, transformed, and loaded to a common data which is harmonized in terms of *structure and content* across data sources. The content of the tables and columns of the data in the CDM must be mapped to a set of predefined concepts from a common vocabulary or set of vocabularies. Semantic harmonization incorporates Syntactic harmonization.



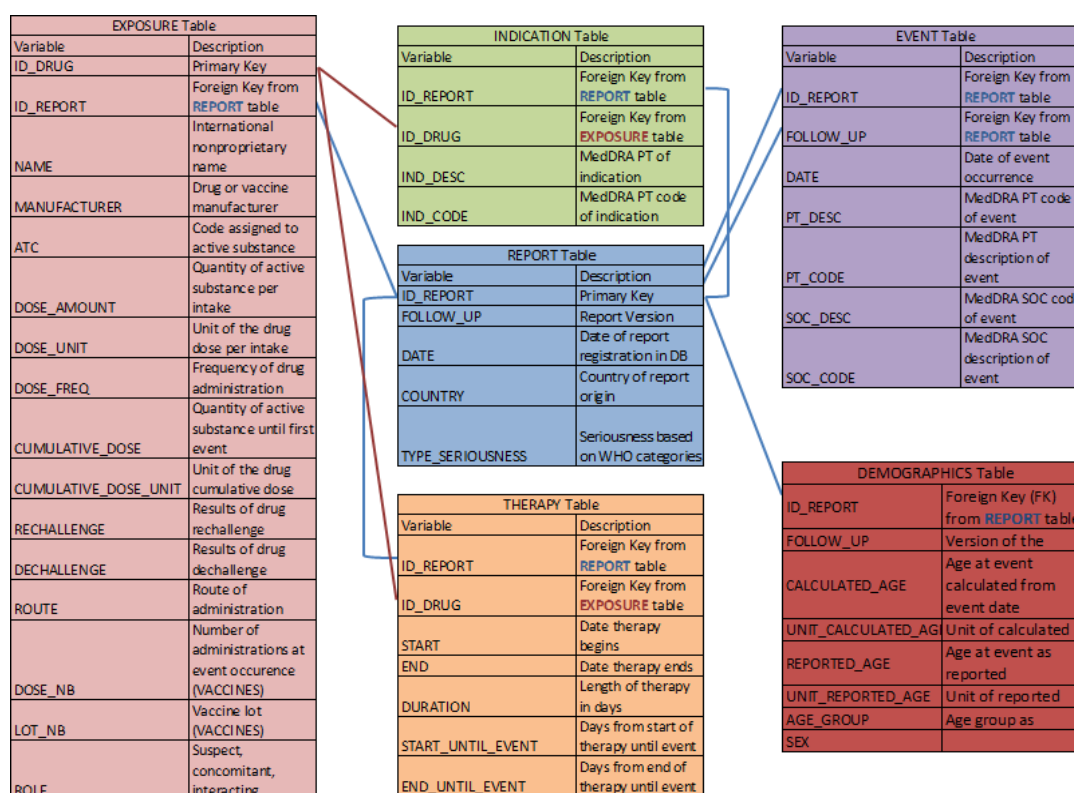
## 2.2 Existing Common Data Models for secondary re-use of data collected for other purposes

While there are countless study-specific common data models designed for one-time use, common data models designed for reuse within a network or community of researchers take on only a limited number of forms, each with one or two quintessential examples of the form in common usage. These are described below.

### 2.2.1 Common Data Models for Spontaneous reporting systems databases

Spontaneous reporting systems databases are those maintained typically by public health institutions for the reporting of suspected adverse events following exposure to a drug or vaccine. Patients, healthcare workers, and others can submit reports of exposures and adverse events to these systems. While each system may develop its own data model, and no one model exists across data sources, the International Conference on Harmonisation (ICH) has developed a guideline for transmission of individual case safety reports for the content and structure of data elements to be included in spontaneous reports (International Conference on Harmonisation, <http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html>).

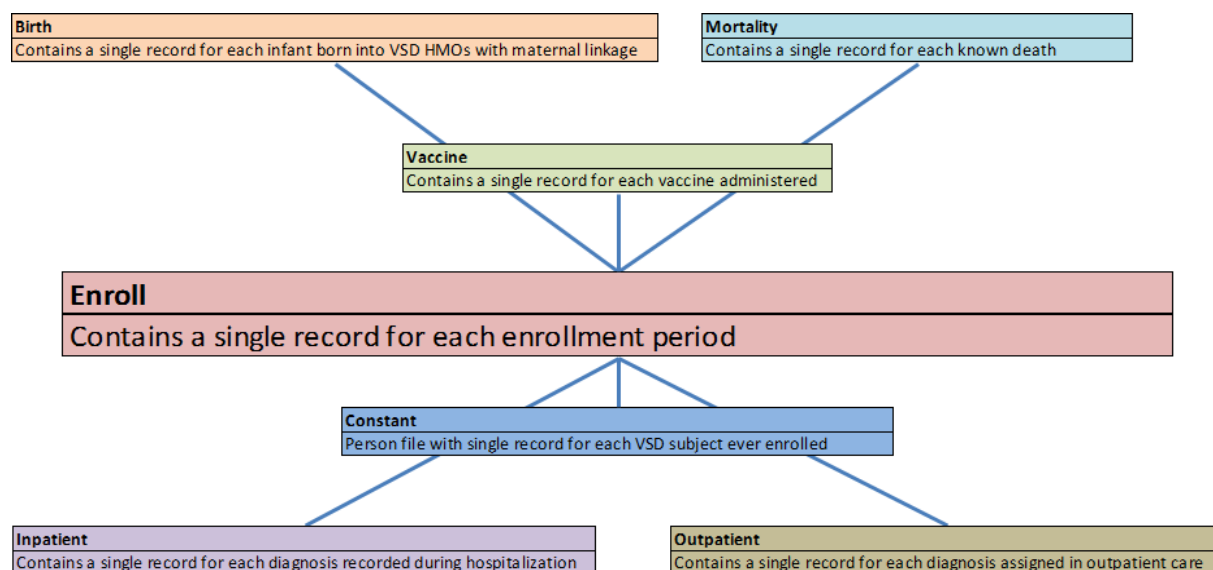
The European Medicines Agency has adopted this standard for the EudraVigilance database to which reports associated with all drugs licensed in Europe must be submitted. While no one common data model exists across all spontaneous reporting systems databases, the ICH guideline may have the potential to be exploited as the basis of a CDM. Additionally, recent work in the Global Research in Pediatrics (GRiP) consortium has developed a limited and semantically harmonized CDM for spontaneous reporting systems databases (5) comprising Report (Source of the report and reporter), Exposure (Suspected drug exposures), Indication (Indication of the reported drug), Event (Reported events and outcomes), Therapy (Therapy details for reported drugs), and Demographics (Subject demographics).



**Figure 1 - GRiP Common Data Model for Spontaneous Reporting System Databases(5).**

## 2.2.2 Vaccine Safety Datalink

The Vaccine Safety Datalink project is a collaboration of the United States healthcare organizations which links data on over 8 million subjects in order to quickly deploy vaccine safety studies against a purpose-built CDM. The CDM employed by the Vaccine Safety Datalink (VSD) is an example of a protocol-independent semantically harmonized common data model. While the CDM is protocol-independent, it is limited in scope in that it is designed solely for the study of vaccine coverage, safety, and efficacy. Therefore, only a limited set of variables relevant to vaccine safety are extracted, transformed, and loaded (ETL) to the CDM. The CDM comprises the following tables: Patient (Demographics and enrollment), Vaccination History (Vaccination dates, types, and manufacturers), Medical Visits (Healthcare encounters and diagnoses), Mortality (Death data), and Birth and Pregnancy (Pregnancy and birth data on mother and child). It has shown utility to address vaccine safety concerns rapidly. However, the semantic harmonization of this CDM together with its limited scope makes it unsuitable for use in addressing other study questions.



**Figure 2 - Vaccine Safety Datalink Common Data model**  
<https://www.cdc.gov/vaccinesafety/pdf/vsd-data.pdf>

Data in the VSD CDM is collected by health care organizations for the purpose of reimbursement and maintenance of electronic health records. Within the VSD system, tools for data analysis are deployed against dynamic data files which are updated on a weekly basis. Analysis methods developed within the VSD system include rapid cycle analysis, maximized sequential probability ratio test <sup>1</sup>, and the case-centred approach<sup>2</sup> but no open source tools to conduct these analyses have been made available. Quality checks within the VSD system include validation of diagnostic codes and other less well-defined quality checks. Records are retained at the patient level and mapped to standardized coding systems for vaccine types and manufacturers, diagnoses, procedures, and causes of death.

### 2.2.3 United States FDA Sentinel

The Sentinel CDM is an example of a CDM which is protocol-independent with a mixture of syntactic and semantic harmonization, dependent upon CDM table and column. The Sentinel Common Data Model is a product of the United States Food and Drug Administration Sentinel Initiative (<https://www.sentinelinitiative.org/>) and comprises the following tables: Enrollment (periods of health plan enrollment), Demographic (demographic characteristics), Dispensing (outpatient pharmacy dispensing), Encounter (healthcare encounters), Diagnosis (in and outpatient diagnoses), Procedure (in and outpatient procedures), Death (Death records), Cause of Death (Causes of death related to a death record), Laboratory Result (Results of laboratory tests), Vital Signs (Results of measurements), Inpatient Pharmacy (Inpatient drug administrations), Inpatient Transfusion (Inpatient transfusion administration), and Mother-Infant Linkage (Linkage between mothers and live-born infants).

<sup>1</sup> Lieu TA, Kulldorff M, Davis RL, Lewis EM, Weintraub E, Yih K *et al.* Real-time vaccine safety surveillance for the early detection of adverse events. *Med Care* 2007. doi:10.1097/mlr.0b013e3180616c0a.

<sup>2</sup> Fireman B, Lee J, Lewis N, Bembom O, Van Der Laan M, Baxter R. Influenza vaccination and mortality: Differentiating vaccine effects from bias. *Am J Epidemiol* 2009. doi:10.1093/aje/kwp173.

Administrative Data						Clinical Data	
Enrollment	Demographic	Dispensing	Encounter	Diagnosis	Procedure	Lab Result	Vital Signs
Patient ID	Patient ID	Patient ID	Patient ID	Patient ID	Patient ID	Patient ID	Patient ID
Enrollment Start & End Dates	Birth Date	Dispensing Date	Service Date(s)	Service Date(s)	Service Date(s)	Result & Specimen Collection Dates	Measurement Date & Time
Drug Coverage	Sex	National Drug Code (NDC)	Encounter ID	Encounter ID	Encounter ID	Test Type, Immediacy & Location	Height & Weight
Medical Coverage	Zip Code	Days Supply	Encounter Type and Provider	Encounter Type and Provider	Encounter Type and Provider	Logical Observation Identifiers Names and Codes (LOINC®)	Diastolic & Systolic BP
Medical Record Availability	Etc.	Amount Dispensed	Facility	Diagnosis Code & Type	Procedure Code & Type	Etc.	Tobacco Use & Type
			Etc.	Principal Discharge Diagnosis	Etc.		Etc.

Registry Data			Inpatient Data		Mother-Infant Linkage Data
Death	Cause of Death	State Vaccine	Inpatient Pharmacy	Inpatient Transfusion	Mother-Infant Linkage
Patient ID	Patient ID	Patient ID	Patient ID	Patient ID	Mother ID
Death Date	Cause of Death	Vaccination Date	Administration Date & Time	Administration Start & End Date & Time	Mother Birth Date
Source	Source	Admission Date	Encounter ID	Encounter ID	Encounter ID & Type
Confidence	Confidence	Vaccine Code & Type	National Drug Code (NDC)	Transfusion Administration ID	Admission & Discharge Date
Etc.	Etc.	Provider	Route	Transfusion Product Code	Child ID
		Etc.	Dose	Blood Type	Child Birth Date
			Etc.	Etc.	Mother-Infant Match Method
					Etc.

**Figure 3 - Sentinel Common Data Model**

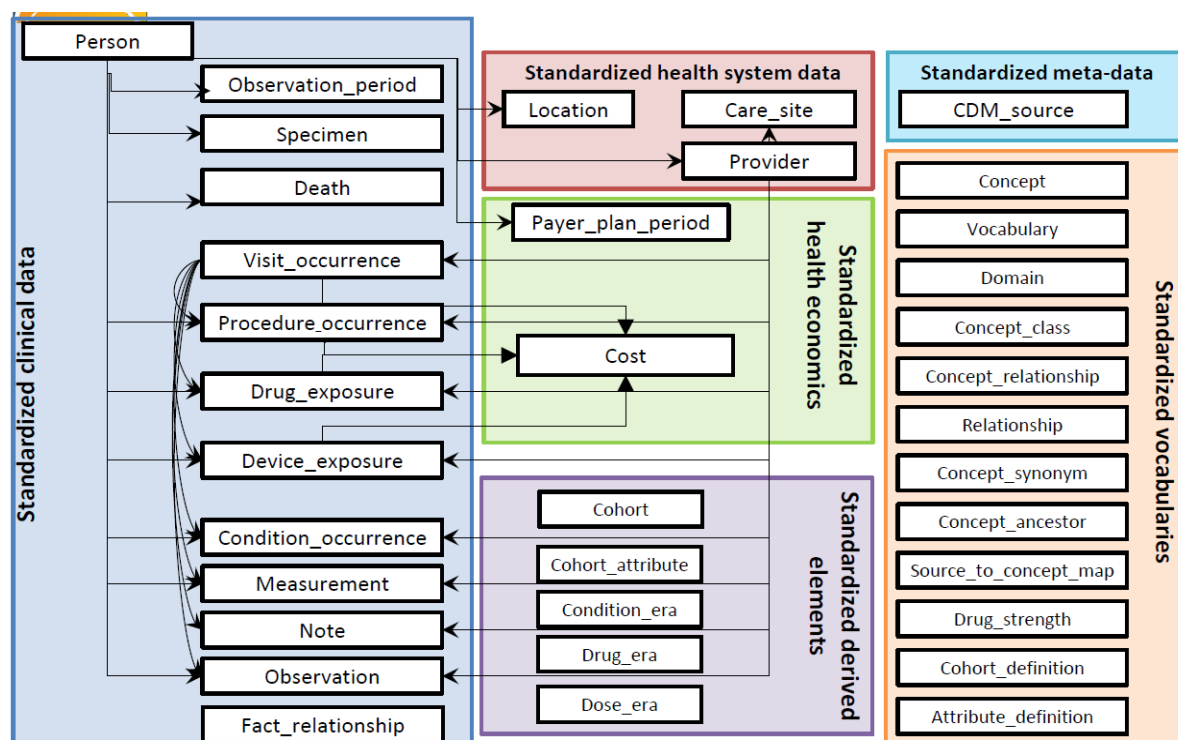
(<https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model>).

Data in the Sentinel CDM is developed for the United States and is primarily administrative and claims data from health insurers, collected for reimbursement purposes. Sentinel maintains a secure portal through which standardized programs are disseminated to data partners and results are shared but it does not provide open-source tools. Tools for data analysis include closed-source routine querying tools and study-specific analysis scripts. Sentinel conducts extensive data checks based upon pre-defined measures including completeness, formats, logical relationships, distributions, and trends over time. Records are retained at the patient level and linked across tables by a unique patient ID. Source data is harmonized to a common vocabulary for a subset of variables but for the most part the Sentinel CDM retains source data in its original format.

