

OMOPartnership



Observational Medical Outcomes Partnership (OMOP)   
Research Lab

Common Data Model (CDM)

Specification

**DRAFT**

Version 1.2

May 18, 2009

Table of Contents

[1. Introduction 5](#_Toc230437740)

[Problem Description 5](#_Toc230437741)

[The Role of the Common Data Model 5](#_Toc230437742)

[Design Principles 6](#_Toc230437743)

[Design Approach 9](#_Toc230437744)

[2. Conceptual Data Model 11](#_Toc230437745)

[The Terminology Dictionary 11](#_Toc230437746)

[CONCEPT 12](#_Toc230437747)

[CONCEPT METADATA 12](#_Toc230437748)

[CONCEPT RELATIONSHIP 13](#_Toc230437749)

[The Common Data Model 15](#_Toc230437750)

[3. Logical Data Model 17](#_Toc230437751)

[Additional Design Principles 17](#_Toc230437752)

[Logical Entity-Relational Diagram 19](#_Toc230437753)

[4. Logical Entities and Attributes 20](#_Toc230437754)

[PERSON 20](#_Toc230437755)

[Business Rules 21](#_Toc230437756)

[Example of Loaded Table 22](#_Toc230437757)

[DRUG\_EXPOSURE 23](#_Toc230437758)

[Business Rules 24](#_Toc230437759)

[Example of Loaded Table 25](#_Toc230437760)

[DRUG\_ERA 26](#_Toc230437761)

[Business Rules 27](#_Toc230437762)

[Example of Loaded Table 28](#_Toc230437763)

[DRUG\_EXPOSURE\_REF 29](#_Toc230437764)

[Example of Loaded Table 29](#_Toc230437765)

[CONDITION\_OCCURRENCE 30](#_Toc230437766)

[Business Rules 31](#_Toc230437767)

[Example of Loaded Table 32](#_Toc230437768)

[CONDITION\_ERA 34](#_Toc230437769)

[Business Rules 36](#_Toc230437770)

[Example of Loaded Table 37](#_Toc230437771)

[CONDITION\_OCCURRENCE\_REF 38](#_Toc230437772)

[VISIT\_OCCURRENCE 39](#_Toc230437773)

[Business Rules 39](#_Toc230437774)

[Example of Loaded Table 40](#_Toc230437775)

[PROCEDURE\_OCCURRENCE 41](#_Toc230437776)

[Business Rules 42](#_Toc230437777)

[Example of Loaded Table 42](#_Toc230437778)

[PROCEDURE\_OCCURRENCE\_REF 44](#_Toc230437779)

[OBSERVATION 45](#_Toc230437780)

[Business Rules 46](#_Toc230437781)

[Example of Loaded Table 47](#_Toc230437782)

[OBSERVATION\_TYPE 48](#_Toc230437783)

[Example of Loaded Table 48](#_Toc230437784)

[OBSERVATION\_PERIOD 49](#_Toc230437785)

[Business Rules 50](#_Toc230437786)

[Example of Loaded Table 50](#_Toc230437787)

[Appendix A: Drug Exposure Type Codes 51](#_Toc230437788)

[Appendix B: Condition Occurrence Type Codes 52](#_Toc230437789)

[Appendix C: Procedure Occurrence Type Codes 54](#_Toc230437790)

**Document Control**

Change Record

| Date | Author | Version | Change Reference |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contributors

| Name | Organization | Title |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

Reviewers

| Name | Role | Title | Date Reviewed |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Document References

| Document Title | Type of Reference | Document Location |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 1. Introduction

The Observational Medical Outcomes Partnership (OMOP) is a public-private partnership designed to protect human health by improving the monitoring of drugs for safety and effectiveness. The partnership, which began in the fourth quarter of 2008, is conducting a two-year research initiative to determine the contribution and utility of using existing health care databases to identify and evaluate safety issues associated with drugs that are already on the market.

OMOP is funded and managed through the Foundation for the National Institutes of Health, and draws on the expertise and resources of the pharmaceutical industry, academic institutions, non-profit organizations, the Food and Drug Administration (FDA), and other federal agencies. In addition to sponsoring specific research efforts, OMOP is creating a set of tools—such as data models, experimental protocols, and database evaluation tools—that will be placed in the public domain to encourage research by a broad community of scientific investigators. All project results will be made public in accordance with the public health mission of the partnership. These will include comprehensive reports on scientific and technical findings, lessons learned, and peer-reviewed articles on the experimental findings by OMOP’s sponsored investigators.

