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

**Level 1 (integrity of ETL) and Level 2 (internal consistency) Checks**

**Statistical Analysis Plan**

**Version 0.7**

DOCUMENT HISTORY

|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **DATE** | **VERSION** | **DESCRIPTION** |
| Caitlin Dodd | 30 March 2020 | 0.1 | Initial Draft |
| Caitlin Dodd | 09 April 2020 | 0.2 | Addition of CDM, Database descriptions, Incidence rates in statistical analysis section |
| Vjola Hoxhaj, Rutger van den Bor, Caitlin Dodd | 06 May 2020 | 0.3 | Addition of first drafts for level 1 and 2 checks |
| Caitlin Dodd | 08 June 2020 | 0.4 | Conversion of SAP to levels 1 & 2 only, conversion to ConcePTION SAP template |
| Miriam Sturkenboom | 4 August 2020 | 0.4 | review |
| Vjola Hoxhaj | 13 August 2020 | 0.5 | Completion of analysis section for level 1 checks |
| Caitlin Dodd | 20 August 2020 | 0.6 | Completion of analysis section for level 2 checks, incorporation of comments |
| Vjola Hoxhaj | 05 March 2021 | 0.7 | Update of level 1 checks, review |

# Table of contents

[1 Table of contents 3](#_Toc67318417)

[2 List of abbreviations 6](#_Toc67318418)

[3 Responsible parties 6](#_Toc67318419)

[3.1 Main Author(s) of the SAP 6](#_Toc67318420)

[3.2 Reviewers of the document 6](#_Toc67318421)

[4 Amendments and Updates 6](#_Toc67318422)

[5 Introduction 8](#_Toc67318423)

[5.1 Preface 8](#_Toc67318424)

[5.2 Aims and objectives 9](#_Toc67318425)

[6 Study methods 10](#_Toc67318426)

[6.1 General study design 10](#_Toc67318427)

[6.2 Source and study population 10](#_Toc67318428)

[6.3 Data management 12](#_Toc67318429)

[6.3.1 Phase 1) Extraction and transformation of local data (Transformation 1) 12](#_Toc67318430)

[6.3.1.1 Defining the ETL specifications 12](#_Toc67318431)

[6.3.1.2 Performing the ETL 13](#_Toc67318432)

[6.3.2 Phase 2: Transformation of CDM-structured data into harmonized data sets (T2) 13](#_Toc67318433)

[6.3.3 Phase 3: Transformation of harmonized datasets into analytical datasets (T3) 13](#_Toc67318434)

[6.3.4 Phase 4: Local analysis of the analytical datasets by data access providers 13](#_Toc67318435)

[6.3.5 Phase 5: Pooling and visualization of analytical results 13](#_Toc67318436)

[6.3.6 Overview of information sharing and storage 13](#_Toc67318437)

[6.3.6.1 Overview and access to the anDREa platform 14](#_Toc67318438)

[6.3.6.2 File transfer and storage procedures 14](#_Toc67318439)

[6.3.6.3 Analysis of output tables stored in anDREa 15](#_Toc67318440)

[6.4 Data Extraction and Harmonization procedure – Data set descriptions 15](#_Toc67318441)

[6.4.1 Original data (D1) 15](#_Toc67318442)

[6.4.2 Syntactically Harmonized CDM (D2) 15](#_Toc67318443)

[6.4.2.1 Detailed description of CDM 18](#_Toc67318444)

[6.5 Data Analysis 20](#_Toc67318445)

[6.5.1 Missing data 24](#_Toc67318446)

[6.5.2 Statistical analysis 24](#_Toc67318447)

[6.5.3 Data analysis steps per objective: Level 1 24](#_Toc67318448)

[6.5.3.1 METADATA table 24](#_Toc67318449)

[6.5.3.2 STEP 1 to 3 26](#_Toc67318450)

[6.5.3.3 CDM\_SOURCE and INSTANCE 31](#_Toc67318451)

[6.5.3.4 STEP 4 to 5 35](#_Toc67318452)

[6.5.3.4.1 VISIT\_OCCURRENCE table 35](#_Toc67318453)

[6.5.3.4.2 EVENTS table 39](#_Toc67318454)

[6.5.3.4.3 MEDICINES table 43](#_Toc67318455)

[6.5.3.4.4 PROCEDURES table 48](#_Toc67318456)

[6.5.3.4.5 VACCINES table 51](#_Toc67318457)

[6.5.3.4.6 MEDICAL\_OBSERVATIONS table 55](#_Toc67318458)

[6.5.3.4.7 SURVEY\_ID table 59](#_Toc67318459)

[6.5.3.4.8 SURVEY\_OBSERVATIONS table 62](#_Toc67318460)

[6.5.3.4.9 PERSONS table 65](#_Toc67318461)

[6.5.3.4.10 OBSERVATION\_PERIODS table 68](#_Toc67318462)

[6.5.3.4.11 PERSON\_RELATIONSHIPS table 71](#_Toc67318463)

[6.5.3.4.12 PRODUCTS table 73](#_Toc67318464)

[6.5.4 Data analysis steps per objective: Level 2 77](#_Toc67318465)

[6.5.4.1 Event dates before date of birth 77](#_Toc67318466)

[6.5.4.2 Event dates after date of death 78](#_Toc67318467)

[6.5.4.3 Event dates outside observation periods 79](#_Toc67318468)

[6.5.4.4 Observations in a table of interest without a corresponding record in the PERSONS table 80](#_Toc67318469)

[6.5.4.5 Observations associated with a visit\_occurrence\_id which occur before the visit\_start\_date 81](#_Toc67318470)

[6.5.4.6 Observations associated with a visit\_occurrence\_id which occur after the visit\_end\_date 82](#_Toc67318471)

[6.5.4.7 Observations associated with a visit\_occurrence\_id for which the associated person\_id differs from that in the VISIT\_OCCURRENCE table 83](#_Toc67318472)

[6.5.4.8 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 84](#_Toc67318473)

[7 References 85](#_Toc67318474)

[8 Annex 1: Characteristics of the databases 86](#_Toc67318475)

[9 Annex 2: Event and Procedure codes 86](#_Toc67318476)

[10 Annex 3: Drug and Vaccine codes 86](#_Toc67318477)

[11 Annex 4: ConcePTION anDREa SOP 87](#_Toc67318478)

[12 Annex 5: Mock tables 87](#_Toc67318479)

[12.1 Level 1 87](#_Toc67318480)

[**Counts stratified by meaning (2 or more categories)** 87](#_Toc67318481)

[**Counts stratified by meaning (other variables)** 88](#_Toc67318482)

[**Counts stratified by meaning and year (2 or more categories)** 88](#_Toc67318483)

[**Counts stratified by meaning and year (other variables)** 88](#_Toc67318484)

[**Date counts stratified by meaning** 88](#_Toc67318485)

[**Date counts stratified by meaning and year** 88](#_Toc67318486)

[12.2 Level 2 88](#_Toc67318487)

[13 Annex 6: Mock report tables and graphics (html markdown report) 88](#_Toc67318488)

[13.1 Level 1 88](#_Toc67318489)

[13.2 Level 2 98](#_Toc67318490)

**List of tables:**

[Table 1. Data sources included in the data characterization 10](#_Toc66086552)

[Table 2. METDATA table 28](#_Toc66086553)

[Table 3. CDM\_SOURCE table 35](#_Toc66086554)

[Table 4. INSTANCE table 36](#_Toc66086555)

[Table 5. VISIT\_OCCURRENCE table 39](#_Toc66086556)

[Table 6. EVENTS table 42](#_Toc66086557)

[Table 7. MEDICINES table 47](#_Toc66086558)

[Table 8. PROCEDURES table 49](#_Toc66086559)

[Table 9. VACCINES table 52](#_Toc66086560)

[Table 10. MEDICAL\_OBSERVATIONS table 56](#_Toc66086561)

[Table 11. SURVEY\_ID table 59](#_Toc66086562)

[Table 12. SURVEY\_OBSERVATIONS table 62](#_Toc66086563)

[Table 13. PERSONS table 64](#_Toc66086564)

[Table 14. OBSERVATION\_PERIODS table 66](#_Toc66086565)

[Table 15. PERSON\_RELATIONSHIPS table 68](#_Toc66086566)

[Table 16. PRODUCTS table 70](#_Toc66086567)

**List of figures:**

[Figure 1. Steps from original data to results (per database) 11](#_Toc66086830)

[Figure 2. Steps of the data processing step between original data and the analytical dataset 11](#_Toc66086831)

[Figure 3. Data management plan 14](#_Toc66086832)

[Figure 4. Schematic representation of the conception cdmv2.0 17](#_Toc66086833)

[Figure 5. Scheme of level 1 checks 26](#_Toc66086834)

[Figure 6. Data quality workflow 27](#_Toc66086835)

[Figure 7. Decision tree example of how each table is verified in step 1 to 3 of level 1 checks (preliminary checks) 31](#_Toc66086836)

**List of boxes:**

[Box 1. Cdm tables 14](#_Toc66086885)

[Box 2. List of medicines to be extracted 18](#_Toc66086886)

[Box 3. List of events to be extracted 18](#_Toc66086887)

[Box 4. List of measurements to be extracted 19](#_Toc66086888)

# 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 |
| EHR | Electronic Health Record |
| LHS | Learning Healthcare System |
| DQA | Data Quality Assessment |
| FDA | Food and Drug Administration |

# Responsible parties

## Main Author(s) of the SAP

|  |  |  |  |
| --- | --- | --- | --- |
| *Name* | *Institution* | *Role* | *Contribution* |
| Vjola Hoxhaj | UMC Utrecht | PI, Epidemiologist, Programmer | Primary Author |
| Caitlin Dodd | UMC Utrecht | Epidemiologist | Primary Author |
| Rutger van den Boer | UMC Utrecht | Statistician, Programmer | Contributing Author |

## Reviewers of the document

|  |  |  |  |
| --- | --- | --- | --- |
| *Name* | *Institution* | *Role* | *Contribution* |
| Miriam Sturkenboom | UMC Utrecht | WP7 lead | Review |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 still 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”.

In order to make best use of RWD for generation of evidence across many data sources in a scalable rapid and reproducible manner, many groups and consortia have turned to the use of common data models (CDMs) (Schneeweiss et al., 2020; Trifiró et al 2014; Gini et al 2016). 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 a number of 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 datasource ‘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 datasource 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. In order 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 ).

The checks described in this statistical analysis plan are closely based upon the Sentinel checks and describe checks consistent with the Sentinel level 1 and 2 checks. Level 3 and 4 checks will also be conducted in ConcePTION and are described in separate statistical analysis plans for level 3 (data characterization) and 4 (benchmarking). The checks conducted by Sentinel as well as those planned in ConcePTION conform to the Kahn framework in that level 1 checks *conformance* and *completeness*, level 2 checks *plausibility*. Both the Sentinel and ConcePTION data quality processes extend this framework to assess consistency and fitness for purpose in level 3, then proceed to further *plausibility* checks through benchmarking.

## Aims and objectives

To assess integrity of the ETL and internal consistency of the CDM instance for each DAP for iterative refinement of their ETL specifications. Subsequently 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 unreconciled instances of internal inconsistency.

The analyses described in this SAP aim to:

1) check the completeness and correctness of the ETL for each DAP and

2) perform a check of internal consistency within and among tables of the CDM. Following completion of these checks, each DAP will move on to

3) characterize the content of the CDM. This third step is not described here but can be reviewed in the companion statistical analysis plan (**ConcePTION Data Characterization for population-based data sources and collections:** **Level 3 (Data characterization**).

**Specific objectives**

1. To assess the integrity of the ETL process for each DAP.
2. To provide feedback on the integrity of the ETL to the DAP iteratively for the refinement of the DAP’s ETL procedure.
3. To produce high-level characterization of the data which has been ETL’d to the instance of the CDM in terms of presence/absence of CDM tables and columns, missingness in key variables, frequencies of categorical variables and distribution of dates and continuous variables.
4. To assess internal consistency both within and between tables of the CDM instance for each DAP.
5. To provide feedback on the internal consistency of the CDM instance to the DAP iteratively for the refinement of the DAP’s ETL procedure.
6. To produce high-level characterization of the data which has been ETL’d to the instance of the CDM in terms of internal consistency of key variables.

The analyses specified in this statistical analysis plan are intended to aid the DAP in their ETL by identifying potential errors or inconsistencies. Following completion of these checks to the satisfaction of the DAP, the final results will provide a description of high-level data patterns such as missingness in key variables, counts of categorical variables, distribution of dates and continuous variables, and irreconcilable internal consistencies in the data.

# Study methods

## General study design

The study design is a descriptive analysis of the entirety of each DAP’s data characterization study instance of the ConcePTION CDM. Data access providers are asked to extract data on subjects aged 12-55 at any time during the study period of 01 January 1995 – 31 December 2019, but may extract a smaller (women of childbearing age and children only) population. The checks described in this statistical analysis plan will be conducted on the full instance of the CDM extracted by the DAP for the data characterization study.

## Source and study population

The source and study population will vary according to DAP. 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) and all children (0-18). See Table 1 for a detailed description of data sources. DAPs may apply inclusion or exclusion criteria for the data characterization instance of the CDM. This will be described in each DAP’s ETL specifications. No inclusion or exclusion criteria will be applied at the analysis stage for the level 1 and 2 checks described in this statistical analysis plan.

Table 1. Data sources included in the data characterization

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Access Provider | Data Source | EUROmediCAT | Women of Childbearing Age | Men of Paternal Age | Children | Existing mother-Child Linkage |
| 01\_UOSL | Norwegian National Registries | NO | YES | NO | YES | YES |
| 04\_AARHUS | Danish National Registries | NO | YES |  | YES | YES |
| 07\_ULST | EUROmediCAT central registry | YES | YES |  | YES (Neonates) | YES |
| 08\_CHUT | EFEMERIS cohort | NO | YES |  | YES  (0-24 mnths) | YES |
| 08\_CHUT | POMME cohort | NO | NO |  | YES  (0-7.5 years) | UNK |
| 10\_BORDEAUX | BPE | NO | YES |  | YES | YES |
| 11\_UMCG | EUROmediCAT Northern Netherlands | YES | YES |  | YES (Neonates) | YES |
| 13\_PHARMO | PHARMO | NO | YES |  | YES | YES |
| 15\_BIPS |  |  |  |  |  |  |
| 18\_FISABIO | FISABIO | YES | YES |  | YES | YES |
| 19\_SIDIAP | SIDIAP |  |  |  |  |  |
| 20\_FERR | IMER Registry of Congenital anomalies | YES | YES |  | YES (Neonates) | YES |
| 21\_CNR-IFC |  |  |  |  |  |  |
| 22\_ARS | ARS | NO | YES | YES | YES | YES |
| 23\_MESSINA | Caserta claims database | NO | YES |  | YES | YES |
| 24\_MALTA | EUROmediCAT Malta | YES | YES | NO | YES (Neonates) | YES |
| 27\_OTTO |  |  |  |  |  |  |
| 33\_THL |  |  |  |  |  |  |
| 34\_USWAN | SAIL | NO | YES | YES | YES | YES |
| 37\_GSK | CPRD | NO | YES | YES | YES | YES |

## 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 dataset 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** (developed for the ConcePTION project, <https://www.imi-conception.eu/>). An exact specification of how to conduct these processes can be found in section **6.4**. For the current study, the majority of analyses will be conducted using the ConcePTION CDM (**D2, Figure 2**) prior to transformation. For a subset of analyses, transformations (**T2, Figure 2**) may be undertaken to create study variables (**D3, Figure 2**)

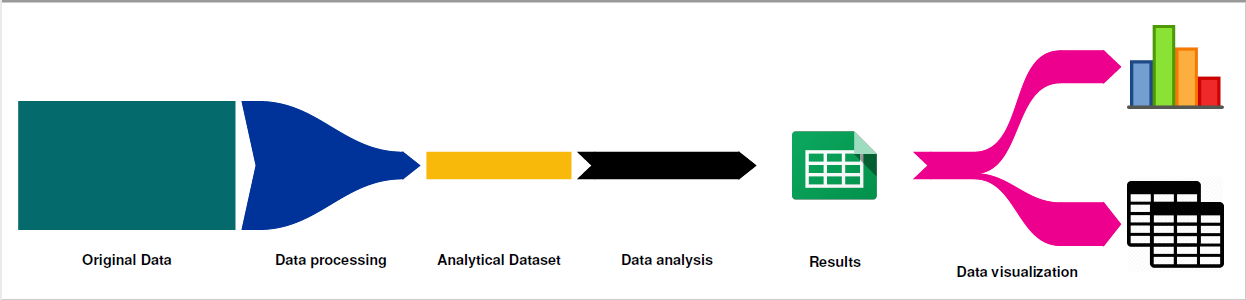


Figure 1. Steps from original data to results (per database)

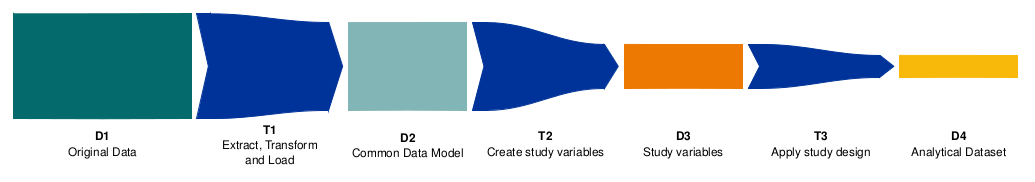


Figure 2. Steps of the Data processing step between original data and the analytical dataset

### Phase 1) Extraction and transformation of local data (Transformation 1)

Extraction of study-specific data on the DAP’s selected study population (minimum study population = individuals captured in a EUROCAT registry, maximum study population = all individual men and women of childbearing age and children aged 0-18) 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.

#### Defining the ETL specifications

The ETL specifications will be defined in a step-wise manner. First, WP7 reviews data dictionaries provided by each data access provider (DAP). Using an ETL specification template based upon the CDM. The ETL specification template can be accessed via this link: <https://docs.google.com/document/d/1SWi31tnNJL7u5jJLbBHmoZa7AvfcVaqX7jiXgL9uAWg/edit>. Each DAP proposes a column or columns to extract to each table and column or columns of the CDM. This will be reviewed by WP7 in collaboration with DAPs in order to finalize the specifications. Specifications may be updated during the ETL process and/or following results from level 1 and 2 checks as described in this SAP.

#### 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 (**Figure 2, T1)**. 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 section **6.5**. against the D2 dataset.

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

Level 1 checks will run directly on the CDM D2 instance.

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

Level 2 analyses will make use of limited transformations including merging, filtering and collapsing. These transformations will result in datasets containing binary indicators for each of the level 2 checks described in section 6.4.4. Detailed descriptions of each transformation are provided in section 6.4.3(**Figure 2, D4**).

### 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. Details of the analyses can be found in section **6.5.** Following completion of both level 1 and 2 checks to the satisfaction of the DAP, aggregate results and masked for anonymity will be created for upload to the anDREa digital research environment.

### 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. (**see Annex 5 for mock tables**).

### Overview of information sharing and storage

Aggregate study results will be uploaded to the anDREa research platform for pooled analysis and visualization. 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. The consortium owns, develops, distributes and promotes the product that is enclosed in its name: The Digital Research Environment (Azure DRE).

The Digital Research Environment (DRE) is a cloud based, globally available research environment where data is stored and organized securely and where researchers can quickly generate workspaces to collaborate in. Within these workspaces, researchers have preinstalled applications at their disposal, as well as the ability to bring own tooling. Globally available and accessible 24/7.

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 4**. All workspace members have deletion rights, but this is highly discouraged for all workspace members as described in the anDREa SOPs (**Annex 4**).

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 4**). 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.

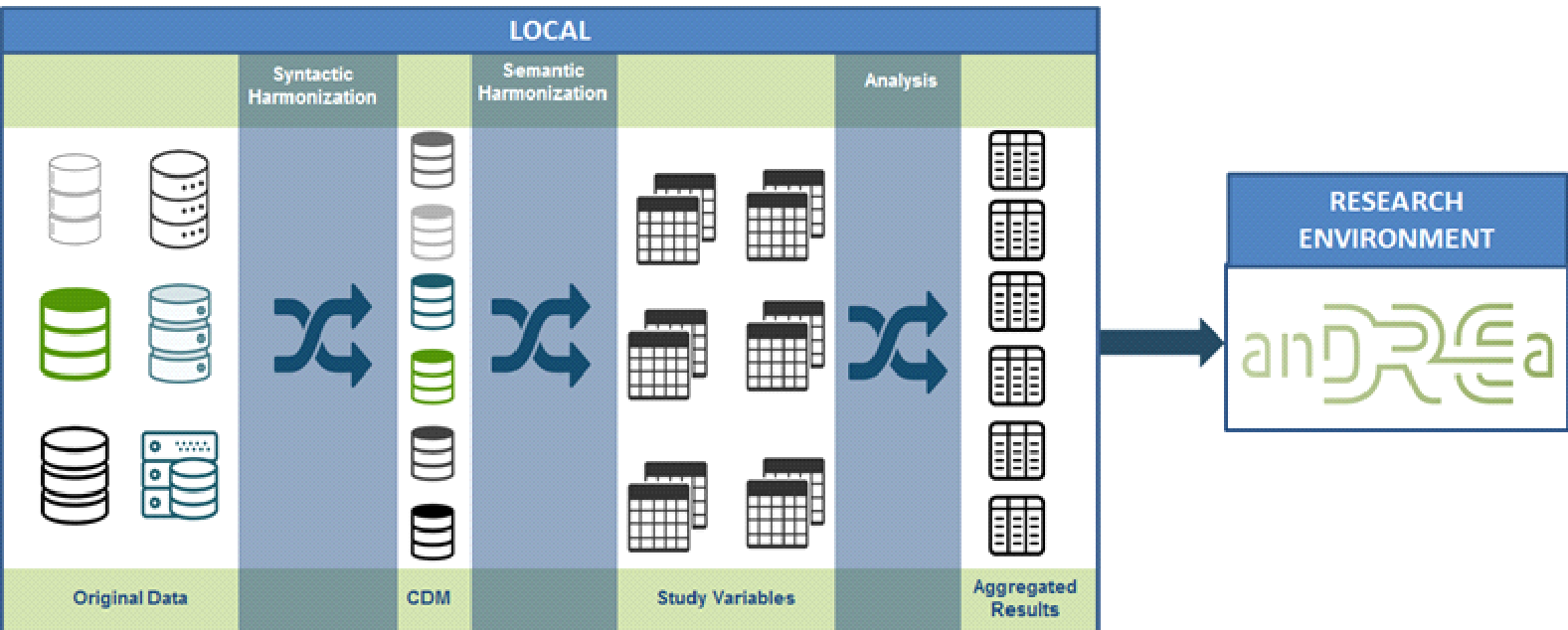


Figure 3. Data Management Plan

#### 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 and 2 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. Dummy tables and graphs are provided in **Annex 5**.