## 2.2.4 OMOP

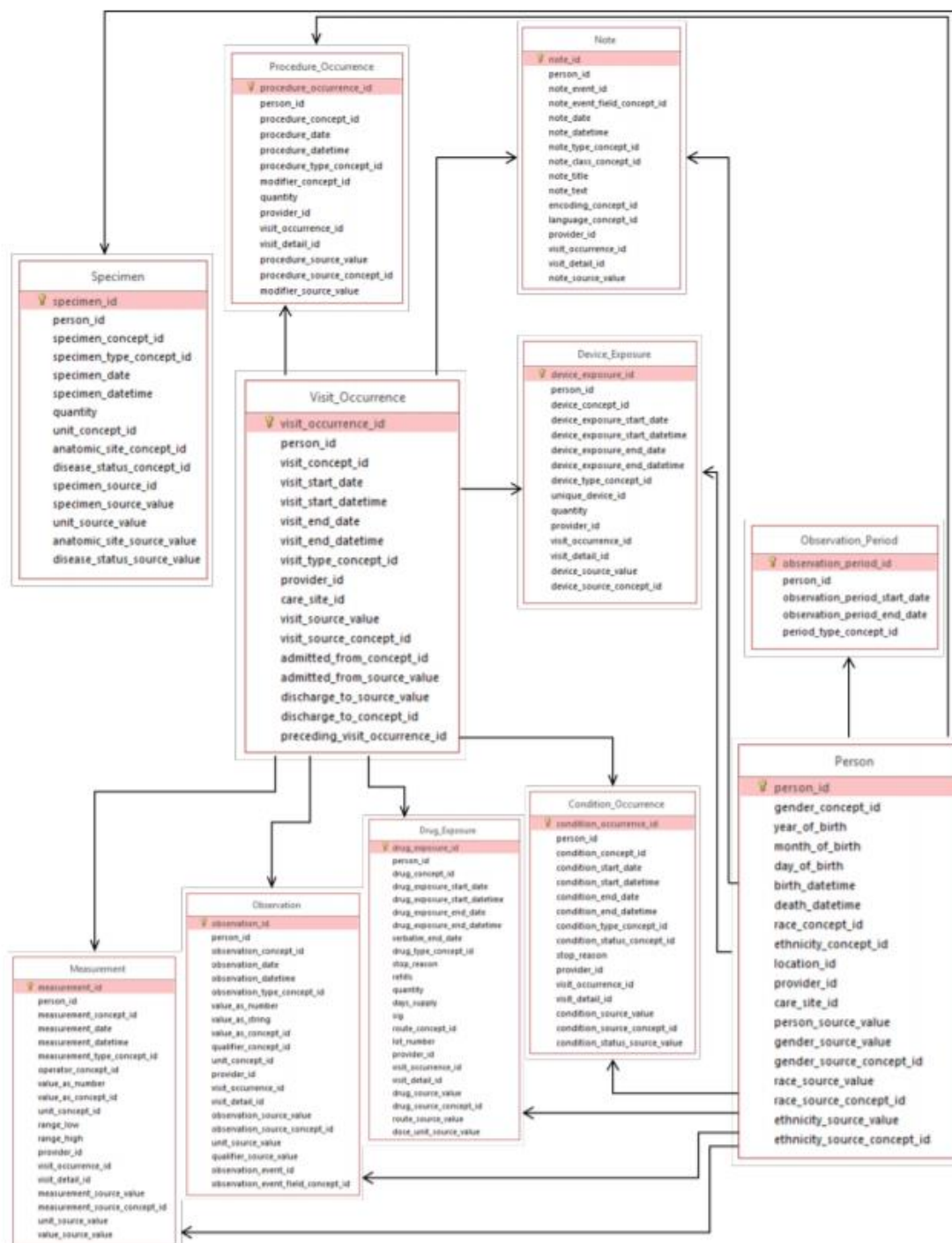
The OMOP Common Data Model is a protocol-independent semantically harmonized CDM. The OMOP CDM is a product of the Observational Medical Outcomes Partnership (OMOP), now OHDSI (Observational Health Data Sciences and Informatics), an international collaboration. The semantic harmonization of this CDM to a common set of vocabularies, terminologies, and coding schemes allows for deployment of analysis scripts against data in the CDM with less extensive definition and construction of variables at the analysis stage. Source data is retained and can be retrieved. The OMOP CDM is extensive. It includes, but is not limited to, the following tables: Person, Observation Period (time periods of observation), Specimen (Biological samples), Death (Causes of death), Visit Occurrence (Outpatient, inpatient, emergency, and long-term care visits), Visit detail (detailed data related to each visit occurrence), Procedure Occurrence (Procedures ordered or carried out), Drug exposure (drug utilization), Device exposure (device utilization), Condition Occurrence (Diagnoses), Measurement (Measurement results), Note (unstructured information), Observation (observations not recorded in other tables), Location (Physical

location of care site), Care Site (Health care units), Provider (healthcare provider), and Drug Era (exposure periods).



**Figure 4 - Graphical representation of the OMOP Common Data Model (v 5.0.1)**  
(<https://github.com/OHDSI/CommonDataModel/wiki>).

Data in the OMOP common data model is generated for reimbursement (claims) and in the process of routine care. Within the OHDSI ecosystem, open-source tools to facilitate distributed network analyses, database characterization, and common statistical analyses are available. Open-source tools for mapping, data source content and structure analysis, and interactive ETL design are available for conversion of raw data to the CDM. Tools for summary statistics and data visualization are available to check data quality and a newly developed data quality dashboard is in beta testing (as of April 2020). The CDM is person-centric (Figure 5) where persons' encounters and care episodes can be identified. Source values are retained in the CDM in each table as a source column and non-standard source data is retained in a separate table.



**Figure 5 – Subset of table relationships in the OMOP Common Data Model**  
[https://github.com/OHDSI/CommonDataModel/blob/master/OMOP\\_CDM\\_v6\\_0.pdf](https://github.com/OHDSI/CommonDataModel/blob/master/OMOP_CDM_v6_0.pdf)).

## 2.2.5 PEDSnet

The PEDSnet Common Data Model is based upon the OMOP CDM with extensions to include data relevant to paediatric investigators such as normalized heights and weights, immunization data, and geocoding. Because it is based upon the OMOP CDM, it contains many of the same tables with the addition of the ADT Occurrence table (admission,



discharge, and transfer events within a clinical visit) and the Immunization table (immunization records).

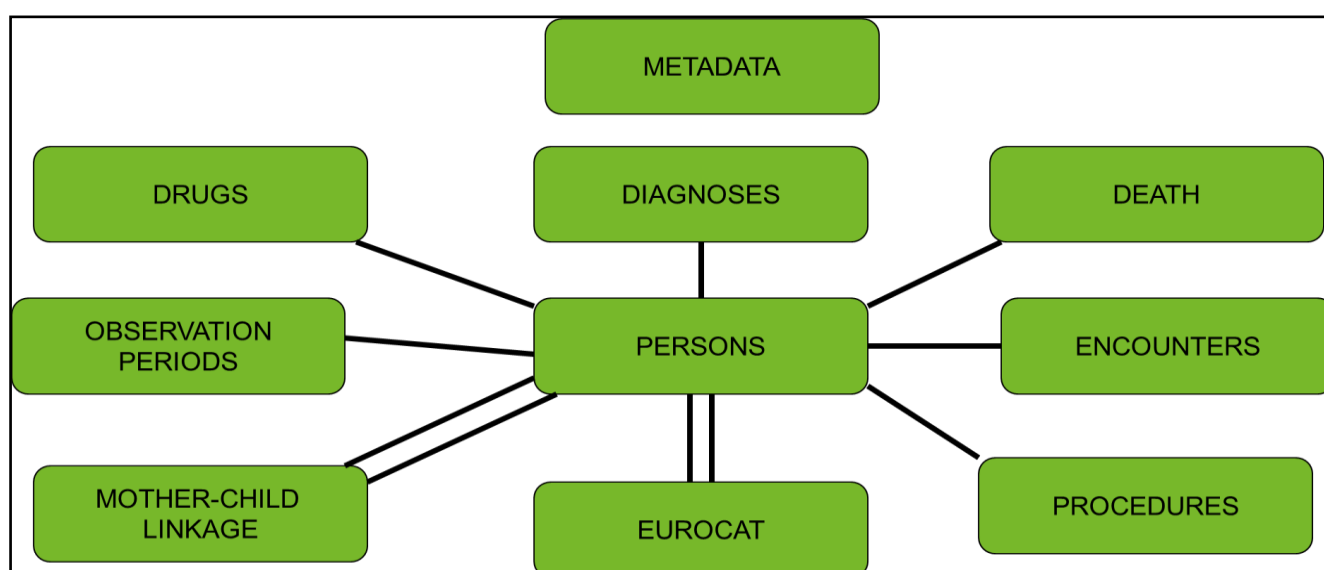
Similar to the OMOP CDM, data captured in the PEDSnet CDM is observational and includes all elements relevant for analysis, including claims and diagnoses. The data is drawn from US children's hospitals. The network includes a coordinating centre to which partner data can be uploaded, but data owners can also choose to keep data local. PEDSnet performs 850 quarterly data checks. ETL errors are assessed and corrected while subsequent data checks include those for fidelity (whether the data reflect the source population) consistency, accuracy, and completeness. Open source tools built to run against data in the OMOP CDM can also be deployed against the PEDSnet CDM, with some modifications if tables unique to the PEDSnet CDM are to be included in analyses.

### 2.2.6 Other common data models

Several other common data models may be created, for specific projects or analyses. In the area of medicines safety in pregnancy, we are aware that EUROLINKCAT is developing a CDM to investigate congenital anomalies (<https://www.eurolinkcat.eu/wp2-buildingresultsrepository>), but this is not publicly available to date.

### 3. ConcePTION CDM for health care data sources v1.0

In the process of developing the protocol for the data characterization study (task 7.6) and algorithm development (task 7.7), a first CDM for routine healthcare data and data which can be linked to population denominator data was developed, based on the models we had used in prior projects. In the protocol, it was specified that data access providers (DAPs) would be asked to extract all available data of relevance for the ConcePTION studies and convert these data into the ConcePTION CDM using their preferred software for syntactic harmonization. Whilst reviewing and obtaining approval for the data characterization protocol, it became apparent that some DAPs could not provide a full extraction of their data source, due to their organisations' policies and GDPR. Therefore, the decision was taken to define the ConcePTION CDM as protocol-based (see definitions in section 2.1). Within the detailed CDM description provided in Appendix 2, this is reflected in the lists of events and procedures requested for extraction into the CDM, for use in data characterization and algorithm development. Figure 6 shows an overall view of the tables of the ConcePTION CDM v1.0.



**Figure 6- Schematic of ConcePTION CDM v1.0**

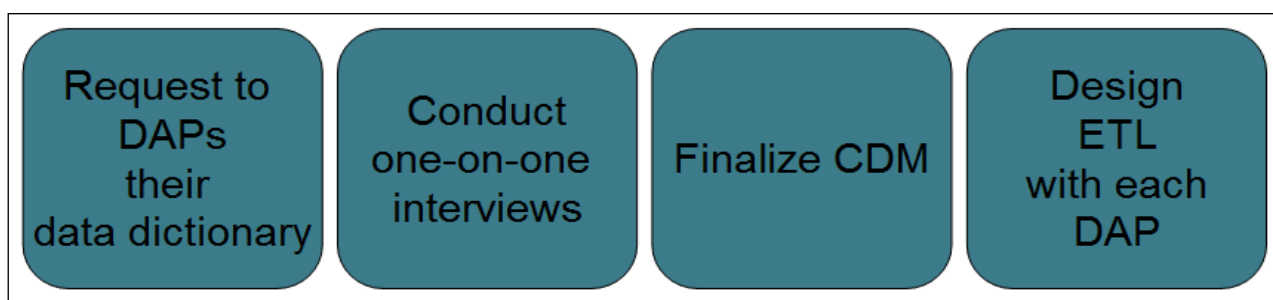


## 4. Processes to create ConcePTION CDMv2.01

### 4.1 Indepth interviews with Data Access Providers

After the first draft was completed, it was deemed necessary to investigate whether the data dictionaries of the data accessed by the ConcePTION DAPs could be faithfully mapped to it.

For this reason, 4 steps were scheduled, as depicted in Figure 1.



**Figure 8 - Pathway to finalization of the CDM and of the ETL of each DAP.**

#### 4.1.1 Request to DAPs their data dictionary

In October 2019 the task ‘provide Data Dictionary’ was launched on the ConcePTION Task Management System, (see Appendix 1 for instructions). The task was closed on 15 November, when 18 DAPs had responded. Later on two additional DAPs were added. The responses fed into the interviews (see 4.1.2) which eventually produced a standardised document (see 4.1.3) which is currently available in the member area of the Project Website and will be included in the Catalogue.

#### 4.1.2 Conduct one-to-one interviews.

The interviews were scheduled from January 2020, see Table 1. In each interview, the DAP was represented by one or more researchers/investigators, and WP7 by two 2 interviewers, one with the role of conductor (one of Miriam Sturkenboom and Caitlin Dodd, UMC; and Rosa Gini and Giuseppe Roberto, ARS), and one with a supporting role (Marianne Cunnington, GSK, Romin Pajouheshnia and Marieke Hollestelle, UMC, and Claudia Bartolini and Olga Paoletti, ARS). Nicolas Thurin, University of Bordeaux, supported the development of the methodology.

**Table 1. Schedule of the interviews**

Code DAP	Name DAP	Acronym (if partner)	Date	WP7
1	University of Oslo (UOSL)	UOSL	8 Jan 10-12	All

4	University of Aarhus		27 Feb 14-16	Caitlin, Romin
5	University of Dundee		28 May 14.30-16.30	Rosa, Olga
7	University of Ulster (ULST)	ULST	28 Jan 14-16	Caitlin, Marianne
8	Centre Hospitalier Universitaire de Toulouse (CHUT)	CHUT	22 Jan 10-12	Rosa, Marianne
10	University of Bordeaux		21-Feb-2020	Rosa. Olga
11	University Medical Center Groningen (UMCG)	UMCG	21 Jan 10-12	Caitlin, Olga
13	PHARMO institute		28 Jan 10-12	Miriam, Claudia
15	Leibniz Institute for Prevention Research and Epidemiology (BIPS)	BIPS	16 Jan 14-16	Giuseppe, Romin
18	Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO)	FISABIO	29 Jan 10-12	Rosa, Olga
19	IDIAP-Jordi Gol		22 Jan 10-12	Caitlin, Romin
20	Università degli Studi di Ferrara – University of Ferrara (FERR)	FERR	11 Feb 10-12	Caitlin, Marianne
21	CNR Tuscany (CNR-IFC)	CNR-IFC	23 Jan 10-12	Rosa, Marianne
22	Agenzia regionale di sanità della Toscana (ARS)	ARS	21 Jan 14-16	Miriam, Romin
23	University of Messina			
24	Malta Congenital Anomalies Registry, Directorate for Health Information and Research		15 Jan 10-12	Caitlin, Marieke
27	Malformation Monitoring Centre Saxony-Anhalt Medical Faculty, Otto-von-Guericke University		22 Jan 14-16	Giuseppe, Marianne
33	National Institute for Health and Welfare, Finland	NIHW	19 Feb 10-12	Rosa, Olga
34	University of Swansea	USWAN	20 Feb 10-12	Rosa, Olga

35	GSK	GSK	24 Apr 10-12	Rosa, Caitlin
----	-----	-----	--------------	---------------

The interview, was conducted by requesting the DAPs to describe each table whose data dictionary they had provided by replying to five questions.

**Table 2. Questions addressed by the DAP during the interview, for each table relevant to ConcePTION that the organization has access to.**

- 1) *What triggers the creation of a record of the table?*
- 2) *Is the table collected for all the population in your database, or only for a sub-population*
- 3) *Can you comment on the completeness and quality of the table? If you don't have formal measurements, feel free to convey the assumptions you commonly make*
- 4) *What is the time span of the table, how often is it refreshed, and what is the lag time between data creation to your organization?*
- 5) *Include other comments you may want to share about this table*
- 6) *Fill out the table below with the names of the variables in this table (as listed in the data dictionary) that you plan to map to the ConcePTION Common Data Model, with a description in English of the meaning of the variable, the name of the classification used (e.g. CIM10, ATC, ..., or national/local) or the description in English of the data dictionary if a small number of values are included in the dictionary, and any comment you may want to share about that variable (when it is missing, or miscoded, or when its content is unreliable)*

Beyond describing the data, each DAP was requested to provide feedback after filling out the Catalogue questionnaire. Finally, the conversation wrapped up some general questions. The complete questionnaire is in Appendix 2.

When the questionnaire was not completed by the end of two hours' conversation, the document was completed offline and emailed.

#### 4.1.3 Finalise interview answer sheets

The finalisation of the documents was scheduled as a task in the Task Management System.

The final versions of the documents were uploaded on the member area of the project website, to allow for ConcePTION investigators to understand the data sources that they plan to use for their studies.

The content of the interviews answer sheets will be formatted and made available on the Catalogue in an interactive manner, to support decisions of investigators on semantic harmonisation of study variables.

## 4.2 Finalize the CDM

### 4.2.1 CDM recommendations following each interview

After each interview, the interviewers would fill out a questionnaire regarding lessons learnt from the interview about the CDM. The questions are indicated in Table 3. The results are in Appendix 3.

**Table 3. Questions collecting input from interviews with respect to how the CDM v 1.0 should be modified**

**Author: ...**

**Date: ...**

**Would there be a need for additional tables?**

...

**Would there be a need for additional columns?**

...

**Would there be a need for additional values for some existing column?**

...

- in particular for provenance?

...

- for type of data source?

...

**What are the coding systems adopted in the DAP?**

- International Coding systems: ...
- Local coding systems: ...

Tables and columns for which the ConcePTION CDMv1.0 did not provide an appropriate accommodation were analysed. Relevant existing models for capture of this data was sought in existing CDMs or standards, to support decisions on how to upgrade to a ConcePTION CDM v2.0 that would validly accommodate all the data described by DAPs during the interviews.

#### 4.2.2 Comparison with existing common data models

A dedicated working group then compared each table of the ConcePTION CDM v1.0 with the OMOP CDM. Each table of the ConcePTION CDMv1.0 was compared to the corresponding table of the OMOP CDM. Subsequently, each table and column of the OMOP CDM was checked to ascertain whether a corresponding table and column with similar meaning was present in the ConcePTION CDM. Tables deemed relevant to ConcePTION which were identified in the process of in-depth DAP interviews (See section 4.2.1) were discussed in the working group. Where possible, tables in existing common data models were used as a basis upon which to define the table for ConcePTION CDMv2.0. For example, vaccination tables in the Vaccine Safety Datalink and the PEDSnet were taken as the basis for defining the vaccinations table in ConcePTION CDMv2.0.

It was decided that, whenever possible, the ConcePTION CDM would use the same names of tables and columns as the OMOP CDM. Moreover, it was decided that whenever a code may be not available but free text fields are available instead (e.g. diagnosis, procedures, indications for drug, drugs themselves), if DAPs have expertise in querying them, they are requested

- 1. To include in the corresponding 'code' column the free text
- 2. To code in the corresponding 'coding system' column the string 'Free text'

#### 4.2.3 Search for standard for medical birth registers

In ConcePTION, a number of DAPs have access to medical birth registries. In the course of DAP interviews, it became clear that these should be incorporated in the ConcePTION CDMv2.0.

