

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

**ConcePTION Data Characterization for population-based data sources and collections:**

**Level 3 Checks**

**Statistical Analysis Plan**

**Version 0.2**

[DOCUMENT HISTORY 4](#_Toc69912901)

[1 List of abbreviations 5](#_Toc69912902)

[2 Responsible parties 5](#_Toc69912903)

[2.1 Main Author(s) of the SAP 5](#_Toc69912904)

[Reviewers of the document 5](#_Toc69912905)

[3 Amendments and Updates 6](#_Toc69912906)

[4. Introduction 7](#_Toc69912907)

[4.1 Preface 7](#_Toc69912908)

[4.2 Use cases 7](#_Toc69912909)

[4.3 Criteria for quality assessment 8](#_Toc69912910)

[4.4 Aims and objectives 10](#_Toc69912911)

[5 Study methods 11](#_Toc69912912)

[5.1 Setting and data sources 11](#_Toc69912913)

[5.2 Source & study population 11](#_Toc69912914)

[5.3 Data management 12](#_Toc69912915)

[5.3.1 Step 1: Extraction & transformation of local data (Transformation T1) 13](#_Toc69912916)

[5.3.1.1 Defining the ETL specification 13](#_Toc69912917)

[5.3.1.2 Performing the ETL 13](#_Toc69912918)

[5.3.2 Phase 2: Transformation of CDM-structured data into harmonized data sets (T2) 14](#_Toc69912919)

[5.3.3 Phase 3: Transformation of harmonized data sets into analytical data sets (T3) 14](#_Toc69912920)

[5.3.4 Phase 4: Local analysis of the analytical datasets by data access providers 14](#_Toc69912921)

[5.3.5 Phase 5: Pooling and visualization of analytical results 14](#_Toc69912922)

[5.3.6 Overview of information sharing and storage 14](#_Toc69912923)

[5.3.6.1 Overview and access to the anDREa platform 15](#_Toc69912924)

[5.3.6.2 File transfer and storage procedures 15](#_Toc69912925)

[5.3.6.3 Analysis of output tables stored in anDREa 16](#_Toc69912926)

[5.4 Data Extraction and Harmonization procedure – Data set descriptions 16](#_Toc69912927)

[5.4.1 Original data (D1) 16](#_Toc69912928)

[5.4.2 Syntactically Harmonized CDM (D2) 16](#_Toc69912929)

[5.5 Analysis 19](#_Toc69912930)

[5.5.1 Populations 19](#_Toc69912931)

[5.5.1.1 Persons and persontime under observation for source and study populations 19](#_Toc69912932)

[5.5.1.1.1 Step 1: Create end and start follow-up in source population 19](#_Toc69912933)

[5.5.1.1.2 Step 2: Create source population table and follow-up 20](#_Toc69912934)

[5.5.1.1.3 Step 3: Creation of the study population based on time censoring 21](#_Toc69912935)

[5.5.1.1.4 Step 4: Create output how many people we lose / have during study follow-up 22](#_Toc69912936)

[5.5.1.1.5 Step 5: Distribution of persontime in study population (No censoring due to events) 22](#_Toc69912937)

[5.5.1.1.6 Step 6: Rounding of birth dates 25](#_Toc69912938)

[5.5.1.1.7 Step 7: Inspection of birth cohorts and time till registration 26](#_Toc69912939)

[5.5.1.1.8 Step 8: Medicines & vaccines counts 27](#_Toc69912940)

[5.5.1.1.9 Step 9: Completeness of medicines information 31](#_Toc69912941)

[5.5.1.1.10 Step 10: Rates of medicines & vaccines use in women of childbearing age 32](#_Toc69912942)

[5.5.1.1.11 Step 11: Pregnancy codes, and rates of codes and medicines use in women with pregnancy codes 33](#_Toc69912943)

[5.5.1.1.12 Step 12: Events & rates 35](#_Toc69912944)

[5.5.1.1.13 Step 13: Health care seeking behaviour 37](#_Toc69912945)

[5.5.1.1.14 Step 14: Lifestyle in women of childbearing age 38](#_Toc69912946)

[5.5.1.1.15 Step 15: Eurocat quality indicator tables 39](#_Toc69912947)

[https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/DQI-List-of-Data-Quality-Indicators-since-2012.pdf 39](#_Toc69912948)

[Annex 1: Medicines and events for data characterization 40](#_Toc69912949)

[Annex 2: Link to pregnancy concepts Matcho and categorization 42](#_Toc69912950)

[Annex 3: Link to event code sheets 43](#_Toc69912951)

# DOCUMENT HISTORY

| **NAME** | **DATE** | **VERSION** | **DESCRIPTION** |
| --- | --- | --- | --- |
| Miriam Sturkenboom | 25 March 2021 | 0.1 | Initial Draft |
| Marianne Cunnington, Carlos Duran | March 28, March 31 |  | Comments |
| Miriam Sturkenboom | April 19 | 0.2 | update |

# List of abbreviations

The following abbreviations are used in this statistical analysis plan:

|  |  |
| --- | --- |
| CDM | Common Data Model |
| ETL | Extract, Transform, and Load |
| DAP | Data Access Provider |
| RWD | Real World Data |
| RWE | Real World Evidence |
| SOP | Standard Operating Procedure |
|  |  |
|  |  |
|  |  |
|  |  |

# Responsible parties

## Main Author(s) of the SAP

|  |  |  |  |
| --- | --- | --- | --- |
| ***Name*** | ***Institution*** | ***Role*** | ***Contribution*** |
| Miriam Sturkenboom | UMC Utrecht | CO-lead WP7 | First draft |
| Vjola HOxhaj | UMC Utrecht | Task lead 7.6 | Contributing author |
|  |  |  |  |

## Reviewers of the document

|  |  |  |  |
| --- | --- | --- | --- |
| ***Name*** | ***Institution*** | ***Role*** | ***Contribution*** |
| Carlos Duran | UMC Utrecht |  | comments |
| Marianne Cunnington | GSK | Co-lead WP7 | comments |
|  |  |  |  |

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

# 

# Amendments and Updates

SAP amendments following first approval:

Overview of SAP Amendments and Updates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number | Date (DDMMMYY) | Section of the SAP | Amendment or update | Reason |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| …. |  |  |  |  |

# Introduction

## Preface

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

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

To make best use of RWD for generation of evidence across many data sources in a scalable rapid and reproducible manner, many groups and consortia have turned to the use of common data models (CDMs) (Schneeweiss et al., 2020; Trifiró et al 2014; Gini et al 2016). Data models are structured representations of data elements and their relationships to each other while common data models are simply data models that have been agreed upon and shared by several entities (institutions, databases, researchers, etc.). When converting data to a CDM, those with access to data (data access providers or DAPs) must first extract, transform, and load (ETL) the data to which they have access from its native original format to the agreed upon CDM.

Common data models vary along two axes: 1) the degree to which content is harmonized and 2) their flexibility for use in the conduct of new studies. Along the first axis, CDMs may be structurally (syntactically) harmonized, meaning that data is transformed into a common structure, but the contents remain unchanged, or semantically harmonized, meaning that data is transformed into a common structure and contents are transformed into common concepts. Along the second axis, common data models may be study specific, designed for a set of studies focused on one therapeutic area or one analysis method, or fully reusable for the application of new study questions and designs.

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

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

In a 2012 paper, Kahn et al set out a pragmatic framework for data quality assessment in electronic health record (EHR) -based research. This pragmatic framework includes stages 1) to assess data quality more broadly for each data source in a network to identify whether it is fit for use and 2) to assess data quality and fitness for purpose for a specific use case (Kahn et al, 2012). However, this approach outlines the iterative process of data quality assessment when the goal is to address a specific study question and does not provide guidance for assessment of data quality in a learning healthcare system such as ConcePTION, in which data will be used to address multiple planned and unforeseen study questions.

Currently, both rule-based and subjective frameworks for assessment of data quality exist and are employed by research networks. To describe these frameworks, Callahan et al have adopted the data quality assessment (DQA) terminology developed by Kahn et al. This framework includes *conformance* (compliance with formatting and structural rules), *completeness* (presence of data), and *plausibility* (believability of data as compared to biological constraints and expert knowledge) (Callahan et al, 2017; Kahn et al, 2016).

The US FDA Sentinel system employs a data quality framework which assesses quality on four levels: 1) completeness and validity, 2) accuracy and integrity, 3) consistency of trends over time, and 4) plausibility (https://www.sentinelinitiative.org/about/how-sentinel-gets-its-data).

The Observational Health Data Science and Informatics program (OHDSI), which employs the Observational Medical Outcomes Partnership (OMOP) CDM, currently addresses data quality with a more subjective framework than that employed in Sentinel and that planned for ConcePTION. OHDSI partners who have converted their data to the OMOP CDM have access to tools such as the Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) Heel data quality program, as well as a newly developed data quality dashboard. ACHILLES assesses integrity of the ETL to the OMOP CDM and alerts data access providers to potential ETL errors, while the data quality dashboard provides results based upon the Kahn Framework at the table, column, and concept level and applies cut-offs to alert the user whether a check has been passed or failed (<https://github.com/OHDSI/Achilles>, <https://github.com/OHDSI/DataQualityDashboard>).