## 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 (see **Figure 1** and **Figure 2** for a schematic overview).

### 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.

### 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. |

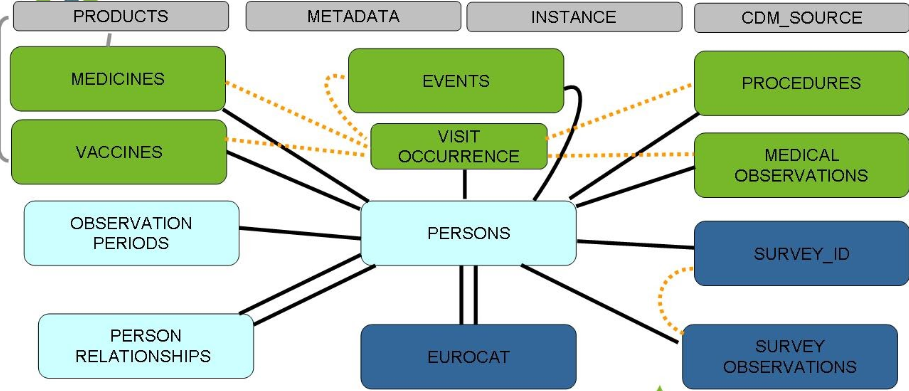


Figure 4. Schematic representation of the ConcePTION CDMv2.0

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-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/u/0/d/1vPZwzQyjXlmmE1vvx3r1Jkw3Juz2DLjU9dKgEo8MijE/htmlview#](https://docs.google.com/spreadsheets/u/0/d/1vPZwzQyjXlmmE1vvx3r1Jkw3Juz2DLjU9dKgEo8MijE/htmlview)

#### Detailed description of CDM

A detailed description of the individual tables of the CDM is presented below.

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

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

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

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

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

**OBSERVATION\_PERIODS**: This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. All fields to be filled for each person in the study population and their periods of follow-up as well as the provenance (source) of the data on follow-up.

**MEDICINES:** This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. We ask that all drugs in the drug classes listed below with a date of dispensing or prescription within the study period, to be extracted. See **Annex 3** for a listing of requested drug classes:

Box 2. List of medicines to be extracted

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

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

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

**EVENTS:** This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. We would like to ask data access providers to extract diagnosis codes for the following events (diagnoses), when there is any occurrence during the study period. See **Annex 2** for a complete listing of event codes.

Box 3. List of events to be extracted

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

**PROCEDURES:** This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. A procedure is a course of action intended to achieve a result in the delivery of care. We would like to ask data access providers to extract the following procedures. A code list for these procedures is included in **Annex 2**:

*List procedures here*

**MEDICAL\_OBSERVATIONS:**  This is a *mandatory* table for all non-EUROCAT DAPs and a *non-mandatory* table for EUROCAT DAPs. We would like to ask data access providers to extract the following measurements and observations if available:

Box 4. List of measurements to be extracted

|  |
| --- |
| Folic acid use  Smoking status  Alcohol use  Educational level  Last menstrual period  Breastfeeding status  Breastfeeding exclusivity  Breastfeeding duration  BMI and/or its components  Socio-economic status and/or proxies of SES |

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

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

**SURVEY\_ID:** This is a *non-mandatory* table. This table should be filled by those DAPs choosing to fill the corresponding SURVEY\_OBSERVATIONStable.

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

## Data Analysis

As a first step, all files are loaded into the R environment. The following data management steps will be conducted in the R environment, meaning that the original source files will not be affected. Level 1 and level 2 checks will be conducted against the CDM instance D2.

*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.

The level 1 checks are divided in 5 major steps:

**Step 1: Check the table formatting**

* Check if all rows of the .csv files in the working directory contain the correct number of fields.
* Check if all variables are present irrespective of their content.
* Check if variables are written in lowercase.
* Check for presence of mandatory variables as according to the CDM.
* Check for presence of non-mandatory variables by comparing between the table of interest and the information recorded in the METADATA table.
* Check presence of vocabularies for specific variables.
* Assess formats for all values and compare to a list of acceptable formats which have been filled out in the METADATA table.

**Step 2: Missing data analysis**

* Tabulate missingness in all variables, overall and by calendar year (in the tables that contain a date variable).
* Missing data will be further stratified by meaning (in the tables that contain a meaning variable).
* Missing data overall will be displayed using bar charts for each CDM table and reported as counts and percentages.
* Missing data stratified by meaning or calendar year will be displayed using line charts for each CDM table and reported as counts and percentages.
* Missing data stratified by meaning and calendar year will be displayed using heat maps for each CDM table and reported as counts and percentages.

**Step 3: Dates check**

* Check if dates are in the correct format (8 characters).
* Check if date variables contain allowable values:

Year: 1995-present (exception for dates that represent end of follow up where years in the future will be allowed.)

Month: 01-12

Day: 01-31

**Step 4: Check conventions and construct frequency tables of other and categorical variables.**

* Check if the table of interest contains any duplicate rows.
* Check that all conventions for the table of interest have been adhered to.
* Construct frequency tables of categorical variables, overall and by calendar year (when the table of interest contains a date variable).
* All frequency tables will be stratified by meaning when the table of interest contains a meaning variable.
* Results will be reported separately for variables with 2 or more categories.
* The results will be displayed graphically with bar charts or line charts.

**Step 5: Distribution of continuous variables and date variables**

* For continuous variables mean, median, interquartile range, skewness and kurtosis will be reported.
* Distribution of date variables will be reported as counts of dates overall and by calendar year. All results will be stratified by the meaning variable if the table of interest contains one.
* Results will be displayed graphically with bar charts or line charts.

**Subsection:**

* Overview table with information on how counts in step 4 and 5 are being calculated.

**Stopping criteria**

Steps 1, 2, and 3 will be run without stopping criteria. For these checks, a report listing all deviations and warnings for all CDM tables in the instance will be produced.

* Variable name not present(presence check) or not written correctly (lowercase check).
* Missingness of mandatory variables.
* Inconsistency between presence of non-mandatory variables in the METADATA table and table of interest.
* Missingness of vocabularies for particular variables of the table of interest in the METADATA table.
* Inconsistency between list of values listed in the METADATA table and actual values in the table of interest.
* Not allowable values or wrong format for date variables.

For Step 4, the script will stop when the following criteria is met:

* Non-adherence to conventions (conditions will depend on the table of interest).

Step 5 contains no stopping criteria.

**Procedure**

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.

All tables will undergo the 5 steps mentioned above except for the EUROCAT, METADATA, CDM SOURCE and, INSTANCE tables. The level 1 checks will start with a report on the METADATA table and then continue with the step 1 to 3. After all issues identified, are fixed DAPs can continue with CDM\_SOURCE and INSTANCE (when present) and step 4 and 5 for all available tables. Based on the results every DAP will proceed with the next tables. The EUROCAT, CDM\_SOURCE and, INSTANCE tables will undergo step1 to step 3. The CDM\_SOURCE and INSTANCE (when present) and the METADATA tables will undergo a specific script.

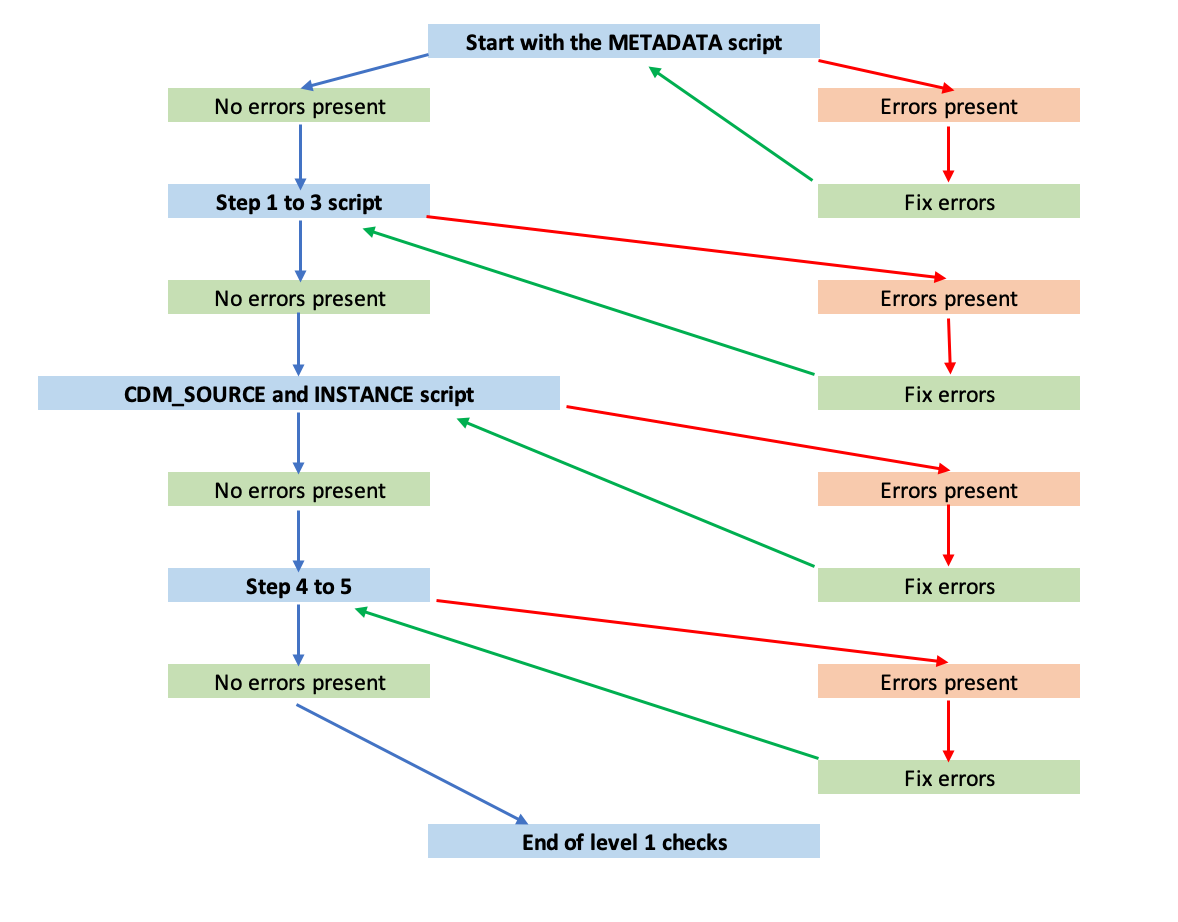


Figure 5. Scheme of level 1 checks

*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. Examples of this type of check include: observations occurring before birth date, observations occurring after a recorded death date, parents aged 12 years old or younger etc.*

Level 2 checks include eight checks:

1. Event dates before birth date
2. Event dates after date of death
3. Event dates outside observation periods
4. Subjects observed in a table of interest without a corresponding record in the PERSONS table
5. Observations associated with a *visit\_occurrence\_id* which occur before the *visit\_start\_date*
6. Observations associated with a *visit\_occurrence\_id* which occur after the *visit\_end\_date*
7. Observations associated with a *visit\_occurrence\_id* for which the associated *person\_id* differs from that in the VISIT\_OCCURRENCE table
8. 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

Following completion of level 1 and 2 checks, results will be reviewed with DAPs. Only after these errors have been resolved to the satisfaction of the DAPs will quality checking proceed to level 3 (Figure 4, also See companion SAP: **ConcePTION Data Characterization for population-based data sources and collections: Level 3 (Data characterization**)).

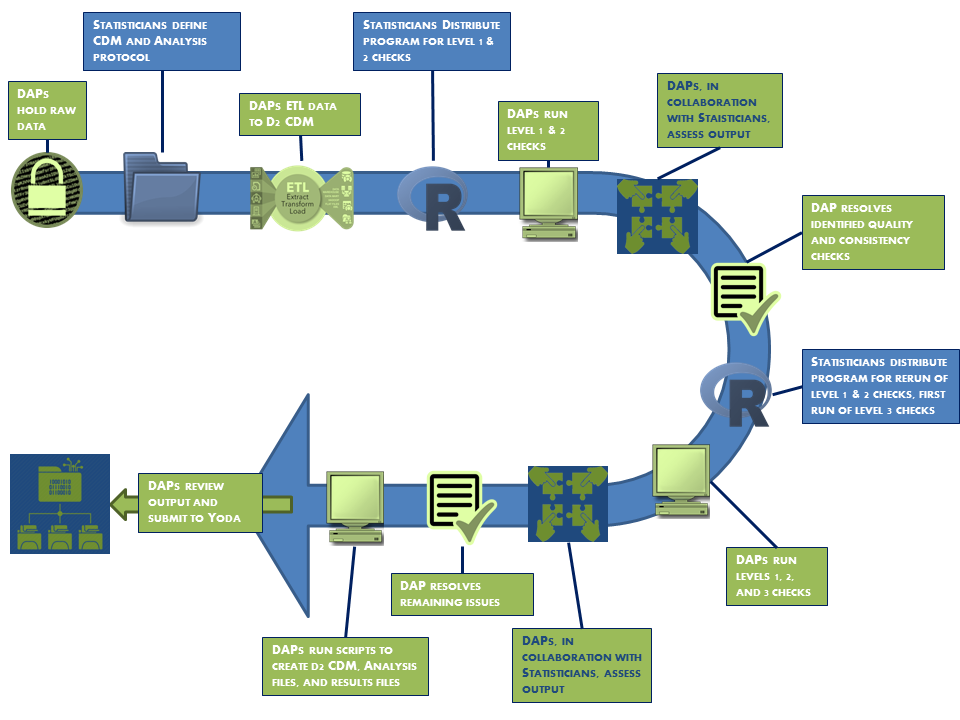


Figure 6. Data quality workflow

### Missing data

Minimal imputation for missing dates of birth and death will be conducted (see section 6.4.3). Additional imputations will not be conducted, as an objective of the current study is to measure and report missingness.

### Statistical analysis

The Level 1 and 2 analyses are descriptive only and do not apply any statistical modeling. Detailed descriptions of each descriptive analysis are provided in **section 6.5.3** below.

### Data analysis steps per objective: Level 1

#### METADATA table

Table 2. METADATA table

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

**The METADATA table will undergo the following checks:**

**Presence in the directory**

* If present, “METADATA table is present in the working directory.” will be printed in the report.
* If there are two or more files named METADATA in the directory the report will show "There are more than two files that contain the pattern METADATA in your working directory. There should be only one. Fix the error and run the script again. The script stops (First stopping criteria).
* If there are no files named METADATA in the directory the report will show "There are no files that contain the pattern METADATA in your working directory. Fix the error and run the script again." The script stops (First stopping criteria).

**Number of fields checks**

* Each row of the .csv file will be checked if contains the correct number of fields. If one of the rows of the .csv contains less or more fields the report will show "This row in your .csv file has missing or extra fields. Fix the error and run the script again.", the number of rows of the .csv, the number of rows being imported and the number of the row of the .csv file where the error is located. The script stops (Second stopping criteria).
* If no errors are found nothing will be reported.

**Presence of variables irrespective of uppercase/lowercase and content**

* If all variables are present the report will show "All variables are present in the METADATA table."
* If a variable is missing the report will show name of variable accompanied by the comment "This variable is missing in your METADATA table. Add it and run the script again." The script stops (Third stopping criteria).

**Check if variables are written in lowercase**

* If variables are written in lowercase the report will show "All variables are written in lowercase."
* If variables fail this check the report will show the name of the variable accompanied by the comment "This variable contains one or more letters in uppercase. Fix the error and run the script again." The script stops (Fourth stopping criteria).

**Allowable values**

*type\_of\_metadata* is allowed to take values:

a. presence\_of\_table

b. presence\_of\_column

c. list\_of\_values

*tablename* is allowed to take values:

a. Only names of tables from the Common Data Model

* If this check is satisfied the report will show "The METADATA table contains only allowable values."
* If this check fails the error will be reported accompanied by the comment "Not allowable value present." The script stops (Fifth stopping criteria).

**Check conventions**

a. For the variables *type\_of\_metadata*=`presence\_of\_table`, *tablename* and *values* should be complete and *columnname* should be empty.

b. For the variables type\_of\_metadata=`presence\_of\_column` and *type\_of\_metadata*=`list\_of\_values`, *tablename, columnname* and *values* should be complete.

* If conventions are met the report will show "All conventions are satisfied."
* If conventions are not met then the report will show “Convention no. is not satisfied.” and the reason why the convention is not met. The script stops (Sixth stopping criteria).

**Consistency check**

* Check if there are any inconsistencies between the information in the METADATA table and the working directory. If a mandatory table is missing either in the directory or the METADATA table it will be colored red.
* If there are no inconsistencies the report will show "All tables mentioned in the METADATA are present in the directory."

**CDM tables present**

A list of the tables present in both the METADATA and the working directory will be given. This can be used as a guide for the steps 4 to 5.

**Subpopulations**

If subpopulations are present in the data, this information will be provided as in.

**Output folder: METADATA**

* list\_of\_presnt\_tables.csv (result of step 8)
* subfolder Variables, where inside there are two files: non\_mandatory\_variables.csv and vocabulary\_variables.csv. These files can be used as guides to help correct the METDATA table if errors regarding variables are present.
* subpopulations.csv
* .html markdown report file(found one level up of this folder, in the output directory)

#### STEP 1 to 3

The level 1 step 1 to 3 checks will be performed simultaneously, for each table present in the CDM instance (*Routine Healthcare Data*: VISIT\_OCCURRENCE, EVENTS, MEDICINES, PROCEDURES, VACCINES, and MEDICAL\_OBSERVATIONS; *Curated Tables*: PERSONS, OBSERVATION\_PERIODS, PERSON\_RELATIONSHIPS; *Surveillance*: EUROCAT, SURVEY\_ID, SURVEY\_OBSERVATIONS; and *Metadata*: PRODUCTS, INSTANCE, and CDM\_SOURCE), except the METADATA table.

For each table, the checks will start with the following:

1. A check of the directory in which CDM .csv files have been stored to identify any filenames containing the name of the CDM table.
2. Check if the rows of the .csv files in the working directory contain correct number of fields according to each CDM table specifications.
3. Check if all variables are present as names irrespective of lowercase/uppercase or their content.
4. Check if all variables names are written in lowercase.
5. Check if variables are imported according to the format specified in the CDM.

See Decision Tree below, depicting an example for the VISIT\_OCCURRENCE table. Only when reaching the bottom left node should the program move on to the next steps for the table of interest.

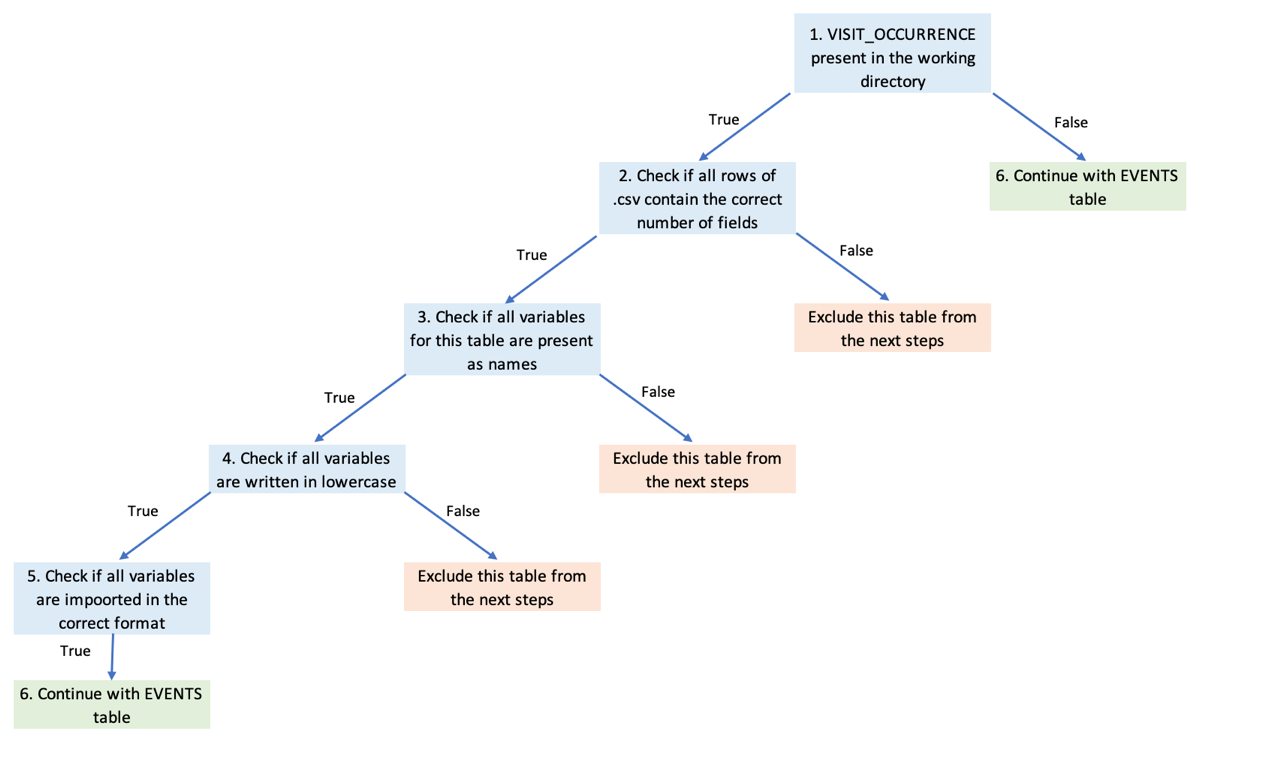


Figure 7. Decision tree example of how each table is verified in step 1 to 3 of level 1 checks (preliminary checks)

Results tables from preliminary checks if errors are found:

* Number of fields check

The results table contains the name of the CDM table, the name of the file located in the directory, the number of rows of the .csv file, the number of fields that are loaded, the number of the first row where the error occurred, and error set to TRUE. If no errors, the following message will be printed:

“All files in the directory contain the correct number of fields”.

* Presence of variables

The results table contains the name of the CDM table, the name of the file located in the directory, the name of the variable, the presence set to FALSE. If no errors, the following message will be printed:

“All variables are present as names for all files in the directory.”

* Lowercase check

The results table contains the name of the CDM table, the name of the file located in the working directory, the name of the variable, the lowercase set to FALSE. If no errors the following message will be printed:

“All variables are written in lowercase for all files in the working directory.”

* Format check

The results table contains the name of the CDM table, the name of the file located in the working directory, the name of the variable, the format according to the CDM, the loaded format, the error set to TRUE. If no errors, the following message will be printed:

“All variables are loaded according to the format of the CDM.”

Following preliminary checks explained above, Level 1(Step 1 to Step 3) checks will proceed for each table in the **Routine Healthcare Data**, **Curated Data, Surveillance Data** and **Metadata** sections of the ConcePTION CDM as follows:

**Table formatting:**

* Check for presence of mandatory variables.

Condition: There should be at least one observation in the mandatory variables.

If there are deviations from this rule, a message describing the deviation, together with the name of the CDM table and the name of the variable is printed to the markdown report. The message will be described as follows:

“This variable is completely empty.”

If there are no deviations for all tables than the message “All mandatory variables are present.” will be printed to the report.

If all tables present in the working directory had an error in at least one the preliminary checks explained in figure 7, they are excluded and this check cannot be performed. The report will show the following message:

"This check might not be performed in case all tables included in the directory had a previous error in fields check, presence of variable or lowercase checks. Fix the problem first and then run the script again."

* Check for presence/absence of non-mandatory variables listed as present/absent in the METADATA table.

Condition:

If listed as present in METADATA table, the variables should have at least one observation.

If listed as not present in METADATA table, the variable should be completely empty.

If any inconsistencies are discovered between the information in the METADATA table and the table of interest either of the following messages, together with the name of the CDM table and the variable name will be printed to the report:

"Present in the METADATA table, but missing in the current table. If variable should be missing, set values for *presence\_of\_column* in the METADATA table to No."

"Present in the current table, but missing in the METADATA table. Fix the METADATA table."

If there are no deviations for all tables than the message "The presence of all non-mandatory variables for each CDM table is according to the METADATA table." will be printed to the report.