In order to develop specifications for a Medical Birth Registry CDM table, a comparison across the Medical Birth Registry tables of DAPs was done. The structure of the Norwegian birth registry was used as a reference since it is well organized and divides all information into categories such as:

- Identification number
- A-demographic data, mother, father
- B-pregnancy and maternal health, mother's health prior to pregnancy, mother's socio-economic status, smoking habits during pregnancy, ART (assisted reproductive technology)
- C-childbirth, position/induction/interventions, complications during delivery, anesthesia/analgesia, placenta/umbilical cord/amniotic fluid, maternal complications *post-partum*
- D- child, mortality, child health and neonatal diagnosis, congenital anomalies,

#### Procedure:

- A spreadsheet containing all variables per birth registry was created.
- For each medical birth registry (Norway, Denmark, England and Wales, Finland, Italy, Netherlands) a category (based on the Norwegian categories) was assigned for each variable.

- Variables of different registries were sorted per category
- If two variables of different registries were very similar, they were positioned in the same row.

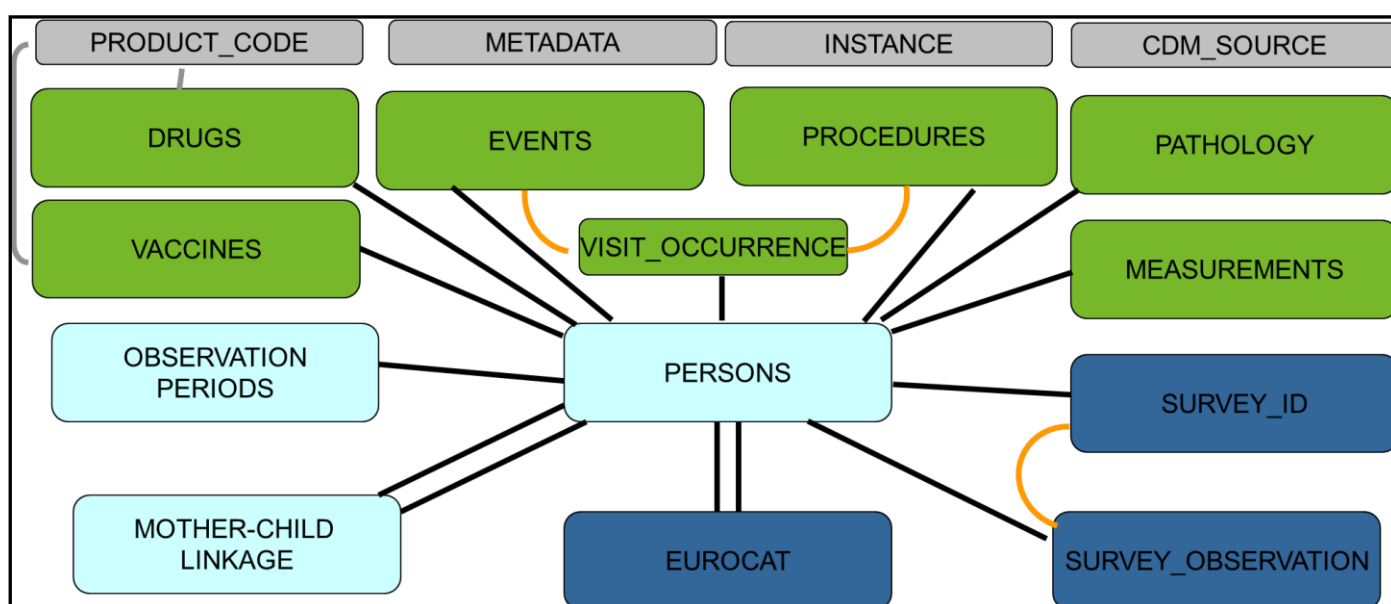
## Conclusion

Due to:

- the complexity of the present birth registries
- loss of information (a large number of different variables between registries was present)

it was decided that birth registries would not be harmonised, but rather incorporated in the CDM in their original format. Two new tables, named SURVEY\_OBSERVATIONS and SURVEY\_ID, were incorporated in the ConcePTION CDMv2.0 for this purpose (see next subsection).

After a series of full and half-day workshops, a new version of the ConcePTION CDM was created and is shown in Figure 8.



**Figure 8 – Schematic of the ConcePTION CDMv2.0**

### 4.2.4 Update to ConcePTION CDM v2.0

Details on the ConcePTION CDM v2.0 and on the discussion that lead to its finalisation is provided in Appendix 3. We refer here to the main decisions. The tables of the CDM were divided into four categories: Metadata (depicted in grey in figure 8), Routine healthcare data (depicted in green in figure 8), Curated tables (depicted in light blue in figure 8), and Surveillance tables (depicted in dark blue in figure 8).

#### Metadata tables

- The metadata table from v1 has been split in three: two are equal to corresponding tables in the OMOP CDM, the third, INSTANCE, is meant to document which local data that is mapped to the current instance of the ConcePTION CDM.
- PRODUCT\_CODE was added as an additional metadata table meant to document medicinal-product specific data which is linked to the MEDCINES and VACCINES tables.

### **Curated tables**

- PERSONS has been classified as a derived table, with one row per person who is included in the instance; variables recorded here are stable at the date of instance creation. The DEATH table is discarded, and causes of death are included in the EVENT table
- OBSERVATION\_PERIODS has been classified as a derived table. It has multiple rows per person corresponding to each period during which the person was considered as under observation according to the DAP.
- MOTHER\_CHILD\_LINKAGE remains as described in ConcePTION CDM v1.

### **Routine Healthcare data**

- DRUGS and VACCINES were separated: the former would collect dispensed and prescribed medicines the latter dispensing's, prescriptions, or administrations of vaccines.
- PROCEDURES and MEASUREMENTS: the decision was made to separate procedures (such as surgeries, or diagnostic procedures, rehabilitation procedures, therapeutical procedures) from measurements.
- PATHOLOGY: data from pathology registry was accommodated in a specific table
- VISIT\_OCCURRENCE: In line with the principle of adhering to OMOP conventions, it was decided that the CDM should incorporate a table to record visit occurrences.

### **Surveillance tables**

This new section of the CDM was created to accommodate tables that are identified as registries (e.g. birth registries) or as surveillance or as surveys. Beyond EUROCAT, which remained unchanged from ConcePTION CDM v1.0, the SURVEY\_ID table was added in combination with SURVEY\_OBSERVATION, partly replicating the choice of the OMOP CDM. The former would collect a single row per subject recorded in a entry, the latter would record in a 'entity-attribute-value' fashion all the information available on that subject in the same entry

## **4.2.5 Update to ConcePTION CDM v2.01**

Following further discussion among the CDM workgroup along with interactions with DAPs and design of the ETL template (see section 5), minor revision of the ConcePTION CDM v2.0 led to the current version, named ConcePTION CDM v2.01 Changes to v2.01 from v2.0 include the following:

- The 'DRUGS' table was renamed to MEDICINES for consistency with ConcePTION preferred terminology.
- Based on a request from WP1 regarding study designs such as sibling controls as well as input from DAPs regarding availability of family links in addition to maternal-

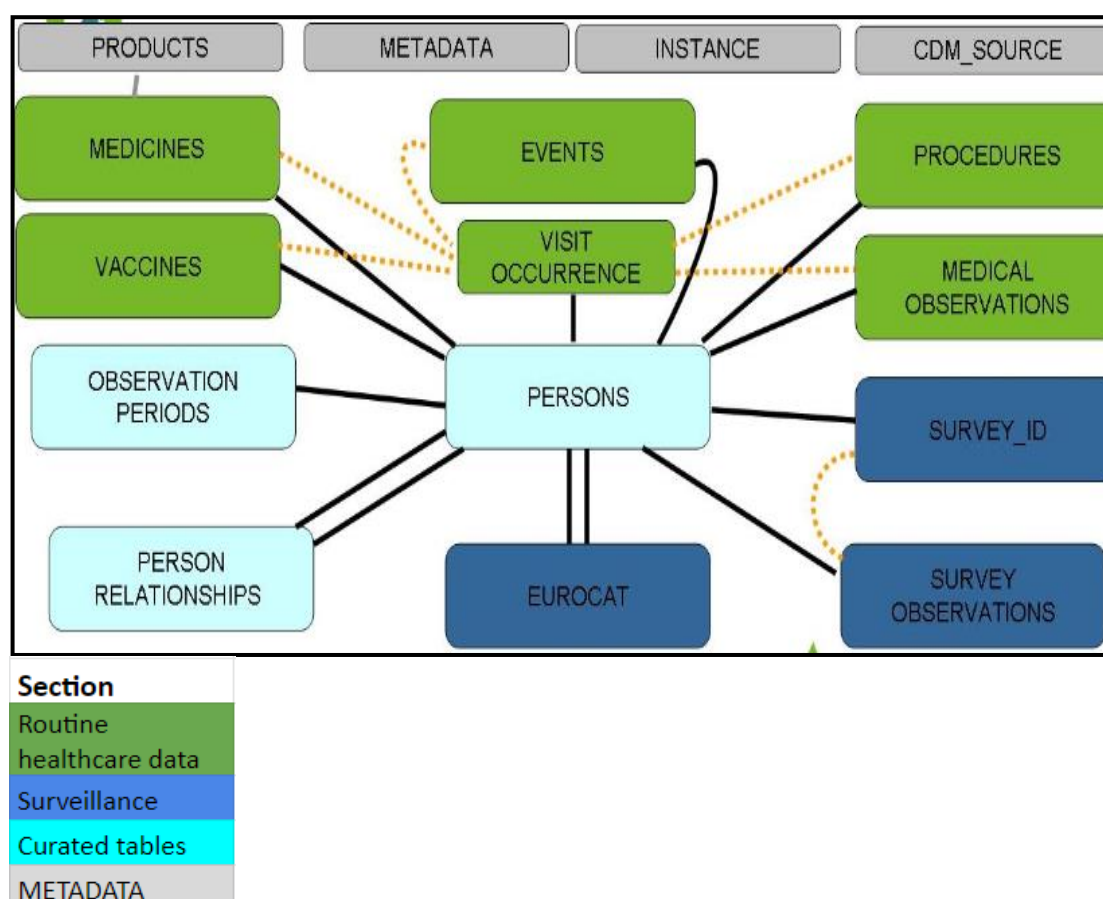


child linkage, the MOTHER\_CHILD table was reformulated as the PERSON\_RELATIONSHIP table. The updated structure of the table allows for recording of familial and household relationships beyond the relationship between mother and child. The MEASUREMENTS table was reformulated as the MEDICAL\_OBSERVATIONS table to allow for capture of observations in addition to measurements made during healthcare encounters such as educational and smoking status.

- Thanks to this choice, the PATHOLOGY table could be removed as its content is incorporated in MEDICAL\_OBSERVATIONS

#### 4.2.6 ConcePTION\_CDM v2.01: tables

The diagram of the tables of the ConcePTION CDM v2.01 is in Figure 9. As in the previous version, tables are classified in four sections: Metadata (four tables, depicted in grey in figure 9), Routine healthcare data (six tables, depicted in green in figure 9), Curated tables (three tables, depicted in light blue in figure 9), and Surveillance tables (three tables, depicted in dark blue in figure 9).



**Figure 9. Schematic of the ConcePTION CDMv2.01**

The full description of the ConcePTION CDM v2.01 can be accessed via this link: <https://drive.google.com/file/d/1hc-TBOfEzRBthGP78ZWla13C0RdhU7bK/view?usp=sharing>

Each table is specified with 4 sections



- In Figure 10 the specification of one of the tables is reproduced.

**Figure 10. Example of specification of a table of the ConcePTION CDM v2.01**

Section	Table	Role
Routine healthcare data	VISIT_OCCURRENCE	This table contains a summary description of the visits during which records of EVENTS, PROCEDURES, but possibly also MEDICAL_OBSERVATIONS or VACCINES or MEDICATIONS were recorded. This serves both to collect visit-level information, and to enable grouping sets of records that were recorded concurrently
	EVENTS	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.
	MEDICINES	This table collects data on drug prescriptions, dispensing or administrations occurred during routine healthcare.

	PROCEDURES	This table collects procedures administered during routine healthcare. Can be a surgery, or a diagnostic procedure, a rehabilitation procedure, a therapeutical procedure...
	VACCINES	This table collects dispensations or administrations of vaccines.
	MEDICAL_OBSERVATIONS	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
Surveillance	EUROCAT	This table collects surveillance data on congenital anomalies, following the EUROCAT standard
	SURVEY_ID	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
	SURVEY_OBSERVATIONS	List of observations in a survey
Curated tables	PERSONS	This table records persons that are to enter analysis of this instance of the CDM
	OBSERVATION_PERIODS	Periods during which data is collected in the data source for this person. This table contributes to defining the data source population.
	PERSON_RELATIONSHIPS	For any person, this table collects the pairing with the identifier of mother or of other relationships that may be available
Metadata	PRODUCTS	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.
	CDM_SOURCE	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
	METADATA	This table contains some general information about how the local 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 tables that are expected to be findable can indeed be found; and that the coding systems that are observed in the data are indeed those listed here)
	INSTANCE	This table displays the list of the tables and columns of the local 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 local data dictionary that are included in the current ETL document.
--	--	---

#### 4.2.7 ConcePTION CDM v2.01: vocabulary

In parallel with development of the ConcePTION CDM v2.0 and v2.01, a set of vocabularies was developed. The current version of the vocabularies is available via this link: [https://docs.google.com/spreadsheets/d/1idAEKC440rkIYlXCSRmEVgEPj\\_UouUI-l3kxNCpJt3U/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1idAEKC440rkIYlXCSRmEVgEPj_UouUI-l3kxNCpJt3U/edit?usp=sharing)

The vocabularies include the following:

Vocabulary	Purpose	CDM Tables
specialty_of_visit_vocabulary	coding system of the specialty	VISIT_OCCURRENCE
status_at_discharge_vocabulary	vocabulary of outcome of the visit	VISIT_OCCURRENCE
meaning_of_visit	meaning of the visit record	VISIT_OCCURRENCE
origin_of_visit	origin of the visit record	VISIT_OCCURRENCE
event_record_vocabulary	Vocabulary to which the 'code_event' belongs to; or, if the record contains 'free_text_event', this column contains the indication 'free_text'	EVENTS
meaning_of_event	meaning of the event record	EVENTS
origin_of_event	origin of the event record	EVENTS
code_indication_vocabulary	Vocabulary to which the 'code_indication' belongs	MEDICINES
vx_type	vaccine type as defined by antigens and components	VACCINES
vx_dose	Dose, particularly for childhood vaccines (1, 2, 3, Booster, etc)	VACCINES
vx_manufacturer	Name of vaccine manufacturer	VACCINES

meaning_of_vx_record	nature of the original record having originated the vaccine record	VACCINES
disp_amount_drug_unit	Unit characterizing the quantity or drug dispensed or administrated	MEDICINES
meaning_of_drug_record	nature of the original record having originated the drug record	MEDICINES
origin_of_drug_record	origin of the original record having originated the drug record	MEDICINES, VACCINES
prescriber_type	Indicates the speciality of the physician or professional who prescribed the drug	MEDICINES
procedure_code_vocabulary	Vocabulary to which the 'procedure_code' belongs to	PROCEDURES
meaning_of_procedure	meaning of the procedure record	PROCEDURES
origin_of_procedure	origin of the procedure record	PROCEDURES
mo_record_vocabulary	Vocabulary to which the 'mo_code' belongs to	MEDICAL_OBSERVATIONS
mo_meaning	nature of the original record having originated the medical observation record	MEDICAL_OBSERVATIONS
mo_unit	unit characterizing the measurement recorded	MEDICAL_OBSERVATIONS, SURVEY_OBSERVATIONS
mo_origin	origin of the original record having originated the medical observation record	MEDICAL_OBSERVATIONS
meaning_of_survey	The meaning of this survey for this person	SURVEY_ID
race	race of the person	PERSONS
country_of_birth	country of birth of the person	PERSONS

sex_at_instance_creation	Sex of the person in the moment when in the instance of the CDM is created	PERSONS
quality	A judgement on the quality of the variables recorded in this table	PERSONS
op_meaning	represents the semantic of the record	OBSERVATION_PERIODS
op_origin	represents what mechanism originated the record	OBSERVATION_PERIODS
origin_of_relationship	where the information about the relationship comes from	PERSON_RELATIONSHIPS
meaning_of_relationship	Which type of relationship there is between the mother and the person	PERSON_RELATIONSHIPS
method_of_linkage	How the linkage was performed	PERSON_RELATIONSHIPS
box_size_unit	Unit of measure characterizing the box size (e.g. tablets or injections) or the total quantity (e.g. ml, g)	PRODUCTS
drug_form	Characterize the form of the product unit	PRODUCTS
route_of_administration	Characterize the route of administration of the product unit	PRODUCTS
data_access_provider_code	Code of this DAP organization in the ConcePTION coding system	CDM_SOURCE
data_access_provider_name	Name of the DAP organization	CDM_SOURCE
cdm_version	version of the ConcePTION CDM vocabulary this instance conforms to.	CDM_SOURCE
cdm_vocabulary_version	version of the ConcePTION CDM this instance conforms to.	CDM_SOURCE

type_of_metadata	There are different types of metadata that are recorded, they may be associated with a table or a table/column, or other	METADATA
------------------	--	----------

In particular, in each table of the ConcePTION CDM (except the Metadata tables) specific items collect the *meaning* and the *origin* of the data. The vocabularies will be updated as DAPs proceeds through ETL specifications (see section 5).