This document describes the design of—and the rationale behind—one of the aforementioned tools, the OMOP Common Data Model (CDM). The remainder of this introductory chapter describes the CDM and its place in the larger OMOP tool set. Subsequent chapters of this document describe how the OMOP project team designed the CDM, and how OMOP researchers will use the CDM to develop and evaluate new, data-driven research methods for drug safety surveillance.

## Problem Description

One of OMOP’s goals is to define processes that can be used to assess the feasibility and utility of using observational data to identify and evaluate associations between drugs and health-related conditions. To facilitate its methodological research, the Partnership will evaluate the performance of various analytical methods for identifying drug-outcome associations across multiple disparate observational data sources (administrative claims and electronic health records). OMOP will partner with a number of different organizations with observational data to undertake this research, including licensing data that can be housed centralized in the OMOP Research Core and collaborating with data providers as a distributed network.

To facilitate this research, OMOP needs to develop a common structure and framework for organizing and standardizing observational data. Such is the role of the Common Data Model in the OMOP pilot infrastructure.

## The Role of the Common Data Model

The Common Data Model, combined with a method for standardizing its content (via a Terminology Dictionary, described below) will ensure that research methods can be systematically applied to produce meaningfully comparable results.

No single observational data source is likely to be sufficient to meet all expected drug safety analysis needs, so there is interest in assessing the feasibility and utility of analyzing multiple data sources concurrently. The CDM, however, is not intended to be an integration point for multiple source data sets. Rather, OMOP researchers will create a separate CDM instance for each source data set. Analysis results from disparate sources can be brought together to facilitate comparisons and synthesis of the aggregated findings.

All analysis methods and code (e.g., SAS, SQL, or R programs) used to execute OMOP research protocols will be developed for the Common Data Model, with the express purpose of enabling a common set of procedures to be applied to (i.e., to be “portable” across) each participating data source. OMOP intends to test the feasibility of both distributed and centralized network architectures to enable analyses across disparate observational data sources. All participating data sources will be transformed into the Common Data Model structure and Terminology Dictionary standards, regardless of where the data reside either logically (e.g., in multiple databases) or physically (e.g., in multiple geographies).

## Design Principles

The OMOP Common Data Model intends to facilitate observational analyses of disparate health care databases, including, but not necessarily limited to, administrative claims and EHRs. Observational research will be conducted to identify and evaluate associations between drug exposure and condition occurrence. Specific Health Outcomes of Interest (HOIs) may be defined by clinical events (e.g., diagnoses, observations, procedures, etc.) in predefined temporal relationships.

The CDM must include all observational data elements that are relevant to identifying drug exposures and defining condition occurrence. However, the model does not necessarily need to provide a mechanism for archiving all observational data elements. For example, cost information—which is a major component of administrative claims data, but which may not play a prominent role in identifying associations between drug exposures and conditions—may not have a place in the CDM.

The CDM design documented herein was guided by six design principles.

**Design Principle 1: The OMOP Common Data Model must accommodate all observational data elements that the partnership wishes to collect, including, but not necessarily limited to, those data elements relevant to identifying drug exposures, condition occurrences, and other clinical observations.**

**Design Principle 2: In designing the CDM, the OMOP should not “reinvent the wheel.” The CDM design should leverage, where reasonable and appropriate, the learning inherent in industry-leading data modeling efforts, such as those associated with the HL7 RIM, the HIMSS EHR Definitional Model, the i2b2 Hive framework, the HMORN Virtual Data Warehouse, and others.**

**Design Principle 3: The CDM design must allow each datum to be standardized on a common vocabulary wherever possible by** **relating (i.e., mapping) to the appropriate corresponding standard health care concept in the Terminology Dictionary.**

**Design Principle 4: The CDM design should anticipate the existence of an ideal Terminology Dictionary that maps each source datum to one and only one standard health care concept. However, the CDM design should remain valid if the mapping of a source datum to multiple standard health care concepts should be required.**

CDM Design Principle 3 implies the existence of a Terminology Dictionary that assigns to-be-standardized values from source data sets to standard health care concepts. Ideally, each unique CDM datum that must be standardized to the Terminology Dictionary will have its best match to exactly one of the Terminology Dictionary’s standard health care concepts. Therefore, there will be a many-to-one relationship between the to-be-standardized data elements in the CDM and the standard health care concepts in the Terminology Dictionary.