If all tables present in the working directory had an error in at least one the preliminary checks explained in figure 7, they are excluded and this check cannot be performed. The report will show the following message:

"This check might not be performed in case all tables included in the directory had a previous error in fields check, presence of variable or lowercase checks. Fix the problem first and then run the script again.”

* Check for presence of vocabularies.

Check if vocabularies for all variables accompanied by a vocabulary are filled out in the METADATA table.

Condition: All vocabulary variables that have at least one observation should be listed in the *list\_of\_values* variable in the METDATA table.

The results will be presented with either of the messages depending on presence/absence of vocabularies together with the name of the CDM table and the name of the variable:

"Values about this variable are present in the METADATA table even though the variable is completely empty in the current table."

"The variable is present in the current table but no information about values of this variable is provided in the METADATA table. Fix the METADATA table."

If all tables present in the working directory had an error in at least one the preliminary checks explained in figure 7, they are excluded and this check cannot be performed. The report will show the following message:

"This check might not be performed in case all tables included in the directory had a previous error in fields check, presence of variable or lowercase checks. Fix the problem first and then run the script again."

* Check list of values for variables with finite number of categories.

Condition: If vocabularies are present, allowable values should be listed in the METADATA table for the corresponding variable.

If there are deviations from what is specified in the METADATA table, a table describing the deviation will be printed to the markdown report. This table will contain the name of the CDM table, the name of the .csv file, the name of the variable, the values found on the .csv file and the values listed in the metadata table.

The CDM table name will be color coded to show different scenarios of deviations:

**red:** If a table is not present in the working directory, but there is information about this table in the METADATA table.

**orange:** If a table is present in the working directory and the vocabulary variable contains at least one observation, but there is no information about this variable in the METADATA table.

**blue:** If a table is present in the working directory and there is information about this variable in the METADATA table, but the information doesn’t match.

If all tables present in the working directory had an error in at least one the preliminary checks explained in figure 7, they are excluded and this check cannot be performed. The report will show the following message:

"This check might not be performed in case all tables included in the directory had a previous error in fields check, presence of variable or lowercase checks. Fix the problem first and then run the script again."

**Missing data analysis**

* Overall missing data counts

For all variables missing data counts and percentages will be reported. The results table will contain the name of the CDM table, the variable name, count, total and percentage of missing. For easier navigation, the name of the table and the name of the variable will be colored green in presence of missing data.

Counts are calculated as number of missing observations, where total is the number of rows of the table of interest. If counts or totals smaller than five are present, those will be replaced by “<5” in the results table.

Visually the missing data will be represented with bar charts for each of the tables of the CDM. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

* Missing data counts stratified by the meaning variable

For tables that contain a meaning variable, the missing data analysis will be stratified by this variable. Tables that are not included in this analysis are CDM\_SOURCE, EUROCAT, INSTANCE, PERSONS, PRODUCTS.

The results table will contain the name of the CDM table, the variable name, the meaning variable, count, total and percentage of missing. For easier navigation, the name of the table and the name of the variable will be colored green in presence of missing data.

Counts are calculated as number of missing observations for each variable with a recorded meaning, where total is the number of total observations of the variable with a recorded meaning. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the missing data will be represented with line charts for each of the tables of the CDM, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

* Missing data counts stratified by calendar year

Tables that will undergo this analysis are tables that contain a date variable, but that do not contain a meaning variable, and specifically the PERSONS table.

The year variable is retrieved from the *year\_of\_birth*.

The results table will contain the name of the CDM table, the variable name, the year, count, total and percentage of missing. For easier navigation, the name of the table and the name of the variable will be colored green in presence of missing data. If years before 1995 or years in the future are present in the data those will be colored red.

Counts are calculated as number of missing observations for each variable in a particular year, where total is the number of total observations of the variable in a particular year. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the missing data will be represented with line charts for each of the tables of the CDM, and color coded by the name of variables. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If years before 1995 or years in the future are present in the data those will not be plotted in the graphs. The same goes for counts or totals that are smaller than five.

* Missing data counts stratified by meaning and by calendar year

For tables that contain a meaning variable and a date variable, the missing data analysis will be stratified by both these variables. Tables that are not included in this analysis are CDM\_SOURCE, EUROCAT, INSTANCE, PERSONS, PERSON\_RELATIONSHIPS, PRODUCTS.

In case two date variables are present, for stratification purposes the one that is complete will be used. If both dates are present then the following choice will be made:

MEDICINED: *date\_dispensing* will be used

VACCINES: *vx\_admin\_date* will be used

The results table will contain the name of the CDM table, the variable name, the meaning variable, the year, count, total and percentage of missing. For easier navigation, the name of the table and the name of the variable will be colored green in presence of missing data. If years before 1995 or years in the future are present in the data those will be colored red.

Counts are calculated as number of missing observations for each variable with a recorded meaning in a particular year, where total is the number of total observations of the variable with a recorded meaning in a particular year. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the missing data will be represented with heatmaps for each of the tables of the CDM and each of the meanings present. If years before 1995 or years in the future are present in the data those will not be plotted in the graphs. The same goes for counts or totals that are smaller than five.

**Dates check**

Tables that will undergo this check are variables that contain a date variable, as well as the PERSONS table. The results for the PERSONS table will be presented separately.

* Check for correct format (8 characters).

All date variables will undergo the format check. If an error is found, those date will be excluded from the second and third step of the dates check(allowable values and counts of future dates).

* Check for allowable values for year, month, and day.

Dates that passed the first check with success will undergo the second step. First dates will be checked against allowable values, which are as follow:

year: 1995-present

month:01-12

day:01-31

* Counts of future dates.

A comparison will be made between the date variable and the present dates, and counts of future dates will be reported.

The results table will contain the name of the CDM table, the name of the .csv file, the variable name, error format, error year, error month, error day and future dates. For easier navigation, the error year will be colored blue if dates before 1995 and in the future are present, red if only dates before 1995 are present, green if years in the future only are present and black if no errors are identified. If the error year is smaller than future dates that means that the year part of the date is equal to the present but the day or month can be in the future. Only dates that have errors will be printed in the report, otherwise the following message will be printed "There are no errors in format and allowable values for date variables.".

If all tables present in the working directory had an error in at least one of the preliminary checks explained in figure 5, they are excluded and this check cannot be performed. The report will show the following message:

"This check might not be performed in case all tables included in the directory had a previous error in fields check, presence of variable or lowercase checks. Fix the problem first and then run the script again."

The results table for the PERSONS table will contain the name of the CDM table, the name of the .csv file, the variable name, error format, error month, error day and future dates. If the year part of the date variable is being checked, the error month and error day take values N/A which means Not Applicable and so on.

If the PERSONS table is not present or had an issue in the preliminary checks(except the third check) the following message will be printed:

"The PERSONS table either is not provided or had an error in fields check, presence of variable or lowercase check."

**Output folder: STEP1to3**

* step1to3\_missing\_overall.csv (result of step 2,i)
* step1to3\_missing\_meaning.csv (result of step 2,ii)
* step1to3\_missing\_year.csv (result of step 2,iii)
* step1to3\_missing\_meaning\_year.csv (result of step 2,iv)
* subfolder Masked:
  + 1. step1to3\_missing\_overall\_masked.csv(small counts and totals are replaced)
    2. step1to3\_missing\_meaning\_masked.csv(small counts and totals are replaced)
    3. step1to3\_missing\_year\_masked.csv(small counts and totals are replaced)
    4. step1to3\_missing\_meaning\_year\_masked.csv(small counts and totals are replaced)
* .html markdown report file (found one level up of this folder, in the output directory)

#### CDM\_SOURCE and INSTANCE

Table 3. CDM\_SOURCE table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CDM\_SOURCE** | **Metadata** | In this table, a high-level, machine-readable description of the instance of the CDM is contained. The scripts of the studies that are deemed to run on this instance will use this information to tailor some choices to the specific DAP and data source. | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| data\_access\_provider\_code | Yes | Code of this DAP organization in the ConcePTION coding system | Character | see corresponding vocabulary table |
| data\_access\_provider\_name | Yes | Name of the DAP organization | Character | see corresponding vocabulary table |
| data\_source\_name | Yes | Name of the DAP datasource whose subset populates this instance of the CDM (if any) | Character |  |
| data\_dictionary\_link | No | link to a source where the data dictionary of the original data source can be found | Character |  |
| etl\_link | No | link to a source where the current version of the ETL document of this data source can be found | Character |  |
| cdm\_vocabulary\_version | Yes | version of the ConcePTION CDM this instance conforms to. | Character | see corresponding vocabulary table |
| cdm\_version | No | version of the ConcePTION CDM vocabulary this instance conforms to. | Character | see corresponding vocabulary table |
| instance\_number | Yes | Sequential number of the instances of the CDM that the DAP data\_access\_provider\_code has created on date\_creation date from the datasource data\_source\_name it has access to | Integer |  |
| date\_creation | Yes | date when this CDM instance is populated | Character yyyymmdd |  |

Table 4. INSTANCE table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **INSTANCE** | **Metadata** | This table displays the list of the tables and columns of the origin data dictionary that are mapped to the instance of the CDM, together with date of last update (both in terms of when the data was accessed by the DAPs, and when the data was actually recorded and can be considered complete). This is to be used, together with a machine-readable version of the ETL, to match the inclusion of the study population and the creation of the study variables to the actual data loaded in the CDM instance. The list is restricted to tables and columns of the origin data dictionary that are included in the current ETL document. | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| source\_table\_name | Yes | Table of the local dictionary that is used in ETL | Character | see corresponding vocabulary table |
| source\_column\_name | Yes | Column of the local dictionary that is used in ETL | Character |  |
| included\_in\_instance | Yes | Specify whether this column of this table was used to populate this specific instance of the CDM | Character | yes/no |
| date\_when\_data\_last\_updated | only if included\_in\_instance='yes' | Date when the DAP last received this column | Character yyyymmdd | N/A=not applicable |
| since\_when\_data\_complete | only if included\_in\_instance='yes' | Date since when the DAP considers this column to contain complete data | Character yyyymmdd | N/A=not applicable |
| up\_to\_when\_data\_complete | only if included\_in\_instance='yes' | Date up to when the DAP considers this column to contain complete data | Character yyyymmdd | N/A=not applicable |
| restriction\_in\_values | only if included\_in\_instance='yes' | Whether the data uploaded in the CDM were selected based on values of this column | Character | yes/no/N/A |
| list\_of\_values | No | List of values of this column that are included in this instance of the CDM | Character |  |
| restriction\_condition | No | Condition involving this column that restricted the data uploaded to the instance of the CDM | Character |  |

**Presence Check**

The first step is to check whether the CDM\_SOURCE and INSTANCE tables are present in the directory. The results will be shown on a table that contains the name of the CDM table, presence Yes/No, error TRUE/FALSE and a comment explaining the error.

The checks look into three different scenarios:

* More than one file matching the name of the CDM\_SOURCE or INSTANCE.

In this case for both tables the presence is set to Yes, error to TRUE and the following comment will be displayed.

“There are more than two files that contain the pattern CDM\_SOURCE in your working directory. There should be only one. Fix the error and run the script again."”

“There are more than two files that contain the pattern INSTANCE in your working directory. There should be only one. Fix the error and run the script again.”

* No files matching the name of CDM\_SOURCE or INSTANCE.

In this case the presence will be set to No, the error to TRUE for the CDM\_SOURCE table since is mandatory and FALSE for the INSTANCE table since is not mandatory. The comments that will be printed are:

“There are no files that contain the pattern CDM\_SOURCE in your working directory. Fix the error and run the script again.”

“The INSTANCE table is not present in the working directory.”

* Only one file that contains the pattern CDM\_SOURCE or INSTANCE.

In this case presence will be set to Yes, error to FALSE and the comments that will be displayed are as follows:

“CDM\_SOURCE is present in the working directory.”

“INSTANCE table is present in the working directory.”

**Number of fields check**

If the INSTANCE table is provided, the fields checks will be performed for both tables. If an error is found the script will stop(Stopping criteria).

The results table contains the name of the CDM table, the name of the .csv file, the number of rows of the .csv, the number of rows currently being imported, the first row of the .csv file where the error was encountered, and error set to TRUE. If no error is found nothing will be reported.

**Check conventions**

The conventions for both tables, if the INSTANCE table is present will be checked.

The conventions for the CDM\_SOURCE table are as follows:

This table contains only one row.

The variable *instance\_number* is expected to be always 1.

The conventions for the INSTANCE table are as follows:

The values of *date\_when\_data\_last\_updated* are expected to be the same for all columns of the same table.

The values *of since\_when\_data\_complete* are expected to be the same for all columns of the same table.

The values of *up\_to\_when\_data\_complete* are expected to be the same for all columns of the same table.

If *included\_in\_instance* is set to `yes` then *restricion\_in\_values* should to be filled out.

If *restriction\_in\_values* is set to `no` then *list\_of\_values* should be empty.

*If restriction\_in\_values* is set to `no` then *restriction\_condition* should be empty.

If there are any deviations, those will be printed in the report. The failed convention is accompanied by the comment “Convention no. is not satisfied.” together with the explanation why.

If no deviations are found, the following message will be printed:

“All conventions are satisfied.”

**Information about DAP and CDM version**

The information present in the CDM\_SOURCE table will be printed to the report.

**INSTANCE table**

If present, the INSTANCE table will be printed to the report.

**Output folder: CDM\_SOURCE and INSTANCE**

* CDM\_SOURCE.csv (a copy of the CDM\_SOURCE table)
* .html markdown report file (found one level up of this folder, in the output directory)

#### STEP 4 to 5

Subsequently, checks described in steps 4 and 5 will be conducted for each table as follows:

**Routine healthcare data**

##### **VISIT\_OCCURRENCE table**

Table 5. VISIT\_OCCURRENCE table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **VISIT\_OCCURRENCE** | **Routine healthcare data** | This table contains a summary description of the visits during which records of EVENTS, PROCEDURES, but possibly also MEDICAL\_OBSERVATIONS or VACCINES or MEDICINES were recorded. This serves both to collect visit-level information, and to enable grouping sets of records that were recorded concurrently | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes | A foreign key to the person in PERSONS table | Character |  |
| visit\_occurrence\_id | Yes | Visit identifier | Character |  |
| visit\_start\_date | Yes | Date when the visit starts, or, if it is just a one-day visit, date of the visit | Character yyyymmdd |  |
| visit\_end\_date | No | Date when the visit ends (only for visits that may last more than one day, such as a hospital admission) | Character yyyymmdd |  |
| specialty\_of\_visit | No | Specialty of the visit, or if this is a hospital admission, specialty of the discharge ward | Character |  |
| specialty\_of\_visit\_vocabulary | No | Coding system of the specialty | Character | see corresponding vocabulary table |
| status\_at\_discharge | No | outcome of the visit | Character |  |
| status\_at\_discharge\_vocabulary | No | vocabulary of outcome of the visit | Character | see corresponding vocabulary table |
| meaning\_of\_visit | Yes |  | Character | see corresponding vocabulary table |
| origin\_of\_visit | Yes | table source name that originated the visit record | Character | see corresponding vocabulary table |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The VISIT\_OCCURRENCE table has 1 convention that can be checked by the script. This convention will be mentioned in the beginning of this section of the report and is as follows:

* Every record of this table refers to a different visit (*visit\_occurrence\_id* is a primary key). This primary key should not be duplicated.

If deviations are found the following message will be printed in the report.

"Convention 1 is not satisfied. There is(are) *no.* duplicated visit\_occurrence\_id(s)."

If no deviations are found the report will show:

“All conventions are satisfied.”

**Counts of categorical variables**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the year part of *visit\_start\_date*) counts will be tabulated. In both cases the results will be stratified by *meaning\_of\_visit*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “VISIT\_OCCURRENCE”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories when stratifying by meaning will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results for variables with two or more categories when stratifying by meaning and year will be displayed by line charts for each of the variables, grouped by the meaning variable and color coded by the vocabulary variable. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Visually the results for other variables when stratifying by meaning and year will be displayed by line charts for each of the variables and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*specialty\_of\_visit\_vocabulary*: number of complete observations per category

*status\_at\_discharge\_vocabulary*: number of complete observations per category

*meaning\_of\_visit*: number of complete observations per category

*origin\_of\_visit*: number of complete observations per category

* Other variables:

*person\_id*: number of unique persons included in this table

*visit\_occurrence\_id*: number of unique visit identification number

*specialty\_of\_visit*: number of complete observations

*status\_at\_discharge*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

There are no continuous variables in the VISIT\_OCCURRENCE table.

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable that is being counted. In both cases the results will be stratified by *meaning\_of\_visit*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “VISIT\_OCCURRENCE”, name of the variable, meaning variable, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results when stratifying by meaning will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results when stratifying by meaning and year will be displayed by line charts for each of the variables, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *visit\_start\_date*: number of complete observations
* *visit\_end\_date*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: VISIT\_OCCURRENCE**

* visits\_meaning\_2categories.csv
* visits\_meaning\_other.csv
* visits\_meaning\_dates.csv
* visits\_meaning\_year\_2categories.csv
* visits\_meaning\_year\_other.csv
* visits\_meaning\_year\_dates.csv
* Masked subfolder:
* visits\_meaning\_2categories\_masked.csv
* visits\_meaning\_other\_masked.csv
* visits\_meaning\_dates\_masked.csv
* visits\_meaning\_year\_2categories\_masked.csv
* visits\_meaning\_year\_other\_masked.csv
* visits\_meaning\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

##### **EVENTS table**

Table 6. EVENTS table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| EVENTS | Routine healthcare data | This table contains a summary description of the visits during which records of EVENTS, PROCEDURES, but possibly also MEDICAL\_OBSERVATIONS or VACCINES or MEDICINES were recorded. This serves both to collect visit-level information, and to enable grouping sets of records that were recorded concurrently | | |
| Variable | Mandatory | Description | Format | Vocabulary |
| person\_id | Yes | A foreign key to the person in "person" table who experimented the event | Character |  |
| start\_date\_record | Yes | Start date of the visit that led to the recording of the event code of free text | Character yyyymmdd |  |
| end\_date\_record | No | End date of the visit that led to the recording of the event code of free text | Character yyyymmdd |  |
| event\_code | Yes, unless ‘event\_free\_text’ is filled in | Code characterizing the event according to the vocabulary defined in event\_record\_vocabulary | Character |  |
| event\_record\_vocabulary | Yes | Vocabulary to which the event\_code belongs to; or, if the record contains event\_free\_text , this column contains the indication 'free\_text' | Character | see corresponding vocabulary table |
| text\_linked\_to\_event\_code | No | If in the original record the code is modified by a text, include this text here | Character |  |
| event\_free\_text | No | Use this cell if in the record there is no code, just a text | Character |  |
| present\_on\_admission | No | Indicates the presence of the event at the start of the visit or hospital admission | Character | yes/no |
| laterality\_of\_event | No | Laterality of the event | Character | right/left/both |
| meaning\_of\_event | Yes | This is a ConcePTION classification of the nature of the original record associated with this event | Character | see corresponding vocabulary table |
| origin\_of\_event | Yes | table source name that originated the event record | Character | see corresponding vocabulary table |
| visit\_occurrence\_id | No | A foreign key linking this record to the VISIT\_OCCURRENCE table | Character |  |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The EVENTS table has 3 conventions that can be checked by the script. These conventions will be mentioned in the beginning of this section of the report and are as follows:

* For each record either an *event\_code* or an *event \_free\_text* must be present.
* *Event\_record\_vocabulary* should be filled out every time (no missing data allowed).
* *Event\_record\_vocabulary* needs to be assigned to free\_text\_*original\_language (e.g., free\_text\_italian)* in case of *event\_free\_text* record.

If deviations are found the following messages will be reported:

"Convention 1 is not satisfied. There is(are) no. observation(s) that have missing values in both event\_code and event\_free\_text."

Depending on the situation you can get one of the following messages if convention 2 is violated:

"Convention 2 is not satisfied. There is(are) *no.* observation(s) that have missing data on event\_record\_vocabulary and there is(are) *no.* observation(s) that have data on event\_record\_vocabulary when neither event\_code or event\_free\_text is complete."

"Convention 2 is not satisfied. There is(are) *no.* observation(s) that have missing data on event\_record\_vocabulary."

"Convention 2 is not satisfied. There is(are) *no.* observation(s) that have data on event\_record\_vocabulary when neither event\_code or event\_free\_text is complete."

"Convention 3 is not satisfied. There is(are) *no.* observation(s) where event\_record\_vocabulary is not equal to free\_text when event\_free\_text is filled out."

If no deviations are found the report will show:

“All conventions are satisfied.”

**Counts of categorical variables**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the year part of *start\_date\_record*) counts will be tabulated. In both cases the results will be stratified by *meaning\_of\_event*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “EVENTS”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories when stratifying by meaning will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results for variables with two or more categories when stratifying by meaning and year will be displayed by line charts for each of the variables, grouped by the meaning variable and color coded by the vocabulary variable. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Visually the results for other variables when stratifying by meaning and year will be displayed by line charts for each of the variables and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*event\_record\_vocabulary*: number of complete observations per category

*present\_on\_admission*: number of complete observations per category

*laterality\_of\_event*: number of complete observations per category

*meaning\_of\_event*: number of complete observations per category

*origin\_of\_event*: number of complete observations per category

* Other variables:

*person\_id*: number of persons included in this table

*event\_code*: number of complete observations

*event\_free\_text*: number of complete observations

*text\_linked\_to\_event\_code*: number of complete observations

*visit\_occurrence\_id*: number of unique visit id

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

There are no continuous variables in the EVENTS table.

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable that is being counted. In both cases the results will be stratified by *meaning\_of\_event*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “EVENTS”, name of the variable, meaning variable, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results when stratifying by meaning will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results when stratifying by meaning and year will be displayed by line charts for each of the variables, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *visit\_start\_date*: number of complete observations
* *visit\_end\_date*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: EVENTS**

* events\_meaning\_2categories.csv
* events\_meaning\_other.csv
* events\_meaning\_dates.csv
* events\_meaning\_year\_2categories.csv
* events\_meaning\_year\_other.csv
* events\_meaning\_year\_dates.csv
* Masked subfolder:
* events\_meaning\_2categories\_masked.csv
* events\_meaning\_other\_masked.csv
* events\_meaning\_dates\_masked.csv
* events\_meaning\_year\_2categories\_masked.csv
* events\_meaning\_year\_other\_masked.csv
* events\_meaning\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

##### **MEDICINES table**

Table 7. MEDICINES table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MEDICINES** | **Routine healthcare data** | This table collects data on drug prescriptions, dispensings or administrations occurred during routine healthcare. | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes | A foreign key to the person in PERSONS table | Character |  |
| medicinal\_product\_id | Yes | Foreign key. The medicinal\_product\_id should be a unique identifier of a specific medicinal product. | Character |  |
| medicinal\_product\_atc\_code | Yes | ATC classification system code attributed to the medicinal product. | Character |  |
| date\_dispensing | Yes, unless 'date\_prescription' is populated | Date when the medicinal product that lead to the recording was dispensed or administrated to the patient | Character yyyymmdd |  |
| date\_prescription | Yes, unless 'date\_dispensing' is populated | Date when the medicinal product that lead to the recording was prescribed | Character yyyymmdd |  |
| disp\_number\_medicinal\_product | No | Number of dispensed units of medicinal\_product\_id. | Numeric |  |
| presc\_quantity\_per\_day | No | Prescribed quantity of medicinal product to be taken daily. | Numeric |  |
| presc\_quantity\_unit | No | Unit of measure of the prescribed daily quantity. | Character | see corresponding vocabulary table |
| presc\_duration\_days | No | Number of days of medication as prescribed. | Numeric |  |
| product\_lot\_number | No | An identifier assigned to a particular quantity or lot of medicinal product from the manufacturer. | Character |  |
| inidication\_code | No | Single identifier of a condition/indication for which the medicinal product was prescribed/dispensed. | Character |  |
| indication\_code\_vocabulary | No | Coding system referring to indication code | Character | see corresponding vocabulary table |
| meaning\_of\_drug\_record | Yes | nature of the original record having originated the drug record | Character | see corresponding vocabulary table |
| origin\_of\_drug\_record | Yes | table source name that originated the drug record | Character | see corresponding vocabulary table |
| prescriber\_speciality | No | Profile of the healthcare professional who has prescribed the medicinal product. | Character | see corresponding vocabulary table |
| visit\_occurrence\_id | No | Identifier of the prescription. A foreign key linking this record to the VISIT\_OCCURRENCE table, indicating the visit where the drug was prescribed or dispensed. | Character |  |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The MEDICINES table has 1 convention that can be checked by the script. These conventions will be mentioned in the beginning of this section of the report and is as follows:

* In case original record are prescription data *date\_prescription* has to be populated, if they are dispensing data then *date\_dispensing* has to be populated.