## 5. Extract, Transform, and Load template and process for specification

The documents described in 4.1.3 are an in-depth description of the data sources accessed by the ConcePTION DAPs. Based on those documents, ETL specification documents are created by each DAP. The ETL specification documents are described in this section.

### 5.1 ETL template

A standard template was developed to describe the process of extracting data, transforming, and loading (ETL) from the local data source (**origin tables**) into the ConcePTION CDM (**target tables**)

The template is structured as follows

**Header:** name of the DAP, name of the data source, authors, date and version of the document

**Section 1:** list and short description of the tables of the version of the ConcePTION CDM that is the target of the ETL (this section is the same across all DAPs)

**Section 2:** list and short description of the origin tables of the data source that are being transformed and loaded to the CDM

**Section 3:** is composed by two tables

- From origin tables to target tables: for each origin table, list target tables that it will populate
- From target tables to origin tables: based on the previous table, for each table of the CDM (target table) the list of origin tables that are populating it

**Section 4:** this is the core of the ETL document. It has a subsection for each table of the CDM. In the subsection of a target table, for all the origin tables that feed it there are two elements

- The description of the rule that generates records of the target table from records of the origin table
- a specification table: The '**Target column**' contains the list of the CDM columns of the target table you are specifying (in the example below, EVENTS); for each of them, the DAP must specify the name(s) of the origin column(s) that will feed the target column, in the '**Origin column**', and/or the rule that will create the content, in **Rule**; the rule may be a simple string. To decide which column(s) goes where, and which rules they should adopt, DAPs should use the description of the target table contained in the CDM table, and in particular the 'description' and 'conventions'

specifications. A shaded background indicates that the values they set for that column must belong to the CDM vocabulary. If DAPs do not find a value that fits their data well, they are invited to include new values in a structured process. In Table 4 an example of an empty table for the EVENTS target table

**Annex:** in this section some general concepts are specified, such as ‘data source’ or ‘data source population’.

**Table 4. Example of a specification table**

Target table: EVENTS			
Origin table:			
Target column	Origin column	Rule	Notes
person_id			
visit_occurrence_id			
visit_start_date			
visit_end_date			
specialty_of_visit			
specialty_of_visit_vocabulary			
status_at_discharge			
status_at_discharge_vocabulary			
meaning_of_visit			
origin_of_visit			

The complete template is included in Appendix 6.

## 5.2 ETL process specification for the ConcePTION DAPs

The request to complete their ETL specification document was launched as a task for 20 DAPs in May 2020, with deadline in June and is ongoing at the time of this deliverable. WP7 is supporting the task with drop-in sessions and through one-on-one conversations through the Task Management System.

### 5.3 ETL documents

The ETL documents will be uploaded to the member area of the project website and in a dedicated area of the ConcePTION Catalogue.

## 6. Semantics superimposed on the ConcePTION CDM

### 6.1 Template of a Statistical Analysis Plan

Appendix 7 contains the template of a statistical analysis plan. Section 7.3 of the template describes the CDM and is ready to accommodate the study-specific harmonised dataset that is built on top of the CDM.

‘Building the harmonised dataset’ is the operation represented as ‘T2: create study variables’ in the figure below

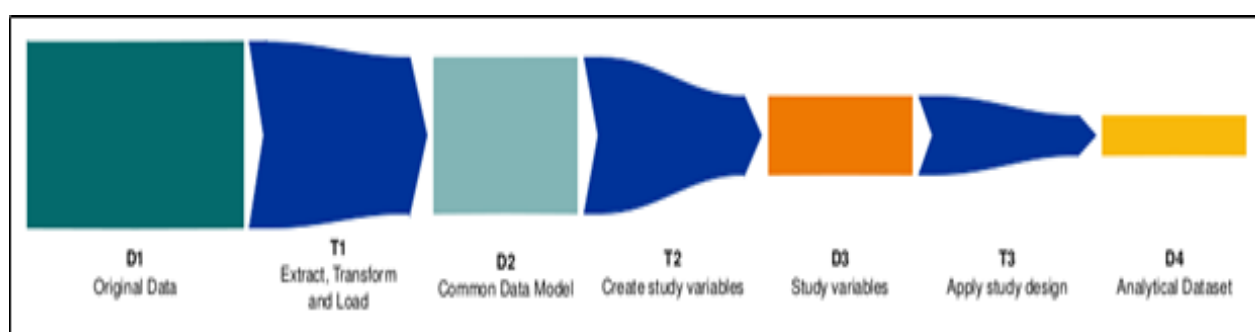


Figure 11. Representation of the data processing steps (adapted from Gini et al, 2016)

When study variables are built for a study, the semantics of the heterogeneous data recorded in the CDM needs to be interpreted to create harmonised data items.

The next sections describe how this step is designed (section 6.2) and implemented in modular programming (section 6.3)

### 6.2 Design study variables

#### 6.2.1 Study variables based on *Surveillance* or *Curated* tables

In the case of study variables based on *Surveillance* and on *Curated* tables of the CDM, harmonisation follows the guidelines set out in the Maelstrom guidelines(6). We created a template that adapts the guidelines to the ConcePTION CDM, see an example in Box 1 below. In the template, the unit of observation of the variable is described, together with its name, meaning and vocabulary; then the rule to derive the variable is specified for each data source, based on the description contained in the interview answer sheet.



**Box 1.Example of specification of a variable based on the *Surveillance* tables**

## 1. Describe its name, unit of observation, meaning and vocabulary

Example:

**Name:** PREVIOUS\_PREG**Unit of observation:** a woman with a current pregnancy**Meaning:** is 'YES' if the woman had had previous pregnancies, 'NO' if she had not, 'UNKNOWN' if unknown**Vocabulary:**

YES = the woman had had previous pregnancies before the current one,

NO = she had not,

UNKNOWN = unknown

## 2. Collect from all the origin tables of SURVEY\_OBSERVATIONS in all the data sources the columns that can be used to retrieve the variable of interest, and list them in the table below, together with the rule to obtain the desired variable(s) from them; it is possible that the variable is created from different origin tables for different subpopulations (e.g. pregnancies that end in delivery, in spontaneous abortions, in induced abortions)

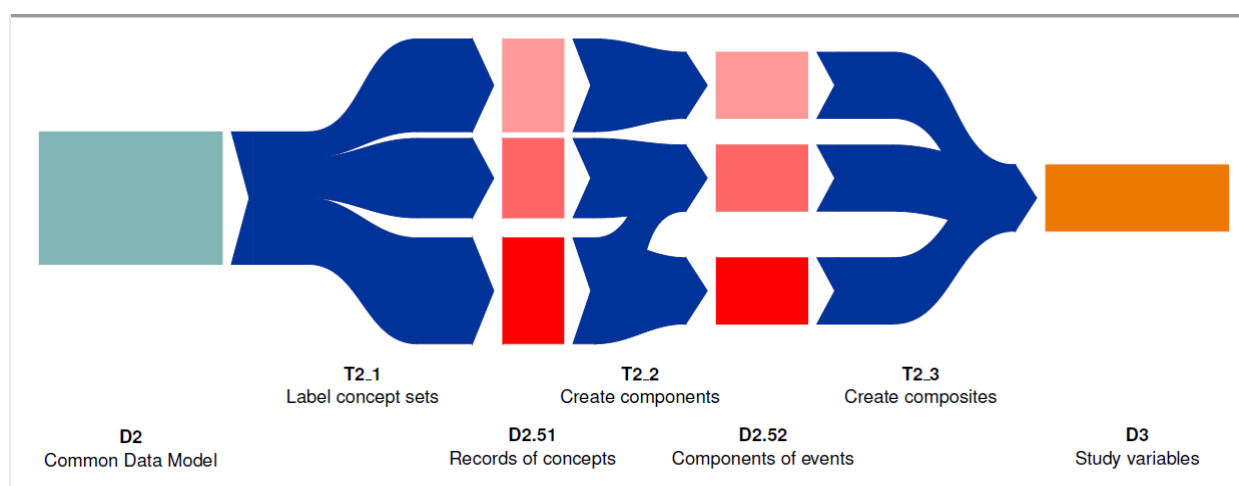
Example

DAP/datasource	tablename	column	values	rule	subpopulation
01_UOSL	MBRN	PARITET	all	set the variable =YES if this number is >=1, NO otherwise UNKNOWN if it is missing	Pregnancies that end after 12 weeks of gestational age
04_AARHUS	MBR	Tidligere_provokerede_aborter	all	Sum the 3 rows and set the variable =YES if the resulting number is >=1, NO otherwise	Pregnancies that end in childbirth
	MBR	Tidligerekejsersnit_i_danmark	all		
	MBR	Tidligere spontane aborter	all		
07_ULST	EUROCAT	TOTPREG	all	set the variable =YES if this number is >=1, NO otherwise	Pregnancies that end in delivery of at least one child with at least 1 congenital anomaly
08_CHUT_EFEMERIS	NAISSANCE	J8_GESTITE	all	set the variable =YES if this number is >=2, NO otherwise	Pregnancies that have started, including those ending in live births but also spontaneous abortion or still birth, or TOPFA
22_ARS	CAP	CONCEP	all	Set the variable =YES if this is 1, NO if this is 2, UNKNOWN if this is 99 or missing	Pregnancies that end in delivery
22_ARS	ABS	NATVIVI	all	Set the variable = UNKNOWN if at least one of the 4 rows contain '99' or missing; otherwise sum the 4	Pregnancies that end in
22_ARS	ABS	NATMORTI	all		

22_ARS	ABS	ABORTI	all	rows and set the variable = 'YES' if the result is >=1, NO, if it is 0	spontaneous abortions
22_ARS	ABS	IVG	all		
22_ARS	IVG	NATVIVI	all	Set the variable = UNKNOWN if at least one of the 4 rows contain '99' or missing; otherwise sum the 4 rows and set the variable = 'YES' if the result is >=1, NO, if it is 0	Pregnancies that end in induced abortions
22_ARS22_ARS	IVG	NATMORTI	Allall		
22_ARS22_ARS	IVG	ABORTI	Allall		
22_ARS	IVG	IVG	all		

### 6.2.2 Study variables based on *Routine healthcare data tables*

In case of variables based on *Routine healthcare data tables* of the CDM, the approach is listed in figure 12



**Figure 12. Step T2 of Figure 11 expanded for *Routine healthcare data tables* of the CDM**

For each variable, three steps are recommended

- T2.1: The diagnostic/therapeutic/procedure codes used to define it are searched in the CDM and the corresponding records are extracted and labelled. As recommended, in order to be identified in a harmonised way across coding systems, codes are grouped semantically in concept sets; for diagnostic codes this is based on the Unified Medical Language System (UMLS) using the Codemapper tool(7, 8).
- T2.2: the concept set datasets are then manipulated at the study subject level, by identifying a pattern (e.g. 'at least one record during a lookback period of 720 days from the index date'); such study-subject level variables are called 'components'
- T2.3: the components are combined, using logical combinations (AND, OR, AND NOT) or using thresholds.

## 6.3 Modular programming of study variables

The templates described in the previous subsection are embedded in modular programs, using existing R libraries or custom-built R packages. As of today, the MDBSTools package is released and available in the GitHub account of ARS (<https://github.com/ARS-toscana/pharmacoepi-repo-public>), containing two R functions.

- CreateConceptSetDataset: this package embeds the template described in 6.2.2 or the output of the Codemapper tool to enable step T2.1 described in section 6.2.3
- MergeFilterAndCollapse: this package embeds step T2.2 described in section 6.2.3

## 7. CDM work for WP2

### 7.1 CDM development

Several discussions were held on the development of the CDM and its implementation in WP2. The implications of the implementation are not yet clear to everyone. Over the past months, attention has been paid to possible solutions like the Data Shield application and the implications of the use of this approach were discussed.

### 7.2 Multiple CDMs

There is a clear distinction between the nature of Spontaneous Reporting Systems (SRS), and the other data characteristics like prospective data collections and data from Teratology Information Services. For the first type of data, there is already a well-established data format (ICH-E2B-R3) and infrastructure for data-exchange (ESTRI-gateway) and central storage (Eudravigilance). From the discussions, it followed that EFPIA partners should stick to this infrastructure and data format for legal reasons. For the other more diverse data types, a dedicated CDM will be developed. With help of WP7 existing data structures will be analysed and the CDM will be designed over the next months. WP7 proposes the same steps as taken for the healthcare data sources.

1. Provide data dictionaries
2. Conduct interview
3. Develop CDM
4. Create ETL design

### 7.3 CDE and CDM

Based on the deliverables of task 2.3: the developments of Core Data Elements required for the prospective collection and follow-up of exposed pregnancies (CDEs), the WP2 CDM will be developed. Since the CDEs have already been defined and accepted, we decided to

study the possibility to populate the CDE fields based on the information collected by various DAPs. If that is possible, it is likely that a CDM could be constructed to support this process. The ability to convert the data into the CDE field, as a proxy for the CDM to be developed is studied in two pilot studies.

#### **7.4 Pilot study 1 - Conversion of data from different type of data sources**

In this study, the possibility to populate the various CDE fields from four DAPs (two EFPIA, two academia) was studied. Participating DAPs were Novo Nordisk (registry) and Novartis (registry), UKTIS and Lareb (DAP for SRS, pREGnant, NLTIS, EURAP). Combined analysis showed that CDE elements in the DAPs were not available (14%), to be derived (25%), available without conversion (50%) and unknown-detailed analysis needed (11%). Large differences between the various DAPs exist. Data present in the EFPIA registries will be easier to convert as compared to the TIS and SRS data.

#### **7.5 Pilot study 2 - Possibility to convert from SRS data**

For the SRS data, for instance those available at PV centre Lareb, it is currently unknown which proportion of the data can be converted to fit the CDE (and thus in the CDM to be developed) and which characteristics predict the quality for conversion. The “quality” being defined as whether the nature of the spontaneous reports allows for this conversion in terms of completeness of data. The primary objective of this study is to analyse the quality of spontaneous reports on exposure to drugs used by pregnant women received by Lareb for conversion towards selected elements of the CDE. The secondary objective for this study is to analyse which characteristics of the reports predict the quality. Results of this pilot study will be available by the end of May 2020. The study might also be a starting point for additional studies into the quality of the data for task 2.5.2.

In summary, the WP2 DAPs do not have a clear view yet on the way the CDM may be implemented and the technical, financial and legal reasons involved. The different type of data sources to be used in WP2 differs in the opportunity to allow for populating the information in the CDE. Over the next months, the data models of the DAPs need to be compared and the CDM should be finalised.

## **8. References**

1. Trifiro G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Intern Med*. 2014;275(6):551-61.
2. Gini R, Sturkenboom MCJ, Sultana J, Cave A, Landi A, Pacurariu A, et al. Different Strategies to Execute Multi-Database Studies for Medicines Surveillance in Real-World Setting: A Reflection on the European Model. *Clin Pharmacol Ther*. 2020.
3. Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf*. 2011;20(1):1-11.

4. Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, et al. Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. EGEMS (Wash DC). 2016;4(1):1189.
5. Dodd C, Pacurariu A, Osokogu OU, Weibel D, Ferrajolo C, Vo DH, et al. Masking by vaccines in pediatric drug safety signal detection in the EudraVigilance database. Pharmacoepidemiol Drug Saf. 2018;27(11):1249-56.
6. Fortier I, Raina P, Van den Heuvel ER, Griffith LE, Craig C, Saliba M, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. 2017;46(1):103-5.
7. Avillach P, Mouglin F, Joubert M, Thiessard F, Pariente A, Dufour JC, et al. A semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European eu-ADR project. Stud Health Technol Inform. 2009;150:190-4.
8. Becker BFH, Avillach P, Romio S, van Mulligen EM, Weibel D, Sturkenboom M, et al. CodeMapper: semiautomatic coding of case definitions. A contribution from the ADVANCE project. Pharmacoepidemiol Drug Saf. 2017;26(8):998-1005.

## Appendix 1. Instructions for task “provide data dictionary”

Task: provide data dictionary

Requestor: XX

Start date: XX

Default time to completion: XX

Dear ConcePTION Data Access Provider,  
you are now requested to provide the data dictionaries of the databases you can access from your organization. The information you provide will be used as a preparation to have a 1-on-1 call with you, where we can have a detailed discussion on the content, and origin of the data, and how to map it to the ConcePTION Common Data Model.

### What you should do



Please collect the following information

1. A short document describing at a high level your organization and the databases your organization has access to and that you plan to include in the Data Characterisation study: just describe the underlying population, the possible reasons why a person enters or exits the database, who is the original data collector, the purpose why this data is collected, and the reason why your organization is entitled to access it; please mention also when the data collection started; find in the attached file **example\_of\_short\_description\_of\_the\_databases** an example.
2. The data dictionary of the databases you have access to; you are free to provide your local documentation, in your original language; in particular, please provide the original names of tables and of variables. Please feel free to send a larger set of tables/variables with respect to those you actually plan to use for the Data Characterisation study. During the 1-on-1 call that we will schedule after this task, we will ask you to indicate which tables/columns you are actually planning to map to the ConcePTION CDM for the Data Characterisation study. Find in the attached file **example\_of\_data\_dictionary.zip** an example.