The EUROCAT network, a partner in ConcePTION, employs a semantically harmonized CDM, focused on a set of variables specific to assessment of congenital anomalies, and uses a thorough and well-described data quality assessment process for each member of the network. The indicators benchmark case ascertainment and assess accuracy of diagnosis, data completeness, timeliness of data availability, and availability of denominator data (Loane et al, 2011; https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/DQI-List-of-Data-Quality-Indicators-since-2012.pdf).

## Use cases

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

1. Assessing medication use in women of childbearing age(12-55 years old) and in pregnancy: this requires that age and sex is known of the population members, as well as pregnancy status.
2. Calculation of incidence rates of events: follow-up needs to be long enough

2. Calculation of incidence rates of events during pregnancy and before/after: this requires that the onset and ending of a pregnancy is known, as well as the events that occur before/ during and after pregnancy; follow-up needs to be long enough

3. Assessing severity of specific maternal conditions: this requires that healthcare use, or disease severity markers/measures are available

4. Assessing prenatal and antenatal outcomes in relation to drug exposure for signal generation and signal evaluation: this requires that pregnancy duration is known, follow-up is available and the relevant outcomes and exposures can be measured as well as confounding factors

## Criteria for quality assessment

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

The IMI-GetReal project, in their final report, Advancing Evidence Generation for New Drugs: IMI GetReal’s Recommendations on Real-World Evidence recommended:

● All stakeholders should collaborate to develop and publish minimum requirements for the integrity and quality of RWD sources used to generate real world evidence submitted for decision making.

● Regulators, Health Technology Assessment (HTA) bodies, payers, researchers and the pharmaceutical industry should collaborate to characterise RWD sources and understand their strengths and weaknesses.

● Characterise barriers to access of RWD.

● Identify and promote efforts to catalogue RWD sources.

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

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

● Derived from data source of demonstrated good quality

● Valid (internal and external validity)

● Consistent (across countries/data sources)

● Adequate (e.g., precision, adequate range of characteristics of population covered, dose and duration of treatment, length of follow-up)

The FDA guidance for Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies using Electronic Healthcare Data Sets describe under the section ‘appropriateness of data sources’ that investigators should describe historical accessibility and appropriateness of data.

The description about historical accessibility should include:

● How long the data source has been available to the research community;

● How often this data source has been used for pharmacoepidemiologic safety studies;

● The capability of the selected data source to validate the outcome and other study

elements (e.g., exposures, key covariates, inclusion/exclusion criteria) based on the safety

question;

● References for any relevant publications, including validation studies of safety outcomes

of interest in the proposed study that are captured in the database.

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

The Sentinel Initiative has implemented the FDA guidance and does quality assessment of the Sentinel data sources upon each refresh. Approximately 1,200 data checks are evaluated during each Data Partners data refresh.

A recent paper by Johnson et al. also advocated for a formal framework to assess data quality in healthcare data, a Healthcare Data Quality Framework. The items and domains can be visualized with a heatmap or radar graph. A data quality ontology was described which provides rigorous definitions and can automate the computation of data quality measures. Johnson made a literature overview of different quality dimension/measure definitions that are useful for ConcePTION.

Data quality in ConcePTION will be characterized in a transparent manner according to the procedures developed in the IMI-ConcePTION project. This process will proceed iteratively and in collaboration with each data access provider.

*Level 1 data checks review the completeness and content of each variable in each table of the D2 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 D2CDM 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 date variables to assess any rounding will be constructed.

The level 1 checks will start with a report on the METADATA table. Within the METADATA table of the CDM, a check for presence of the table of interest in the instance will be conducted. If the table is not present, print a notification of its absence to the report.

Diagram

Description automatically generated

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

1. Check the table formatting including number of fields of the .csv files, variable names, presence of mandatory variables, comparison between the information in the METADATA table and presence of non-mandatory variables, acceptable vocabularies and formats of finite values variables.
2. Missing data analysis overall and stratified by calendar year. All analyses will be stratified by the meaning variable.
3. Dates checks including checking the format of the date variable as well as check for non-allowable values.
4. Check conventions and construct frequency tables for all finite values variables. The frequency tables will be stratified by the meaning variable and will be shown overall and by calendar year.
5. Distribution of continuous variables and date variables.

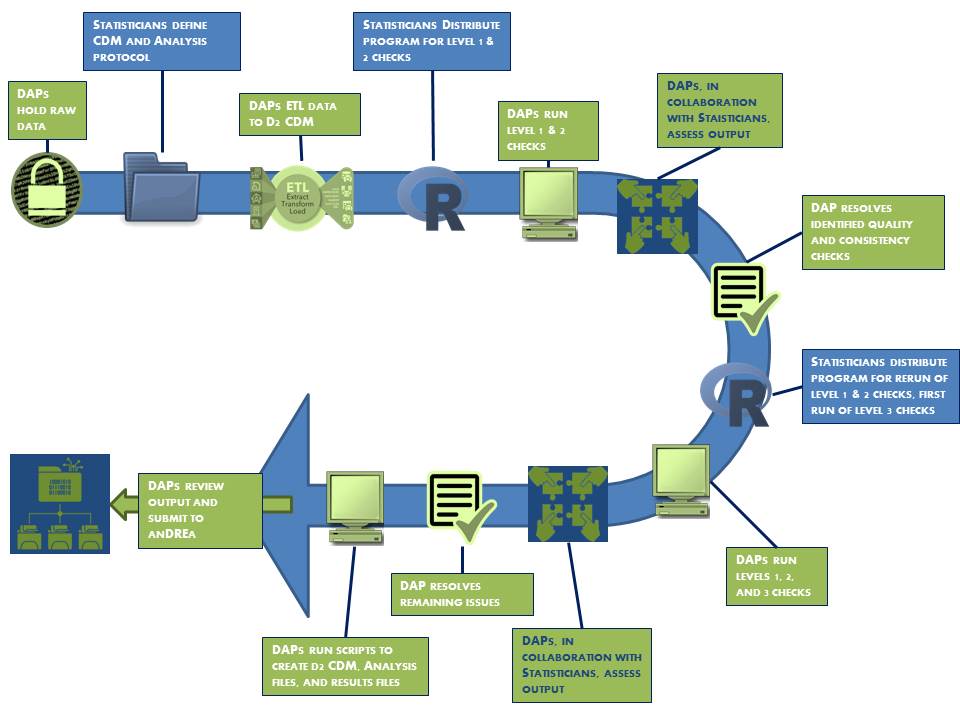
Each DAP will be responsible for running the script to complete the Level 1 checks. An R Markdown report describing results of the checks for each table of the CDM will be produced. After addressing any issues identified in Level 1 checks, DAPs may rerun the script and inspect the results. This may proceed iteratively until the DAP declares the ETL to be complete and correct. An example R Markdown report produced using simulated data will be included as an annex to the study report.

*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 the level 2 checks, we assess records with event dates before birth date, event dates after date of death,e vent dates outside observation periods, subjects observed in a table of interest without a corresponding record in the PERSONS table, observations associated with a *visit\_occurrence\_id* which occur before the *visit\_start\_date* , observations associated with a *visit\_occurrence\_id* which occur after the *visit\_end\_date*, observations associated with a *visit\_occurrence\_id* for which the associated *person\_id* differs from that in the VISIT\_OCCURRENCE table, subjects indicated in PERSON\_RELATIONSHIPS as the parent of a child with a birthdate less than 12 years prior to the recorded birthdate of the associated child.

Each DAP will be responsible for running the script to complete the Level 2 checks. An R Markdown report describing results of the checks for each table of the CDM will be produced. After addressing any issues identified in Level 2 checks, DAPs may rerun the script and inspect the results. This may proceed iteratively until the DAP declares the ETL to be sufficiently complete and correct.

.Only after these errors have been resolved to the satisfaction of the DAPs will quality checking proceed to level 3 Statistical analyses plans for level 1 and 2 checks have been shared before. In this SAP we describe the level 3 checks.



**Figure 1 Data Quality Pipeline**

## Aims and objectives

To describe the *Level 3 data checks for benchmarking between data sources and against external sources.*

Level 3 checks will quantify population and person time in each data source for the source and study population as a whole as well as for subpopulations of interest. Additionally, counts of codes extracted to identify each event and exposure of interest will be calculated overall and by calendar year. Finally, codes will be grouped into concept sets based upon Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) as identified using the Codemapper tool (Becker et al., 2017). Counts and rates of each concept set will be calculated overall and by calendar year. Characterization summaries based upon level 3 checks will be included as an annex to the study report. Medicines and vaccine records and consumption rates will be described and assessed for trends over time. Pregnancies will be described and compared to Europeristat indicators. Prevalence rates of events and medication and vaccine exposure in pregnant women will be reported.

External benchmark data will be incidence rates and prevalence rates of disease that have been obtained from the literature and are listed in the event definition form. Incidence rates from literature will be presented together with incidence rates estimated for the current study in the final study report. 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.