If deviations are found the following messages will be reported:

“Convention 1 is not satisfied. There is(are) no. observation(s) that have missing values in both date\_prescription and date\_dispensing.”

If no deviations are found the report will show:

“All conventions are satisfied.”

**Counts of categorical variables**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the year part of *date\_dispensing* when *date\_prescription* is missing or both are present, and *date\_prescription* when *date\_dispensing* is missing) counts will be tabulated. In both cases the results will be stratified by *meaning\_of\_drug\_record*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “MEDICINES”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories when stratifying by meaning will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results for variables with two or more categories when stratifying by meaning and year will be displayed by line charts for each of the variables, grouped by the meaning variable and color coded by the vocabulary variable. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Visually the results for other variables when stratifying by meaning and year will be displayed by line charts for each of the variables and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*presc\_quantity\_unit*: number of complete observations per category

*indication\_code\_vocabulary*: number of complete observations per category

*meaning\_of\_drug\_record*: number of complete observations per category

*origin\_of\_drug\_record*: number of complete observations per category

*prescriber\_speciality*: number of complete observations per category

* Other variables:

*person\_id:* number of persons included in this table

*visit\_occurrence\_id:* number of unique visit id

*medicinal\_product\_id:* number of complete observations

*medicinal\_product\_atc\_code:* number of complete observations

*disp\_number\_medicinal\_product:* number of complete observations

*presc\_quantity\_per\_day:* number of complete observations

*presc\_duration\_days:* number of complete observations

*product\_lot\_number:* number of complete observations

*inidication\_code:* number of complete observations

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

All results will be stratified by year (according to the year part of *date\_dispensing* when *date\_prescription* is missing or both are present, and *date\_prescription* when *date\_dispensing* is missing) and the ATC code (truncated to the third level).

*disp\_number\_medicinal\_product*: mean, median, interquartile range, skewness

*presc\_quantity\_per\_day* stratified by *presc\_quantity\_unit*: mean, median, interquartile range, skewness

*presc\_duration\_days*: mean, median, interquartile range, skewness

Visually it will be represented by a line graph for each variable.

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable that is being counted. In both cases the results will be stratified by *meaning\_of\_drug\_record*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “MEDICINES”, name of the variable, meaning variable, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results when stratifying by meaning will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results when stratifying by meaning and year will be displayed by line charts for each of the variables, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *date\_prescription*: number of complete observations
* *date\_dispensing*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: MEDICINES**

* medicines\_meaning\_2categories.csv
* medicines\_meaning\_other.csv
* medicines\_meaning\_dates.csv
* medicines\_meaning\_year\_2categories.csv
* medicines\_meaning\_year\_other.csv
* medicines\_meaning\_year\_dates.csv
* medicines\_disp\_num\_medicinal\_product\_dist.csv (if available)
* medicines\_presc\_quantity\_per\_day\_dist.csv (if available)
* medicines\_presc\_duration\_days\_dist.csv (if available)
* Masked subfolder:
* medicines\_meaning\_2categories\_masked.csv
* medicines\_meaning\_other\_masked.csv
* medicines\_meaning\_dates\_masked.csv
* medicines\_meaning\_year\_2categories\_masked.csv
* medicines\_meaning\_year\_other\_masked.csv
* medicines\_meaning\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

##### **PROCEDURES table**

Table 8. PROCEDURES table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PROCEDURES** | **Routine healthcare** | This table collects procedures administered during routine healthcare. Can be a surgery, or a diagnostic procedure, a rehabilitation procedure, a therapeutical procedure. | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes |  | Character |  |
| procedure\_date | Yes |  | Character yyyymmdd |  |
| procedure\_code | Yes |  | Character |  |
| procedure\_code\_vocabulary | Yes |  | Character | see corresponding vocabulary table |
| visit\_occurrence\_id | No | A foreign key linking this record to the VISIT\_OCCURRENCE table | Character |  |
| meaning\_of\_procedure | Yes |  | Character | see corresponding vocabulary table |
| origin\_of\_procedure | Yes | table source name that originated the procedure record | Character | see corresponding vocabulary table |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The PROCEDURES table does not contain any conventions that can be checked by the script.

**Counts of categorical variables**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the year part of *procedure\_date*) counts will be tabulated. In both cases the results will be stratified by *meaning\_of\_procedure*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “PROCEDURES”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories when stratifying by meaning will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results for variables with two or more categories when stratifying by meaning and year will be displayed by line charts for each of the variables, grouped by the meaning variable and color coded by the vocabulary variable. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Visually the results for other variables when stratifying by meaning and year will be displayed by line charts for each of the variables and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*procedure\_code\_vocabulary*: number of complete observations per category

*meaning\_of\_procedure*: number of complete observations per category

*origin\_of\_procedure*: number of complete observations per category

* Other variables:

*person\_id*: number of unique persons included in this table

*procedure\_code*: number of complete observations

*visit\_occurrence\_id*: number of unique visit id

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

There are no continuous variables in the PROCEDURES table.

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable *procedure\_date*. In both cases the results will be stratified by *meaning\_of\_procedure*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “PROCEDURES”, name of the variable, meaning variable, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results when stratifying by meaning will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results when stratifying by meaning and year will be displayed by line charts for each of the variables, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *procedure\_date*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: PROCEDURES**

* procedures\_meaning\_2categories.csv
* procedures\_meaning\_other.csv
* procedures\_meaning\_dates.csv
* procedures\_meaning\_year\_2categories.csv
* procedures\_meaning\_year\_other.csv
* procedures\_meaning\_year\_dates.csv
* Masked subfolder:
* procedures\_meaning\_2categories\_masked.csv
* procedures\_meaning\_other\_masked.csv
* procedures\_meaning\_dates\_masked.csv
* procedures\_meaning\_year\_2categories\_masked.csv
* procedures\_meaning\_year\_other\_masked.csv
* procedures\_meaning\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

##### **VACCINES table**

Table 9. VACCINES table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **VACCINES** | **Routine healthcare data** | This table collects dispensations or administrations of vaccines. | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes |  | Character |  |
| vx\_record\_date | Yes, if vx\_admin\_date is missing |  | Character yyyymmdd |  |
| vx\_admin\_date | Yes, if vx\_record\_date is missing |  | Character yyyymmdd |  |
| vx\_atc | Yes, if vx\_type is missing |  | Character |  |
| vx\_type | Yes, if vx\_atc is missing |  | Character | see corresponding vocabulary table |
| vx\_text | No |  | Character |  |
| product\_code | No | external key to PRODUCTS | Character |  |
| origin\_of\_vx\_record | Yes | table source name that originated the vaccine record | Character | see corresponding vocabulary table |
| meaning\_of\_vx\_record | Yes |  | Character | see corresponding vocabulary table |
| vx\_dose | No | Dose, particularly for childhood vaccines (1, 2, 3, Booster, etc.) | Character | see corresponding vocabulary table |
| vx\_manufacturer | No | Name of vaccine manufacturer | Character |  |
| vx\_lot\_num | No |  | Character |  |
| visit\_occurrence\_id | No | external key to VISIT\_OCCURRENCE | Character |  |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The VACCINES table has 2 conventions that can be checked by the script. These conventions will be mentioned in the beginning of this section of the report and is as follows:

* At least one of *vx\_record\_date* and *vx\_admin\_date* must be populated. If both are available both should be populated.
* At least one of *vx\_atc* and *vx\_type* date must be populated. If both are available both should be populated.

If deviations are found the following messages will be reported:

"Convention 1 is not satisfied. There is(are) *no.* observation(s) that have missing values in all variables vx\_record\_date and vx\_admin\_date."

"Convention 2 is not satisfied. There is(are) *no.* observation(s) that have missing values in all variables vx\_atc and vx\_type."

If no deviations are found the report will show:

“All conventions are satisfied.”

**Counts of categorical variables**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the year part of *vx\_admin\_date* when *vx\_record\_date* is missing or both are present, and *vx\_record\_date* when *vx\_admin\_date* is missing) counts will be tabulated. In both cases the results will be stratified by *meaning\_of\_vx\_record*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “VACCINES”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories when stratifying by meaning will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results for variables with two or more categories when stratifying by meaning and year will be displayed by line charts for each of the variables, grouped by the meaning variable and color coded by the vocabulary variable. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Visually the results for other variables when stratifying by meaning and year will be displayed by line charts for each of the variables and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*vx\_type*: number of complete observations per category

*meaning\_of\_vx\_record*: number of complete observations per category

*origin\_of\_vx\_record*: number of complete observations per category

*vx\_dose*: number of complete observations per category

* Other variables:

*person\_id*: number of unique persons included in this table

*vx\_atc*: number of complete observations

*vx\_text*: number of complete observations

*product\_code*: number of complete observations

*vx\_manufacturer*: number of complete observations

*vx\_lot\_num*: number of complete observations

*visit\_occurrence\_id*: number of unique visit id

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

There are no continuous variables in the VACCINES table.

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable that is being counted. In both cases the results will be stratified by *meaning\_of\_vx\_record*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “VACCINES”, name of the variable, meaning variable, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results when stratifying by meaning will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results when stratifying by meaning and year will be displayed by line charts for each of the variables, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *vx\_admin\_date*: number of complete observations
* *vx\_record\_date*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: VACCINES**

* vaccines\_meaning\_2categories.csv
* vaccines\_meaning\_other.csv
* vaccines\_meaning\_dates.csv
* vaccines\_meaning\_year\_2categories.csv
* vaccines\_meaning\_year\_other.csv
* vaccines\_meaning\_year\_dates.csv
* Masked subfolder:
* vaccines\_meaning\_2categories\_masked.csv
* vaccines\_meaning\_other\_masked.csv
* vaccines\_meaning\_dates\_masked.csv
* vaccines\_meaning\_year\_2categories\_masked.csv
* vaccines\_meaning\_year\_other\_masked.csv
* vaccines\_meaning\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

##### **MEDICAL\_OBSERVATIONS table**

Table 10. MEDICAL\_OBSERVATIONS table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MEDICAL\_OBSERVATIONS** | **Routine healthcare data** | This table collects observations recorded during routine healthcare. Can be a result from a laboratory test, or a physical measurement, but also level of education, or sex, or a pathology report | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes |  | Character |  |
| mo\_date | Yes |  | Character yyyymmdd |  |
| mo\_code | No |  | Character |  |
| mo\_record\_vocabulary | No |  | Character | see corresponding vocabulary table |
| mo\_source\_table | No |  | Character | see corresponding vocabulary table |
| mo\_source\_column | No |  | Character |  |
| mo\_source\_value | Yes |  | Character |  |
| mo\_unit | No |  | Character | see corresponding vocabulary table |
| mo\_meaning | Yes |  | Character | see corresponding vocabulary table |
| mo\_origin | Yes | table source name that originated the record | Character | see corresponding vocabulary table |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The MEDICAL\_OBSERVATIONS table does not have any conventions that can be checked by the script.

**Counts of categorical variables**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the year part of *mo\_date*) counts will be tabulated. In both cases the results will be stratified by *mo\_meaning*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “MEDICAL\_OBSERVATIONS”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories when stratifying by meaning will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results for variables with two or more categories when stratifying by meaning and year will be displayed by line charts for each of the variables, grouped by the meaning variable and color coded by the vocabulary variable. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Visually the results for other variables when stratifying by meaning and year will be displayed by line charts for each of the variables and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*mo\_record\_vocabulary*: number of complete observations per category

*mo\_source\_table*: number of complete observations per category

*mo\_unit*: number of complete observations per category

*mo\_meaning*: number of complete observations per category

*mo\_origin*: number of complete observations per category

* Other variables

*person\_id*: number of unique persons included in this table

*mo\_code*: number of complete observations

*mo\_source\_column*: number of complete observations

*mo\_source\_value*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

The distribution of *mo\_source\_value* will be displayed only for numeric values of this variable:

*mo\_source\_value*: mean, median, interquartile range, skewness, kurtosis

Results are stratified by *mo\_source\_table*, *mo\_source\_column*, and *mo\_unit* and year (retrieved from the year part of *mo\_date*).

Visually it will be presented by a line graph.

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable *mo\_date*. In both cases the results will be stratified by *mo\_meaning*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “MEDICAL\_OBSERVATIONS”, name of the variable, meaning variable, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results when stratifying by meaning will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results when stratifying by meaning and year will be displayed by line charts for each of the variables, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *mo\_date*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: MEDICAL\_OBSERVATIONS**

* medical\_observations\_meaning\_2categories.csv
* medical\_observations\_meaning\_other.csv
* medical\_observations\_meaning\_dates.csv
* medical\_observations\_meaning\_year\_2categories.csv
* medical\_observations\_meaning\_year\_other.csv
* medical\_observations\_meaning\_year\_dates.csv
* medical\_observations\_mo\_source\_value\_dist.csv (if available)
* Masked subfolder:
* medical\_observations\_meaning\_2categories\_masked.csv
* medical\_observations\_meaning\_other\_masked.csv
* medical\_observations\_meaning\_dates\_masked.csv
* medical\_observations\_meaning\_year\_2categories\_masked.csv
* medical\_observations\_meaning\_year\_other\_masked.csv
* medical\_observations\_meaning\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

**Surveillance**

##### **SURVEY\_ID table**

Table 11. SURVEY\_ID table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SURVEY\_ID** | **Surveillance** | This table contains a summary description of the survey during which records of SURVEY\_OBSERVATIONS were recorded. This serves both to collect survey-level information, and to enable grouping sets of records that were recorded concurrently | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes | Person whose information is collected in this survey | Character |  |
| survey\_id | Yes | Identifier of the survey | Character |  |
| survey\_date | Yes | Date when the survey is recorded | Character yyyymmdd |  |
| survey\_meaning | Yes | The meaning of this survey for this person | Character | see corresponding vocabulary table |
| survey\_origin | Yes | table source name that originated the record | Character | see corresponding vocabulary table |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The SURVEY\_ID table has 1 convention that can be checked by the script. This convention will be mentioned in the beginning of this section of the report and is as follows:

* Every record of this table refers to a different survey for a different subject in a different date: survey\_id + person\_id + survey\_date is a primary key.

If deviations are found the following messages will be reported:

"Convention 1 is not satisfied. There is(are) *no.* observation(s) that have duplicated combination between person\_id, survey\_id and survey\_date."

If no deviations are found the report will show:

“All conventions are satisfied.”

**Counts of categorical variables.**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the year part of *survey\_date*) counts will be tabulated. In both cases the results will be stratified by *survey\_meaning*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “SURVEY\_ID”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories when stratifying by meaning will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results for variables with two or more categories when stratifying by meaning and year will be displayed by line charts for each of the variables, grouped by the meaning variable and color coded by the vocabulary variable. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Visually the results for other variables when stratifying by meaning and year will be displayed by line charts for each of the variables and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*survey\_meaning*: number of complete observations per category

*survey\_origin*: number of complete observations per category

* Other variables

*person\_id*: number of unique persons included in this table

*survey\_id*: number of unique survey id

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

There are no continuous variables in the SURVEY\_ID table.

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable *survey\_date*. In both cases the results will be stratified by *survey\_meaning*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “SURVEY\_ID”, name of the variable, meaning variable, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results when stratifying by meaning will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results when stratifying by meaning and year will be displayed by line charts for each of the variables, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *survey\_date*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: SURVEY\_ID**

* survey\_id\_meaning\_2categories.csv
* survey\_id\_meaning\_other.csv
* survey\_id\_meaning\_dates.csv
* survey\_id\_meaning\_year\_2categories.csv
* survey\_id\_meaning\_year\_other.csv
* survey\_id\_meaning\_year\_dates.csv
* Masked subfolder:
* survey\_id\_meaning\_2categories\_masked.csv
* survey\_id\_meaning\_other\_masked.csv
* survey\_id\_meaning\_dates\_masked.csv
* survey\_id\_meaning\_year\_2categories\_masked.csv
* survey\_id\_meaning\_year\_other\_masked.csv
* survey\_id\_meaning\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

##### **SURVEY\_OBSERVATIONS table**

Table 12. SURVEY\_OBSERVATIONS table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SURVEY\_OBSERVATIONS** | **Surveillance** | List of observations in a survey | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes |  | Character |  |
| so\_date | Yes |  | Character |  |
| so\_source\_table | Yes |  | Character | see corresponding vocabulary table |
| so\_source\_column | Yes |  | Character |  |
| so\_source\_value | Yes |  | Character |  |
| so\_unit | No |  | Character | see corresponding vocabulary table |
| so\_meaning | Yes |  | Character | see corresponding vocabulary table |
| so\_origin | Yes | table source name that originated the record | Character | see corresponding vocabulary table |
| survey\_id | Yes |  | Character |  |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The SURVEY\_OBSERVATIONS table has 1 convention that can be checked by the script. This convention will be mentioned in the beginning of this section of the report and is as follows:

* Every record of this table must refer to a record in SURVEY\_ID.

If the convention is not satisfied than “Convention 1 is not satisfied. Take caution!”.

**Counts of categorical variables**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the year part of *so\_date*) counts will be tabulated. In both cases the results will be stratified by *so\_meaning*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “SURVEY\_OBSERVATIONS”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories when stratifying by meaning will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results for variables with two or more categories when stratifying by meaning and year will be displayed by line charts for each of the variables, grouped by the meaning variable and color coded by the vocabulary variable. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Visually the results for other variables when stratifying by meaning and year will be displayed by line charts for each of the variables and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*so\_source\_table*: number of complete observations per category

*so\_unit*: number of complete observations per category

*so\_meaning*: number of complete observations per category

*so\_origin*: number of complete observations per category

* Other variables:

*person\_id*: number of persons included in this table

*so\_source\_column*: number of complete observations

*so\_source\_value*: number of complete observations

*survey\_id*: number of unique survey id

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

The distribution of *so\_source\_value* will be displayed only for numeric values of this variable:

*so\_source\_value*: mean, median, interquartile range, skewness, kurtosis

Results are stratified by *so\_source\_table*, *so\_source\_column*, and *so\_unit* and year (retrieved from the year part of *so\_date*).

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable *so\_date*. In both cases the results will be stratified by *so\_meaning*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “SURVEY\_OBSERVATIONS”, name of the variable, meaning variable, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results when stratifying by meaning will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results when stratifying by meaning and year will be displayed by line charts for each of the variables, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *so\_date*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: SURVEY\_OBSERVATIONS**

* survey\_observations\_meaning\_2categories.csv
* survey\_ observations\_meaning\_other.csv
* survey\_ observations\_meaning\_dates.csv
* survey\_ observations\_meaning\_year\_2categories.csv
* survey\_ observations\_meaning\_year\_other.csv
* survey\_ observations\_meaning\_year\_dates.csv
* surcey\_observations\_so\_source\_value\_dist.csv (if available)
* Masked subfolder:
* survey\_ observations\_meaning\_2categories\_masked.csv
* survey\_ observations\_meaning\_other\_masked.csv
* survey\_ observations\_meaning\_dates\_masked.csv
* survey\_ observations\_meaning\_year\_2categories\_masked.csv
* survey\_ observations\_meaning\_year\_other\_masked.csv
* survey\_ observations\_meaning\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

**Curated tables**

##### **PERSONS table**

Table 13. PERSONS table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PERSONS** | **Curated tables** | This table records persons that are to enter analysis of this instance of the CDM | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes |  | Character |  |
| day\_of\_birth | No |  | character dd |  |
| month\_of\_birth | No |  | character mm |  |
| year\_of\_birth | Yes |  | character yyyy |  |
| day\_of\_death | No |  | character dd |  |
| month\_of\_death | No |  | character mm |  |
| year\_of\_death | Yes |  | character yyyy |  |
| sex\_at\_instance\_creation | Yes | Sex of the person in the moment when in the instance of the CDM is created | Character | see corresponding vocabulary table |
| race | No |  | Character |  |
| country\_of\_birth | No |  | Character |  |
| quality | No | A judgement on the quality of the variables recorded in this table | Character | see corresponding vocabulary table |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The PERSONS table has 1 convention that can be checked by the script. These conventions will be mentioned in the beginning of this section of the report and is as follows:

* Every person\_id has a single row, meaning that person\_id is a primary key.

If the convention is not satisfied than “Convention no. is not satisfied. There is(are) no. of duplicates person\_id.”

**Counts of categorical variables**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the *year\_of\_birth*) counts will be tabulated. In case both count and total are both equal to zero, the following messages will be printed in top of the results table:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular year.”

The results table will contain the name of the table “PERSONS”, name of the variable, year variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories will be displayed by bar charts and color coded by the vocabulary variable for each of the years. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*sex\_at\_instance\_creation*: number of complete observations per category

*quality*: number of complete observations per category

* Other variables:

*person\_id*: number of persons included in this table

*race*: number of complete observations

*country\_of\_birth*: number of complete observations

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

There are no continuous variables in the PERSONS table.

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable that is being counted. In case both count and total are both equal to zero, the following messages will be printed in top of the results table:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular year.”

The results table when will contain the name of the table “PERSONS”, name of the variable, year, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. Six indicator variables will be created to assess counts for different scenarios:

* dmy\_b: day, month and year of birth present
* my\_b: day of birth missing, and month and year of birth present
* dy\_b: month of birth missing, and day and year of birth present
* dmy\_d: day, month and year of death present
* my\_d: day of death missing and month and year of death present
* dy\_d: month of death missing, and day and year of death present

Visually the results will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *day\_of\_birth*: number of complete observations
* *month\_of\_birth*: number of complete observations
* *year\_of\_birth*: number of complete observations
* *day\_of\_death*: number of complete observations
* *month\_of\_death*: number of complete observations
* *year\_of\_death*: number of complete observations
* *dmy\_b*: number of complete observations
* *my\_b*: number of complete observations
* *dy\_b*: number of complete observations
* *dmy\_d*: number of complete observations
* *my\_d*: number of complete observations
* *dy\_d*: number of complete observations

Total: Number of total observations in a particular year.

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: PERSONS**

* persons\_year\_2categories.csv
* persons\_year\_other.csv
* persons\_year\_dates.csv
* Masked subfolder:
* persons\_year\_2categories\_masked.csv
* persons\_year\_other\_masked.csv
* persons\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

##### **OBSERVATION\_PERIODS table**

Table 14. OBSERVATION\_PERIODS table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **OBSERVATION\_PERIODS** | **Curated tables** | Periods during which data is collected in the datasource for this person. This table is a starting point to define the study population of all studies based on this instance | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes |  | Character |  |
| op\_start\_date | Yes |  | Character yyyymmdd |  |
| op\_end\_date | Yes |  | Character yyyymmdd |  |
| op\_origin | Yes | represents the source table that originated the record | Character | see corresponding vocabulary table |
| op\_meaning | Yes | represents the semantic of the record | Character | see corresponding vocabulary table |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The OBSERVATION\_PERIODS table does not have any conventions that can be checked by the script.