Whenever one of the documents is ready, please send it to via the Task Management System, as an attachment to a message in the task page.

### How to resolve the task

After you have sent all the documents as attachments in the Task Management System, please access it, enter the page of this task and click on the tab Actions > Resolve, see the Figure below

RT for helpdesk.imi  

New task in management Search...

Display History Basics People Dates Links Jumbo Reminders Actions ☆ ⌚

Reply  
Comment  
Forward  
Open It  
**Resolve**  
Reject  
Delete  
Untake

↑ Dates

Created: Mon Aug 19 16:45:51 2019  
Starts: Mon Aug 19 16:00:00 2019  
Started: Not set  
Last Contact: Not set  
Due: Sun Sep 15 12:00:00 2019  
Closed: Not set  
Updated: Mon Aug 19 16:45:51 2019 by The RT System itself

↑ Links

Graph Gantt Chart

Depends on: (Create)

Depended on by: (Create) • 516: Demonstration Task [new] (Nobody in particular)

Parents: (Create)

Children: (Create)

Refers to: (Create) • 516: Demonstration Task [new] (Nobody in particular)

Referred to by: (Create)

Create Depends on Task in management

Show all quoted text — Show full headers

Reply Comment Forward

### How to ask to reassign the task to someone else

If you wish that another person in your organization execute this task, please write to me in the Task Management System or by replying to this email and I will redirect the task to this person. Please note that this person needs to be registered in the Task Management System.

### How to ask to reschedule the deadline

Please note that I set a deadline for this task. In case you need an extended deadline, don't hesitate to contact me in the Task Management System or by replying to this email. However please note that we need the protocol to be finalized by next week or it won't be possible for many DAPs to submit it in time.

Please contact me with any questions you have regarding this task: you can do so either through the Task Management System or by replying to this email.



## Appendix 2. Details of the ConcePTION CDM v1.0

The ConcePTION CDM v1.0 for electronic health care data comprised the following tables:

**Metadata** – contains information about the data source that describes the data, and can be used to develop characterization programs based upon presence or absence of CDM tables.

**Person** – contains stable information on a person: date of birth, sex at birth, ethnicity.

**ObservationPeriods** – contains information on the follow-up periods for each person with multiple observation periods per person possible.

**Drugs** – contains information on medicines and vaccines prescribed or dispensed to a person.

**Events** – contains information on events characterised by a date and a code belonging to a coding system for each diagnosis, sign, symptom. Each record will contain information on its coding system and on its provenance.

**Encounter** - contains data on the encounter. If it is hospital admission: length of stay, ward of admission, specialty of unit. If it is a visit: type of visit and specialty of the physician or healthcare professional.

**Procedures** – contains information regarding procedures (including measurements) characterised by a date and by a description with a result/outcome and units of measurement if applicable.

**Death** - contains records of death from any source including medical records, death registries, hospital discharge records, etc.

**MotherChild** - contains identifiers for mother-infant linkage and data on methodology used to link each dyad. Also includes data on fathers if available.

**EUROCAT** – contains the EUROCAT table, if one is maintained by the participating database.

Tables provided below provide a high-level description of each CDM table.

**Meta-Data:** In order to have automated procedures to look at the CDM and program, DAPs are asked to fill out the following meta-data table in the following format:

### Metadata

Variable	Format	Mandatory
Data_source_name	Character	Yes
Data_access_provider	Character	Yes
DateDrawDown	Character yyyyymmdd	Yes
DateLastUpdate	Character yyyyymmdd	Yes

Provenances	Character	Yes
IsPresORDisp	Character	Yes
IsIndication	Binary	Yes
IsEUROCAT	Binary	Yes
IsDeath	Binary	Yes
IsEncounters	Binary	Yes
IsVaccineRecords	Binary	Yes
IsProcedure	Binary	Yes
IsDiagnoses	Binary	Yes

**Person:** all fields to be filled for the study population (M/F up to and including 55 years of age)

### Person

Variable	Format	Mandatory
PersonID	Character	Yes
DateBirth	Character yyyyymmdd	Yes
SexAtBirth	String	Yes
Ethnicity	String	No
CountryOfBirth	String	No

**Observation Periods:** All fields for each person in the study population and its periods of follow-up as well as the provenance of the data on follow-up variables.

### Observation Periods

Variable	Format	Mandatory
PersonID	Character	Yes
Start_followup	Character yyyyymmdd	Yes
End_followup	Character yyyyymmdd	Yes
Provenance	Character	Yes

### Drugs

Variable	Format	Mandatory
PersonID	Local code string	yes
DateDrug	Character yyyyymmdd	Yes
DrugCode	Character	Yes
CodeTypeDrug	Character	Yes
Vactype	Character	No
BrandDrug	Character	Yes
AmountDrug	Numeric	Yes
Amount_Unit	Character	Yes
StrengthIngredient_1	Numeric	Yes
StrengthIngredient_2	Numeric	No
Units_day	Character	No
DDD_value	Numeric	No
ProductCode	Character	Yes

DoseRecordedVaccine	Character	No
CodeIndication	Character	No
CodeTypeIndication	Character	No
IsType	Character	Yes
Provenance	Character	Yes
Prescriber	Character	No

### Events

Variable	Format	Mandatory
PersonID	Character	Yes
DateEvent	Character yyyymmdd	Yes
CodeEvent	Character	Yes
CodeTypeEvent	Character	Yes
Provenance	Character	Yes

**Encounters:** Encounters are all medical visits in medical health care facilities (hospitals, GP, outpatient clinics).

### Encounters

Variable	Format	Mandatory
PersonID	Character	Yes
EncounterID	Character	Yes
StartDate	Character yyyymmdd	Yes
EndDate	Character yyyymmdd	Yes
Provenance	Character	Yes

**Procedures:** A procedure is a course of action intended to achieve a result in the delivery of care. A procedure with the intention of determining, measuring, or diagnosing a patient condition or parameter is also called a measurement or test.

### Procedures

Variable	format	Mandatory
PersonID	Local code string	Yes
DateProcedure	Character yyyymmdd	Yes
CodeProcedure	Character	Yes
CodeTypeProcedure	Character	Yes
MeasurementResult	Character	No
ResultUnits	Character	No
Provenance	Character	Yes

### Death:

For each person who has a death record in any of the data sources, one or more records may be filled with information on the date, cause and provenance of the data. Note: one

person may have more sources of death data. Each source should include the provenance (origin). DAPs may also have a cause of death in one database, and the date in another. If a year, month, or date associated with a record of death is unavailable, the record should still be included with a missing date.

### Death

Variable	Format	Mandatory
PersonID	Local code string	Yes
DateDeath	Character yyyymmdd	Yes
CauseOfDeath	Cause of death	No
IsUnderlyingCause	Character	No
CodeType	Character	No
Provenance	Character	Yes

### MotherChild:

In ConcePTION the mother child linkage is very important. For those who can identify the father, the father's identifier should also be provided. DAPs should provide linkage information for children (0-18 years), their mothers, and their fathers if available using the following format:

### MotherChild

Variable	Format	Mandatory
MPersonID	Character	Yes
FPersonID	Character	No
CPersonID	Character	Yes
InfantProvenance	Character	Yes
MValidLink	Character	Yes
FValidLink	Character	No

## Appendix 3. Details of ConcePTION CDM v2.0

### Metadata tables

- The metadata table from v1 has been split in three: two are equal to corresponding tables in the OMOP CDM, the third, INSTANCE, is meant to document in a machine-readable way the local data that is mapped to the current instance of the ConcePTION CDM.
- PRODUCT\_CODE was added as an additional metadata table meant to document medicinal-product specific data which is linked to the MEDICINES and VACCINES tables. This should be used as much as possible, especially for combination products (product with 2 or more active ingredients). In case product code is absent for a combination product two alternatives are possible
  - Generate an ad-hoc product code
  - Generate two different rows, each one corresponding to a single ingredient (ATC5 level), including strength and regimen.

### Curated tables

- PERSONS has been classified as a derived table, with one row per person who is included in the instance; variables recorded here are stable at the date of instance creation. The DEATH table is discarded, and causes of death are included in the EVENT table. In the PERSON table, a unique date of death, date of birth and sex are created by the DAP according to an algorithm stored in the ETL specification document. Dates will be represented by three separate fields (year, month, date) of which only year is mandatory. It is intended that persons who are not in PERSON but are in the data should never be considered in any study. We leave race, but we will double check if this is collected by any DAP, and if no we will discard it.
- OBSERVATION\_PERIODS has been classified as a derived table. It has multiple rows per person corresponding to each period during which the person was considered as under observation according to the DAP.
  - Information about the same person may refer to different levels of inclusion in the data
    1. Inclusion in the underlying population
    2. Inclusion in the data source population
    3. Inclusion in the instance population (which is recorded at the instance level in the INSTANCE table)
- MOTHER\_CHILD\_LINKAGE remains as described in ConcePTION CDM v1. It is a derived table describing linkages between mothers and their infants using the algorithm employed by the DAP.

### Routine Healthcare data

- DRUGS. Dispensed and prescribed medicines will be recorded in this table. Strength will be encoded here in data sources where it is recorded as prescribed (for both prescribed and dispensed medicines)

- **VACCINES:** Dispensing's, prescriptions, or administrations of vaccines. Vaccines have been kept separate from drugs in order to allow for vaccine-specific data such as dose number, vaccine type at the antigen level if available, and lot number.
- **EVENTS** records observations pertaining the data domain of diagnosis, which also encompasses symptoms. We would like to keep two dates for the case of diagnosis recorded during hospital admissions, to be practical during data processing. Two dates are envisioned for each record 'start date' and 'end date', but they will probably be both compiled only if the event is recorded during a hospitalisation.
- **PROCEDURES.** The decision was made to separate procedures (such as surgeries, or diagnostic procedures, rehabilitation procedures, therapeutical procedures) from measurements. Procedures which are coded with diagnostic codes are recorded in the 'EVENT' table, and then when they are included in a concept set they will pertain to a 'special' data domain and queried accordingly.
- **MEASUREMENTS:** The decision was made to separate measurements from procedures. Procedures producing a result with or without units may be stored in the MEASUREMENTS table if this is applicable to the data held by a DAP.
- **PATHOLOGY:** data from pathology reports are characterised by information about topography (site of tumour) and morphology of the sample, as well as by textual description; this was deemed incompatible with the structure envisioned for the MEASUREMENTS table and a different table was therefore envisioned
- **VISIT\_OCCURRENCE**  
In line with the principle of adhering to OMOP conventions, it was decided that the CDM should incorporate a table to record visit occurrences for those data sources in which observations occurring during a defined encounter such as a hospitalization could be linked with each other.

## Surveillance tables

- **EUROCAT:** The EUROCAT table remains unchanged from ConcePTION CDM v1.0. It is a well-defined and widely used structured data table and will remain in its original format for those DAPs which have access to this table.
- **SURVEY\_ID:** The SURVEY\_ID table was added in combination with SURVEY\_OBSERVATION. This table records subjects recorded in surveillance data which has been mapped to the SURVEY\_OBSERVATION table.
- **SURVEY\_OBSERVATION:** The SURVEY\_OBSERVATION table was added in combination with SURVEY\_ID. This decision was motivated by availability of pregnancy registries in a number of data sources held by DAPs as well as the availability of unique surveillance-based tables such as records of well-child visits in early childhood with enriched data on growth, nutrition, and development. The diversity of these tables led to a decision to include them in the CDM in their original format.





## Appendix 4. Answer sheet for interview with DAP

This template must be filled out by the interviewers around one week before the interview, by

- (1) filling out all the spots marked in yellow,
- (2) replicating the questionnaire in Stage 2 for every table submitted by the DAP
- (3) adapting questions in Stage 4 in such a way that questions asked in Stage 2 are not repeated.

The resulting document must be sent to the DAP one week before the interview for the DAP to fill out the questions in Stage 2 and be prepared to address questions in Stage 4

### Answer sheet

#### Interview to bridge between Data Dictionary and ConcePTION CDM

The **main objective** of the interview is to explore and obtain a documented understanding of the local data that has the potential of being used in ConcePTION, in order to support the design of the Extract, Transformation and Load (ETL) procedure to the ConcePTION CDM, and to facilitate correct interpretation of the results

#### Date

*Day, time*

#### Interviewers

- Conductor: XXXXX
- Assistant: XXXXX

#### DAP

Organization: XXXX

#### People:

- XXXXX
- XXXXX

### Main stages

1. Introduction of the main objective
2. Go through the answers to the 'table questionnaire' for the DC tables, focus on capturing assumptions and 'what is not' in the data
3. Please access the Catalogue questionnaire and fill it out; mark in this sheet the questions that are not clear to you
4. Please check question in a specific list may have been missed during Stage 2 or in the Catalogue
5. Is there any additional data you would like to describe?

## Stage 1: Introduction of the main objectives

## Stage 2: Table questionnaire

### Table XXXX (repeat for each table)

- 1) *What triggers the creation of a record of the table?*
- 2) *Is the table collected for all the population of your database, or only for a subpopulation?*
- 3) *Can you comment on the completeness and quality of the table? If you don't have formal measurements, feel free to convey the assumptions you commonly make*
- 4) *What is the time span of the table, how often it is refreshed, and which is the lag time between the data creation and the time when the data has the potential of being available to your organization?*
- 5) *Include other comments you may want to share about this table*
- 6) *Fill out the table below with the names of the variables of this table (as listed in the data dictionary) that you plan to map to the ConcePTION Common Data Model, with a description in English of the meaning of the variable, the name of the classification used (e.g. CIM10, ATC, ..., or national/local) or the description in English of the data dictionary if a small number of values are included in the dictionary, and any comment you may want to share about that variable (when it is missing, or miscoded, or when its content is non reliable)*

Original name	Meaning	Data dictionary in English (if useful)	Comment

**Stage 3:** Please access the Catalogue questionnaire and fill it out; mark in this sheet the questions that are not clear to you

**Stage 4: please check what of the following information may have been missed during Stage 2 or in the Catalogue**

- Dates
  - Do you have exact or approximated date of birth? If you have multiple and conflicting recordings if this information, how do you resolve the ambiguity?
  - How do you define date of entrance/exit from the database?
  - Do you have exact or assumed date of death? From which source(s)? With which delay? If you have multiple and conflicting recordings if this information, how do you resolve the ambiguity?
- From which tables are the various pregnancy outcomes captured, and with which algorithm? And in each case, what is the algorithm to define start and end of pregnancy??
  - Live birth
  - Still birth
  - Termination for undefined reasons
  - Termination for fetal anomaly
  - Termination for other medical reason
  - Spontaneous abortion
- Is pregnancy captured in alternative ways wrt to its outcomes? (e.g. from hospital admissions during pregnancy, from specialist or primary care visits during pregnancy...). How do you estimate start and end date of pregnancy in those cases?
- From which tables is breastfeeding captured, and with which algorithm?
- Drugs is there any additional table where information referred to utilisation of medications is collected?
  - Dispensing? or prescription?
  - In the case of dispensing data, are there specific national rules to know?
  - Refill?
  - Drugs dispensed in community pharmacies for domiciliary use/ dispensed in hospital pharmacies for outpatient use / prescribed to outpatients for ambulatory administration / inpatients/prescribed by general practitioners/prescribed by specialists/therapeutic procedures?
  - Brand?
  - Batch number?
  - Diagnoses –
    - mental care
    - exemptions from copayment
    - pathology register
    - disease register
    - birth register
    - other?
- Diagnostic procedures
  - Coding system?

- Prescribed/dispensed?
  - Results? (bioimaging/ procedures)
- Rehabilitation procedures
- Ethnicity

**Stage 5: Is there any additional data you would like to describe?**

## Appendix 5. Results from analysis of interviews for the purpose of updating the CDM

### Would there be a need for additional tables?

#### 08\_CHUT

In EFEMERIS (CHUT) many children (and the corresponding pregnancies) are observed at 8 days, 9 months and 24 months, with structured questionnaires. It would be possible to ETL this information to the Measurements table, but I would suggest to a different solution: **create a separate table Questionnaire**. The main motivation is that in a questionnaire the date when the information is recorded is definitely different from the day when the event took place, so the semantics of 'date' is different. The second motivation is that a questionnaire is by definition a primary data source, and this is a second semantic difference. Moreover, we may want to capture characteristics of the questionnaire in the CDM somehow, and 'squeezing' the information in the columns that exist already, or create additional columns that would be empty for rows which are not answers to questionnaires, sound inappropriate to me.