# Study methods

## Setting and data sources

All relevant population-based data sources (data sources that capture person-time of follow-up of a defined dynamic or fixed population during which medication use and/or events can be observed/linked) which aim to participate in one or more of the demonstration projects in ConcePTION and are willing to participate in the ConcePTION database characterization (see table 1).

## Source & study population

The source and study population will vary according to DAP. The source population is the population captured by the data base instance in the CDM.

The study population will comprise at the minimum the EUROmediCAT table of a DAP and at the maximum women and men of child-bearing age (12-55 years old) and all children (0-18 years old) and capture data during the study period. The study population will be followed from the latest of the moment of registration of the person in the data source, birth date or start of the study period (1-1-1995) until the earliest of the following dates: death, reaching the age of end of follow-up (56) for data characterization, last data draw down, or leaving the data source.

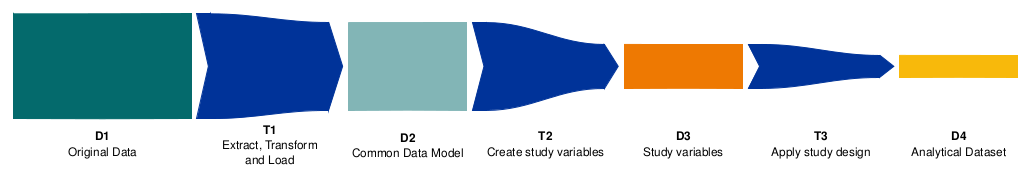
See Table 1 for a detailed description of data sources.

**Table 1. Data sources initially agreeing to participate in the ConcePTION data characterization**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Country** | **area** | **source pop. Size (million)** | **Total Births captured per year** | **Type of data sources** |
|
| **Population based data sources** | | | | |
| **Italy** | Tuscany | 3.7 | 25 | Record linkage of claims data and registries |
| Caserta | 0.9 | 0.06 | Record linkage of claims |
| Emilia Romagna | 4.4 | 35 | Record linkage of claims data and registries |
| **Norway** | Entire country | 5.4 | 60 | Record linkage of claims data and registries |
| **Netherlands** | Sample | 4.4 | 15 | Record linkage of claims data and registries |
| **Denmark** | Entire country | 5.6 | 60 | Record linkage of claims data and registries |
| **UK** | Scotland | 5 | 50 | Record linkage of claims & GP data and registries |
| Wales | 3.7 | 33 | Record linkage of claims & GP data and registries |
| **Spain** | Catalunya | 5.8 | 40 | Record linkage of claims & GP data and registries |
| Valencia | 5 | 50 | Record linkage of claims & GP data and registries |
| **Finland** | Entire country | 5.6 | 60 | Record linkage of claims data and registries |
| **France** | Entire country | 66 | 700 | Claims data |
| Haute Garonne | 1.4 | 10 | Cohort & linkage to claims data |
| **Germany** | sample | 16 | 100 | Claims data |
| **Pregnancy surveillance data & case reports** | | | | |
|  |  |  | **# cases reported** |  |
| **Multiple countries** | EUROMEDICAT | Approx. 75 million | 750 000 | Congenital anomaly registries in EUROCAT surveillance |
| **Netherlands** | PREGNANT | 170.000 births a year | 4,899 | Population based pregnancy register |

## 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 (see figure 2). An exact specification of how to conduct these processes can be found below. For the level 3 checks, the majority of analyses will be conducted using the ConcePTION CDM (**D2, Figure 2**) prior to transformation. For a subset of analyses, transformations (**T2)** may be undertaken to create study variables (**D3, Figure 2**)



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

### 

### Step 1: Extraction & transformation of local data (Transformation T1)

Extraction of study-specific data on the DAP’s selected study population from the original databases into a common data model (CDM) will be conducted locally by each participating database, using detailed extract, transformation and load (ETL) specifications. The common data model version needs to be specified. Most recent to date is 2.2

<https://docs.google.com/spreadsheets/d/1hc-TBOfEzRBthGP78ZWIa13C0RdhU7bK/edit#gid=413205035>

#### Defining the ETL specification

The ETL specifications are conducted in a step-wise manner. First, data dictionaries provided by each data access provider (DAP) will be requested and reviewed. Second, DAPs are asked to fill out the ETL specification template: <https://docs.google.com/document/d/1SWi31tnNJL7u5jJLbBHmoZa7AvfcVaqX7jiXgL9uAWg/edit>. Each DAP maps to the CDM target tables the column or columns from its original tables. This ETL design is reviewed by WP7 in collaboration with DAPs to finalize the specifications. Specifications may be updated during the ETL process and/or following results from level 1 and 2 checks.

#### Performing the ETL

Each database may use software of their choice to perform the extraction, transformation, and loading (ETL) of data into the ConcePTION CDM, based upon the ETL specifications. This CDM allows 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 the Level 1 and 2 SAP. Level 1 and 2 scripts have recently been updated to the CDM version 2.2.

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

Level 1, 2 and 3 checks will run directly on the CDM D2 instance.

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

Level 3 analyses will make use of limited transformations including merging, filtering and collapsing.

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

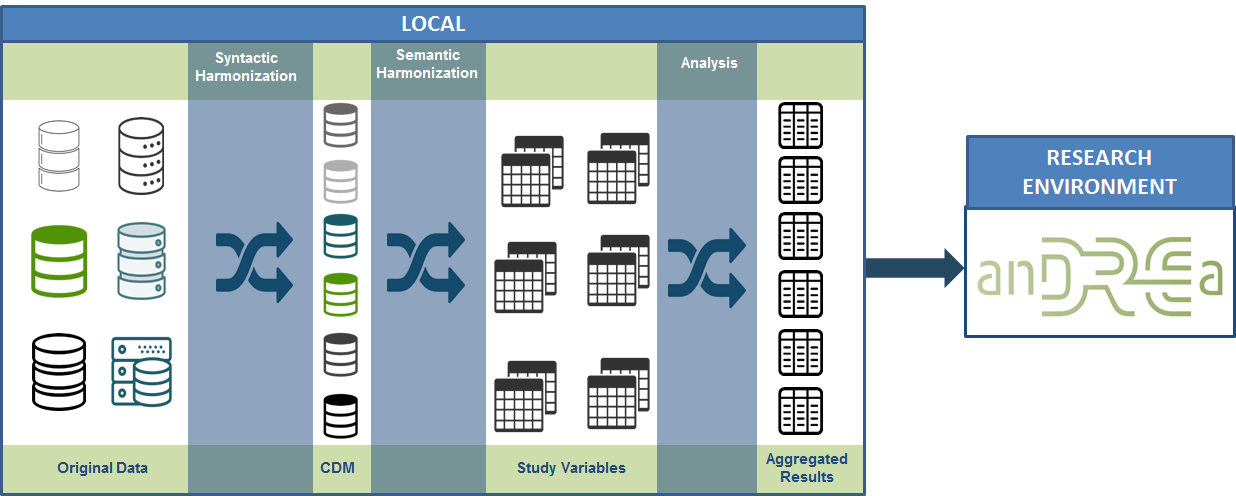
Following ETL of local data to the data characterization study instance of the ConcePTION CDM (D2), scripts to conduct ETL integrity (level 1) checks will be run locally by each DAP against data in the CDM. Once DAPs have resolved issues identified by level 1 checks, they will proceed with internal consistency (level 2) checks. Following completion of both level 1 and 2 checks to the satisfaction of the DAP, aggregate results will be uploaded to the anDREa digital research environment (mydre.org). Once level 1 and 2 checks are complete, the level 3 checks can be run. Aggregated information will be transferred to the DRE.

### 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 on an analysis-by-analysis level for pooled analyses and visualization of the results.

### Overview of information sharing and storage

The DRE is made available through UMCU <https://mydre.org/>. The DRE is a cloud based, globally available research environment where data are stored and organized securely and where researchers can collaborate (<https://www.andrea-consortium.org/azure-dre/>).



**Figure 3 Data management plan**

An informational video is available here: <https://www.andrea-consortium.org/about-andrea/>

#### Overview and access to the anDREa platform

Access to the anDREa research platform is granted at the level of ‘workspace’. A workspace is simply a file system similar to a file system on a standard PC. This file system can be accessed from one or more virtual machines within the workspace. Two-factor authentication for workspace access via each user’s mobile phone number is employed to limit access to workspace members only. Access to workspaces within the anDREa research platform is granted at two levels: owner and researcher. Researchers are able to upload files and have read and write access to all files within a workspace. Owners have all rights granted to researchers plus the right to download and to approve or deny download requests from users with researcher rights.

#### File transfer and storage procedures

All Data Characterization workspace members may upload files using drag-and-drop or browse features to access files on their local machine. Standard operating procedures (SOPs) including documentation to accompany each uploaded file have been developed. The template for this ‘readme’ documentation file is available in **Annex 1**. All workspace members have deletion rights, but this is highly discouraged for all workspace members as described in the anDREa SOPs (**Annex 1**).