**Counts of categorical variables**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Overall and by calendar year (according to the year part of *op\_start\_date*) counts will be tabulated. In both cases the results will be stratified by *op\_meaning*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “OBSERVATION\_PERIODS”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories when stratifying by meaning will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results for variables with two or more categories when stratifying by meaning and year will be displayed by line charts for each of the variables, grouped by the meaning variable and color coded by the vocabulary variable. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Visually the results for other variables when stratifying by meaning and year will be displayed by line charts for each of the variables and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*op\_meaning*: number of complete observations per category

*op\_origin*: n number of complete observations per category

* Variables with less than 2 categories:

*person\_id*: number of unique persons included in this table

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

There are no continuous variables in the OBSERVATION\_PERIODS table.

**Distribution of dates**

Overall and by calendar year counts will be tabulated. The year will be retrieved from the year part of the date variable that is being counted. In both cases the results will be stratified by *op\_meaning*. In case both count and total are both equal to zero, the following messages will be printed in top of the results table for analysis stratified by meaning, and by meaning and year, respectively:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category and year.”

The results table when stratifying by meaning will contain the name of the table “OBSERVATION\_PERIODS”, name of the variable, meaning variable, count, total and percentage. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results when stratifying by meaning will be displayed by line charts and color coded by the variable name. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

The results table when stratifying by meaning and year will contain an extra year variable compared with the previous results table. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table. If years before 1995 or future years are present in the data those will be colored red.

Visually the results when stratifying by meaning and year will be displayed by line charts for each of the variables, and color coded by the meaning variable. If you want to remove a meaning from the graph, you can do that by clicking in the meaning variable you do not need. If counts and totals smaller than 5, years before 1995 or in the future are present in the data, those will not be plotted in the graphs.

Date counts by year will be calculated as follows:

* *op\_start\_date*: number of complete observations
* *op\_end\_date*: number of complete observations

Total: Number of total observations with a recorded meaning (when stratifying by meaning) and number of total observations with a recorded meaning in a particular year(when stratifying by meaning and year).

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: OBSERVATION\_PERIODS**

* observation\_periods\_meaning\_2categories.csv
* observation\_periods\_meaning\_other.csv
* observation\_periods\_meaning\_dates.csv
* observation\_periods\_meaning\_year\_2categories.csv
* observation\_periods\_meaning\_year\_other.csv
* observation\_periods\_meaning\_year\_dates.csv
* Masked subfolder:
* observation\_periods\_meaning\_2categories\_masked.csv
* observation\_periods\_meaning\_other\_masked.csv
* observation\_periods\_meaning\_dates\_masked.csv
* observation\_periods\_meaning\_year\_2categories\_masked.csv
* observation\_periods\_meaning\_year\_other\_masked.csv
* observation\_periods\_meaning\_year\_dates\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

##### **PERSON\_RELATIONSHIPS table**

Table 15. PERSON\_RELATIONSHIPS table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PERSON\_RELATIONSHIPS** | **Curated table** | For any person, this table collects the pairing with the identifier of mother or of other relationships that may be available | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| person\_id | Yes |  | Character |  |
| related\_id | Yes |  | Character |  |
| origin\_of\_relationship | Yes | where the information about the relationship comes from | Character | see corresponding vocabulary table |
| meaning\_of\_relationship | Yes | Which type of relationship there is between the mother and the person | Character | see corresponding vocabulary table |
| method\_of\_linkage | Yes | How the linkage was performed | Character | see corresponding vocabulary table |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The PERSON\_RELATIONSHIPS table does not have any conventions that can be checked by the script.

**Counts of categorical variables.**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Counts stratified by *meaning\_of\_relationship* will be tabulated. In case both count and total are both equal to zero, the following message will be printed in top of the results table:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular meaning category.”

The results table will contain the name of the table “PERSON\_RELATIONSHIPS”, name of the variable, meaning variable, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories will be displayed by bar charts and color coded by the vocabulary variable for each of the meanings. If you want to remove a vocabulary from the graph, you can do that by clicking in the vocabulary variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the name of the variable. If you want to remove a variable from the graph, you can do that by clicking in the variable you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*origin\_of\_relationship*: number of complete observations per category

*meaning\_of\_relationship*: number of complete observations per category

*method\_of\_linkage*: number of complete observations per category

* Other variables

*person\_id*: number of unique persons included in this table

*related\_id*: number of unique records filled in

Total: Number of total observations with a recorded meaning.

**Step 5: Distribution of continuous variables and dates**

**Distribution of continuous variables**

There are no continuous variables in the PERSON\_RELATIONSHIPS table.

**Distribution of dates**

There are no date variables in PERSON\_RELATIONSHIPS table.

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: PERSON\_RELATIONSHIPS**

* person\_relationship\_meaning\_2categories.csv
* person\_relationship\_meaning\_other.csv
* Masked subfolder:
* person\_relationship\_meaning\_2categories\_masked.csv
* person\_relationship\_meaning\_other\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

**Metadata tables**

##### **PRODUCTS table**

Table 16. PRODUCTS table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PRODUCTS** | **Metadata table** | This table collects the information associated to each marketed product that may have been prescribed, dispensed or administered to a patient. It contains one row per product | | |
| **Variable** | **Mandatory** | **Description** | **Format** | **Vocabulary** |
| medicinal\_product\_id | Yes | Foreign key. The medicinal\_product\_id should be a unique identifier of a specific medicinal product. | Character |  |
| medicinal\_product\_name | Yes | Any substance or combination of substances, which may be administered for treating or preventing disease, with the view to making a medical diagnosis or to restore, correct or modify physiological functions. | Character |  |
| unit\_of\_presentation\_type | No | Qualitative term describing the discrete countable entity in which a pharmaceutical product or manufactured item is presented, in cases where strength or quantity is expressed referring to one instance of this countable entity. | Character | see corresponding vocabulary table |
| unit\_of\_presentation\_num | No | Number of unit of presentation type within a medicinal product. | Numeric |  |
| administration\_dose\_form | No | Pharmaceutical dose form for administration to the patient. | Character | see corresponding vocabulary table |
| administration\_route | No | Route of administration of the pharmaceutical product. | Character | see corresponding vocabulary table |
| medicinal\_product\_atc\_code | Yes | Unique standardized identification code from the ATC classification system from WHO associated to the medicinal product. | Character |  |
| subst1\_atc\_code | No | Unique standardized identification code from the ATC classification system from WHO associated to the active priciple. | Character |  |
| subst2\_atc\_code | No | Unique standardized identification code from the ATC classification system from WHO associated to the active priciple. | Character |  |
| subst3\_atc\_code | No | Unique standardized identification code from the ATC classification system from WHO associated to the active priciple. | Character |  |
| subst1\_amount\_per\_form | No | Quantity of the first active principle contained in the medicinal product | Numeric |  |
| subst2\_amount\_per\_form | No | Quantity of the second active principle contained in the medicinal product | Numeric |  |
| subst3\_amount\_per\_form | No | Quantity of the third active principle contained in the medicinal product | Numeric |  |
| subst1\_amount\_unit | No | Unit of measure of the quantity of the first active principle contained in the medicinal product | Character | see corresponding vocabulary table |
| subst2\_amount\_unit | No | Unit of measure of the quantity of the second active principle contained in the medicinal product | Character | see corresponding vocabulary table |
| subst3\_amount\_unit | No | Unit of measure of the quantity of the third active principle contained in the medicinal product | Character | see corresponding vocabulary table |
| subst1\_concentration | No | Strength or quantity contained into a single unit of presentation or dose form. | Numeric |  |
| subst2\_concentration | No | Strength or quantity contained into a single unit of presentation or dose form. | Numeric |  |
| subst3\_concentration | No | Strength or quantity contained into a single unit of presentation or dose form. | Numeric |  |
| subst1\_concentration\_unit | No | Unit of measure of the strength or quantity by which a particular type of unit of presentation or dose form is described. | Character | see corresponding vocabulary table |
| subst2\_concentration\_unit | No | Unit of measure of the strength or quantity by which a particular type of unit of presentation or dose form is described. | Character | see corresponding vocabulary table |
| subst3\_concentration\_unit | No | Unit of measure of the strength or quantity by which a particular type of unit of presentation or dose form is described. | Character | see corresponding vocabulary table |
| concentration\_total\_content | No | Total content of a single unit such as particular type of pharmaceutical unit of presentation or dose form. | Numeric |  |
| concentration\_total\_content\_unit | No | Unit of measure of the concentration total content. | Character | see corresponding vocabulary table |
| medicinal\_product\_manufacturer | No | Name of the manufactured of the pharmaceutical product. | Character |  |

**Step 4: Convention and counts of categorical variables**

**Duplicated rows**

Firstly, the data will be checked for duplicated information. If there are duplicated rows present, then the following message will be displayed:

"There are *no.* duplicated rows in the data. Take caution when interpreting counts."

If no duplicated rows are identified the report will show:

"There are no duplicated rows in the data."

**Convention check**

The PRODUCTS table does not have any conventions that can be checked by the script.

**Counts of categorical variables.**

Results will be reported separately for variables with two or more categories due to the need to stratify by the values of the variable.

Counts stratified by *medicinal\_product\_atc\_code* truncated to the third level will be tabulated. In case both count and total are both equal to zero, the following message will be printed in top of the results table:

“Attention! There is(are) *no.* row(s) with a zero value for both count and total. Those will not be displayed in the tables or graphs. This happens when a variable is completely missing for a particular ATC category.”

The results table will contain the name of the table “PRODUCTS”, name of the variable, atc\_code, count, total and percentage. In case of results for variables with two or more categories an extra vocabulary variable will be present in the results table, which reflects the different categories of the variable. If counts and totals smaller than 5 are present in the data, those will be replaced by “<5” in the results table.

Visually the results for variables with two or more categories will be displayed by line charts and color coded by the atc\_code variable for each of the variables. If you want to remove an ATC code from the graph, you can do that by clicking in the ATC code you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

In case of other variables, the results will be displayed by line charts and color coded by the atc\_code variable. If you want to remove an ATC code from the graph, you can do that by clicking in the ATC code you do not need. If counts and totals smaller than 5 are present in the data, those will not be plotted in the graphs.

* Variables with 2 or more categories

*unit\_of\_presentation\_type*: number of complete observations per category

*administration\_dose\_form*: number of complete observations per category

*administration\_route*: number of complete observations per category

*subst1\_amount\_unit*: number of complete observations per category

*subst2\_amount\_unit*: number of complete observations per category

*subst3\_amount\_unit*: number of complete observations per category

*subst1\_concentration\_unit*: number of complete observations per category

*subst2\_concentration\_unit*: number of complete observations per category

*subst3\_concentration\_unit*: number of complete observations per category

*concentration\_total\_content\_unit*: number of complete observations per category

* Other variables:

*medicinal\_product\_id*: number of complete observations

*medicinal\_product\_name*: number of complete observations

*unit\_of\_presentation\_num*: number of complete observations

*medicinal\_product\_atc\_code*: number of complete observations

*subst1\_atc\_code*: number of complete observations

*subst2\_atc\_code*: number of complete observations

*subst3\_atc\_code*: number of complete observations

*subst1\_amount\_per\_form*: number of complete observations

*subst2\_amount\_per\_form*: number of complete observations

*subst3\_amount\_per\_form*: number of complete observations

*subst1\_concentration*: number of complete observations

*subst2\_concentration*: number of complete observations

*subst3\_concentration*: number of complete observations

*concentration\_total\_content*: number of complete observations

*medicinal\_product\_manufacturer*: number of complete observations

**Step 5. Distribution of continuous variables and dates**

**Distribution of continuous variables**

All distribution will be stratified by the ATC code truncated to the third level.

*unit\_of\_presentation\_num* stratified by *unit\_of\_presentation\_type* and *medicinal\_product\_atc\_code* (truncated to third level): mean, median, interquartile range, skewness, kurtosis

*subst1\_amount\_per\_form* stratified by *subst1\_atc\_code*(truncated to third level): and *subst1\_amount\_unit:*

mean, median, interquartile range, skewness, kurtosis

*subst2\_amount\_per\_form* stratified by *subst2\_atc\_code*(truncated to third level): and *subst2\_amount\_unit:*

mean, median, interquartile range, skewness, kurtosis

*subst3\_amount\_per\_form* stratified by *subst3\_atc\_code*(truncated to third level): and *subst3\_amount\_unit:*

mean, median, interquartile range, skewness, kurtosis

*concentration\_total\_content* stratified by *medicinal\_product\_atc\_code*(truncated to third level): and *concentration\_total\_content\_unit*

**Distribution of dates**

There are no date variables in the PRODUCTS table.

**Calculation**

An overview on how counts on step 4 and step 5 are calculated.

**Output folder: PRODUCTS**

* products\_atc\_2categories.csv
* products\_atc\_other.csv
* products\_unit\_of\_presentation\_num\_dist.csv (if available)
* products\_subst1\_amount\_per\_form\_dist.csv (if available)
* products\_subst2\_amount\_per\_form\_dist.csv (if available)
* products\_subst3\_amount\_per\_form\_dist.csv (if available)
* products\_concentration\_total\_content\_dist.csv (if available)
* Masked subfolder:
* products\_atc\_2categories\_masked.csv
* products\_atc\_other\_masked.csv
* .html markdown report file (found one level up of this folder, in the output directory)

### Data analysis steps per objective: Level 2

#### Event dates before date of birth

The result of this check is the proportion of non-missing values of a given date variable which takes place before the person’s reported date of birth, overall and stratified by the meaning variable of the table of interest.

Counts and proportions of records for which the indicator variable indicating that the observed date comes before the date of birth, overall and stratified by the meaning variable, are calculated for each date variable in the CDM instance (with exception to the OBSERVATION\_PERIODS, CDM\_SOURCE and INSTANCE tables). Tables that undergo this check are: VISIT\_OCCURRENCE, EVENTS, MEDICINES, PROCEDURES, VACCINES, MEDICAL\_OBSERVATIONS, SURVEY\_ID, and SURVEY\_OBSERVATIONS.

The numerator of this check is the number of event dates in the table of interest occurring before the date of birth recorded for the *person\_id* in the PERSONS table, overall and stratified by the meaning variable.

The denominator of this check is the number of rows of the table created after merging the PERSONS table with the table of interest by *person\_id* (keeping only *person\_id* present in the PERSONS table) and after removing persons with missing or wrong format of *date\_of\_birth* and date variable from the table of interest.

The input for this check includes the D2 *PERSONS* table (for the date of birth) and the D2 table containing the date of interest (e.g., *MEDICINES*), and the name of the column of interest (e.g., *date\_dispensing*). Steps for this check are described below:

* In PERSONS, *date\_of\_birth* is created as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **day\_of\_birth** | **month\_of\_birth** | **year\_of\_birth** | **date\_of\_birth** |
| NA | mm | yyyy | 01-mm-yyyy |
| dd | NA | yyyy | dd-01-yyyy |
| NA | NA | yyyy | 01-01-yyyy |
| dd | mm | NA | NA |
| dd | NA | NA | NA |
| NA | mm | NA | NA |
| NA | NA | NA | NA |

* The table containing the date of interest is merged to the PERSONS table by *person\_id*, keeping only *person\_id* values present in the PERSONS table.
* Rows for which the values of the date variable (from the table of interest) and/or *date\_of\_birth* are either missing or in an incorrect format are removed from the table.
  + The number of observations excluded are recorded for subsequent reporting.
* For observations, where the meaning variable is missing is transformed to ‘Not entered’.
* The number of rows of the resulting table will be used as the denominator for the proportion.
* An indicator variable is added to the resulting merged table. This variable equals 1 when the observed date comes before the date of birth, 0 otherwise.
* The number of rows where the indicator variable is equal to 1 will be used as the numerator.

The results table when stratifying by the meaning variable in the table of interest, contains the CDM table name, the meaning variable, the number of issues, total and percentage.

The overall results table contains the CDM table name, number of issues, total and percentage.

The warnings table contains the CDM table name, the name of the file in the directory (d), and warning showing the number of rows for which the analysis could not be performed.

Visually the analysis, stratified by the meaning variable is represented by a reversed line graph for each of the combinations name of CDM table: meaning variable.

#### Event dates after date of death

The result of this check is the proportion of non-missing values of a given date variable which takes place after the person’s reported date of death, overall and stratified by the meaning variable of the table of interest.

Counts and proportions of records for which the indicator variable indicating that the observed date comes after the date of death, overall and stratified by the meaning variable, are calculated for each date variable in the CDM instance (with exception to the OBSERVATION\_PERIODS, CDM\_SOURCE and INSTANCE tables). Tables that undergo this check are: VISIT\_OCCURRENCE, EVENTS, MEDICINES, PROCEDURES, VACCINES, MEDICAL\_OBSERVATIONS, SURVEY\_ID, and SURVEY\_OBSERVATIONS.

The numerator of this check is the number of event dates in the table of interest occurring after the date of death recorded for the *person\_id* in the PERSONS table, overall and stratified by the meaning variable.

The denominator of this check is the number of rows of the table created after merging the PERSONS table with the table of interest by *person\_id* (keeping only *person\_id* present in the PERSONS table) and after removing persons with missing or wrong format of *date\_of\_death* and date variable from the table of interest.

The input for this check includes the D2 *PERSONS* table (for the date of birth) and the D2 table containing the date of interest (e.g., *MEDICINES*), and the name of the column of interest (e.g., *date\_dispensing*). Steps for this check are described below:

* In PERSONS, *date\_of\_death* is created as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **day\_of\_death** | **month\_of\_death** | **year\_of\_death** | **date\_of\_death** |
| NA | mm | yyyy | 31-mm-yyyy |
| dd | NA | yyyy | dd-12-yyyy |
| NA | NA | yyyy | 31-12-yyyy |
| dd | mm | NA | NA |
| dd | NA | NA | NA |
| NA | mm | NA | NA |
| NA | NA | NA | NA |

* The table containing the date of interest is merged to the PERSONS table by *person\_id*, keeping only *person\_id* values present in the PERSONS table.
* Rows for which the values of the date variable (from the table of interest) and/or *date\_of\_death* are either missing or in an incorrect format are removed from the table.
  + The number of observations excluded are recorded for subsequent reporting.
* For observations, where the meaning variable is missing is transformed to ‘Not entered’.
* The number of rows of the resulting table will be used as the denominator for the proportion.
* An indicator variable is added to the resulting merged table. This variable equals 1 when the observed date comes after the date of death, 0 otherwise.
* The number of rows where the indicator variable is equal to 1 will be used as the numerator.

The results table when stratifying by the meaning variable in the table of interest, contains the CDM table name, the meaning variable, the number of issues, total and percentage.

The overall results table contains the CDM table name, number of issues, total and percentage.

The warnings table contains the CDM table name, the name of the file in the directory (d), and warning showing the number of rows for which the analysis could not be performed.

Visually the analysis, stratified by the meaning variable is represented by a reversed line graph for each of the combinations name of CDM table: meaning variable.

#### Event dates outside observation periods

The result of this check is the proportion of non-missing values of a given date variable which takes place outside of any of the reported observation periods for a given person, overall and stratified by the meaning variable of the table of interest.

Counts and proportions of records for which the indicator variable indicating that the observed date falls before *op\_start\_date* or after *op\_end\_date*, overall and stratified by the meaning variable, are calculated for each date variable in the CDM instance (with exception to the CDM\_SOURCE and INSTANCE tables). Tables that undergo this check are: VISIT\_OCCURRENCE, EVENTS, MEDICINES, PROCEDURES, VACCINES, MEDICAL\_OBSERVATIONS, SURVEY\_ID, and SURVEY\_OBSERVATIONS.

The numerator of this check is the number of event dates in the table of interest occurring outside of observation period dates recorded in the OBSERVATION\_PERIODS table, overall and stratified by the meaning variable.

The denominator of this check is the number of rows created after merging the OBSERVATION\_PERIODS table with the table of interest by *person\_id* (keeping only *person\_id* present in the OBSERVATION\_PERIODS table) and after removing persons with wrong format of both *op\_start\_date* and *op\_end\_date,* and of the date variable from the table of interest and missing values of *op\_start\_date* and date of interest.

The input for this check includes the D2 *OBSERVATION\_PERIODS* table and the D2 table containing the date of interest (e.g., *MEDICINES*), and the name of the column of interest (e.g., *date\_dispensing*). Steps for this check are described below:

* In OBSERVATION\_PERIODS, replace missing *op\_end\_date* with present date (if *op\_end\_date* is missing, the persons is still being followed).
* The table containing the date of interest is merged to the OBSERVATION\_PERIODS table by *person\_id*, keeping only *person\_id* values present in the OBSERVATION\_PERIODS table.
* Rows for which the values of the date variable (from the table of interest) and/or *op\_start\_date and op\_end\_date* in an incorrect format are removed from the table as well as observations with missing values of date variable (from table of interest) and/or *op\_start\_date*.
  + The number of observations excluded are recorded for subsequent reporting.
* For observations, where the meaning variable is missing is transformed to ‘Not entered’.
* The number of rows of the resulting table will be used as the denominator for the proportion.
* An indicator variable is added to the resulting merged table. This variable equals 1 when the observed date comes before *op\_start\_date* or after *op\_end\_date*, 0 otherwise.
* The number of rows where the indicator variable is equal to 1 will be used as the numerator.

The results table when stratifying by the meaning variable in the table of interest, contains the CDM table name, the meaning variable, the number of issues, total and percentage.

The overall results table contains the CDM table name, number of issues, total and percentage.

The warnings table contains the CDM table name, the name of the file in the directory (d), and warning showing the number of rows for which the analysis could not be performed.

Visually the analysis, stratified by the meaning variable is represented by a reversed line graph for each of the combinations name of CDM table: meaning variable.

#### Observations in a table of interest without a corresponding record in the PERSONS table

The result of this check is the proportion of rows with a *person\_id* in a given table but without any corresponding record associated with this *person\_id* in the PERSONS table, overall and stratified by the meaning variable of the table of interest.

Counts and proportions of records for which the indicator variable indicating that the *person\_id* observed in the table of interest is not observed in the PERSONS table, overall and stratified by the meaning variable, are calculated for each table in the CDM instance.

The numerator of this check is the number of rows with a *person\_id* in the table of interest without a record corresponding to the *person\_id* in the PERSONS table, overall and stratified by the meaning variable.

The denominator is the number of rows in the table of interest.

The input for this check includes the D2 *PERSONS* table and the D2 table of interest (e.g., EVENTS). Steps for this check are described below:

* The table of interest is merged to the PERSONS table by *person\_id*, keeping all *person\_id* values present in the table of interest. Both *person\_id* columns from the PERSONS table and the table of interest are present in the merged table.
* For observations, where the meaning variable is missing is transformed to ‘Not entered’.
* An indicator variable is added to the resulting merged table. This variable equals 1 when the *person\_id* is available from the table of interest but missing in the PERSONS table, 0 if present in both.
* The number of rows where the indicator variable is equal to 1 will be used as the numerator.

The results table when stratifying by the meaning variable in the table of interest, contains the CDM table name, the meaning variable, the number of issues, total and percentage.

The overall results table contains the CDM table name, number of issues, total and percentage.