Consider the example in which source data set A indicates that patient B, in the context of hospital visit C, had a discharging diagnosis represented by ICD-9-CM diagnosis code 410.01, which means “Acute Myocardial Infarction, Anterolateral Wall, Initial Episode of Care.” In this example, the CDM must associate with patient B at least two pieces of information: the diagnosis itself (i.e., the source datum value), and the fact that the diagnosis was a discharging diagnosis (i.e., the source datum value type). As will be explained later in this document, the notion of a “value/value type pair” is a central theme of the design.

Source data set A represents the diagnosis “Acute Myocardial Infarction, Anterolateral Wall, Initial Episode of Care” using ICD-9-CM diagnosis code 410.01. However, another source data set may represent this same diagnosis in a completely different way (e.g., using a different coding system, as a text description, etc.) The Terminology Dictionary will contain one single standard concept, having concept code C123, which means “Initial Episode of Care of Acute Myocardial Infarction in Anterolateral Wall.” Furthermore, the Terminology Dictionary will map to this concept (i.e., standardize) all of the various source-specific representations of this diagnosis, including ICD-9-CM diagnosis code 410.01 from data source A. Queries against a CDM instance will use the standard concept code for this diagnosis rather than its source-specific representation to ensure selection of all patients with this diagnosis regardless of how the data were originally represented in any source data set. That is, by standardizing our query to the Terminology Dictionary, we ensure that the query will be portable to any CDM instance that has also been standardized to the Terminology Dictionary.

Continuing the previous example, the CDM must capture the diagnosis (i.e., value), and also that the diagnosis was a primary outpatient diagnosis (i.e., value type) rather than from an inpatient claim, in a standardized way. To this end, in addition to providing standard concept code C123 to represent the specific diagnosis, the Terminology Dictionary must also provide a standard way of referencing the concept of a primary outpatient diagnosis. To achieve this, the Terminology Dictionary will provide a single concept that means Discharging Diagnosis, and map to it (i.e., standardize) all of the various source-specific representations thereof.

|  |  |  |  |
| --- | --- | --- | --- |
| **Patient** | **Standardized**  **Value** | **Standardized**  **Value Type** | **Translation:** |
| B | C123 | C345 | Patient B has “Acute Myocardial Infarction, Anterolateral Wall, Initial Episode of Care” as a Discharging Diagnosis |

A query for all patients with this discharging diagnosis — that is portable to any CDM instance that has been standardized to the Terminology Dictionary — might resemble the following.

SELECT [Patient] FROM [Table]

WHERE [Standardized Value] = ‘C123’

AND [Standardized Value Type] = ‘C345’;

**Design Principle 5: The CDM design should discourage the use of Protected Health Information (PHI), except where necessary to conduct analyses to protect the public health.**

Observational analyses should be able to be supported by a CDM that minimizes the use of PHI. Such protections would ensure analysis results can inform public health interests without jeopardizing patient privacy. For this reason, CDM tables that correspond to identifiable entities (e.g., Person) should not include columns for HIPAA-recognized identifiers, such as names, patient identification numbers, addresses, telephone numbers, and dates of birth. Only those data elements required to facilitate analysis of drug safety issues should be captured in the CDM, including visit dates, prescription details, and enrollment information. Year of birth can be used as a minimally sufficient surrogate to measure age, acknowledging that this may limit the utility of the model for studying drug effects in infants.

**Design Principle 6: The CDM design, and the databases that instantiate it, must be usable. Of primary importance is the ability of the CDM design to provide a user with the data that he requires for his research. Of secondary importance is the ability of the CDM design to provide a user with data in the manner (i.e., format) that he prefers.**

The CDM design must ultimately be intuitive, not overly complex, and otherwise “researcher-friendly.” Researchers who find it difficult to understand the CDM design will find it difficult to formulate an accurate and efficient query against a CDM instance. And since CDM queries are the starting point for many data-driven research methods, an unwieldy and unintuitive Common Data Model design will effectively undermine the OMOP mission.