Download of data and results is possible only after making a request to those workspace members with an ‘Owner’ role. As specified in the ConcePTION anDREa SOPs, it is highly advised that the principal investigator take on the owner role, and assign this role as well to a second trusted collaborator (**Annex 1**). As described in the SOPs, the request must be accompanied with an explanation of the necessity of the request.

All files, together with their accompanying readme files will be stored in the associated workspace of the anDREa Platform for the duration of the ConcePTION project, following which they will be archived on a secure server.

#### Analysis of output tables stored in anDREa

All data management steps and individual database-level analyses will be conducted locally by the individual DAPs. Following completion of level 1, 2 and 3 checks to the satisfaction of each DAP, aggregated results tables will be uploaded to the anDREa platform for pooled analysis and visualization. Within the virtual machines of the workspace, standard statistical analysis and desktop software are available to all workspace members. Study statisticians will pool and visualize aggregate results within the platform using analysis tools of their choice. These analyses will be primarily conducted using R Markdown and/or R Shiny.

## Data Extraction and Harmonization procedure – Data set descriptions

The common data model to be employed has been developed based upon the principals of minimum information loss and maximum transparency in derivation of study variables. Each data set (**D**) and each transformation step (**T**) is described below.

### 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 these data from its original structure to analysis ready datasets and results are described below.

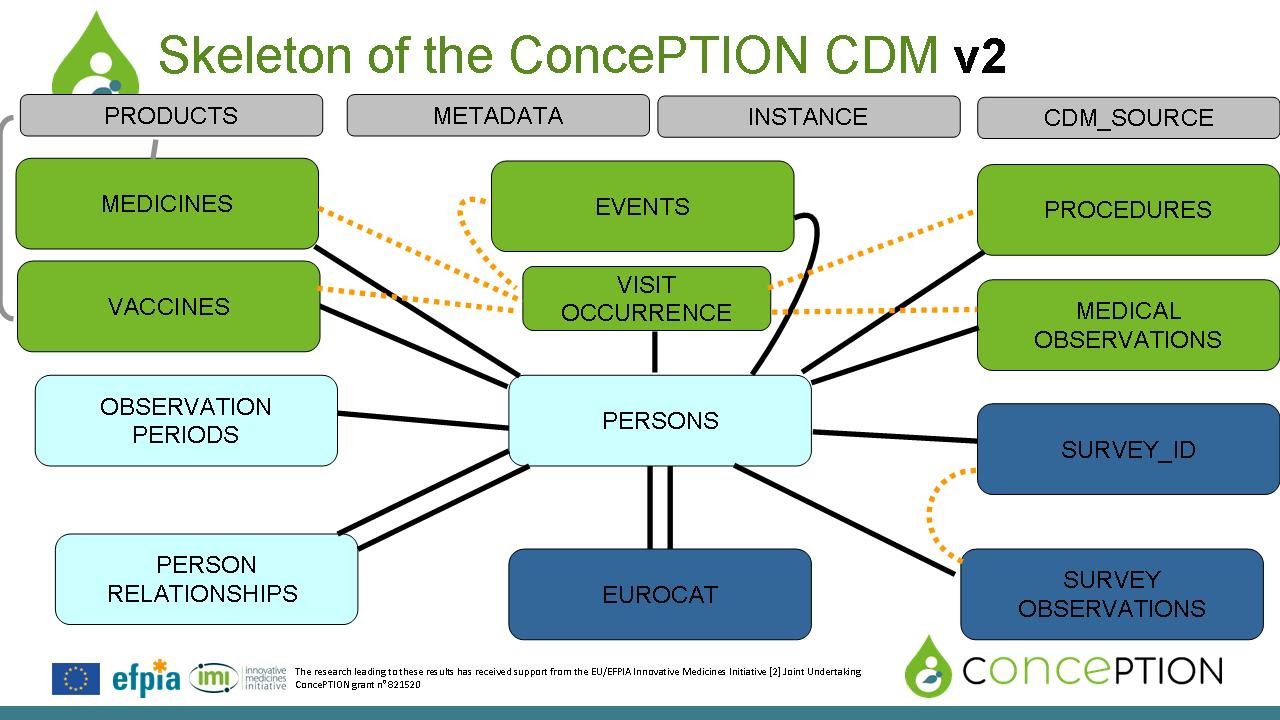
### Syntactically Harmonized CDM (D2)

All original data (for the study population selected by each DAP and 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. This common data model will share the same structure across data access providers but the content of each variable will be minimally modified. In other words, the CDM is syntactically (in terms of structure) rather than semantically (in terms of meaning) harmonized. Data sources may use their preferred software to conduct the ETL. ETL scripts should be retained locally.

The CDM tables to be used for the data characterization study 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.  PRODUCTS  Listing of national product codes for medicinal products. Contains a product ID foreign key to the MEDICINES and VACCINES table. The PRODUCTS table contains detailed data on products at the package level.  METADATA  The METADATA table contains indicators which can act as machine readable guides for code written against the CDM. For instance, whether data in the MEDICINES table represents prescription or dispensing.  INSTANCE  The INSTANCE table contains data on the specific instance of the ConcePTION CDM, such as tables and columns from source data which have been included.    CDM\_SOURCE  Contains high-level metadata describing the source data for the current instance such as the name of the source, data access provider, and date of last update.  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.  PERSONS  One row of data per subject present in the data and meeting inclusion criteria for the CDM instance at any point during the study period. Data on each subject includes sex at the date of the instance creation, day of birth, month of birth, year of birth, and one day of death, month of death, year of death(these may be derived using DAP-specific rules).  OBSERVATION\_PERIODS  One row per period during which a subject is present in the data source. One subject can have one or more observation periods. This may be based upon registration in a geographical area, registration in a GP practice, presence in a registry, etc.  PERSON\_RELATIONSHIPS  Contains one row of data for each relationship between two persons identifiable in the database. This relationship may be parent-child, sibling, or shared household status.  ROUTINE HEALTH CARE DATA TABLES  Routine health care data tables capture data observed in the course of routine health care in hospitals, GP offices, pharmacies, outpatient clinics, etc.  VISIT\_OCCURRENCE  Contains an identifier of a visit to allow for linkage of diagnoses, procedures, dispensings, etc in the same visit if this information is available in a data source.  EVENTS  Contains data on events indicated by a diagnosis code or free text. It contains one row per diagnosed event.  MEDICINES  One record per prescription or dispensing. Contains data required to estimate duration of exposure. Linkage to PRODUCTS table to access data on drugs at the package level.  PROCEDURES  Contains data on procedures ordered or completed. For those procedures with an associated result, results and units are recorded in the MEDICAL\_OBSERVATIONS table. It contains one row per procedure.  VACCINES  Contains data on vaccinations with one row per vaccine. Data on dose number for childhood vaccines and manufacturer are accommodated by this table.  MEDICAL\_OBSERVATIONS  Contains observations recorded during routine health care. Can be a result from a laboratory test, or physical measurement, a pathology report, even socio-economic status, smoking etc.  SURVEILLANCE TABLES  Surveillance tables contain data collected for purposes beyond routine health care either for surveillance of specific events or for recording of detailed information related to a unit of observation such as a pregnancy or chronic illness.  EUROCAT  Contains the EUROCAT or EUROmediCAT (a subset of the EUROCAT) table for those data access providers which have access to this standard table.  SURVEY\_ID  Contains metadata on observations contained in the SURVEY\_OBSERVATIONS table and allows for linkage between mothers and infants captured in a medical birth registry.  SURVEY\_OBSERVATIONS  Contains one row per observation in any survey or registry data table – such as a medical birth registry, well child program database, cancer registry, etc. |



**Figure 3. Schematic representation of the ConcePTION CDMv2.2**

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-TBOfEzRBthGP78ZWIa13C0RdhU7bK/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-I3kxNCpJt3U/edit?usp=sharing>

## Analysis

The level 3 checks will focus on assessing populations, event rates, code counts, medicine and vaccine use, as well as incidence, prevalence and user rates that may be compared with existing indicators as described in the Data Characterization protocol.

The logic is:

1. Description of population and observation time in source and study population
2. Description of information on medicines, vaccines, events, observations and EUROCAT tables during the follow-up period
3. Calculation of rates

### Populations

In this analysis the aim is to assess the stability of the population and the attrition.

The data will be displayed in two reports, specifically:

1. Study\_population – contains information about the distribution of population in the source and study population, the difference between these two populations, the distribution of person time and the lost person time when going from source to study population etc
2. Dates – contains information about the distribution by year and month of op\_start\_date and op\_end\_date(as retrieved by the OBSERVATION\_PERIODS table of the ConcePTION CDM) and strat\_follow\_up and end\_follow\_up(calculated from op\_start\_date and op\_end\_date based on the rules explained below)

The analysis will be performed overall or separately for different subpopulations (as defined in the METDATA table of the ConcePTION CDM by the data provenance).

#### Persons and person time under observation for source and study populations

Operations using D2 CDM tables to define start and end of follow-up for each subject.

**Decisions:**

* If a person has multiple overlapping observation periods, those will be concatenated into one (step 1).
* If a person re-enters in the data source (i.e. having multiple entries for the same person in the OBSERVATION\_PERIODS table), only the last observation period for person time will be kept (step 2).