The warnings table contains the CDM table name, the name of the file in the directory (d), and warning showing the number of rows for which the analysis could not be performed.

Visually the analysis, stratified by the meaning variable is represented by a reversed line graph for each of the combinations name of CDM table: meaning variable.

#### Observations associated with a visit\_occurrence\_id which occur before the visit\_start\_date

The result of this check is the proportion of observations with an associated visit\_occurrence\_id which occur before the visit\_start\_date, overall and stratified by the meaning variable of the table of interest.

Counts and proportions of records for which the indicator variable indicating that the observation date observed in the table of interest occurs before the *visit\_start\_date*, overall and stratified by the meaning variable, are calculated for each date variable in the CDM instance (with exception to the OBSERVATION\_PERIODS, SURVEY\_ID, SURVEY\_OBSERVATIONS, CDM\_SOURCE and INSTANCE tables). Tables that undergo this check are: EVENTS, MEDICINES, PROCEDURES, VACCINES, and MEDICAL\_OBSERVATIONS.

The numerator of this check is the number of event dates in the table of interest associated with a *visit\_occurrence\_id* occurring before the *visit\_start\_date* associated with the same *visit\_occurrence\_id* in the VISIT\_OCCURRENCE table, overall and stratified by the meaning variable of the table of interest.

The denominator of this check is the number of rows associated with a *visit\_occurrence\_id* in the table of interest.

The input for this check includes the D2 *VISIT\_OCCURRENCE* (person\_id, *visit\_occurrence\_id*, *visit\_start\_date*) table and the D2 table containing the date of interest (e.g., *MEDICINES*), and the name of the columns of interest (e.g., *date\_dispensing, person\_id, visit\_occurrence\_id*). Steps for this check are described below:

* The table containing the date of interest is merged to the VISIT\_OCCURRENCE table by *person\_id* and *visit\_occurrence\_id,* keeping only *visit\_occurrence\_id* in the table of interest.
* Observations with either missing or incorrect format of the date variable (table of interest) and *visit\_start\_date* are removed.
  + The number of observations excluded are recorded for subsequent reporting.
* For observations, where the meaning variable is missing is transformed to ‘Not entered’.
* The number of rows of the resulting table will be used as the denominator for the proportion.
* An indicator variable is added to the resulting merged table. This variable equals 1 when the observation date observed in the table of interest occurs before the *visit\_start\_date*, 0 otherwise.
* The number of rows where the indicator variable is equal to 1 will be used as the numerator.

The results table when stratifying by the meaning variable in the table of interest, contains the CDM table name, the meaning variable, the number of issues, total and percentage.

The overall results table contains the CDM table name, number of issues, total and percentage.

The warnings table contains the CDM table name, the name of the file in the directory (d), and warning showing the number of rows for which the analysis could not be performed.

Visually the analysis, stratified by the meaning variable is represented by a reversed line graph for each of the combinations name of CDM table: meaning variable.

#### Observations associated with a visit\_occurrence\_id which occur after the visit\_end\_date

The result of this check is the number of observations with an associated visit\_occurrence\_id which occur after the visit\_end\_date, overall and stratified by the meaning variable of the table of interest.

Counts and proportions of records for which the indicator variable indicating that the observation date observed in the table of interest occurs after the visit\_end\_date, overall and stratified by the meaning variable, are calculated for each date variable in the CDM instance (with exception to the OBSERVATION\_PERIODS, SURVEY\_ID, SURVEY\_OBSERVATIONS, CDM\_SOURCE and INSTANCE tables). Tables that undergo this check are: EVENTS, MEDICINES, PROCEDURES, VACCINES, and MEDICAL\_OBSERVATIONS.

The numerator of this check is the number of event dates in the table of interest associated with a *visit\_occurrence\_id* occurring after the *visit\_end\_date* associated with the same *visit\_occurrence\_id* in the VISIT\_OCCURRENCE table, overall and stratified by the meaning variable of the table of interest.

The denominator of this check is the number of observations associated with a *visit\_occurrence\_id* in the table of interest.

The input for this check includes the D2 *VISIT\_OCCURRENCE* (person\_id, *visit\_occurrence\_id*, *visit\_end\_date*) table and the D2 table containing the date of interest (e.g., *MEDICINES*), and the name of the columns of interest (e.g., *date\_dispensing, person\_id, visit\_occurrence\_id*). Steps for this check are described below:

* The table containing the date of interest is merged to the VISIT\_OCCURRENCE table by *person\_id* and *visit\_occurrence\_id,* keeping only *visit\_occurrence\_id* in the table of interest.
* Observations with either missing or incorrect format of the date variable (table of interest) and *visit\_end\_date* are removed.
  + The number of observations excluded are recorded for subsequent reporting.
* For observations, where the meaning variable is missing is transformed to ‘Not entered’.
* The number of rows of the resulting table will be used as the denominator for the proportion.
* An indicator variable is added to the resulting merged table. This variable equals 1 when the observation date observed in the table of interest occurs before the *visit\_end\_date*, 0 otherwise.
* The number of rows where the indicator variable is equal to 1 will be used as the numerator.

The results table when stratifying by the meaning variable in the table of interest, contains the CDM table name, the meaning variable, the number of issues, total and percentage.

The overall results table contains the CDM table name, number of issues, total and percentage.

The warnings table contains the CDM table name, the name of the file in the directory (d), and warning showing the number of rows for which the analysis could not be performed.

Visually the analysis, stratified by the meaning variable is represented by a reversed line graph for each of the combinations name of CDM table: meaning variable.

#### Observations associated with a visit\_occurrence\_id for which the associated person\_id differs from that in the VISIT\_OCCURRENCE table

The result of this check is the number of observations with an associated *visit\_occurrence\_id* which occur in association with a *person\_id* other than that associated with the *visit\_occurrence\_id* in the VISIT\_OCCURRENCE table, overall and stratified by the meaning variable of the table of interest.

Counts and proportions of records for which the indicator variable indicating that the *person\_id* associated with a *visit\_occurrence\_id* in the VISIT\_OCCURRENCE table is not equal to the *person\_id* associated with the *visit\_occurrence\_id* in the table of interest, overall and stratified by the meaning variable, are calculated for each table in the CDM instance. Tables that undergo this check are: EVENTS, MEDICINES, PROCEDURES, VACCINES, and MEDICAL\_OBSERVATIONS.

The numerator of this check is the number of observations in the table of interest with an associated *visit\_occurrence\_id* which occur in association with a *person\_id* other than that associated with the *visit\_occurrence\_id* in the VISIT\_OCCURRENCE table, overall and stratified by the meaning variable of the table of interest.

The denominator of this check is the number of rows in the table of interest.

The input for this check includes the D2 *VISIT\_OCCURRENCE* (person\_id, *visit\_occurrence\_id*) table and the D2 table of interest (e.g., *MEDICINES*), and the name of the columns of interest (e.g., *person\_id, visit\_occurrence\_id*). Steps for this check are described below:

* In the table of interest, the observations with missing *visit\_occurrence\_id* will be removed.
  + The number of observations excluded are recorded for subsequent reporting.
* In the table of interest and VISIT\_OCCURRENCE, a combination variable of *person\_id* and *visit\_occurrence\_id* is created. In VISIT\_OCCURRENCE table, if *person\_id* or *visit\_occurrence\_id* is missing the combination variable is set to missing.
* The table of interest is merged to the VISIT\_OCCURRENCE table by the combination variable*,* keeping only the combinations present in the table of interest. Both combination variables from the table of interest and the VISIT\_OCCURRENCE are present in the merged table.
* For observations, where the meaning variable is missing is transformed to ‘Not entered’.
* The number of rows of the resulting table will be used as the denominator for the proportion.
* An indicator variable is added to the resulting merged table. This variable equals 1 whenever the combination variable from the table of interest is present but missing for the VISIT\_OCCURRENCE table, 0 if it is present for both.
* The number of rows where the indicator variable is equal to 1 will be used as the numerator.

The results table when stratifying by the meaning variable in the table of interest, contains the CDM table name, the meaning variable, the number of issues, total and percentage.

The overall results table contains the CDM table name, number of issues, total and percentage.

The warnings table contains the CDM table name, the name of the file in the directory (d), and warning showing the number of rows for which the analysis could not be performed.

Visually the analysis, stratified by the meaning variable is represented by a reversed line graph for each of the combinations name of CDM table: meaning variable.

#### 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

The result of this check is the number and proportion of records indicated in the *PERSON\_RELATIONSHIPS* table as having the relationship of parent (meaning\_of\_relationship = birth\_mother, adoptive\_mother, genetic\_mother, undefined\_mother, adoptive\_father, genetic\_father, undefined\_father) to a child, with a birth date less than 12 years prior to the recorded birth date of the associated child, overall and stratified by the *meaning\_of\_relationship*.

Counts and proportions of records for which the indicator variable indicating that the subject present in the *PERSON\_RELATIONSHIPS* table as having the relationship of parent to a child has a birth date less than 12 years prior to the recorded birth date of the associated child will be calculated for the *PERSON\_RELATIONSHIPS* table, overall and stratified by the *meaning\_of\_relationship*.

The numerator of this check is the number of observations in the PERSON\_RELATIONSHIPS table indicating a parent-child relationship for which the person indicated as the parent has a date of birth in the PERSONS table less than 12 years prior to the date of birth in the PERSONS table associated with the person identified as his or her child, overall and stratified by the *meaning\_of\_relationship*.

The denominator of this check is the number of parent-child relationships (number of rows) in the PERSON\_RELATIONSHIPS table.

The input for this check includes the D2 *PERSONS* (*person\_id*, *day\_of\_birth, month\_of\_birth, year\_of\_birth)* table and the D2 PERSON\_RELATIONSHIPS and the name of the columns of interest (*person\_id, related\_id, meaning\_of\_relationship*). Steps for this check are described below:

* In PERSONS, *date\_of\_birth* is created as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **day\_of\_birth** | **month\_of\_birth** | **year\_of\_birth** | **date\_of\_birth** |
| NA | mm | yyyy | 01-mm-yyyy |
| dd | NA | yyyy | dd-01-yyyy |
| NA | NA | yyyy | 01-01-yyyy |
| dd | mm | NA | NA |
| dd | NA | NA | NA |
| NA | mm | NA | NA |
| NA | NA | NA | NA |

* In the *PERSON\_RELATIONSHIPS* table, observations not indicating a parental relationship are removed.
* The *PERSONS* table is merged to the *PERSON\_RELATIONSHIPS* table, by
* Rename the birthdate associated with the person\_id to date\_of\_birth\_person. Rename the birthdate associated with the related\_id to date\_of\_birth\_related.
* For observations, where the meaning variable is missing is transformed to ‘Not entered’.
* The number of rows of the resulting table is used as the denominator for the proportion.
* An indicator variable is added to the resulting merged table. This variable equals 1 whenever the date\_of\_birth\_related is less than twelve years prior to the date\_of\_birth\_person, 0 otherwise.
* The number of rows where the indicator variable is equal to 1 will be used as the numerator.

The results table when stratifying by the meaning variable in the table of interest, contains the CDM table name, the meaning variable, the number of issues, total and percentage.

The overall results table contains the CDM table name, number of issues, total and percentage.

The warnings table contains the CDM table name, the name of the file in the directory (d), and warning showing the number of rows for which the analysis could not be performed.

Visually the analysis, stratified by the meaning variable is represented by a reversed line graph for each of the combinations name of CDM table: meaning variable.

**Output folder: Level 2**

# References

1. Schneeweiss, Sebastian, et al. "Choosing among common data models for real‐world data analyses fit for making decisions about the effectiveness of medical products." Clinical Pharmacology & Therapeutics 107.4 (2020): 827-833.

2. Trifirò, Gianluca, et al. "Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how?." Journal of internal medicine 275.6 (2014): 551-561.

3. Gini, Rosa, 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 4.1 (2016).

4. Brown, Jeffrey, Michael Kahn, and Sengwee Toh. "Data quality assessment for comparative effectiveness research in distributed data networks." Medical care 51.8 0 3 (2013): S22.

5. Kahn, Michael G., et al. "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data." Egems 4.1 (2016).

6. Kahn, Michael G., et al. "A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research." Medical care 50 (2012).

7. Kahn, Michael G., et al. "Transparent reporting of data quality in distributed data networks." Egems 3.1 (2015).

8. Wang, Richard Y., and Diane M. Strong. "Beyond accuracy: What data quality means to data consumers." Journal of management information systems 12.4 (1996): 5-33.

9. Friedman, Charles P., Adam K. Wong, and David Blumenthal. "Achieving a nationwide learning health system." Science translational medicine 2.57 (2010): 57cm29-57cm29.

10. Friedman, Charles, et al. "Toward a science of learning systems: a research agenda for the high-functioning Learning Health System." Journal of the American Medical Informatics Association 22.1 (2015): 43-50.

11. Grossman, Claudia, Brian Powers, and J. Michael McGINNIS. Digital infrastructure for the learning health system: the foundation for continuous improvement in health and health care: workshop series summary. Washington, DC: National Academies Press, 2011.

12. Callahan, Tiffany J., et al. "A comparison of data quality assessment checks in six data sharing networks." eGEMs 5.1 (2017).

13. Qualls, Laura Goettinger, et al. "Evaluating foundational data quality in the national patient-centered clinical research network (PCORnet®)." eGEMs 6.1 (2018).

14. Huser, Vojtech, et al. "Multisite evaluation of a data quality tool for patient-level clinical data sets." eGEMs 4.1 (2016).

15. Loane, Maria, et al. "Paper 3: EUROCAT data quality indicators for population‐based registries of congenital anomalies." Birth Defects Research Part A: Clinical and Molecular Teratology 91.S1 (2011): S23-S30.

# Annex 1: Characteristics of the databases

**Table A1.1: Characteristics of the Databases**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

# Annex 2: Event and Procedure codes

*To be added*

# Annex 3: Drug and Vaccine codes

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

# Annex 4: ConcePTION anDREa SOP

*To be added*

# Annex 5: Mock tables

## Level 1

***Results Data Sets***

**RESULTS dummy tables**

In this section is displayed how the results tables from step 2, step 4 and, step 5 will look.

**STEP 2**

**Overall missing data analysis**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| table\_name | variable\_name | count | total | percentage |
| name of CDM table | name of variable |  |  |  |

**Overall missing data analysis stratified by meaning**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| table\_name | variable\_name | meaning | count | total | percentage |
| name of CDM table | name of variable | category of meaning variable |  |  |  |

**Missing data analysis stratified by year**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| table\_name | variable\_name | year | count | total | percentage |
| name of CDM table | name of variable | year part of date variable |  |  |  |

**Missing data analysis stratified by year and meaning**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| table\_name | variable\_name | meaning | year | count | total | percentage |
| name of CDM table | name of variable | category of meaning variable | year part of date variable |  |  |  |

**STEP 4**

**Counts stratified by meaning (2 or more categories)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| table\_name | variable\_name | meaning | vocabulary | count | total | percentage |
| name of CDM table | name of variable | category of meaning variable | category of variable |  |  |  |

**Counts stratified by meaning (other variables)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| table\_name | variable\_name | meaning | count | total | percentage |
| name of CDM table | name of variable | category of meaning variable |  |  |  |

**Counts stratified by meaning and year (2 or more categories)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| table\_name | variable\_name | meaning | vocabulary | year | count | total | percentage |
| name of CDM table | name of variable | category of meaning variable | category of variable | year part of date variable |  |  |  |

**Counts stratified by meaning and year (other variables)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| table\_name | variable\_name | meaning | year | count | total | percentage |
| name of CDM table | name of variable | category of meaning variable | year part of date variable |  |  |  |

**STEP 5**

**Date counts stratified by meaning**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| table\_name | variable\_name | meaning | count | total | percentage |
| name of CDM table | name of date variable | category of meaning variable |  |  |  |

**Date counts stratified by meaning and year**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| table\_name | variable\_name | meaning | year | count | total | percentage |
| name of CDM table | name of date variable | category of meaning variable |  |  |  |  |

## Level 2

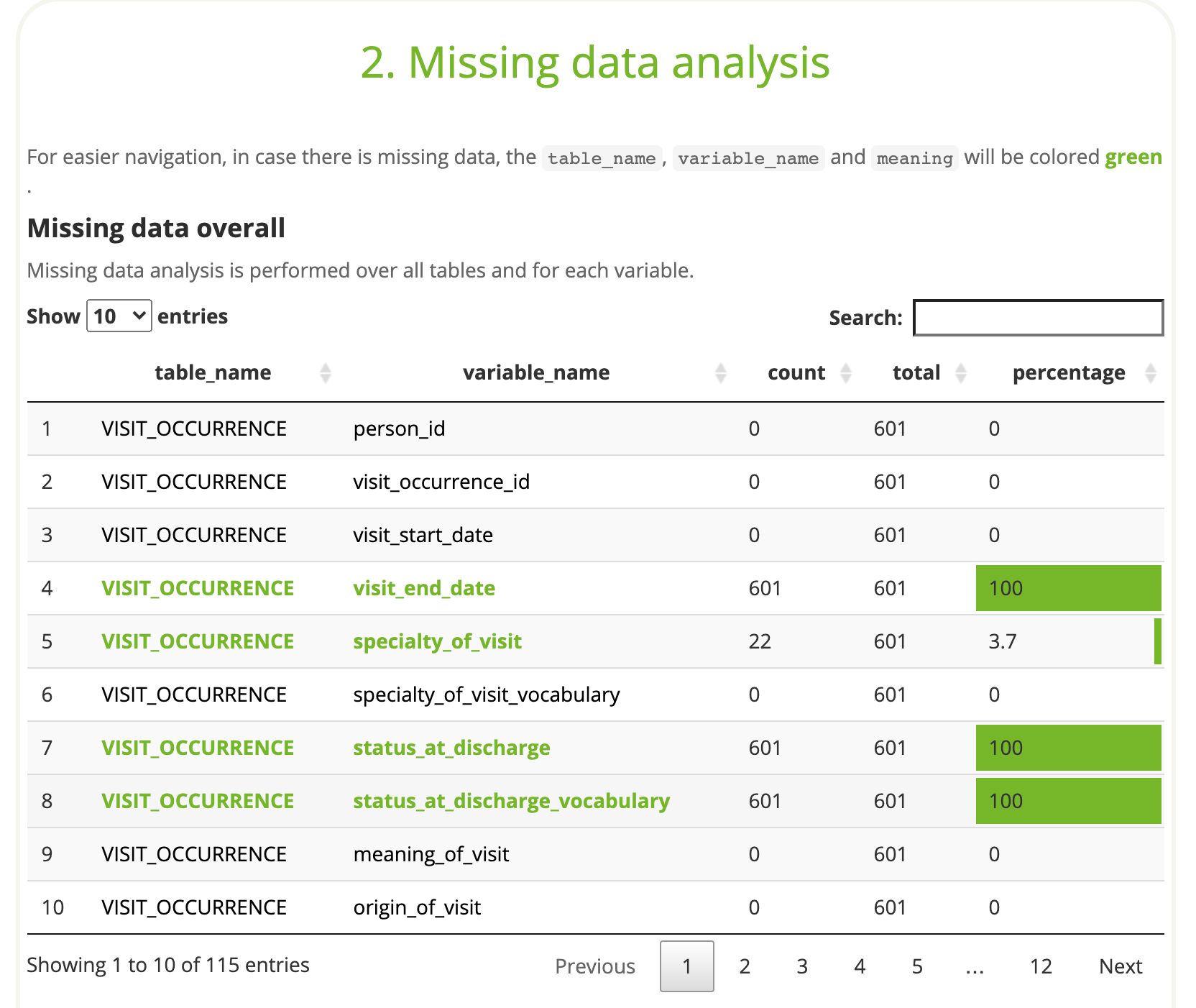
***Results Data Sets***

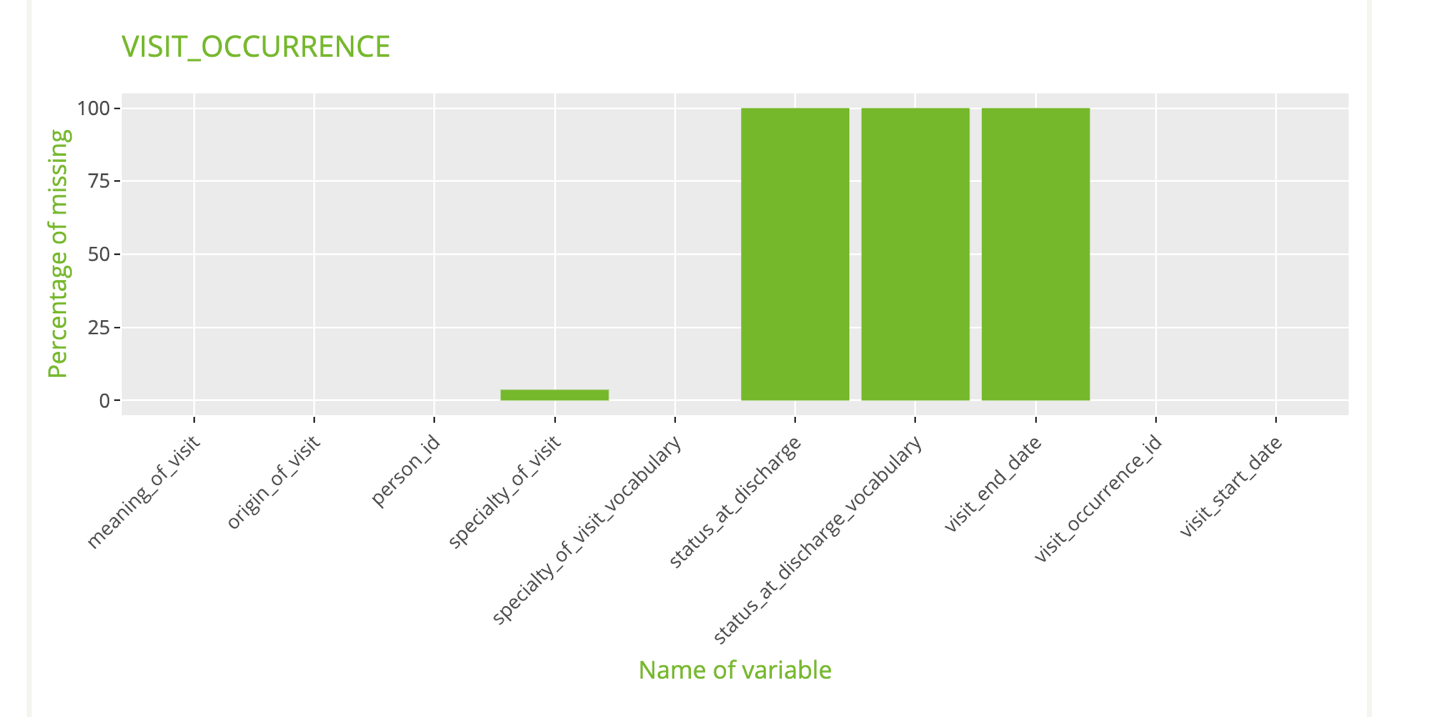
*To be added*

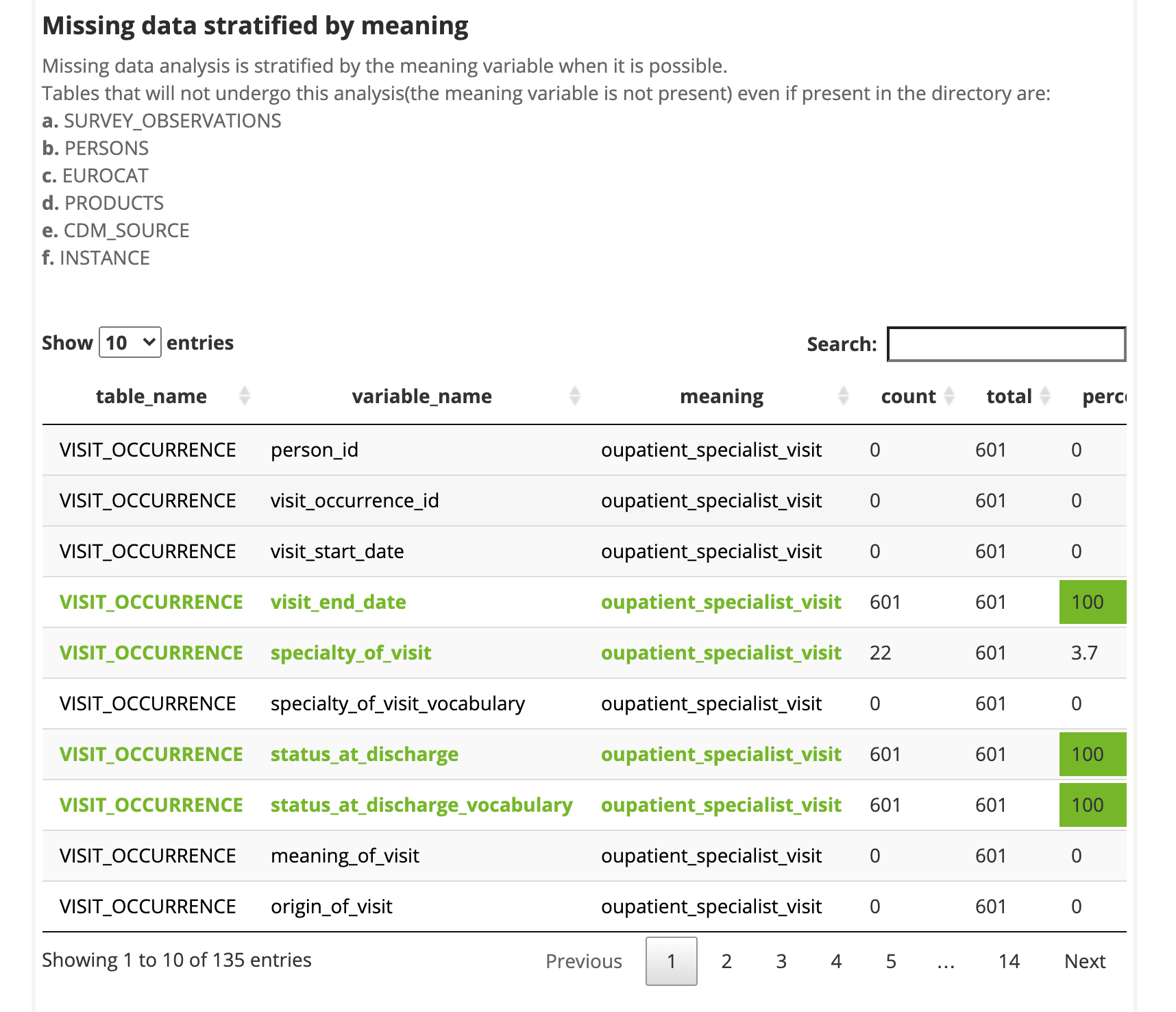
# Annex 6: Mock report tables and graphics (html markdown report)