Original name	Meaning	Data dictionary in English (if useful)	Comment
QUESTION		Free text	
QUESTION_SUMMARY	Just one-two words to represent the content of the question		
QUESTION_CODE	Identifies questions across different DAPs/questionnaire	Will only be available for some questions (e.g. BIRTH WEIGHT)	
EXTERNAL_KEY	External key to a meta questionnaire table, with date and place of collection?		
DATE	Date when the information was collected		
WHO	Who is answering	Physician, other healthcare provider, patient, parent...?	
ANSWER	If the answer is free text	Free text	
ANSWER_CODE	If the answer is a code		

ANSWER_DESCRIPTION	A description of the answer code		
ANSWER_CODING	Which is the data dictionary of the answer, if any		
ANSWER_VALUE	If the answer is a measurement: value		
ANSWER_UNIT	If the answer is a measurement: unit		

## 15\_BIPS

Should we consider Exemption from healthcare services payment as additional table or added as additional category among provenances? In BIPS data from this table will be not used to feed the CDM because of local inaccuracy and wide lag time of the information recorded. However, I expect that it should be taken in consideration for other data sources that has this table and actually use the info recorded in it (e.g. ARS)

## 18\_FISABIO

Often in ConcePTION, the DAP obtains datasets from another organization(s), which is linking large population-based tables with a specific cohort (e.g. the EUROCAT table; the birth registry; or a pregnancy registry), and cutting only the rows of the large tables whose subjects match the cohort.

Moreover the tables can be cut per

- Columns available
- Rows available, selected by
  - o Timespan around a certain date or in a certain calendar period
  - o Families of codes

The same DAP may access larger or stricter datasets according to the protocol or other conditions (type of funding...?). This is the case of FISABIO and others.

Therefore, this information should be captured somehow in each specific instance of the CDM, so there should be a place in the CDM data model to capture this: in the PERSON table? or in the METADATA table? or in an additional table? And how?

## 19\_SIDIAP

Possibly 'Observations'. This is triggered by the fact that SIDIAP collects gestational age and breastfeeding as an observation. In SIDIAP and in other data sources, these types of observations may be difficult to capture as procedures/measurements as they may not have an associated procedure code.

SIDIAP provides additional justification for an additional ‘Survey’ or ‘Questionnaire’ table. SIDIAP includes a ‘Healthy child program’ table containing longitudinal growth and development data on all children in the national health program. Much of the data in this table is in free text fields.

## **20\_FERR**

Currently the table for capturing EUROCAT and EUROMediCAT data in the ConcePTION CDM is the EUROCAT table. This is a slightly larger table than the EUROMediCAT table. The EUROMediCAT table is a subset of the EUROCAT table. We can keep the EUROCAT table in the CDM or limit it to the EUROMediCAT table.

## **21\_CNR-IFC**

The EUROCAT table may have some local additional columns, this is the case in CNR-IFC: should we include them in the EUROCAT table? Or as separate rows in a ‘questionnaire’ table (see suggestion in 08\_CHUT)?

## **24\_MALTA**

Not in the CDM. However, there seems to be a need for something like a ‘meta-meta data’ table to record items like year of digitization, public vs. private sector, availability of abortion services, etc. Relevant to EUROCAT sources is whether women have access to abortions (In Malta they do not) as this will impact rates of both TOPFA and observed anomalies.

Additional metadata of interest is whether data is drawn solely from public sector care (as is the case in Malta) or from public and private sector care, and whether the populations served in each are likely to differ in important ways.

Capture of cases seems to differ from one EUROCAT registry to another. For EUROCAT registries, it may be beneficial to record whether minor & major cases are captured (EUROCAT policy only requires capture of major anomalies <https://eu-rd-platform.jrc.ec.europa.eu/eurocat/data-collection/guidelines-for-data-registration>).

**NOTE:** Malta is able to link to a mortality registry so we may consider including this in their ETL.

## **27\_Otto**

Should we consider to use a separate table for questionnaire-based data sources? Or should we consider questionnaires as a provenance category only?



## **Would there be a need for additional columns?**

### **08\_CHUT**

Columns to describe answers to questionnaires need to be created (either in the new table as suggested above, or in the 'measurement' table)

Add a column to all tables to indicate the originating table where the info in the row was stored (as described in the catalogue)? Each original table is mapped to one (or more than one?) provenance and/or type of data source

### **19\_SIDIAP**

- Possibly number of packages dispensed (prescription table)
- Possibly for identification of primary vs. secondary diagnoses in hospital discharge data.

### **21\_CNR-IFC**

There is something we may want to capture in the EUROCAT table: different partners may have different rules, methodologies, institutional networks... to fill out the same columns. For instance CNR-IFC has a network of contact points in each relevant unit in each Tuscan hospital, and this person is in charge of filling out the questionnaire, possibly by interviewing the mother and accessing the medical records. Some fields however are coded by CNR-IFC based on free text fields. Should we capture that, and how?

### **24\_MALTA**

The source of the case is not included in the EUROCAT table. Cases can be ascertained through medical record review or reporting. Possibly provenance of gestational age data. This can be from ultrasound, LMP, maternal self-report, etc and this choice is not recorded. Cases are also validated using different sources such as hospital records, imaging, ECHO reports. The data source(s) used for validation would be interesting to have.

### **27\_Otto**

Add a column to all tables to indicate the originating table/provenance where the info in the row was stored is probably a more flexible solution than adding a table for questionnaire-based data sources. This malformation registry, for instance, contains data from parents' self-reported information or, in case parents' consent is not provided, basic mandatory info on the malformed baby are reported by the concerned physician also using available medical records.

**Would there be a need for additional values for some existing column?**

## **08\_CHUT**

Both questions and answers to the questionnaire need to be recorded somehow in the target table, the coding of the answers is in the original data, but how should we best capture the question? We may both capture a synthetic name and the text describing it?

## **15\_BIPS**

Exemption from healthcare services payment

- **in particular for provenance?**

## **08\_CHUT**

The date of LMP I CHUT\_EFEMERIS is collected during pregnancy whenever the woman accesses a pharmacy. I suggest we **label this provenance** and compare this measure of LMP with the measure obtained at pregnancy outcome. This may inform the validity of the LMP in other databases (and specifically in BPE).

## **11\_UMCG**

If we decide to include gestational age as a measurement, the provenance for this should be included.

## **23\_ARS**

Hospital discharge records with a 'death' as discharge cause

## **24\_MALTA**

- for gestational age (source/method for this data)
- **for type of data source?**

## **08\_CHUT**

What is the hierarchy between type of data source and provenance? Are they independent or is one a reclassification of the other?

## **11\_UMCG**

We will not get the underlying data from EUROCAT registries directly, but parental questionnaire may be relevant for some data sources (this could be additional support for the proposed 'survey' table).

## **24\_MALTA**

Possibly 'Stimulated reporting' or something to that effect

## **27\_Otto**

Questionnaire-based registry?

### What are the coding systems adopted in the DAP?

- **International Coding systems:** ICD10, ATC, ICD9ISCO-88, OMIM, BPA (stands for British Paediatric Association coding system which provides a more granular information compared to ICD10)
- **Local coding systems:**
  - 08\_CHUT\_EFEMERIS CIP (Presentation (?) Identifying Code: marketing authorisation in France?),
  - 11\_UMCG 'Wooncode' from central bureau of statistics to identify region of residence of the mother (EUROCAT variable RESIDMO)
  - 15\_BIPS: ICD10 GM, Operation and Procedure Code (OPS), EBM for accounted treatment during encounters
  - 18\_FISABIO: SKU\_PRODUCTO (Spanish coding system for drugs?)
  - 19\_SIDIAP: ECAP (diagnoses as entered by healthcare professionals)
  - 21\_CNR-IFC: same as ARS and FERR

## Notes

### 11\_UMCG

There may be the need for a structured way to store meta-data outside of the CDM. For example, the NNL EUROCAT registry requires parental consent, which is different from other EUROCAT registries. From 2010 they have been allowed to record minimal information without consent. This is captured in the interview document but no formal template for recording of this type of meta-data exists. Other examples of this type of meta-data: Switch from ICD-9 to ICD-10 in 2002

## Appendix 6. ETL template v1.0



Template of ETL Specification  
v1.0

***ConcePTION: template of an ETL specification***

***Author of the template:*** Rosa Gini, Miriam Sturkenboom, Caitlin Dodd, Vjola Hoxhaj, Nicolas Thurin, Giuseppe Roberto.

***Version of the template:*** 0.1

***Date:*** 31 March 2020

***Version of the template:*** 0.2

***Date:*** 15 April 2020

***Version of the template:*** 0.3

***Date:*** 20 April 2020

***Version of the template:*** 0.4

***Date:*** 21 April 2020

***Version of the template:*** 0.6

***Date:*** 30 April 2020

***Version of the template:*** 1.0

***Date:*** 3 May 2020

## Instructions to complete the ETL specification of a data source

- Create a copy of this file and delete the preamble, up to the instruction pages
- Edit page 1 and include the name of your organization and your name(s), as well as date and version
- Edit section 2 by adding the list of the origin tables in your data source
- Edit Section 3 by indicating:
  - For each origin table of your data source, which target table(s) of the ConcePTION CDM it feeds
  - Conversely, for each target table of the ConcePTION CDM, which origin tables of your data source are loaded in that table
- Edit Section 4 by creating for each target table as many subsections as the origin tables you are loading in it. For each table, origin table, indicate which records of the origin table are loaded in the target, indicate whether one or more target records are created, and add a specification table, see below the instructions.

## Specification tables

A specification table is associated to a target table (of the CDM) and an origin table (of your local data source) as follows

Target table: EVENTS			
Origin table:			
Target column	Origin column	Rule	Notes
person_id			
visit_occurrence_id			
visit_start_date			
visit_end_date			
specialty_of_visit			
specialty_of_visit_vocabulary			
status_at_discharge			
status_at_discharge_vocabulary			
meaning_of_visit			
origin_of_visit			

The '**Target column**' contains the list of the CDM columns of the target table you are specifying (in this example, EVENTS); for each of them, you must specify the name(s) of your origin column(s) that will feed the target column, in the '**Origin column**', and/or the rule that will create the content, in **Rule**; the rule may be a simple string.



To decide which column(s) goes where, and which rules you should adopt, please use the description of the target table contained in the CDM table specifications [at this link](#) (also in the member area of the Project Website [at this link](#)), and in particular the ‘description’ and ‘conventions’ specifications.

A shaded background indicates that the values you set for that column must belong to the CDM vocabulary: to pick one of the allowed values, visit the CDM vocabulary specifications [at this link](#) (also in the member area of the Project Website [at this link](#)). If you don’t find a value that fits your data well, please read the subsection ‘Suggest updates to the ConcePTION CDM vocabulary’ below.

### Suggest updates to the ConcePTION CDM vocabulary

If the target column with CDM vocabulary does not list a value that fits your data well, please access [this document](#) to suggest updates to the CDM vocabulary. Open the tab corresponding to the vocabulary you suggest to update, see for instance below the tab of ‘meaning\_of\_visit’ in the VISIT\_OCCURRENCE table.

Table: VISIT_OCCURRENCE					
Variable: meaning_of_visit					
value	description	Comments	who_added	from_which _DAP	from_which_table
hospitalisation	hospitalisation with an assigned bed				
hospitalisation_not_overnight	hospitalisation where there is no overnight stay foreseen	may not be deduced by start-end dates because it may actually last longer			
outpatient_specialist_visit	visit with a specialist, outside of a hospitalisation				

Add the new value in **value**, describe it in **description**, and indicate your name, your DAP and the origin table that triggered your suggestion.



# concePTION

## SAFETY EVIDENCE ECOSYSTEM

### ConcePTION: ETL specification

**DAP:** ...

**Data source:** ...

**Version of this document:** ...

**Author of this document:** ...

**Date of this document:** ...

#### Contents

Introduction

1. The ConcePTION CDM
2. The data dictionary of this data source
3. Course of action
  - 3.1. Origin tables and their target tables
  - 3.2. Target tables and their origin tables
4. Target tables and their origin tables: actions and specification tables
  - 4.1. Routinary healthcare data
    - VISIT\_OCCURRENCE
    - EVENTS
    - MEDICINES
    - PROCEDURES
    - VACCINES
    - MEDICAL\_OBSERVATIONS
  - 4.2. Surveillance
    - EUROCAT
    - SURVEY\_ID
    - SURVEY\_OBSERVATIONS
  - 4.3. Curated tables
  - 4.4. Metadata

Annex. General concepts

## Introduction

This document describes the procedure to Extract, Transform and Load (ETL) an origin data source to the Conception CDM, the target source.

The document has two purposes

1. It serves as a guidance for the programmers who implement the ETL specifications into a computer program
2. It serves as a reference for investigators to understand the origin of the data they find in the CDM, to design their study and to interpret their results.

This document refers to the following sources

- The ConcePTION CDM table specifications document, last version, which is available in the Project Website [at this link](#) in the member area
- The ConcePTION CDM vocabulary specifications document, last version, which is available in the Project Website [at this link](#) in the member area
- The description and data dictionary of the origin data source, which are available in the Project Website, in the DAP's folder of [this page](#) of the member area.

## 1. The ConcePTION CDM v2.01

The ConcePTION CDM v 2.01 is composed by the following tables

A) Routine healthcare data

- VISIT\_OCCURRENCE
- EVENTS
- MEDICINES
- PROCEDURES
- VACCINES
- MEDICAL\_OBSERVATIONS

B) Surveillance

- EUROCAT
- SURVEY\_ID
- SURVEY\_OBSERVATIONS

C) Curated tables

- PERSONS
- OBSERVATION\_PERIODS
- PERSON\_RELATIONSHIPS

D) Metadata

- PRODUCTS
- CDM\_SOURCE
- METADATA
- INSTANCE

The tables and the vocabulary of the derived variables are described, respectively, at [this link](#) and at [this link](#).

Some details are specified in the following Table

Section	Table	Role
Routine healthcare data	VISIT_OCCURRENCE	This table contains a summary description of the visits during which records of EVENTS, PROCEDURES, but possibly also MEDICAL_OBSERVATIONS or VACCINES or MEDICATIONS were recorded. This serves both to collect visit-level information, and to enable grouping sets of records that were recorded concurrently
Routine healthcare data	EVENTS	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.
Routine healthcare data	MEDICINES	This table collects data on drug prescriptions, dispensings or administrations occurred during routine healthcare.
Routine healthcare data	PROCEDURES	This table collects procedures administered during routine healthcare. Can be a surgery, or a diagnostic procedure, a rehabilitation procedure, a therapeutical procedure...
Routine healthcare data	VACCINES	This table collects dispensations or administrations of vaccines.
Routine healthcare data	MEDICAL_OBSERVATIONS	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
Surveillance	EUROCAT	This table collects surveillance data on congenital anomalies, following the EUROCAT standard
Surveillance	SURVEY_ID	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
Surveillance	SURVEY_OBSERVATIONS	List of observations in a survey
Curated tables	PERSONS	This table records persons that are to enter analysis of this instance of the CDM
Curated tables	OBSERVATION_PERIODS	Periods during which data is collected in the datasource for this person. This table contributes to defining the datasource population

Curated tables	PERSON_RELATIONS HIPS	For any person, this table collects the pairing with the identifier of mother or of other relationships that may be available
Metadata	PRODUCTS	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
Metadata	CDM_SOURCE	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 datasource
Metadata	METADATA	This table contains some general information about how the local 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 tables that are expected to be findable can indeed be found; and that the coding systems that are observed in the data are indeed those listed here)
Metadata	INSTANCE	This table displays the list of the tables and columns of the local 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 local data dictionary that are included in the current ETL document.

## 2. The data dictionary of this data source

This data source contains the following tables

The data dictionary of this data source is described in the Project Website, in the DAP's folder of [this page](#) of the member area.



### 3. Course of action

#### 3.1 Origin tables and their target tables

The course of action of the ETL procedure is as follows: for each origin table, all the target tables are populated. The CDM target tables VISIT\_OCCURRENCE and SURVEY\_ID, whenever they are associated to a origin table, must be populated before the other targets, because the identifiers visit\_occurrence\_id and survey\_id, respectively, must be created first, and then reused in the other target tables.

Origin table	First target table	Other target tables
...	...	...
...	...	...
...	...	...
...	...	...

### 3.2 Target tables and their origin tables

As a consequence, the target tables of the CDM are fed by the origin tables as follows

Target table	Origin table(s)
VISIT_OCCURRENCE	...
EVENTS	...
MEDICINES	...
PROCEDURES	...
VACCINES	...
MEDICAL_OBSERVATIONS	...
EUROCAT	...
SURVEY_ID	...
SURVEY_OBSERVATIONS	...
PERSONS	...
OBSERVATION_PERIODS	...
PERSON_RELATIONSHIPS	...
PRODUCTS	...
CDM_SOURCE	...
METADATA	...
INSTANCE	...

The specification tables that illustrate how each source tables must be used to populate the ConcePTION CDM target tables are listed in section 4

## 4. Target tables and their origin tables: actions and specification tables

### 4.1 Tables of healthcare data

#### VISIT\_OCCURRENCE

The origin tables feeding this target CDM table are: ...

For each record

Target table: VISIT_OCCURRENCE			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
visit_occurrence_id			
visit_start_date			
visit_end_date			
specialty_of_visit			
specialty_of_visit_vocabulary			
status_at_discharge			
status_at_discharge_vocabulary			
meaning_of_visit			
origin_of_visit			

## EVENTS

The origin tables feeding this target CDM table are: ...

Target table: EVENTS			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
start_date_record			
end_date_record			
event_code			
event_record_vocabulary			
text_linked_to_event_code			
event_free_text			
present_on_admission			
meaning_of_event			
origin_of_event			
visit_occurrence_id			

## MEDICINES

The origin tables feeding this target CDM table are: ...