The average number of spells in the source and study population will be reported. The average number of spells per person in the study population is expected to be one.

P1

P2

Step 1

Step 2

##### 

##### Step 1: Date variables of interest

The variable start\_study\_date is set to 1 January 1995.

The variable end\_study\_date is set to the earliest date between date creation, recommended end date and 31 December 2021.

**CDM-Source:**

* *date\_creation*: create variable date\_creation as a date field from date\_creation variable
* *recommended\_end\_date*: create variable recommended\_end\_date as a date field from recommended\_end\_date variable

**PERSONS**:

* *birth\_date*: Combine day\_of\_birth, month\_of\_birth, year\_of\_birth to a single date value. Note: if day\_of\_birth and month\_of\_birth are missing but year\_of\_birth is present, impute 1 for the day and 6 for the month of birth, if day\_of\_birth is missing and month and year of birth are present impute 16 as day\_of\_birth, if month\_of\_birth is missing and day and year of birth are present impute 6 as month of birth, set date\_of\_birth to missing if there is no information in all of the dates parts or year of birth is missing.
* *death\_date*. Combine day\_of\_death, month\_of\_death, year\_of\_death to a single date value. Note: if day\_of\_death and month\_of\_death are missing but year\_of\_death is present, impute 1 for the day and 6 for the month of death, if day\_of\_death is missing and month and year of death are present impute 16 as day\_of\_death, if month\_of\_death is missing and day and year of death are present impute 6 as month of death, set date\_of\_death to missing if there is no information in all of the dates parts or year of death is missing.
* Retain indicator that dates are imputed

|  |  |  |
| --- | --- | --- |
| **Present** | **Missing** | **Imputed/Final date** |
| DB-MB-YB | - | DB-MB-YB |
| MB-YB | DB | **16**-MB-YB |
| DB-YB | MB | DB-**6**-YB |
| YB | DB-MB | **1**-**6**-YB |
| DB-MB | YB | - |
| - | DB-MB-YB | - |

\*DB- Day of birth, MB-Month of birth, YB-Year of birth

The same steps as above are taken for creation of death\_date.

##### Step 2: Subpopulation analysis

A subpopulation analysis will be performed when a data source contains data with different provenance. The information on the subpopulation names and the meanings that form a subpopulation will be retrieved from the METADATA table. The analysis will be performed separately for each subpopulation and in the time period where all are present.

Example of this analysis:

The data source contains data from the hospital (HOSP) and primary care (PC). The op\_meaning variable in the OBSERVATION\_PERIODS table will be used as a filtering variable to create the different subpopulations. The analysis will be performed in the data that comes from the hospital only, from primary care only and on the subpopulation that is being fed by both hospital and primary care data (PC\_HOSP).

##### Step 2: Create source population table and follow-up

From PERSONS table, copy all variables to table ***Source\_population***

Link to table OBSERVATION\_PERIODS

In *Source\_population* table create

* + *end\_follow\_up* creation (date field)
  + Replace *end\_follow\_up* with OBSERVATION\_PERIODS (*op\_end\_date*). If *op\_end\_date* is empty replace *end\_follow\_up* with the earliest date between date\_creation and recommended\_end\_date as retrieved by CDM\_SOURCE table.

**Results: Is the source population representative of total population?**

Create a graphic for number of persons in year 2018, age at start 2018, by 5-year age categories and sex from the *source\_population* table

* Compare with United Nations population trees (See example below)[[1]](#footnote-2)



##### Step 3: Creation of the study population based on time censoring

**Definition of follow-up in the study population (may vary per study):**

Use *Source\_population* and copy to new table *Study\_population*

\*\*Note that study population definition is dependent on study protocol\*\*

**Decisions:**

* A lookback period of 365 days is taken into account when calculating the start\_follow\_up date.
* If the subject is 0 years old at op\_start\_date, the lookback period in this case is equal to zero.
* If op\_start\_date is before birth\_date then start\_follow will be replaced with birth\_date.
* If op\_end\_date is later than death\_date, op\_end\_date is replaced with death\_date.
* If op\_end\_date is missing, it is replaced with the end\_study\_date or date when reaching the maximum age of interest (56 years old).
* If a person is followed after his 56th birthday (i.e. op\_end\_date is after date when the person turn 56 years old), op\_end\_date is replaced with date when reaching the maximum age of interest (56 years old).
* *start\_follow\_up*: If a subject is zero years old at op\_start\_date/start\_study\_date(when the latest is equal to op\_start\_date) set to op\_start\_date and when the subjects is not 0 years old at op\_start\_date, set to op\_start\_date + lookback period
* *end\_follow\_up*: The earliest date between death\_date, op\_end\_date, end\_study\_date or date when reaching the maximum age of interest (56 years old).

*Use Study\_population*

Replace *start\_follow\_up* latest of

* date of birth *(birth\_date)*
* first data availability (*op\_start\_date)*
* *start of study period (1-1-1995)*

*Start\_fup\_study* is maximum(1-1-1995, *date\_of\_birth, op\_start\_date\_max*)

Replace *end of follow-up*:

* earliest of death (*Date\_of\_ death)*
* last data draw down (*op\_end\_date\_min)*
* *Age 56 years*

*End\_fup\_study* is minimum(*end\_follow\_up*, *op\_end\_date\_min*, *date\_creation*, (*date\_of\_birth*+365.25\*56)) for not empty dates

Create variable *follow\_up* (*end\_fup\_study-start\_fup\_study*)

**Count and report all subjects with *Study\_population* for *follow\_up* <0**

# number excluded outside of persontime

Check required: verify whether this makes sense

##### Step 4: Create output how many people we lose / have during study follow-up

Calculate output table attrition for each subpopulation (in and exclusions based on applying study period & follow-up restrictions)

* *Source\_population (for subpopulations):* 
  + sum of unique person\_id in *Source\_population* table
  + sum of persontime (*end\_follow\_up-start\_follow\_up*)/365.25
* *Study\_population* 
  + sum of unique person\_id in *Study\_population* table for follow\_up>0
  + sum of persontime (*end\_fup\_study-start\_fup\_study*)/365.25 for follow\_up>0
* Create variable Age\_start= (*Start\_fup\_study-date\_of\_birth*)/365.25 in *Source\_population and in Study\_population*

*Question for output table 1 From source to study time, what selection is made?*

*Role of output table: Good to compare across studies and updates of same DAP as well as selection of study population*

*Output table 1: Selection from source to study time*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| DAP acronym | Age at start of follow\_up | Year start of follow\_up\* | # persons in *source\_ population* | # personyears in *Source\_ population* | # persons in *study\_ population* | # personyears in *Study\_ population* |
|  | 0-4 | <1995 |  |  |  |  |
|  | 5-9 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | 0-4 | 1995 |  |  |  |  |
|  | 5-9 |  |  |  |  |  |
|  | .. |  |  |  |  |  |
|  | 0-4 | 1996 |  |  |  |  |
|  | 5-9 |  |  |  |  |  |
|  | .. |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

\*Note for *Source\_population* year and age at start are calculated with reference to *Start\_follow\_up*

for *Study\_population* year and age at start are calculated with reference to *Start\_fup\_study*

##### Step 5: Distribution of persontime in study population (No censoring due to events)

Input: *Study\_population table (D3)*

Exclude all subjects with *follow\_up*<0

**Create results (D4)**

* Create Distribution (mean, median, range) of sum of follow-up ((*follow\_up*)/365.25) per person, and by age category at start of follow-up

*D4: Output table 2: duration of follow-up by calendar year of start and age patterns*

*Question: do we have enough follow-up across different age groups?*

*Question: how does follow-up length change by year of start?*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DAP** | **Year of start** | **Age** | **total** | **male** | | | **Female** | | |
| Acronym |  | At start of follow-up (yrs) | No. of persons | No. of males  # | Mean follow-up (yrs) | Median follow-up (yrs) | No. of females # | Mean follow-up (yrs) | Median follow-up (yrs) |
|  | 1995 | 0-11 years |  |  |  |  |  |  |  |
|  |  | 12-55 years |  |  |  |  |  |  |  |
|  |  | Total |  |  |  |  |  |  |  |
|  | 1996 | 0-11 years |  |  |  |  |  |  |  |
|  |  | 12-55 years |  |  |  |  |  |  |  |
|  |  | Total |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | … |  |  |  |  |  |  |  |  |
|  | All years |  |  |  |  |  |  |  |  |

* Count person-years over calendar years and month with age attributed to category in which they are

Note: Use CountPersontime function, by calendar year and month) without any event for censoring.