## Design Approach

Design Principle 1 points to a Common Data Model that is flexible. Ideally, the CDM will accommodate any value, of any value type, from any OMOP data source, either present or future. Theoretically, we can imagine the CDM as a single table that can hold any data that we care to put into it. For example:

|  |  |  |
| --- | --- | --- |
| Entity | Value Type \* | Value \*\* |
| … | … | … |
| Patient B | Admission Date | 1/1/2009 |
| Patient B | Admission Source | Via Emergency Department |
| Patient B | Gender | Male |
| Patient B | Year of Birth | 1947 |
| Patient B | Discharge Date | 1/10/2009 |
| Patient B | Discharging Diagnosis | Acute Myocardial Infarction, Anterolateral Wall, Initial Episode of Care |
| Patient C | Admission Date | 1/2/2009 |
| Patient C | Admission Source | Physician Referral |
| Patient C | Gender | Female |
| Patient C | Year of Birth | 1980 |
| Patient C | Discharge Date | 1/5/2009 |
| Patient C | Discharging Diagnosis | Acute Laryngotracheitis, With Obstruction |
| … | … | … |

\* Non-coded values provided here for readability. Actual value types would be standard concept codes from the Terminology Dictionary.

\*\* Non-coded values provided here for readability. Except for dates and years, actual values would be standard concept codes from the Terminology Dictionary.

If designed correctly, a CDM consisting of a single, highly normalized table like the one shown above would place no arbitrary limits on the number or kinds of entities, value types, or values that may be stored in a CDM instance. Such a design would provide complete flexibility for new entities, value types, and values in the future, without requiring changes to the data model itself. That is why this kind of design approach is attractive in research environments, where improvements in the methodology might incur iterative changes in the data representation.

Design Principle 6 (usability, simplicity, and intuitiveness) makes attractive a different kind of data model — one that comprises multiple tables with familiar names that connote real-world entities of interest (e.g., patient, diagnosis, procedure, medication, etc.), and columns with familiar names that connote real-world value types of interest (e.g., medication name, NDC, diagnosis name, ICD-9-CM diagnosis code, etc.) Obviously, such a data model would not be as compact as the “one big table” shown above. Separating the model into multiple tables with potentially many columns each would result in tables that are “wider” (i.e., comprising more columns), but not as “tall” (i.e., comprising fewer rows).

While the second approach ameliorates some problems with the first approach — such as improving the performance of queries that employ entire table scans — it does introduce some problems that the first approach does not have. For example, the second approach may require data model changes to accommodate new data sources. In addition, the second approach has the tendency to produce tables that are “sparse.” That is, for any given row in the table, most of the columns will usually be empty (i.e., NULL). Data sparseness like this is not necessarily a problem in and of itself, but it can contribute to inefficient use of storage by the database management system.

There is no clear winner between the two design approaches. User acceptance, based on usability and intuitiveness, is important, but so is the flexibility to accommodate new and different source data sets in the future. Because the specific data sources may not be known at the time when the design is finalized, minimizing the potential for future data model changes is of paramount importance.

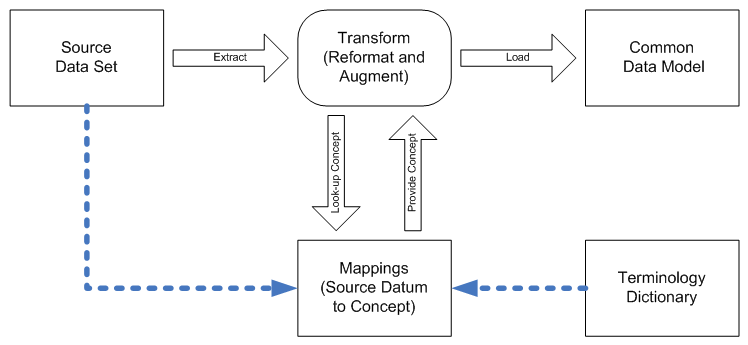
For these reasons, regarding the two design approaches described in this section, the CDM design documented herein aims to achieve “the best of both worlds.” Specifically, the CDM design defines separate table structures for different OMOP data domains (i.e., persons, visits, drugs, conditions, observations, procedures, etc.), but employs the efficient “value/value type” design pattern for those table structures.