Target table: MEDICINES			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
date_dispensing			
date_prescription			
disp_amount_drug			
disp_amount_drug_unit			
presc_units_per_day			
presc_duration			
product_lot_number			
product_code			
product_ATCcode			
code_indication			
code_indication_vocabulary			
meaning_of_drug_record			
origin_of_drug_record			
prescriber_type			
visit_occurrence_id			

## PROCEDURES

The origin tables feeding this target CDM table are: ...

Target table: PROCEDURES			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
procedure_date			
procedure_code			
procedure_code_vocabulary			
visit_occurrence_id			
meaning_of_procedure			
origin_of_procedure			

## VACCINES

The origin tables feeding this target CDM table are: ...

Target table: VACCINES			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
vx_record_date			
vx_admin_date			
vx_atc			
vx_type			
vx_text			
product_code			
origin_of_vx_record			
meaning_of_vx_record			
vx_dose			
vx_manufacturer			
vx_lot_num			
visit_occurrence_id			

## MEDICAL\_OBSERVATIONS

The origin tables feeding this target CDM table are: ...

Target table: MEDICAL_OBSERVATIONS			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
mo_date			
mo_code			
mo_code_vocabulary			
mo_source_table			
mo_source_column			
mo_source_value			
mo_unit			
mo_meaning			
mo_origin			
visit_occurrence_id			



## 4.2 Surveillance

### EUROCAT

The origin tables feeding this target CDM table are: ...

### SURVEY\_ID

The origin tables feeding this target CDM table are: ...

Target table: SURVEY_ID			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
survey_id			
observation_date			
meaning_of_survey			

## SURVEY\_OBSERVATIONS

The origin tables feeding this target CDM table are: ...

Target table: SURVEY_OBSERVATIONS			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
so_date			
so_source_table			
so_source_column			
so_source_value			
so_unit			
survey_id			

### 4.3 Curated tables

#### PERSONS

The origin tables feeding this target CDM table are: ...

Target table: PERSONS			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
date_birth			
date_death			
sex_at_instance_creation			
race			
country_of_birth			
quality			

## OBSERVATION\_PERIODS

The origin tables feeding this target CDM table are: ...

Target table: OBSERVATION_PERIODS			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
op_start_date			
op_end_date			
op_origin			
op_meaning			

## PERSON\_RELATIONSHIPS

The origin tables feeding this target CDM table are: ...

Target table: PERSON_RELATIONSHIPS			
Origin table: ...			
Target column	Origin column	Rule	Notes
person_id			
related_id			
origin_of_relationship			
meaning_of_relationship			
method_of_linkage			

## 4.4 Metadata

### PRODUCTS

The origin tables feeding this target CDM table are: ...

Target table: PRODUCTS			
Origin table: ...			
Target column	Origin column	Rule	Notes
product_code			
full_product_name			
box_size			
box_size_unit			
drug_form			
route_of_administration			
product_ATCcode			
ingredient1_ATCcode			
ingredient2_ATCcode			
ingredient3_ATCcode			
amount_ingredient1			
amount_ingredient2			
amount_ingredient3			
amount_ingredient1_unit			
amount_ingredient2_unit			
amount_ingredient3_unit			
product_manufacturer			

**CDM\_SOURCE**

Fill as follows

Target column	Origin column	Rule	Notes
data_access_provider_code			
data_access_provider_name			
data_source_name			
data_dictionary_link			
etl_link			
cdm_version			
cdm_vocabulary_version			
instance_number			
date_creation			

## METADATA

Fill out the table

type_of_metadata	tablename	columnname	other	values
presence_of_table	VISIT_OCCURRENCE			
presence_of_column	VISIT_OCCURRENCE	visit_end_date		
presence_of_column	VISIT_OCCURRENCE	specialty_of_visit		
presence_of_column	VISIT_OCCURRENCE	specialty_of_visit_vocabulary		
presence_of_column	VISIT_OCCURRENCE	status_at_discharge		
presence_of_column	VISIT_OCCURRENCE	status_at_discharge_vocabulary		
presence_of_table	EVENTS			
presence_of_column	EVENTS	event_code		
presence_of_column	EVENTS	text_linked_to_event_code		
presence_of_column	EVENTS	event_free_text		
presence_of_column	EVENTS	present_on_admission		
presence_of_column	EVENTS	visit_occurrence_id		
presence_of_table	MEDICINES			
presence_of_column	MEDICINES	date_dispersing		
presence_of_column	MEDICINES	date_prescription		
presence_of_column	MEDICINES	disp_amount_drug		
presence_of_column	MEDICINES	disp_amount_drug_unit		
presence_of_column	MEDICINES	presc_units_per_day		
presence_of_column	MEDICINES	presc_duration		
presence_of_column	MEDICINES	product_code		
presence_of_column	MEDICINES	code_indication		
presence_of_column	MEDICINES	code_indication_vocabulary		
presence_of_column	MEDICINES	prescriber_type		



presence_of_column	MEDICINES	visit_occurrence_id		
presence_of_column	MEDICINES	product_lot_number		
presence_of_table	PROCEDURES			
presence_of_column	PROCEDURES	visit_occurrence_id		
presence_of_table	VACCINES			
presence_of_column	VACCINES	vx_record_date		
presence_of_column	VACCINES	vx_admin_date		
presence_of_column	VACCINES	vx_atc		
presence_of_column	VACCINES	vx_type		
presence_of_column	VACCINES	vx_text		
presence_of_column	VACCINES	product_code		
presence_of_column	VACCINES	meaning_of_vx_record		
presence_of_column	VACCINES	vx_dose		
presence_of_column	VACCINES	vx_manufacturer		
presence_of_column	VACCINES	vx_lot_num		
presence_of_column	VACCINES	visit_occurrence_id		
presence_of_table	MEDICAL_OBSERVATIONS			
presence_of_column	MEDICAL_OBSERVATIONS	mo_code		
presence_of_column	MEDICAL_OBSERVATIONS	mo_record_vocabulary		
presence_of_column	MEDICAL_OBSERVATIONS	mo_source_table		
presence_of_column	MEDICAL_OBSERVATIONS	mo_source_column		
presence_of_column	MEDICAL_OBSERVATIONS	mo_unit		
presence_of_column	MEDICAL_OBSERVATIONS	visit_occurrence_id		
presence_of_table	EUROCAT			
presence_of_column	EUROCAT	death_date		
presence_of_column	EUROCAT	datemo		
presence_of_column	EUROCAT	bmi		
presence_of_column	EUROCAT	residmo		
presence_of_column	EUROCAT	totpreg		
presence_of_column	EUROCAT	condisc		
presence_of_column	EUROCAT	firstpre		

presence_of_column	EUROCAT	sp_firstpre		
presence_of_column	EUROCAT	karyo		
presence_of_column	EUROCAT	sp_karyo		
presence_of_column	EUROCAT	gentest		
presence_of_column	EUROCAT	sp_gentest		
presence_of_column	EUROCAT	pm		
presence_of_column	EUROCAT	presyn		
presence_of_column	EUROCAT			
presence_of_column	EUROCAT	premal1		
presence_of_column	EUROCAT	premal2		
presence_of_column	EUROCAT	premal3		
presence_of_column	EUROCAT	premal4		
presence_of_column	EUROCAT	premal5		
presence_of_column	EUROCAT	premal6		
presence_of_column	EUROCAT	premal7		
presence_of_column	EUROCAT	premal8		
presence_of_column	EUROCAT	omim		
presence_of_column	EUROCAT	orpha		
presence_of_column	EUROCAT	assconcept		
presence_of_column	EUROCAT	occupmo		
presence_of_column	EUROCAT	illbef1		
presence_of_column	EUROCAT	illbef2		
presence_of_column	EUROCAT	matdiab		
presence_of_column	EUROCAT	hba1c		
presence_of_column	EUROCAT	illdur1		
presence_of_column	EUROCAT	illdur2		
presence_of_column	EUROCAT	folic_g14		
presence_of_column	EUROCAT	firsttri		
presence_of_column	EUROCAT	drugs1		
presence_of_column	EUROCAT	spdrugs1		
presence_of_column	EUROCAT	drugs2		
presence_of_column	EUROCAT	spdrugs2		
presence_of_column	EUROCAT	drugs3		
presence_of_column	EUROCAT	spdrugs3		
presence_of_column	EUROCAT	drugs4		
presence_of_column	EUROCAT	spdrugs4		
presence_of_column	EUROCAT	drugs5		
presence_of_column	EUROCAT	spdrugs5		
presence_of_column	EUROCAT	extra-drugs		
presence_of_column	EUROCAT	consang		
presence_of_column	EUROCAT	sp_consang		
presence_of_column	EUROCAT	sibanom		
presence_of_column	EUROCAT	sp_sibanom		
presence_of_column	EUROCAT	prevsib		
presence_of_column	EUROCAT	sib1		
presence_of_column	EUROCAT	sib2		

presence_of_column	EUROCAT	sib3		
presence_of_column	EUROCAT	moanom		
presence_of_column	EUROCAT	sp_moanom		
presence_of_column	EUROCAT	faanom		
presence_of_column	EUROCAT	sp_faanom		
presence_of_column	EUROCAT	matedu		
presence_of_column	EUROCAT	socm		
presence_of_column	EUROCAT	socf		
presence_of_table	SURVEY_ID			
presence_of_table	SURVEY_OBSERVATIONS			
presence_of_column	SURVEY_OBSERVATIONS	so_unit		
presence_of_table	PERSONS			
presence_of_column	PERSONS	day_of_birth		
presence_of_column	PERSONS	month_of_birth		
presence_of_column	PERSONS	day_of_death		
presence_of_column	PERSONS	month_of_death		
presence_of_column	PERSONS	race		
presence_of_column	PERSONS	country_of_birth		
presence_of_column	PERSONS	quality		
presence_of_table	OBSERVATION_PERIODS			
presence_of_table	PERSON_RELATIONSHIPS			
presence_of_table	PRODUCTS			
presence_of_column	PRODUCTS	full_product_name		
presence_of_column	PRODUCTS	box_size		
presence_of_column	PRODUCTS	box_size_unit		
presence_of_column	PRODUCTS	drug_form		
presence_of_column	PRODUCTS	route_of_administration		
presence_of_column	PRODUCTS	ingredient1_ATCcode		
presence_of_column	PRODUCTS	ingredient2_ATCcode		
presence_of_column	PRODUCTS	ingredient3_ATCcode		
presence_of_column	PRODUCTS	amount_ingredient1		
presence_of_column	PRODUCTS	amount_ingredient2		
presence_of_column	PRODUCTS	amount_ingredient3		
presence_of_column	PRODUCTS	amount_ingredient1_unit		
presence_of_column	PRODUCTS	amount_ingredient2_unit		
presence_of_column	PRODUCTS	amount_ingredient3_unit		
presence_of_column	PRODUCTS	product_manufacturer		
presence_of_table	CDM_SOURCE			
presence_of_table	METADATA			
presence_of_table	INSTANCE			

list_of_values	VISIT_OCCURRENCE	specialty_of_visit_vocabulary		
list_of_values	VISIT_OCCURRENCE	status_at_discharge_vocabulary		
list_of_values	VISIT_OCCURRENCE	meaning_of_visit		
list_of_values	VISIT_OCCURRENCE	origin_of_visit		
list_of_values	EVENTS	event_record_vocabulary		
list_of_values	EVENTS	meaning_of_event		
list_of_values	EVENTS	origin_of_event		
list_of_values	MEDICINES	disp_amount_drug_unit		
list_of_values	MEDICINES	code_indication_vocabulary		
list_of_values	MEDICINES	meaning_of_drug_record		
list_of_values	MEDICINES	origin_of_drug_record		
list_of_values	MEDICINES	prescriber_type		
list_of_values	PROCEDURES	procedure_code_vocabulary		
list_of_values	PROCEDURES	meaning_of_procedure		
list_of_values	PROCEDURES	origin_of_procedure		
list_of_values	VACCINES	origin_of_vx_record		
list_of_values	VACCINES	meaning_of_vx_record		
list_of_values	VACCINES	vx_dose		
list_of_values	MEDICAL_OBSERVATIONS	mo_code_vocabulary		
list_of_values	MEDICAL_OBSERVATIONS	mo_source_table		
list_of_values	MEDICAL_OBSERVATIONS	mo_source_column		
list_of_values	MEDICAL_OBSERVATIONS	mo_unit		
list_of_values	MEDICAL_OBSERVATIONS	mo_meaning		
list_of_values	MEDICAL_OBSERVATIONS	mo_origin		
list_of_values	SURVEY_ID	survey_meaning		
list_of_values	SURVEY_OBSERVATIONS	so_unit		
list_of_values	PERSONS	sex_at_instance_creation		
list_of_values	PERSONS	race		
list_of_values	PERSONS	country_of_birth		

list_of_values	PERSONS	quality		
list_of_values	PERSON_RELATIONSHIPS	origin_of_relationship		
list_of_values	PERSON_RELATIONSHIPS	meaning_of_relationship		
list_of_values	PERSON_RELATIONSHIPS	method_of_linkage		

## INSTANCE

Fill the table below

source_table_name	source_column_name	included_in_instance	date_when_data_last_updated	since_when_data_complete	up_to_when_data_complete	restrictions_in_values	list_of_values	restriction_condition

## Annex to ETL design. General concepts

### Data Access Provider (DAP)

A Data Access Provider (DAP) is an organization with access to data and with expertise to process and interpret it.

### Datasource, instance, underlying population

A **datasource** is a collection of information pertaining to a population. The population is described by two attributes:

- the **underlying population** is a set of persons based on one or more criteria chosen in the following list
  - Persons who are legally resident in a geographic area (to be specified)
  - Persons who are citizens of a country (to be specified)
  - Persons who are entitled to receive healthcare assistance by an organization (to be specified)
  - Persons who are entitled to receive primary care by a list of practices (to be specified)
  - Persons who have received care (to be specified) by a list of providers (to be specified)
- the **datasource population** is a set of persons in the underlying population that are captured by the data accessed by the DAP. They may either be the whole source population, or a subset, for instance all the persons in the source population who were born with a congenital malformation or delivered a child with a congenital malformation in a given period of time (to be specified).

When a study is requested, a ConcePTION DAP extracts a subset of the information of a subset of its datasource population: this subset is called the **instance** of the datasource. The **instance population** is equal or larger than the study population. The instance is transformed and loaded to the ConcePTION CDM according to the design illustrated in this document. Even though the instance is study-specific, this document is study-independent.

The datasources of ConcePTION are thoroughly described in the Project Website, in the DAP's folder of [this page](#) of the member area.

## **Appendix 7: Template statistical analysis plan based on CDM**

**Title**

**Statistical Analysis Plan**

**Version X.X**



## DOCUMENT HISTORY

NAME	DATE	VERSION	DESCRIPTION

## 1. Table of Contents

...

## 2. List of abbreviations

The following abbreviations are used in this statistical analysis plan:


## 3. Responsible parties

### 3.1 Main Author(s) of the SAP

<b>Name</b>	<b>Institution</b>	<b>Role</b>	<b>Contribution</b>

SAP version	Read and approved by (name)	Role	Signature	Date

## 4 Amendments and Updates

SAP amendments following first approval:

Overview of SAP Amendments and Updates

Number	Date (DDMMYY)	Section of the SAP	Amendment or update	Reason
1				
2				
....				

## 5 Introduction

5.1 Preface

5.2 Purpose of the specific analyses

## 6 Study objectives

## 7 Study methods

7.1 General study design

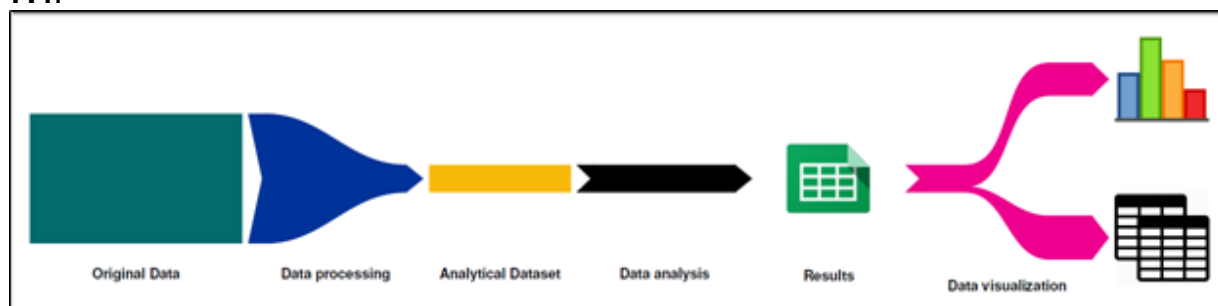
7.1.1 Base population & follow-up

Use graphics to show how the populations selection is done (e.g. repeatit)

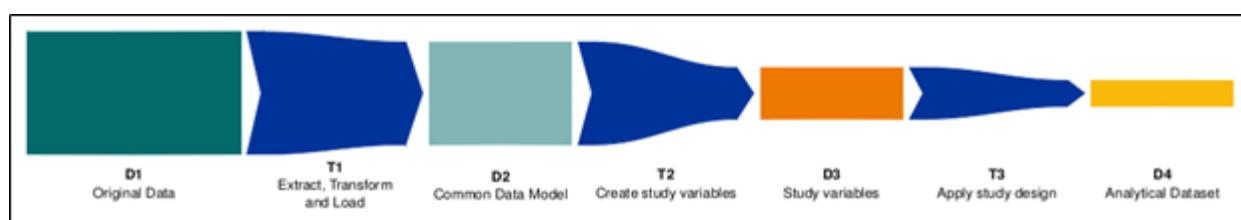
7.2 Data management

This section contains a high-level description of the data management processes required for the study and of the datasets that we will create at different stages of the process of extraction of raw data to creation of a data set for analysis. The process will be divided into 5 phases and 3 transformation steps. A summary schematic of the data management can be found in **Figure 1** and an elaboration of the data processing step can be found in **Figure 2**.