*D4 Output table 3 distribution of person-time by age and sex by calendar year:*

*Question: do we see patterns by age/sex (stability of population) across calendar years?*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **DAP** | **Year** | **Age** | **male** | | **female** | | **Total** |
| Acronym |  | Category of persontime contribution (yrs) | # PY | % (of column total) | # PY | % (of column total) | # PY |
|  | 1995 | 0-9 years |  |  |  |  |  |
|  |  | 10-19 |  |  |  |  |  |
|  |  | 20-29 |  |  |  |  |  |
|  |  | etc |  |  |  |  |  |
|  |  | Total |  |  |  |  |  |
|  | 1996 | 0-9 years |  |  |  |  |  |
|  |  | 10-19 |  |  |  |  |  |
|  |  | 20-29 |  |  |  |  |  |

*Question to statisticians: can we do a test to pick up changes in age distribution over calendar years (based on one reference year)?*

*Question: do we see patterns in person-time contributions across calendar years/months?*

*D4 Output table 4: stability of person-time by month and calendar year of the study period*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DAP** | **Calendar year** | **calendarmonth** | **male** | | **female** | | **Total** | |
| Acronym | year |  | # PT | % (of column total) | # PT | % (of column total) | # PT | % (of column total) |
|  | 1995 | 1 |  |  |  |  |  |  |
|  | 1995 | 2 |  |  |  |  |  |  |
|  | 1995 | 3 |  |  |  |  |  |  |
|  | 1995 | 4 |  |  |  |  |  |  |
|  | Etc. |  |  |  |  |  |  |  |
|  | 1996 | 1 |  |  |  |  |  |  |

*Question to statisticians: can we do a test to pick up changes in distribution over calendar time?*

*Question: do we see patterns in starting dates that may point to irregularities in the data/person migration?*

*D4 Output table 5:* Counts of year and month of *Start\_fup\_study* date in *Study\_population*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Year | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | .. |
|  | Month | No. | No. |  |  |  |  |  |  |  |
|  | 1 |  |  |  |  |  |  |  |  |  |
|  | 2 |  |  |  |  |  |  |  |  |  |
|  | 3 |  |  |  |  |  |  |  |  |  |
|  | 4 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

*Question to statisticians: can we do a test to pick up changes over time? We expect a cluster for month and year of study period*

Make graphic per DAP based on output table 5: x-axis year *Start\_fup\_study*, y axis, month.

*Question: do we see patterns in end\_fup\_study* *dates that may point to irregularities in the data/person migration*

*D4 Output table 6:* Count year and month of *end\_fup\_study* date in *Study\_population*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Year | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | .. |
|  | Month | No. | No. |  |  |  |  |  |  |  |
|  | 1 |  |  |  |  |  |  |  |  |  |
|  | 2 |  |  |  |  |  |  |  |  |  |
|  | 3 |  |  |  |  |  |  |  |  |  |
|  | 4 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

*Question to statisticians: can we do a test to pick up changes over time? We expect a cluster for month and year of end of study period*

Make graphic per DAP based on Output table 6: x-axis year *end\_fup\_study*, y axis, month

##### Step 6: Rounding of birth dates

*Question: can we see patterns of rounding of dates?*

*D4: Output table 7: Counts of months and day of birth* from *date\_of\_birth in Study\_population across all years*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Month of birth | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | .. |
|  | Day of birth | No. | No. |  |  |  |  |  |  |  |
|  | 1 |  |  |  |  |  |  |  |  |  |
|  | 2 |  |  |  |  |  |  |  |  |  |
|  | 3 |  |  |  |  |  |  |  |  |  |
|  | 4 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

*D4: Output table 7: Counts of years and month of birth* from *date\_of\_birth in Study\_population*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | .. |
| Year of birth | No. | No. |  |  |  |  |  |  |  |
| 1990 |  |  |  |  |  |  |  |  |  |
| 1991 |  |  |  |  |  |  |  |  |  |
| 1992 |  |  |  |  |  |  |  |  |  |
| 1993 |  |  |  |  |  |  |  |  |  |
| .. |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

*Conclusion to be drawn based on output tables 7 & 8: are dates of birth rounded and does rounding vary by age?*

##### Step 7: Inspection of birth cohorts and time till registration

Distribution of weeks distance from recorded date of birth until start of follow-up ((*start\_study\_fup-date\_of\_birth*)/7) for person with year of birth>1995 (*after start of study period*)

*Question to be addressed: do we see new-borns quickly in the database?*

*D4 Output table 8: Duration to start follow-up* from birth

|  |  |  |  |
| --- | --- | --- | --- |
| DAP | Distance in weeks | No. | % (of column total with year of birth>1995 |
|  | 0-1 |  |  |
|  | 1-2 |  |  |
|  | 2-3 |  |  |
|  | 3-4 |  |  |
|  |  |  |  |
|  |  |  |  |
|  | Total | Sum |  |

Make a line graphic: x-axis duration in weeks: y-axis %

Number of children present during follow-up from birth (< 8 weeks delay) for duration of follow-up within (((*start\_study\_fup-date\_of\_birth*)/7) for person with year of birth>1995)<8) calculate *((end\_study\_fup-start\_study\_fup)/30.4)*

*Question to be addressed: how long can we follow birth-cohorts?*

*D4: Output table 9: Duration of follow-up from birth in children followed from within 2 months after birth*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DAP | Year of birth | Distance in months | No. | % (of column total for year of birth, anytime) |
|  | 1995 | median |  |  |
|  |  | <6 |  |  |
|  |  | 6-11 |  |  |
|  |  | 12-23 |  |  |
|  |  | 24-35 |  |  |
|  |  | 36-47 |  |  |
|  |  | …. |  |  |
|  | Total | Any time |  |  |
|  | 1996 | median |  |  |
|  |  | <6 |  |  |
|  |  | .. |  |  |
|  |  | Total | Sum |  |

##### Step 8: Medicines and vaccines counts

Input tables

* MEDICINES
* VACCINES
* *Study\_population (if subpopulations are present in the data the analysis will be perfomed for each subpopulation)*

Procedure and the flowchart table creation:

1. The MEDICINES/VACCINES table and the study population dataset are loaded, and the original number of records for each table will be reported.
2. If there are records with a specific meaning to be excluded (as reported in the METADATA), those records are removed, and the number will be reported.
3. The MEDICINES/VACCINES table and the study population dataset are joined by keeping only those records for the subjects in the study population. All subjects that are present in the study population but not in the MEDICINES/VACCINES table are saved and will be used to calculate medicines/vaccine exposure rates. The number of these subjects will be reported.
4. All records that are outside the study period (before start\_follow\_up or after end\_follow\_up) will be excluded. Number of records excluded is reported.
5. All records with empty meaning variable are excluded and reported.
6. All records with values unknown or other for the sex\_at\_creation\_variable, are excluded and the number is reported.
7. All the exclusions will be reported in the flowchart table in the markdown report.

Select from MEDICINES and VACCINES table all the records between *Start\_study\_fup-365 and End\_study\_fup linked on person\_id, in Study\_population and copy resulting records to medicines\_study table*

**Produce following counts in a log so that we can observe the deletions**

1. Count # records in MEDICINES & VACCINES table
2. Count # records in MEDICINES & VACCINES table for *Study\_population* (independent of time)
3. Count # records in *Medicines\_study\_table* for *Study\_population*
4. Count # empty (product\_atc\_code) in *Medicines\_study\_table*
5. Count # substr (product\_atc\_code, 5,2)=’ ‘ in *Medicines\_study\_table (incomplete ATC codes)*
6. Calculate % of incomplete 7 level ATC =e/c
7. Count # substr (product\_atc\_code, 3,4)=’ ‘ in *Medicines\_study\_table(incomplete ATC codes)*
8. Calculate % of incomplete 5 level ATC =g/c

**Add from *Study\_population the variable sex\_at\_instance\_creation to Medicines\_study table and the date\_of\_birth.* Calculate the *age\_at\_study\_start (start\_study\_fup-date\_of\_birth)/365.25***

*Question: are all the drug classes that were asked for present across time?*

*D4 output table 10 Number of prescriptions/dispensings by ATC 1 level in the study population by year of dispensing/prescribing and by meaning\_of\_drug\_record*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DAP | Meaning drug record | Year of *date\_dispensing* or *date\_prescription* | ATC 1 level\* | No. (count of records) | % (of column total by meaning) |
|  | Prescription | 1995 | A |  |  |
|  | Dispensing.. |  | B |  |  |
|  |  |  | C |  |  |
|  |  |  | .. |  |  |
|  |  |  |  |  |  |
|  |  |  | Empty |  |  |
|  |  | **total** |  |  |  |
|  |  | 1996 | A |  |  |
|  |  |  | B |  |  |
|  | B | 1996 | A |  |  |
|  |  |  | B |  |  |
|  |  |  | .. |  |  |
|  |  |  |  |  |  |

*\*All medicines ATC starting with specific letter will be aggregated*

*Question: do we have exposures for females in childbearing age?*

*D4 output table 11 Number of prescriptions/dispensings by ATC 1 level in the* ***female*** *study population of childbearing age 12-55 years (based on age at Start\_study\_fup) by year of dispensing/prescribing and by meaning\_of\_drug\_record*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DAP | Meaning drug record | Year of *date\_dispensing* or *date\_prescription* | ATC 1 level | No. (count pf records) | % (of column total by meaning) |
|  | Prescription | 1995 | A |  |  |
|  | Dispensing.. |  | B |  |  |
|  |  |  | C |  |  |
|  |  |  | .. |  |  |
|  |  |  |  |  |  |
|  |  |  | Empty |  |  |
|  |  | total |  |  |  |
|  |  | 1996 | A |  |  |
|  |  |  | B |  |  |
|  |  | 1996 | A |  |  |
|  |  |  | B |  |  |
|  |  |  | .. |  |  |
|  |  |  |  |  |  |