# 2. Conceptual Data Model

As explained previously within this document, the CDM defines table structures for each of the data domains (e.g., persons, visits, drugs, conditions, observations, procedures, etc.). Loading a CDM instance from a source data set standardizes the data, both in format and in content, to ensure that research methods applied to the CDM instance will be portable to any other CDM instance. As the figure below depicts, the ETL logic that loads the CDM instance from the source data set performs two actions:

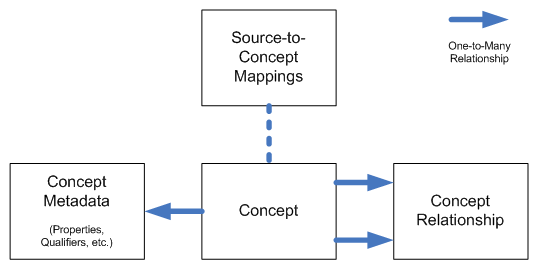
1. Transforms (i.e., reformats) the source data set content to conform to the CDM’s table structures, and
2. Augments the source data with their corresponding concept codes from the Terminology Dictionary.

To achieve this augmentation, the ETL logic uses an individual source datum to perform a look-up on the Terminology Dictionary, finds the standard concept that corresponds to the datum, and loads the concept code into the appropriate table and column of the CDM instance. Predefined mappings between the source data set’s distinct values and the Terminology Dictionary’s standard concepts are required for this look-up.



## The Terminology Dictionary

The Terminology Dictionary is a semantic network containing all of the concepts, their attributes, concept-to-concept relationships, and other metadata necessary to describe the meanings and structures of the data within the CDM. The Terminology Dictionary will accommodate concepts for each of the domains of interest, including drugs, conditions, procedures, visits, and demographics. The following figure depicts its internal organization.



Continuing the example from the previous chapter, the Terminology Dictionary will contain a single standard concept, having concept code C123, that means “Acute Myocardial Infarction, Anterolateral Wall, Initial Episode of Care.” Furthermore, the Terminology Dictionary will map to this concept (i.e., standardize) all of the various source-specific representations of this diagnosis, including ICD-9-CM diagnosis code 410.01 from data source A. The concept itself will be captured in the Concept section of the Terminology Dictionary. The mappings from the concept to its various source-specific representations will be captured in the Concept Metadata section, along with other information.

### CONCEPT

The Concept section of the Terminology Dictionary will contain, at a minimum, a unique identifier (i.e., Concept Code) for each concept (e.g., C123), and a corresponding unique Concept Name (e.g., “Acute Myocardial Infarction, Anterolateral Wall, Initial Episode of Care”). The Concept section may also contain an indicator of each concept’s “Kind” (e.g., event, substance, organism, biological process, chemical or laboratory finding, etc.), to aid in grouping the concepts of a large Terminology Dictionary into more easily manageable sections.

### CONCEPT METADATA

The Concept Metadata section of the Terminology Dictionary captures the mappings between each concept and its various source-specific representations. For example, ICD-9-CM diagnosis code 410.01 from data source A is synonymous with concept C123, which means “Acute Myocardial Infarction, Anterolateral Wall, Initial Episode of Care.” Such source-to-concept mappings, by attaching source-specific synonyms to a concept, are made possible by “properties” and “qualifiers.” Source-to-concept mappings will be developed for several standard coding schemes (i.e., NDC, GPI, ICD9, CPT, and others) to facilitate transformation from source data into the Common Data Model.

Properties capture the characteristics of a concept, other than its unique Concept Code and Concept Name. Examples of concept properties include, but are not limited to, the following.

* Preferred Name
* Semantic Type
* Definition
* Unit of Measure
* Synonym

Each property has an associated value. For example, the value of the Preferred Name property for concept C123 might be “AMI – Initial Episode,” or simply “AMI.” Unlike the Concept Name (in the Concept section), the Preferred Name does not necessarily have to be unique. Standard terminologies and conventions will be applied where possible.

Capturing concept synonyms requires both properties and qualifiers. Qualifiers are property-specific modifiers that express additional information about a property, and so, like properties, also have values. Examples of qualifiers for the Synonym property include, but are not limited to, the following.