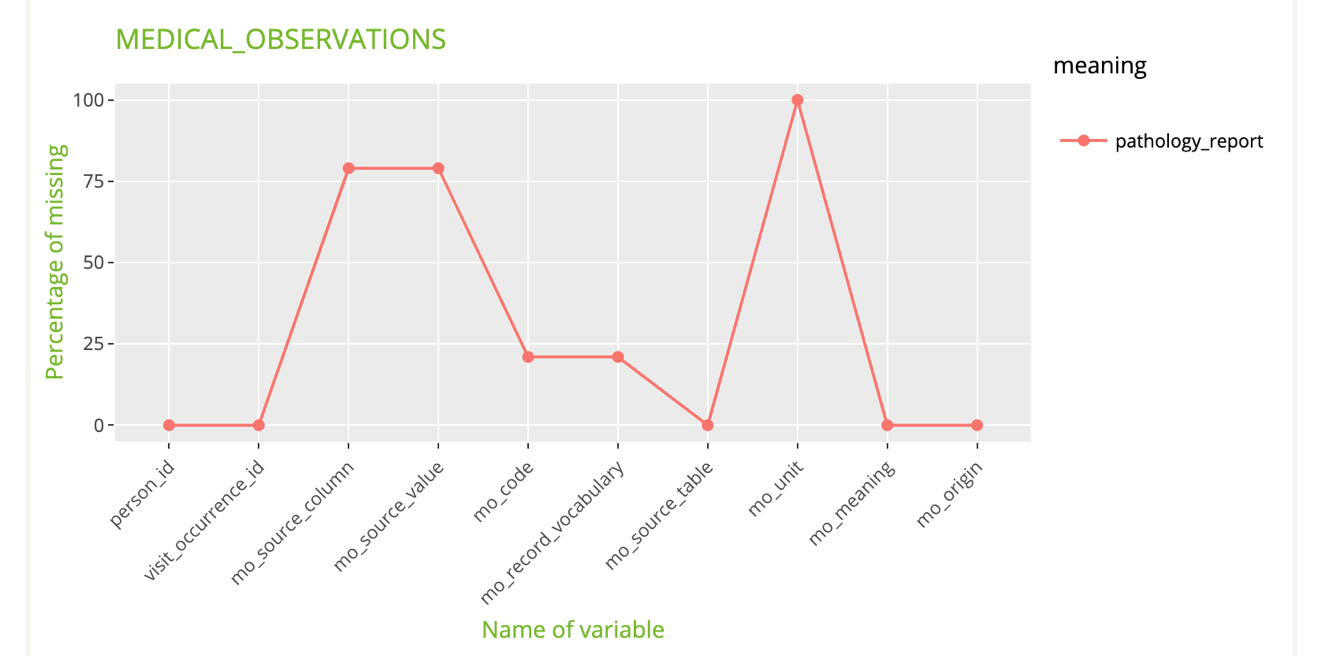
## Level 1

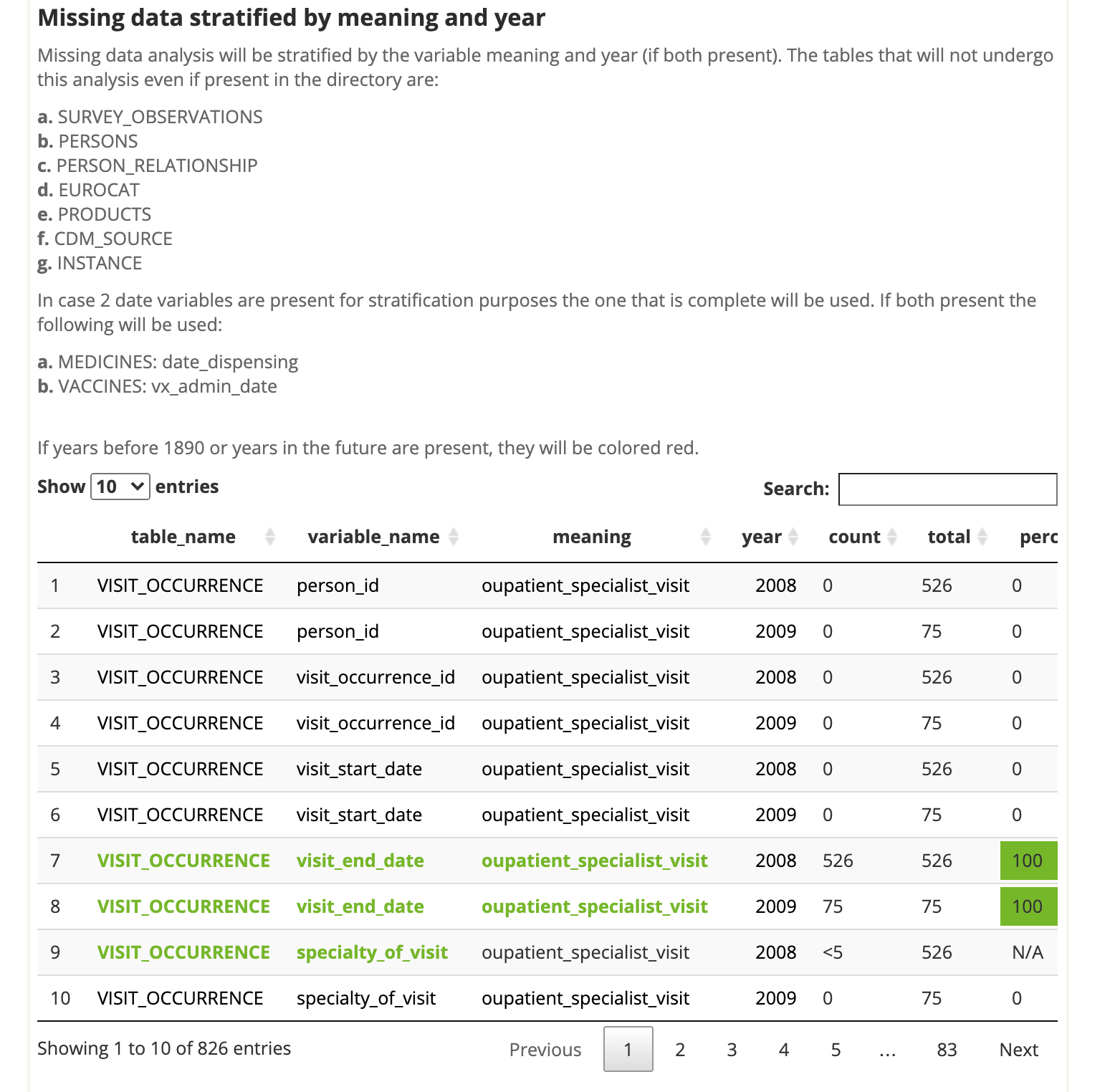
**Step 1 to 3**

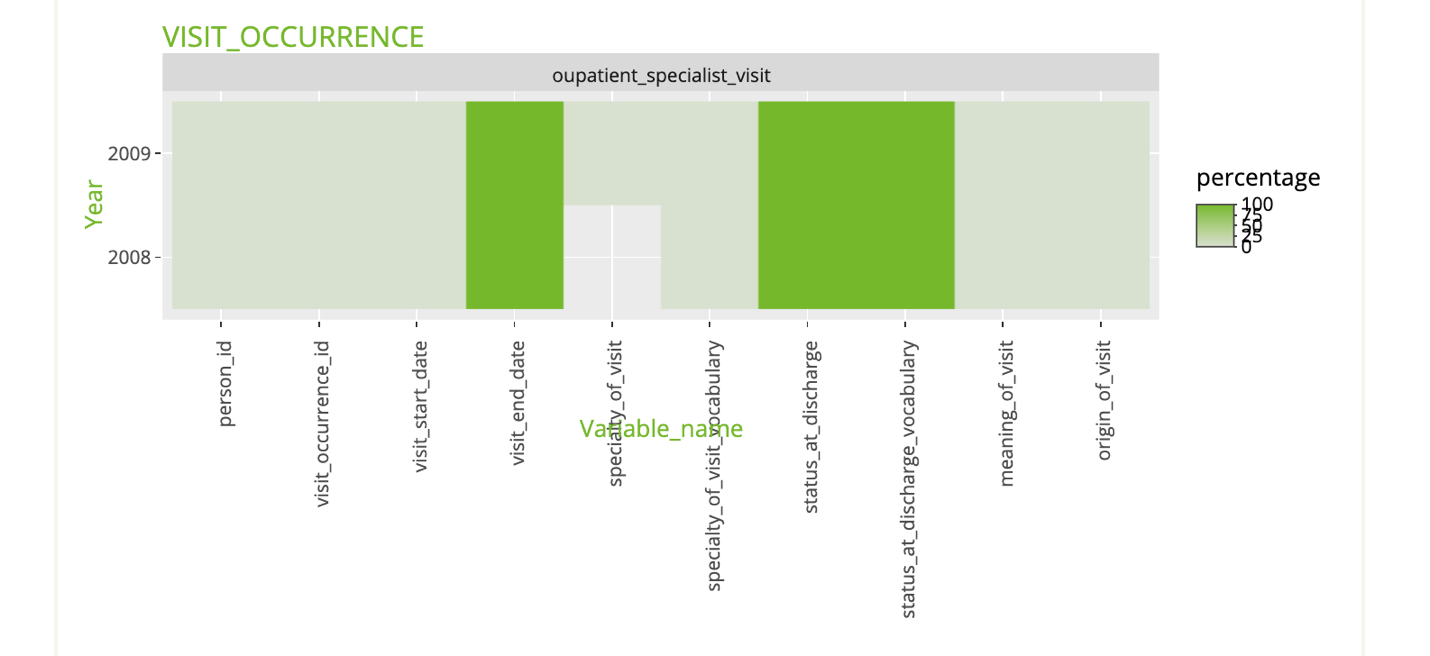
******

******

******

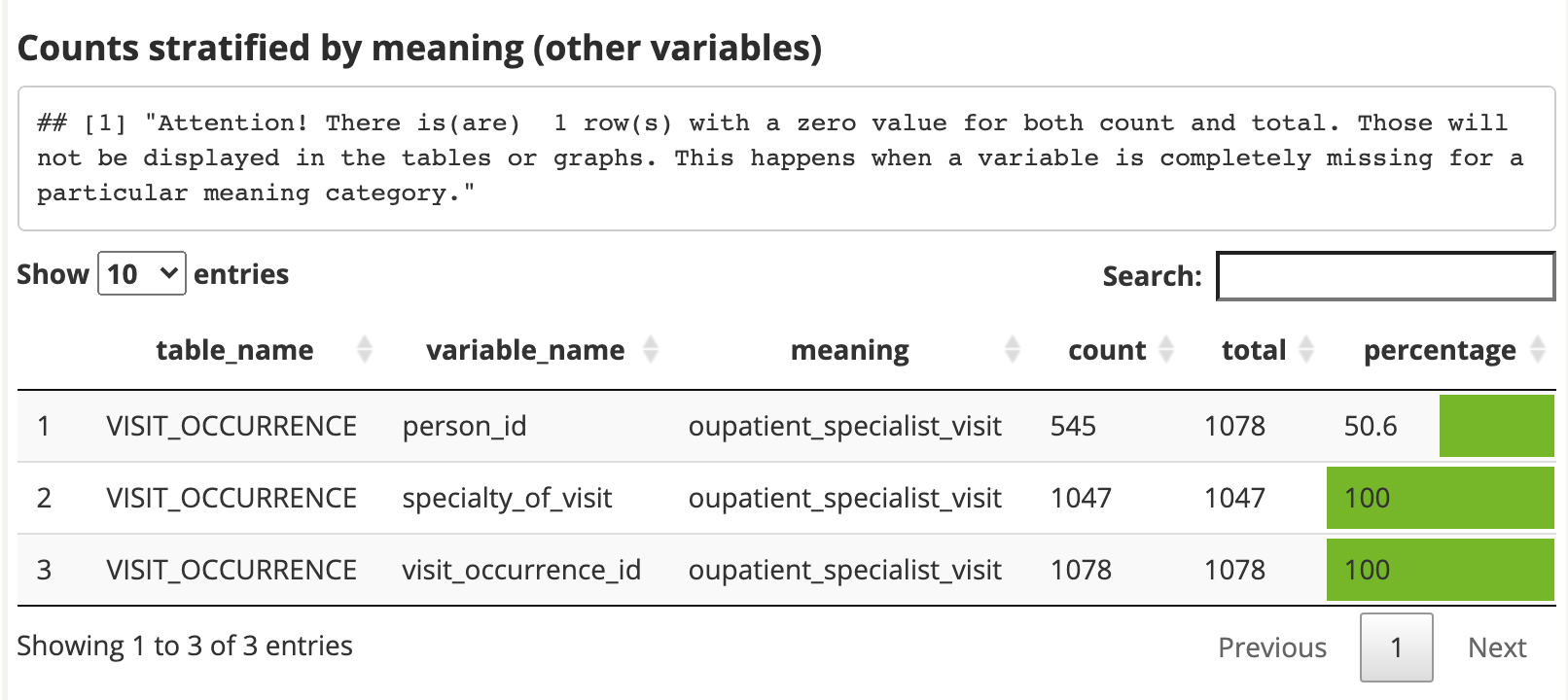
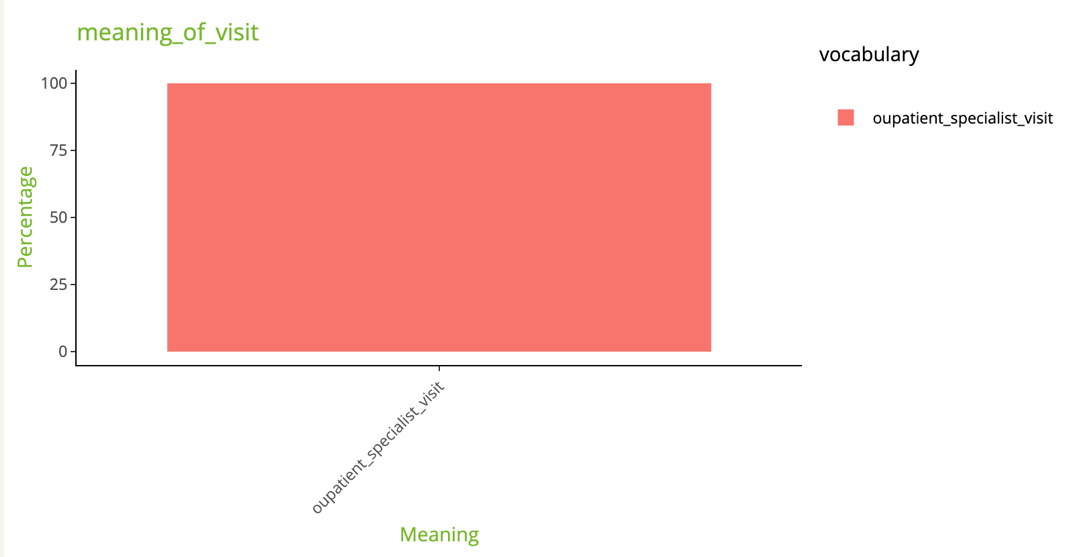
******