Create counts of medicines records based on ATC level and meaning of the drug record, and based on link with Study\_population on *person\_id* calculate how many male and females in the study population do not have a medicines record during follow-up, for those who do calculate mean and median no. of records in *medicines\_study table by meaning, ATC1 and ATC3 and sex*

*Question: number of prescription/dispensings and users in study population*

*D4 output table 12 Number of prescriptions/dispensings by ATC 1 & 3 level in the study population by year of dispensing/prescribing and by meaning\_of\_drug\_record*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Meaning drug record | ATC 1 level | ATC3 level | ATC4  level | No. of records  during follow-up | No. of male users  males during follow-up | Median  No. Rx  Per male user | No. of  Female users during follow-up | Median  No. Rx per female |
|  | Prescription | None |  |  | 0 |  | 0 |  | 0 |
|  | Dispensing.. | A | A01 | A01A |  |  |  |  |  |
|  |  | A | A02 |  |  |  |  |  |  |
|  |  | A | A03 |  |  |  |  |  |  |
|  |  | .. | … |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  | A |  |  |  |  |  |  |  |
|  |  | B |  |  |  |  |  |  |  |
|  |  | .. |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | All | A | A01 |  |  |  |  |  |  |
|  |  | A | A01 |  |  |  |  |  |  |

Create counts of medicines records based on ATC level and meaning of the drug record, and based on link with Study\_population on *person\_id* calculate how many male and females in the study population do not have a medicines record during follow-up, for those who do calculate mean and median no. of records in *medicines\_study table by meaning, ATC3 and ATC7 and sex*

*Question: number of prescription/dispensings and users in study population*

*D4 output table 13 Number of prescriptions/dispensings by ATC 3 & 7 level in the study population by year of dispensing/prescribing and by meaning\_of\_drug\_record for each ATC class*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Meaning drug record | ATC 3 level | ATC7 level | No. of records  during follow-up | No. of male users  males during follow-up | Median  No. Rx  Per male user | No. of  Female users during follow-up | Median  No. Rx per female |
|  | Prescription | None |  | 0 |  | 0 |  | 0 |
|  | Dispensing.. | A01 | A01AA01 |  |  |  |  |  |
|  |  |  | A01AA02 |  |  |  |  |  |
|  |  |  | A01AA03 |  |  |  |  |  |
|  |  | .. | … |  |  |  |  |  |
|  |  | A02 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  | A03 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | B | A |  |  |  |  |  |  |
|  |  | B |  |  |  |  |  |  |
|  |  | .. |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | All | A01 | A01AA01 |  |  |  |  |  |
|  |  | A01 | A01AA02 |  |  |  |  |  |

*Question: number of prescription/dispensings and users in study population*

*D4 output table 13 Number of prescriptions/dispensings by ATC 3 & 7 level in the study population by year of dispensing/prescribing and by meaning\_of\_drug\_record* ***for females of childbearing age (12-55*** *age\_start\_fup)*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | year | Meaning drug record | ATC 3 level | ATC7 level | No. of records  during follow-up | No. of  Female users during follow-up | Median  No. Rx per female |
|  |  |  | None |  | 0 |  | 0 |
|  | 1995 |  | A01 | A01AA01 |  |  |  |
|  |  |  |  | A01AA02 |  |  |  |
|  |  |  |  | A01AA03 |  |  |  |
|  |  |  | .. | … |  |  |  |
|  |  |  | A02 |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  | A03 |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  | A |  |  |  |  |
|  |  |  | B |  |  |  |  |
|  |  |  | .. |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  | All | A01 | A01AA01 |  |  |  |
|  |  |  | A01 | A01AA02 |  |  |  |

##### Step 9: Completeness of medicines information

**Input table**

*Medicines\_study*

*Vaccines\_study*

*Question: What is the completeness of medicines records for calculation of duration, prescribed and indication*

*D4 output table 14 Number of prescriptions/dispensings with incomplete data*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Year | Meaning drug record | ATC 3 level | No. of records in Study\_population | No. medicines records without Indication (code\_indication empty and not empty vocabulary) | | No. medicines records without prescriber type (prescriber\_type empty and not empty voc.) | | No. medicines records without quantity prescribed/dispensed (disp\_amount\_drug empty) | | No. medicines records without unit quantity prescribed/dispensed (disp\_amount\_drug\_unit empty) | |
|  |  |  |  |  | No. | % | No. | % | No. | % | 0 | % |
| ARS | 1995 | dispensing | A01 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | .. |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | All |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

*Question: What is the completeness of vaccine records?*

*D4 output table 15 Number of vaccines with incomplete data*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Year | Meaning drug record | Vaccine type | No. of records in Study\_population | No. vaccine records without dose | | No. vaccine records without brand | |
|  |  |  |  |  | No. | % | No. | % |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  | .. |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  | All |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

##### Step 10: Rates of medicines & vaccines use in women of childbearing age

**Input tables**

*Study\_population*

*Medicines\_study*

*Vaccines\_study*

*List of ATC classes for study drugs of interest (See methods)*

*Question: how many users of medicines /vaccines do we have (needed to calculate power in studies)*

SUM per class of medicines the number of prescriptions/dispensings (Rx) and the users, denominator is total amount of person-years in each calendar year (irrespective of missingness)

*D4 output table 16 Users of medicines/vaccines for all type of medicines meanings for females 12-55 years of age*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **DAP** | **Calendar year** | **# PY** | **medicines class** | **No. of Rx** | **Rx Rate/100 PY** | **No. of users (unique)** | **Users/100 PY** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  | B |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  | .. |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  | All |  |  |  |  |
|  |  |  |  |  |  |  |  |

*D4 output table 17 Users of medicines for all type of medicines for females 12-55 years of age by age*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Calendar year | Age | Drug class | No. of Rx | # PY | Rx Rate/100 PY | No. of users (unique) | Users/100 PY |
|  |  | 12-19 | A03 |  |  |  |  |  |
|  |  | 20-29 |  |  |  |  |  |  |
|  |  | 30-39 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  | B |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  | .. |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  | All |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

##### Step 11: Pregnancy codes, and rates of codes and medicines use in women with pregnancy codes

**Input tables**

*Study\_population*

*Medicines\_study*

*Vaccines\_study*

*List of study drug classes of interest*

Events table

Medical observations table

Survey observations table

Procedures table

*Question: What type of pregnancy codes do we see ?*

Create a *pregnancy\_study* table by using Matcho concepts (see annex) and retrieve relevant data from all potential CDM tables (in blue). The pregnancy\_study table should have the following format and be restricted to the study population

Format pregnancy\_study table

|  |  |  |
| --- | --- | --- |
| Variable | Format | Values |
| Person\_ID | character | Person identifier (unique) /key |
| Date | date | Date listed in the original CDM table |
| Pregnancy\_event | character | Origin  CDM table |
| Category | character | Classification in terms of the type of code  StartPregnancy  OngoingPregnancy  EndofPregnancy/delivery  PostPregnancy |
| Code | character | Original codes/descriptors in CDM table |

Make a figure with the sequence of codes on one line per women (see example below for another type of event)



Color codes should be category of codes for pregnancy

*Question to be addressed in table 18: what are the codes we can find and how many women are pregnant?*

Personyears are totals per year (as above, no censoring needed), numerators stratified by category of pregnancy code.

*D4 output table 18 Codes for pregnancies for females 12-55 years of age*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| DAP | Calendar year | # PY | Category pregnancy code | No. of records | No. of women(unique) with at least one such record | Rate  Women with at least one code  Per 100 PY |
|  | 1995 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | 1996 |  |  |  |  |  |

*Question to be addressed in table 19: what is the rough rate of pregnancy by age? Does this compare to what we know from national statistics?*

Personyears are totals per year (as above), numerators stratified by category of pregnancy code for the first pregnancy code in that year

*D4 output table 19 Pregnancies code for females 12-55 years of age by age*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DAP | Calendar year | age | # PY | No. of women(unique) with at least one such record in that year and age group | Rate  Women with at least one code  Per 100 PY |
|  | 1995 | 12-19 |  |  |  |
|  |  | 20-29 |  |  |  |
|  |  | 30-39 |  |  |  |
|  |  | 40-49 |  |  |  |
|  |  | 50-55 |  |  |  |
|  | 1996 |  |  |  |  |

*Question table 20: how many women with pregnancy code have at least one year of follow-up (able to see end of pregnancy)?*

For all women 12-55 years in the *Study\_population,* with at least one type of pregnancy code in the specific calendar year, we will assess the distance between *end\_study\_fup* and the date of the first pregnancy code in the calendar year.

