# Data characterization of population-based data sources: ConcePTION pipeline

First published: 16/06/2023

Last updated: 22/02/2024



# Administrative details

#### **EU PAS number**

EUPAS50142

#### **Study ID**

107620

#### **DARWIN EU® study**

No

#### **Study countries**

Finland

France

Italy

Norway

Spain

United Kingdom

#### **Study description**

The use of real-world data (RWD) and real-world evidence (RWE) in regulatory decision making is increasing (https://rwe-navigator.eu/use-real-worldevidence/sources-of-real-world-data/). In order to make best use of RWD many data sources in a scalable rapid and reproducible manner, many groups and consortia have turned to the use of common data models (CDMs). This approach aims to transform data from different databases into a common format (i.e data model). This data can then be analyzed using scripts written based on that common data format. In this study, we will develop the pipeline of three-levels quality checks for the ConcePTION CDM which include the evaluation of the Extract-Transform-Load (ETL) process, quantify the completeness of the databases, and assessment of counts and distributions of variables and dates. The characterization study will be based on at least 12 data sources, covering at least 6 European countries: France, Finland, Norway, Italy, UK, and Spain (approx. 36.2 million individuals). After each check an html report visualizing the results will be produced. In the final study report, incidence rates from literature will be presented together with incidence rates estimated for the current study. Discrepancies will be identified and interpreted based upon descriptions of the data source(s), algorithms for identification of events, and design choices including inclusion and exclusion criteria in published studies vs. those employed for this protocol. The output of this study will be a series of scripts openly available on an online repository. All the quality checks reports will undergo through an approval process assessed by the conception quality check advisor and the data access provider representative.

#### **Study status**

Ongoing

## Research institutions and networks

## Institutions

## University Medical Center Utrecht (UMCU)

Netherlands

First published: 24/11/2021

Last updated: 22/02/2024

**Educational Institution** 

igl( Hospital/Clinic/Other health care facility igr)

(ENCePP partner)

Institution

# Julius Center

## Networks

## ConcepTION

First published: 01/02/2024

Last updated: 01/02/2024

Network

## Contact details

## Study institution contact

Vjola Hoxhaj v.hoxhaj@umcutrecht.nl



v.hoxhaj@umcutrecht.nl

Primary lead investigator Vjola Hoxhaj Primary lead investigator

## Study timelines

Date when funding contract was signed Actual: 01/04/2019

Study start date Actual: 08/05/2021

Data analysis start date Planned: 01/02/2023

Date of final study report Planned: 31/01/2024

## Sources of funding

• EU institutional research programme

## More details on funding

IMI ConcePTION

## Study protocol

ConcePTION Data Characterization Indicators\_populationbased\_V1.1\_08NOV2019 (2).pdf(2.22 MB)

ConcePTION Data Characterization Indicators\_populationbased\_Revised\_13Nov2023.pdf(2.83 MB)

# Regulatory

Was the study required by a regulatory body?

No

Is the study required by a Risk Management Plan (RMP)?

Not applicable

# Methodological aspects

# Study type

# Study type list

## Study type:

Non-interventional study

## Main study objective:

To assess integrity of the ETL and internal consistency of the CDM instance for each DAP. To produce high-level characterization results describing the final outcomes of these checks in terms of missingness in key variables, distributions of key variables, and internal inconsistency. To assess the data quality and determine 'fit for purpose' for studies on drugs safety and utilization.

## Study Design

Non-interventional study design

Other

# Population studied

## Age groups

Preterm newborn infants (0 – 27 days) Term newborn infants (0 – 27 days) Infants and toddlers (28 days – 23 months) Children (2 to < 12 years) Adolescents (12 to < 18 years) Adults (18 to < 46 years) Adults (46 to < 65 years)

### Estimated number of subjects

36320000

# Study design details

### Data analysis plan

First, we will construct frequency tables based on categorical data and we will plot the distribution of continuous variables and dates. We will use these outputs to identify outlinear, picks, drops, and trends on the data sources. After we will analyzed the missing data pattern, repeated events and incidence rate. Last, we will check the compliance with conception CDM.

## Data management

## Data sources

### Data source(s)

SAIL Databank The Information System for Research in Primary Care (SIDIAP) ARS Toscana EUROmediCAT central database

#### Data source(s), other

FISABIO Spain, Bourdeux France, NorPD, SIDIAP, Drugs and Pregnancy Finland, Emilia Romagna GPs drug prescription

#### Data sources (types)

Administrative healthcare records (e.g., claims) Electronic healthcare records (EHR) Drug dispensing/prescription data

# Use of a Common Data Model (CDM)

### CDM mapping

No

## Data quality specifications

## **Check conformance**

Unknown

## **Check completeness**

Unknown

## **Check stability**

Unknown

## Check logical consistency

Unknown

## Data characterisation

### Data characterisation conducted

No