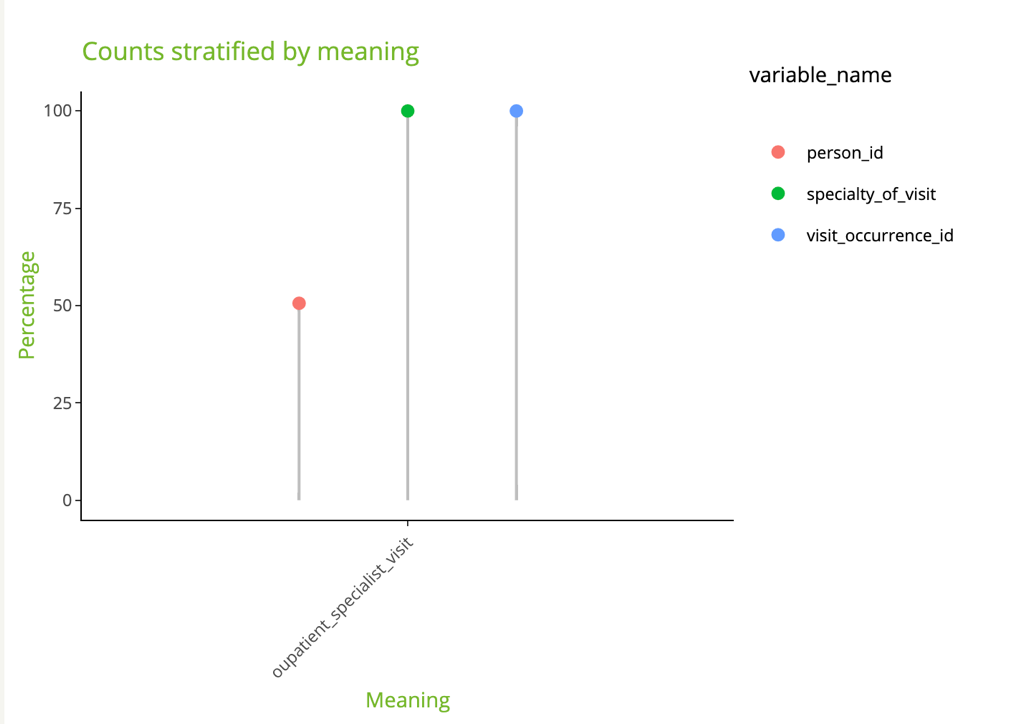
******

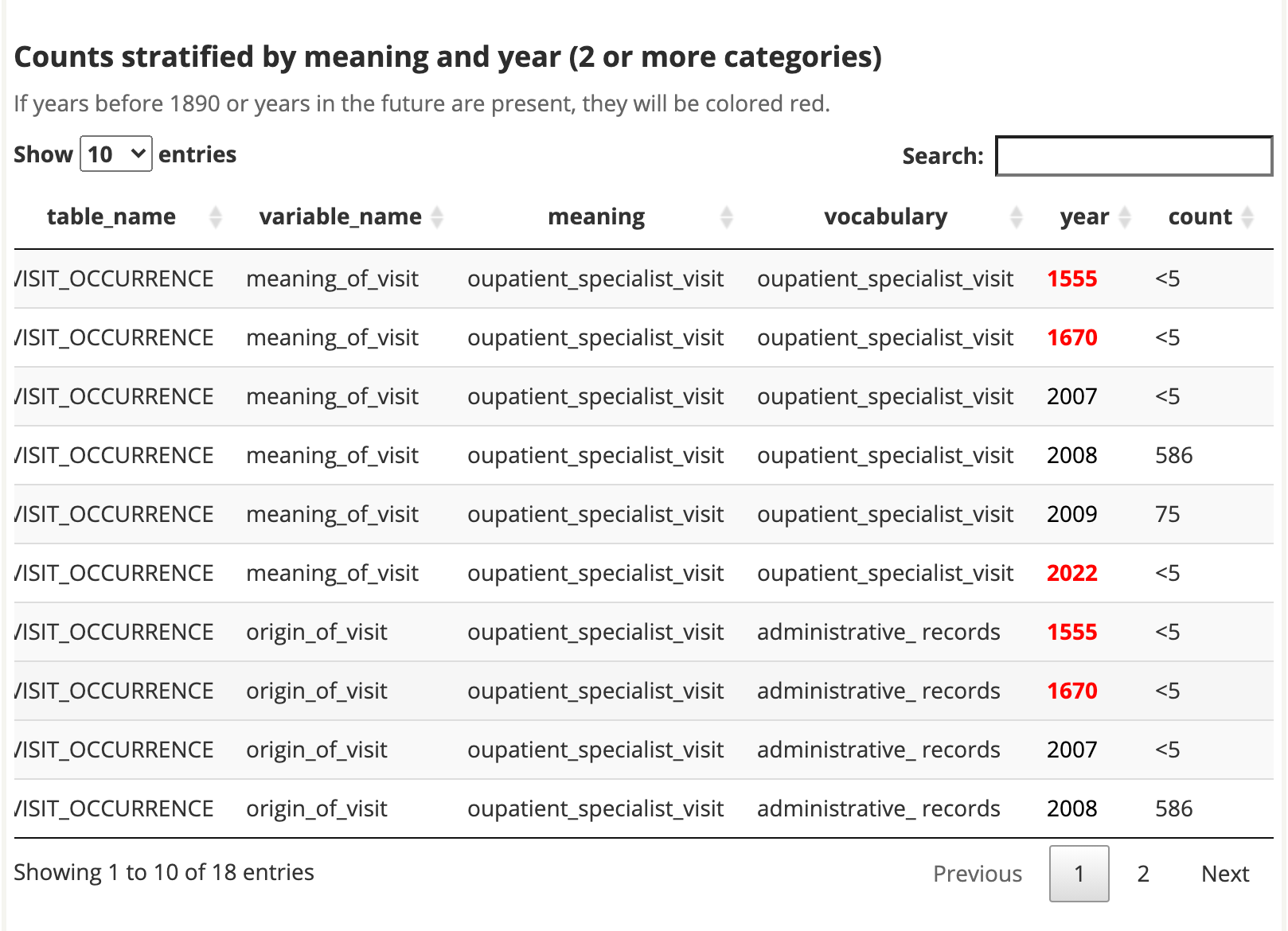
******

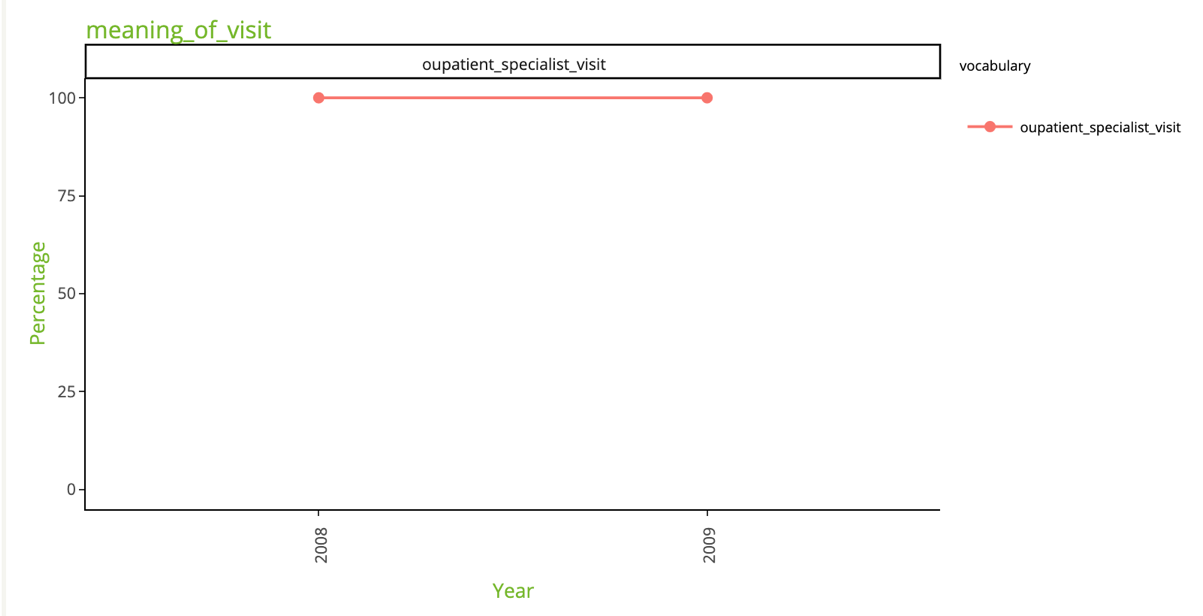
**Step 4 to 5**

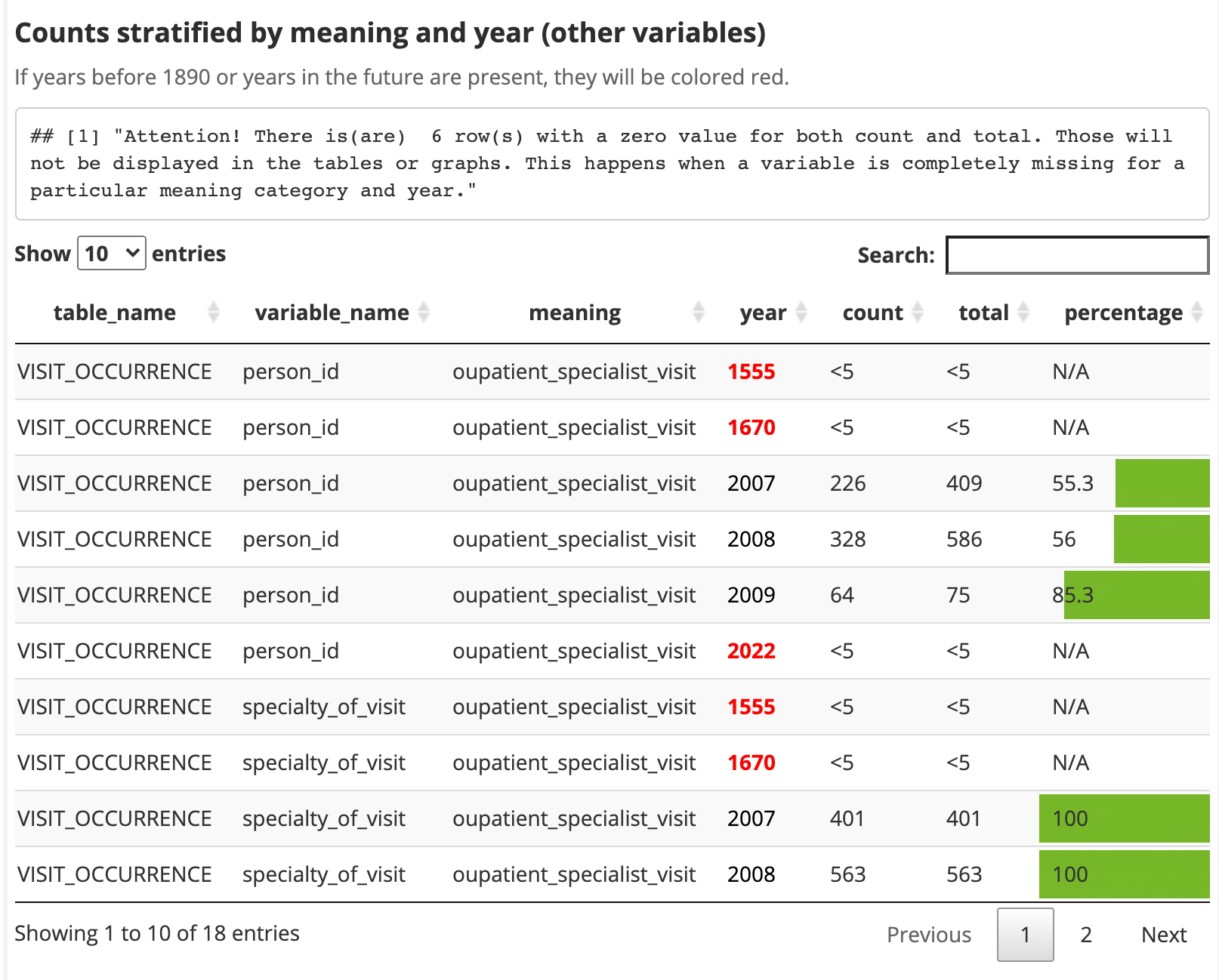
******

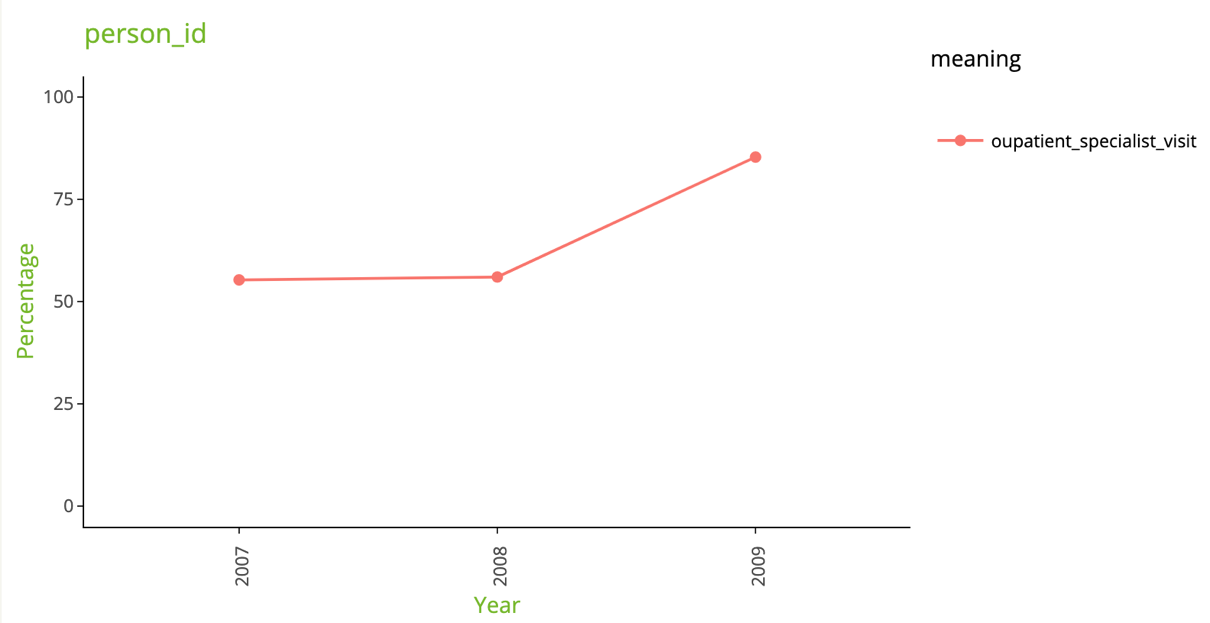
******

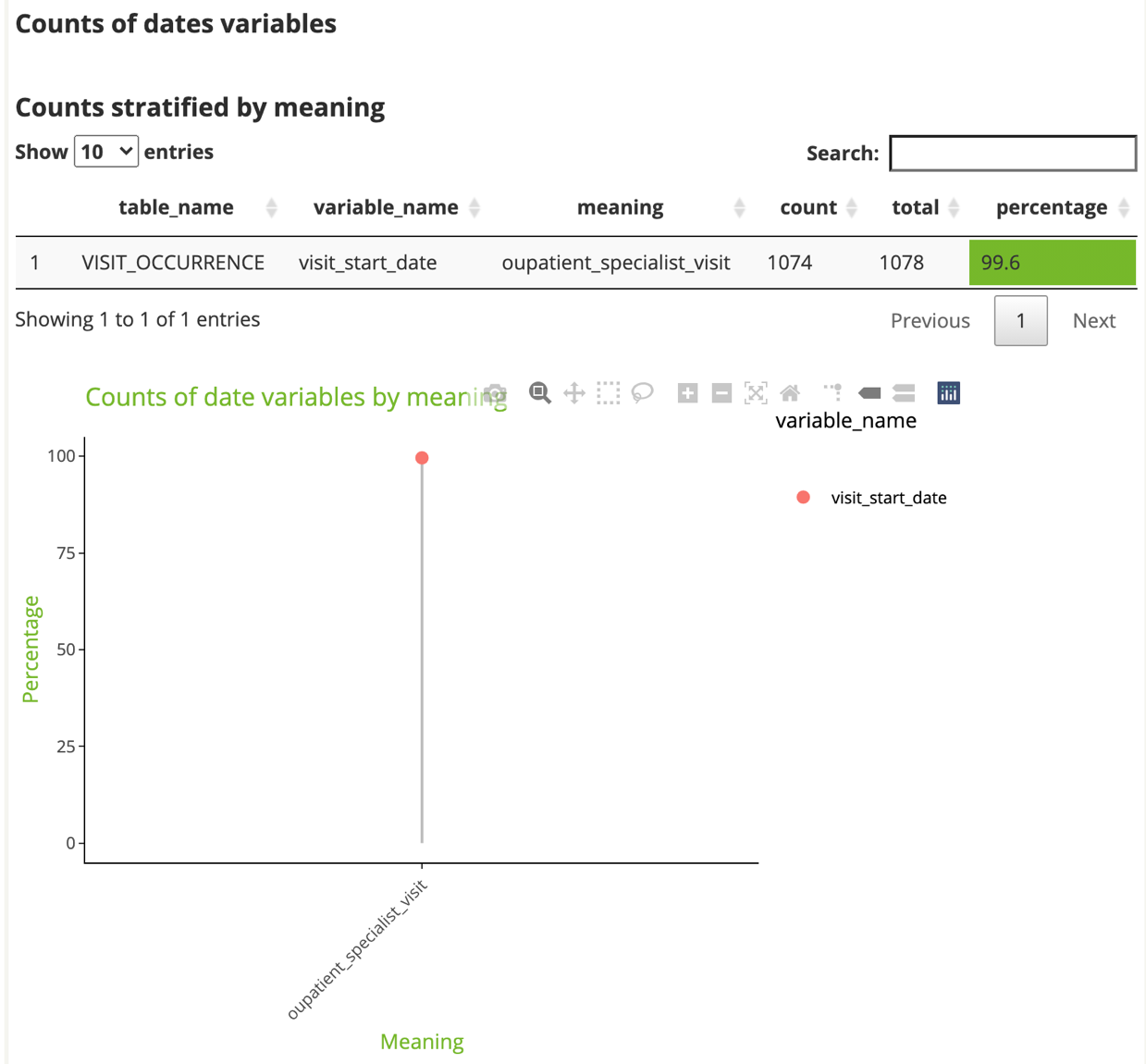
******

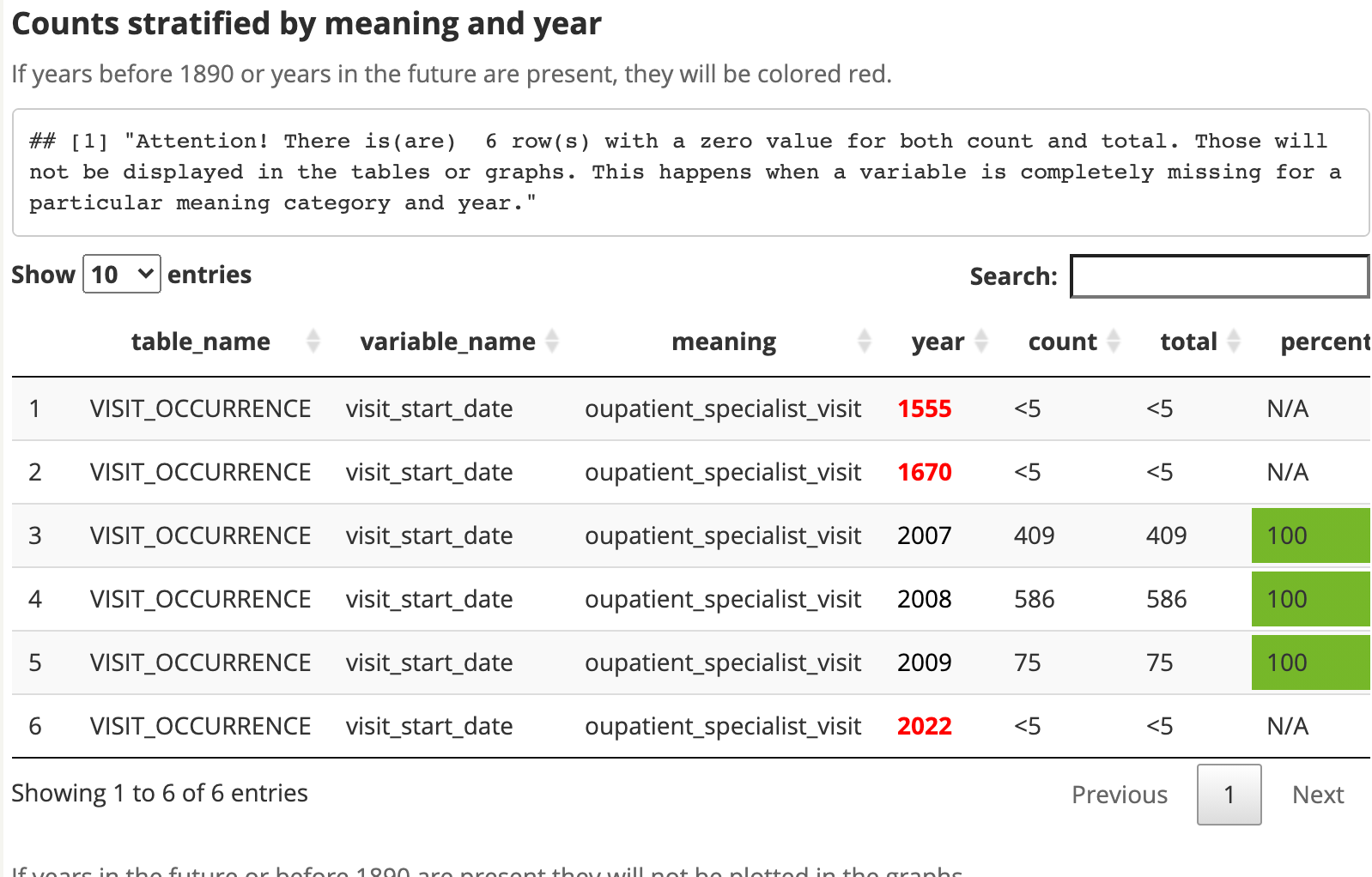
******

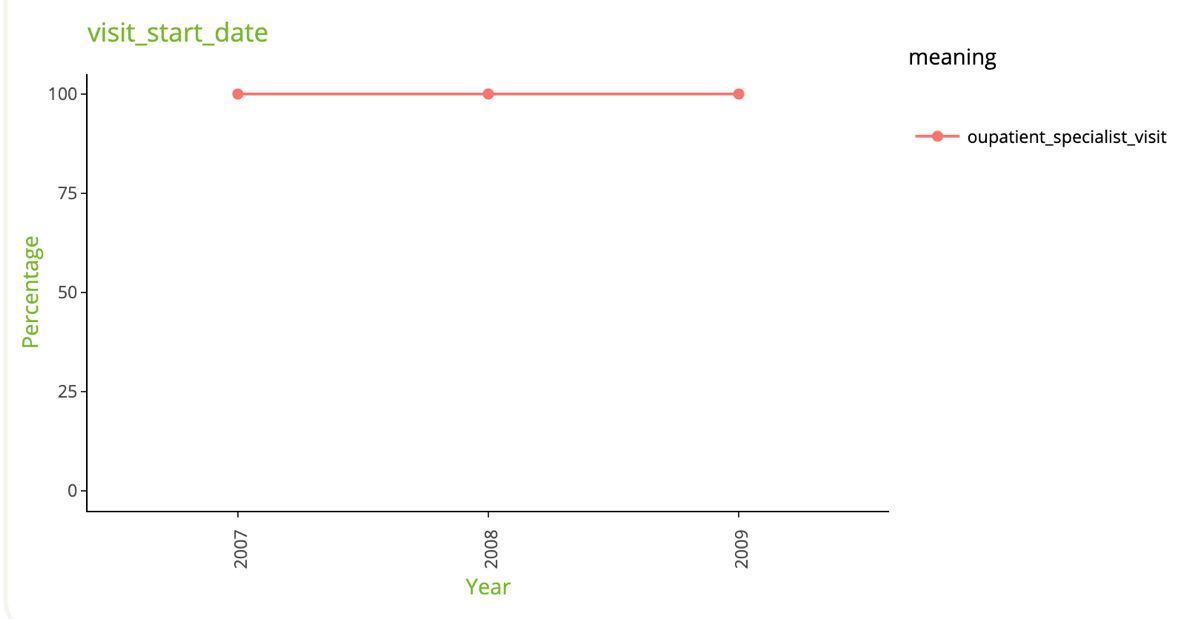
******

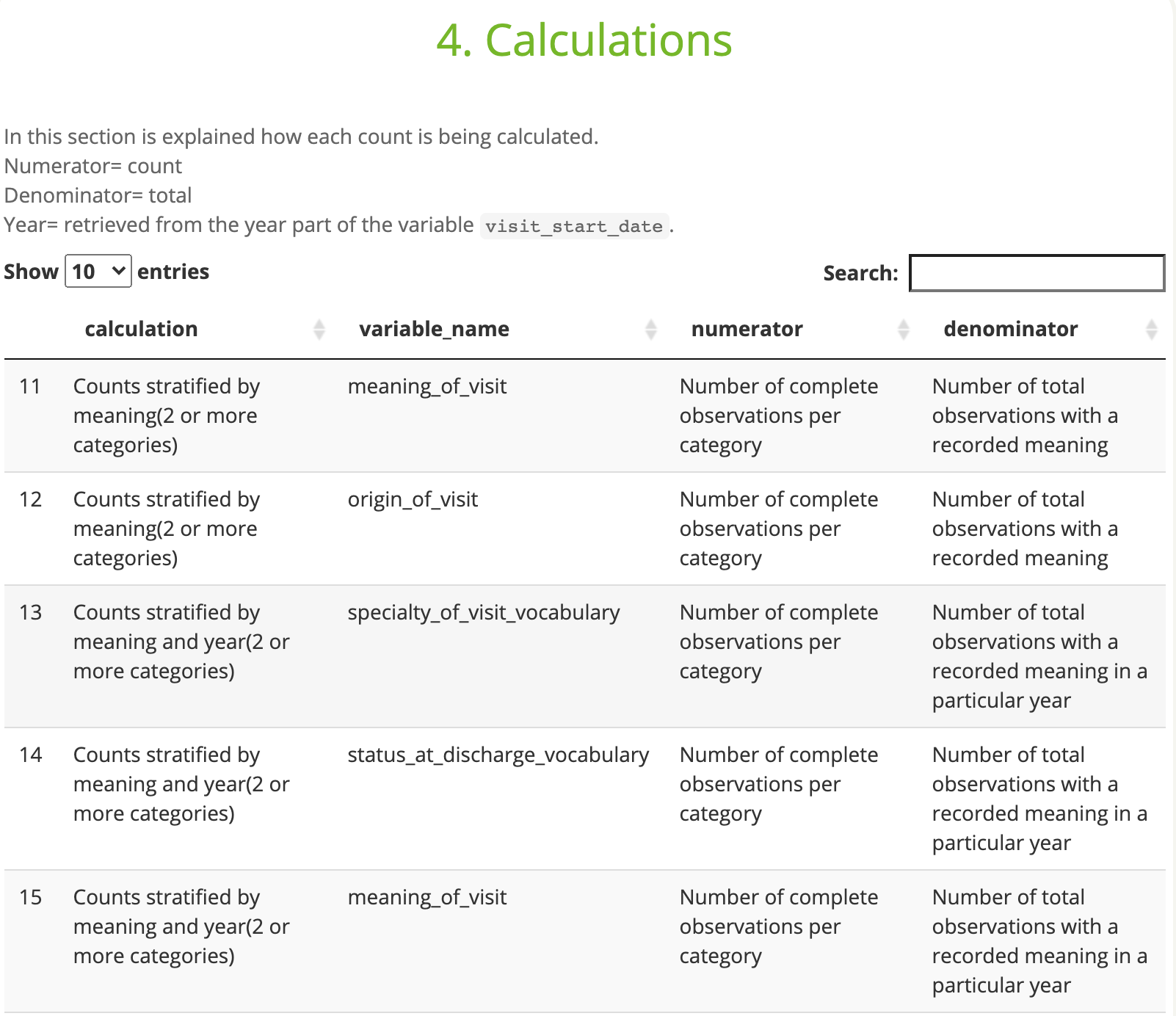
******

******

******

******

******

******

## Level 2

**Example** Check 2.4 Observations in a table of interest without a corresponding record in the PERSONS