*D4 output table 20 Follow-up for women 12-55 years of age with a pregnancy code*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DAP | Calendar year | Category first pregnancy code in that year | No. of women(unique) with at least one such record | Median, range of follow-up | % of women with at least one such record and more than 365 days of follow-up after the first pregnancy code |
|  | 1995 |  |  |  |  |
|  |  |  |  |  |  |

##### Step 12: Events & rates

**Input tables**

*Study\_population*

*Medicines\_study*

*List of event diagnoses codes (codemapper codesheet) see annex 3*

*Pregnancy-study*

Events CDM

Use Events CDM and select for study population all records prior to *start\_study\_fup till end-study-fup* to *events\_study*

The Codemapper output comprises Concept names and concepts (UMLS), the totality of the different concepts is the event

*Question to be addressed in table 21: How many diagnosis records do not have a meaning?*

*Output table 20 Missingness of meanings. By concept for study population 0-55 years of age*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DAP | Calendar year | Concept | No. of events records | No of records with empty meaning | % empty of no. of events |
|  | 1995 |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | 1996 |  |  |  |  |

*Question to be addressed in table 22: What are the code counts of the various event concepts of interest by meaning?*

Use *events\_study* and select all codes appearing in the codesheet for the events of relevance (see annex 3) copy to *events\_study\_interest,* link on code and add from the codesheet the UMLS concept and UMLS concept name

*Output table 22 Code count by meaning for study population 0-55 years of age*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Event | Concept (CUI) | Concept name | Vocabulary | Meaning | Code (truncated) | No. | % |
|  | MS |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

*Question to be addressed in table 23: What are the incidence rates of the various events by meaning?*

Use *events\_study* and select all codes appearing in the codesheet for the events of relevance (see annex 3) copy to *events\_study\_interest,* link on code and add from the codesheet the UMLS concept and UMLS concept name

*D4 Output table 23 Rate of events for study population 0-55 years of age by event, gender, age*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Event | Gender | Age | Meaning | Concept name | No. of cases | PY | Rate per 100,000 PY | 95%CI |
| A | MS | Male | All | All | All |  |  |  |  |
|  |  | Female | All | All | All |  |  |  |  |
|  |  |  |  | A | X |  |  |  |  |
|  |  |  |  | B | X |  |  |  |  |
|  |  |  |  | A | Y |  |  |  |  |
|  |  | Male | 0-11 | All | All |  |  |  |  |
|  |  | Female | 0-11 |  | All |  |  |  |  |
|  |  |  |  |  | X |  |  |  |  |
|  |  |  |  |  | X |  |  |  |  |
|  |  |  | 12-55 |  |  |  |  |  |  |
|  | Epilepsy |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| B |  |  |  |  |  |  |  |  |  |

*Question to be addressed in table 24: What are the incidence rates of the various events by age? This should be compared with published rates*

Calculate incidence rates, censoring on occurrence of first event and censoring if event occurs prior to *end\_study\_fup*, no event in year prior to *start\_study\_fup*

*D4 Output table 24 Rate of selected events (based on one diagnosis codes only) for study population 0-55 years of age by event, gender, age*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Event | Gender | Age (years) | No. of cases | PY | Rate per 100,000 PY | 95%CI |
| A | MS | Male | 0 |  |  |  |  |
|  |  |  | 1-4 |  |  |  |  |
|  |  |  | 5-11 |  |  |  |  |
|  |  |  | 12-19 |  |  |  |  |
|  |  |  | 20-29 |  |  |  |  |
|  |  |  | 30-39 |  |  |  |  |
|  |  |  | 40-49 |  |  |  |  |
|  |  |  | 50-55 |  |  |  |  |
|  |  | Female | 0 |  |  |  |  |
|  |  |  | 1-4 |  |  |  |  |
|  |  |  | 5-11 |  |  |  |  |
|  |  |  | 12-19 |  |  |  |  |
|  |  |  | 20-29 |  |  |  |  |
|  |  |  | 30-39 |  |  |  |  |
|  |  |  | 40-49 |  |  |  |  |
|  |  |  | 50-55 |  |  |  |  |
|  |  |  |  |  |  |  |  |

##### Step 13: Health care seeking behaviour

**Input**

*Study\_population*

VISITS CDM

*Question to be addressed in table 25: How many visits do people have?*

Select from the Visits CDM table the people in the study population during the time of the study (between *end\_study\_fup* and *start\_study\_fup*) to *visits\_study*

*Output table 25 Healthcare visits by meaning for study population 0-55 years of age*

SUM for visits the meaning and users denominator is total amount of person-years in each calendar year (no censoring)

*D4 output table 25 Users of medicines for all type of medicines meanings for females 12-55 years of age*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DAP | Calendar year | Age | # PY | Visit meaning | No. of Visits (sum) | No. of persons with at least one visit in year | Visit Rate  No. of visits/1000 PY |
|  |  | 0 |  | Primary care |  |  |  |
|  |  |  |  | specialist |  |  |  |
|  |  |  |  | .. |  |  |  |
|  |  | 1-4 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  | 5-11 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  | 12-19 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  | 20-29 |  |  |  |  |  |
|  |  | Etc. |  |  |  |  |  |

##### Step 14: Lifestyle in women of childbearing age

**Input**

*Study\_population*

Survey\_OBSERVATIONS

Medical\_observations

*Question to be addressed in table 26: what are the lifestyle behaviors?*

Select from the survey\_observations & medical\_observations the people in the study population during the time of the study (any entry before *end\_study\_fup*) to *lifestyle\_study*

*Output table 26 Lifestyle factors in women of childbearing age*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Factor |  |  |  |  |  |
| DAP | Folic acid use | Calendar year | Age | # women with record | # of women in category | % |
|  |  | 1995 | All |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  | 12-19 |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  | 20-29 |  |  |  |
|  | Smoking record |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | Alcohol abuse record |  |  |  |  |  |
|  | BMI record |  |  |  |  |  |
|  | SES record |  |  |  |  |  |

##### Step 15: Eurocat quality indicator tables

#### <https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/DQI-List-of-Data-Quality-Indicators-since-2012.pdf>

<https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/Pubs-2011-Loane-Report9%283%29-DQI.pdf>

# 

# Annex 1: Medicines and events for data characterization

**Medicines & vaccines**

|  |
| --- |
| ACE Inhibitors/Angiotensin II Receptor Blockers (ARB) (C09)  Analgesics (N02)  Antiasthmatics (R03A)  Antibacterials (J01)  Antidepressants (N06A)  Antiemetics (A04A)  Antiepileptics (N03A)  Antihypertensives (C02)  Antineoplastic agents (L07)  Anti-Parkinson drugs (N04)  Antipsychotics (N05A)  Antivirals (J05)  Vaccines (J07)  Betablockers (C07)  Calcium blockers (C08)  Corticosteroids for systemic use (H02)  Diuretics (C03)  Drugs used in Diabetes (A10)  Endocrine therapy (L02)  Immunostimulants (L03)  Immunosuppressants (L04)  Muscle relaxants (M03)  Other nervous system drugs (N07)  Vaccines (J07) |

# 

**Events**

|  |
| --- |
| Breast Cancer  Depression/ anxiety  Epilepsy  Fetal growth restriction  Gestational Diabetes  Induced terminations of pregnancy - elective  Maternal death  Migraine  Multiple gestation  Multiple sclerosis  Pre-eclampsia  Rheumatoid arthritis  Spontaneous abortions  Stillbirth  Systemic Lupus Erythematosus (SLE)  Termination of Pregnancy for Fetal Anomaly (TOPFA)  Attention Deficit Hyperactivity Disorder (ADHD)  Autism spectrum disorder  Low birth weight  Major congenital anomalies  Microcephaly  Neonatal death  Preterm birth |

**Observations**

Pregnancy identifiers

# Annex 2: Link to pregnancy concepts Matcho and categorization

<https://app.vac4eu.org/codemapper/#/case-definition/EMA-Valproate-Retinoids/Matcho_1> till 34

# Annex 3: event code sheets

<https://drive.google.com/drive/folders/1cPV_vzv3ci3OjAEAVzCWXdF4ibpqKPUz?usp=sharing>

|  |
| --- |
| Event |
| 1. Breast Cancer |
| 1. Depression/ anxiety |
| 1. Epilepsy |
| 1. Fetal growth restriction |
| 1. Gestational Diabetes |
| 1. Induced terminations of pregnancy - elective |
| 1. Maternal death |
| 1. Migraine |
| 1. Multiple gestation |
| 1. Multiple sclerosis |
| 1. Pre-eclampsia |
| 1. Rheumatoid arthritis |
| 1. Spontaneous abortions |
| 1. Stillbirth |
| 1. Systemic Lupus Erythematosus (SLE) |
| 1. Termination of Pregnancy for Fetal Anomaly (TOPFA) |
| 1. Attention Deficit Hyperactivity Disorder (ADHD) |
| 1. Autism spectrum disorder |
| 1. Low birth weight |
| 1. Major congenital anomalies |
| 1. Microcephaly |
| 1. Neonatal death |
| 1. Preterm birth |

1. <http://esa.un.org/wpp/> [↑](#footnote-ref-2)