An exact specification of how to conduct these processes can be found in sections 7.3 and 7.4.



**Figure 1 Steps from original data to results (per database)**



**Figure 2 Steps of the Data processing step (see figure 1) between original data and the analytical dataset**

### 7.2.1 Phase 1) Extraction & transformation of local data (Transformation 1)

Extraction of study-specific data on individuals meeting inclusion criteria from the original databases (**Figure 2**, D1) into a common data model (CDM) will be conducted locally by each participating database, using detailed extract, transformation and load (ETL) specifications.

#### 7.2.1.1 Defining the ETL specifications

The ETL specifications will be defined in a step-wise manner. First, study statisticians will review data dictionaries provided by each data access provider (DAP). Using an ETL specification template based upon the CDM, each DAP will propose a column or columns to extract to each table and column or columns of the CDM. This will be reviewed by study statisticians in collaboration with DAPs in order to finalize the specifications.

#### 7.2.1.2 Performing the ETL

Each database may use software of their choice to perform the extraction, transformation, and loading (ETL) of data into the CDM, based upon the ETL specifications (**Figure 2, T1**). This CDM will serve to restructure source data into a common format (syntactic harmonization) but will not alter the content of the source data. The result of this process will be a syntactically harmonized common data model including all data elements required for this study (**Figure 2, D2**).

In order to check and finalize the ETL, *Level 1* and *Level 2* quality checks of the data in the CDM will be performed iteratively as described in section 7.5.

#### 7.2.2 Phase 2: Transformation of CDM-structured data into harmonized data sets (T2)

Following extraction of local data into the CDM, the source data will be harmonized through the creation of study-specific exposures, events, and covariates using a set of agreed upon algorithms. Draft code lists for exposures and events have been constructed and are provided in Annexes 2 & 3. During the harmonization phase, these codes will be used to construct algorithms based upon the consensus of data providers. These algorithms will then be applied by each data provider to construct events, exposures, and relevant covariates. This will proceed iteratively, with *Level 3* data quality and benchmarking (fingerprinting) taking place in each iteration (see section 7.5). Scripts to conduct this transformation will be written centrally by study statisticians and distributed to data contributing sites. The result of this process will be a semantically harmonized common data model, limited to relevant study variables (**Figure 2, D3**).

#### 7.2.3 Phase 3: Transformation of harmonized data sets into analytical data sets (T3)

Using data which has been semantically harmonized, data access providers will use R scripts (written and tested centrally by study statisticians) to create analysis datasets, which will remain local. These data sets will contain the final study variables and should contain only anonymized data (**Figure 2, D4**). There will be at least one analytical dataset for each of objectives 1-X in this study, per DAP (see Annex 4 for mock tables). Testing will be conducted via independent coding by two statisticians against a test data set which has either been contributed by a participating database or simulated to mimic expected data. Study sites will run these scripts and ensure all documentation (i.e. log files created in R, recording of site-specific modifications to the code, and all versions of the code) are correctly archived in the anDREa platform.

#### 7.2.4 Phase 4: Local analysis of the analytical datasets by data access providers

Following creation of analysis data sets, data access providers will be asked to run analysis scripts locally against the analytical dataset(s). Details of the analyses can be found in section 7.4.3. This will result in creation of aggregated results which can be uploaded by DAPs to the research platform.

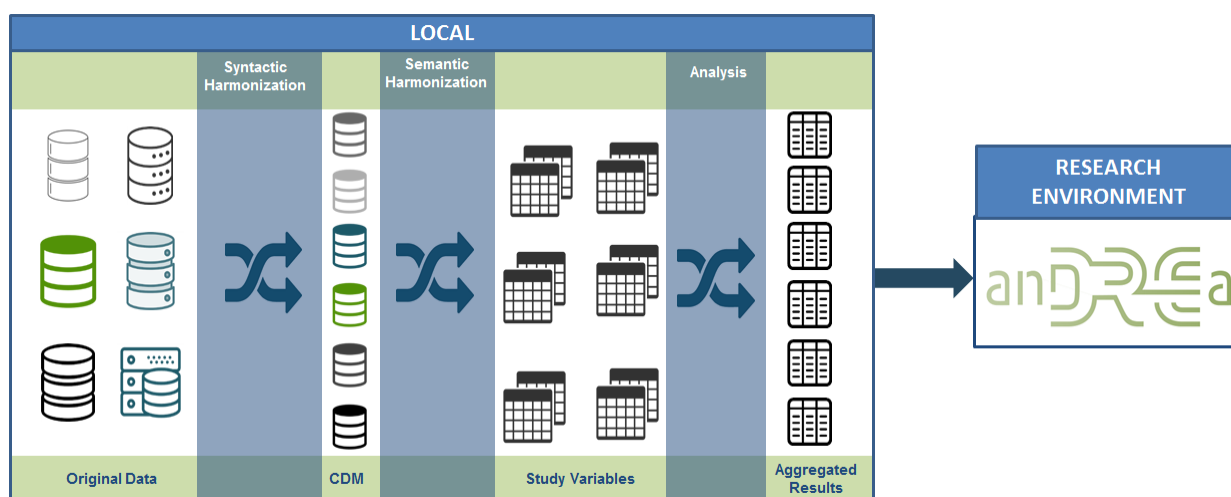
### 7.2.5 Phase 5: Pooling and visualization of analytical results

After quality checks are conducted on the individual results output tables uploaded by each DAP, the uploaded tables will be aggregated into a single analysis table for pooled analyses and visualization of the analytical results (see [Annex 4](#) for mock tables).

### 7.2.6 Overview of information sharing and storage

#### 7.2.6.1 Overview and access to the anDREa platform

The anDREa Research Environment is available through the anDREa consortium, a collaboration between the Dutch University hospitals Radboudumc Nijmegen, Erasmus MC Rotterdam, and UMC Utrecht (<https://www.andrea-consortium.org/>). The Digital Research Environment (DRE) is a cloud based, globally available research environment where data is stored and organized securely and where researchers can collaborate (<https://www.andrea-consortium.org/azure-dre/>).



**Figure 3: data management plan**

#### 7.2.6.2 File transfer and storage procedures

#### 7.2.6.3 Analysis of output tables stored in anDREa

### 7.3 Data Extraction and Harmonization procedure – Data set descriptions

The common data model to be employed has been developed based upon the principles of minimum information loss and maximum transparency in derivation of study variables. Each data set (**D**) and each transformation step (**T**) is described below (see **Figure 1** and **Figure 2** for a schematic overview).

### 7.3.1 Original data (D1)

The original data, meaning those tables available to the data access provider for the purposes of the current protocol, will remain local and unmodified. Processes to transform this data from its original structure to analysis ready datasets and results are described below.

### 7.3.2 Syntactically Harmonized CDM (D2)

All original data (as defined in section 7.3.1) for the study population and study period present in each data source during the study period will be extracted, transformed, and loaded (ETL) into a common data model (CDM) which will be retained locally by each data source. Data sources may use their preferred software to conduct the ETL. ETL scripts should be retained locally and ETL specifications written according to the template in **Annex X**.

The CDM tables to be used for the current project are listed below (**Box 1**):

#### Box 1 CDM tables

##### **METADATA TABLES**

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

##### **METADATA**

This table contains some general information about how the local 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 for quality check (eg check that all the tables that are expected to be findable can indeed be found; and that the coding systems that are observed in the data are indeed those listed here)

##### **INSTANCE**

This table displays the list of the tables and columns of the local 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 local data dictionary that are included in the current ETL document.

##### **CDM\_SOURCE**

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 datasource.

##### **PRODUCTS**

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.

##### **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.**

##### **PERSON**

This table records persons that are to enter analysis of this instance of the CDM.

**OBSERVATION\_PERIODS**

Periods during which data is collected in the datasource for this person. This table contributes to defining the datasource population.

**PERSON\_RELATIONSHIPS**

For any person, this table collects the pairing with the identifier of mother or of other relationships that may be available.

**ROUTINE HEALTH DATA TABLES**

**Routine health care data tables capture data observed in the course of routine health care in hospitals, GP offices, pharmacies, outpatient clinics, etc.**

**VISIT\_OCCURRENCE**

This table contains a summary description of the visits during which records of EVENTS, PROCEDURES, but possibly also MEDICAL\_OBSERVATIONS or VACCINES or MEDICATIONS were recorded. This serves both to collect visit-level information, and to enable grouping sets of records that were recorded concurrently.

**EVENTS**

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.

**MEDICINES**

This table collects data on drug prescriptions, dispensings or administrations occurred during routine healthcare.

**PROCEDURES**

This table collects procedures administered during routine healthcare. Can be a surgery, or a diagnostic procedure, a rehabilitation procedure, a therapeutic procedure.

**VACCINES**

This table collects dispensations or administrations of vaccines.

**MEDICAL\_OBSERVATIONS**

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.

**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.**

**EUROCAT**

This table collects surveillance data on congenital anomalies, following the EUROCAT standard.

**SURVEY\_ID**

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.

**SURVEY\_OBSERVATION**

List of observations in a survey.

Data sources will be requested to extract and fill the following type of data. Text below provides a high-level description of each CDM table. Detailed CDM specifications can be accessed online using this link: <https://drive.google.com/file/d/1hc-TBOtEzRBthGP78ZWla13C0RdhU7bK/view?usp=sharing>.



Additionally, detailed descriptions of vocabularies defined for the CDM can be accessed online using this link:

[https://docs.google.com/spreadsheets/d/1idAEKC440rkIYIxCSRmEVgEPj\\_UouUI-l3kxNCpJt3U/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1idAEKC440rkIYIxCSRmEVgEPj_UouUI-l3kxNCpJt3U/edit?usp=sharing)

#### 7.3.2.1 Detailed description of CDM

A detailed description of the individual tables of the instance of the CDM created for this study is presented below.

- **METADATA:**
- **INSTANCE:**
- **CDM\_SOURCE:**
- **PRODUCTS:**
- **PERSON:**
- **OBSERVATION\_PERIODS:**
- **PERSON\_RELATIONSHIPS:**
- **VISIT\_OCCURRENCE:**
- **EVENTS:**  
*List events here*
- **MEDICINES:**  
*List medicines here*
- **PROCEDURES:**  
*List procedures here*
- **VACCINES:**  
*List vaccines here*
- **MEDICAL\_OBSERVATIONS:**  
*List medical observations here*
- **EUROCAT:**
- **SURVEY\_ID:**
- **SURVEY\_OBSERVATIONS:**

#### 7.3.3 Semantically harmonized CDM (D3)

The semantically harmonized CDM (D3) will be derived from the syntactically harmonized CDM (D2) and will contain the following tables:

*List each D3 table and the steps taken to move from the D2 -> D3 tables*

#### 7.3.4 Analytical Datasets(D4)

*Per study objective: List each D4 table and the steps taken to move from the D3 -> tables*

## 7.4 Data Analysis

### 7.4.1 Missing data

### 7.4.1 Statistical analysis

### 7.4.3 Data analysis steps per objective

## 7.5 Data quality

### 7.5.1 Quality check and benchmarking

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

This is a check conducted in collaboration with Data Access Providers to verify that the extract, transform, and load (ETL) procedure to convert from source data to the D2 CDM has been completed as expected. 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. Distributions of continuous variables and dates will be constructed.

*List study specific checks here*

The level 1 checks proceed as follows for each table of interest in the CDM:

1. Within the METADATA table of the CDM, check for presence of the table of interest in the instance.
2. Verify that the table is present in the directory specified by the DAP. If the table is not present, print a notification of its absence to the report.
3. Verify that mandatory variables are present and contain data. If a mandatory variable is absent or contains only missing data, print a notification of this to the report.
4. Check that all conventions for the table of interest have been adhered to. If a convention is not adhered to, print a notification of this to the report.
5. Check consistency between listed allowable values in the METADATA table and data in the table of interest.
6. Tabulate missingness in all variables, overall and by calendar year.
7. Construct distributions of continuous variables, overall and by calendar year.
8. Construct frequency tables of categorical variables, overall and by calendar year.

Each DAP will be responsible for running the script to complete the Level 1 checks. After addressing any issues identified in level 1 checks, DAPs may rerun the script and inspect the results. This may proceed iteratively until the DAP considers the ETL sufficiently complete and correct.

*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.*

In this check, we will assess records occurring outside of recorded person time (i.e. before birth, after death, outside of recorded observation periods, and outside of visit occurrence dates if applicable). [List study specific checks here](#)

Following completion of level 1 and 2 checks, study statisticians will review results with DAPs and assess any detected errors. Only after these errors have been resolved to the satisfaction of the DAPs will quality checking proceed to level 3.

*Level 3 data checks examine data distributions and trends over time within a DAP's database by examining output by year. For example, 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 person-time in the study population by age and calendar year. We will also calculate incidence of events and exposures of interest by database and calendar year. Counts for each code used to identify exposures or events will be tabulated overall and by calendar year. [List study specific checks here](#)

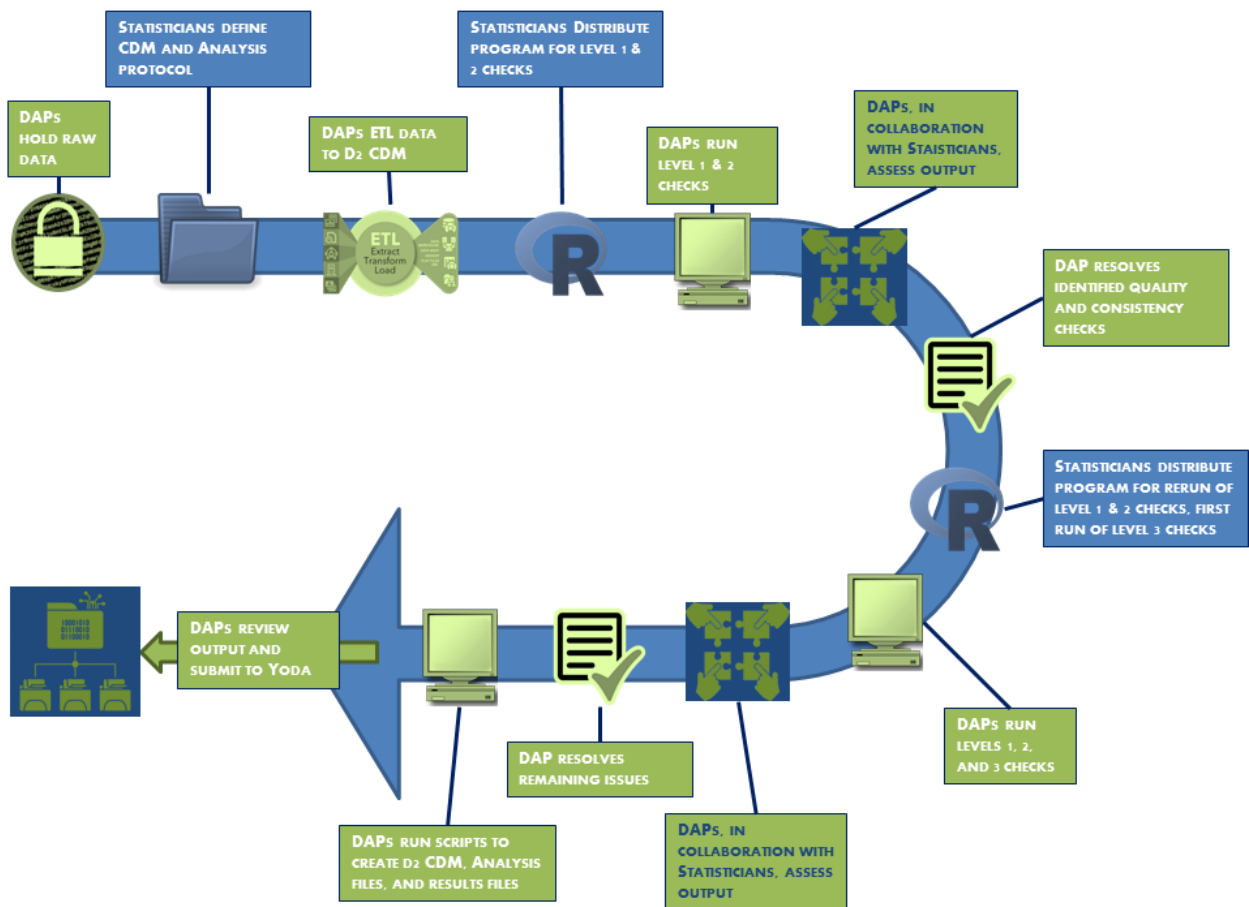


Figure 4. Data quality workflow

## 8 References

### Annex 1: Characteristics of the databases

Table A1.1: Characteristics of the Databases



Annex 2: Event definitions and codes

Annex 3: Drug codes

Annex 4: Mock analysis and results tables

Annex 5: Mock report tables

Annex 6: Additional tables and materials

---