* Synonym Source
* Local Code
* Term Type

Continuing the previous example, we would create a Synonym property with the value “410.01,” and on this property we would establish two qualifiers. The first qualifier, Synonym Source, might have the value “Source A.” The second qualifier, Term Type, would have the value “ICD-9-CM.”

As noted above, the ETL logic that loads the CDM instance from the source data set looks up concepts from the Terminology Dictionary as it transforms (i.e., reformats) the source and concept data to conform to the CDM table structures. The look-up process entails matching the inbound source data against the properties and qualifiers in the Terminology Dictionary to determine which standard Concept Code values are mapped to which source data values. In the preceding example, the ETL logic would match the inbound discharging diagnosis code from data source A, ICD-9-CM code 410.01, to the properties and qualifiers for concept C123, and would place this Concept Code value in the appropriate table and column in the CDM instance.

### CONCEPT RELATIONSHIP

The Concept Relationship section of the Terminology Dictionary captures the types of relationships (such as parent-child) contained within concept hierarchy. The concept hierarchy in the Terminology Dictionary is of special importance to OMOP, because it allows researchers to query a CDM instance for classes of concepts without needing to know the individual concepts that those classes subsume. For example, a researcher should be able to query a CDM instance for all drugs within a specific therapeutic class without needing to know the specific concept codes of each of the drugs that fall within the class, and will be able to query a CDM instance for a particular class of medical conditions without necessarily needing to know which individual diagnoses comprise that condition class. The Terminology Dictionary contains the class of concepts on which to query or analyze, and will transfer the class’ appropriate Concept Code from the Terminology Dictionary to the query and analysis tool.

Within the Concept Relationship section of the Terminology Dictionary is a parent-child table that captures the hierarchical arrangement of concepts. Continuing the previous example, an excerpt from this table might appear as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parent Concept Code** | **Parent Concept Name** | **Child Concept Code** | **Child Concept Name** |
| … | … | … | … |
| C904 | Ischemic Heart Disease | C593 | Acute Myocardial Infarction |
| C904 | Ischemic Heart Disease | C221 | Angina Pectoris |
| C593 | Acute Myocardial Infarction | C123 | Acute Myocardial Infarction, Anterolateral Wall, Initial Episode of Care |
| C593 | Acute Myocardial Infarction | C803 | Acute Myocardial Infarction, Inferoposterior Wall, Subsequent Episode of Care |
| … | … | … | … |

An OMOP researcher may be interested in retrieving data on all patients with a discharging diagnosis of “Acute Myocardial Infarction, Anterolateral Wall, Initial Episode of Care,” in which case he would incorporate Concept Code C123 into the WHERE… clause of his SQL query. If the researcher wishes to expand his query to all AMI patients, then he would replace C123 in his query with C593, which is the parent of C123. Likewise, if the researcher wishes to expand his query even further, to all patients with Ischemic Heart Disease, he would incorporate Concept Code C904 into the WHERE… clause of his SQL query. “Behind the scenes” in the Terminology Dictionary, an ancestor-descendant table keeps track of all concepts that must be returned in such a class-based query. For example, the class-based query for all Ischemic Heart Disease patients (C904) must return all child concepts of C904, regardless of the hierarchical level of the child concept. These are C593, C221, C123, and C803, as well as parent C904 itself. This is also commonly referred to as the “transitive closure” of class concept C904.

|  |  |
| --- | --- |
| Ancestor Concept Code | Descendant Concept Code |
| … | … |
| C904 | C904 |
| C904 | C593 |
| C904 | C221 |
| C904 | C123 |
| C904 | C803 |
| C593 | C593 |
| C593 | C123 |
| C593 | C803 |
| C221 | C221 |
| … | … |

## The Common Data Model

Conceptually, the Common Data Model has eight entities. These are:

1. Person
2. Observation Period (the time at which health care information may be available, which can be used to estimate event rates over time)
3. Drug Exposure (i.e., the association between Person and Drug for a specific time period)
4. Health Outcome of Interest, which may be based on a combination of:
5. One or more medical Condition(s) of the Person
6. One or more Clinical Observations about the Person (e.g., laboratory test results)
7. One or more Medical Procedures that were administered to the Person

8. One or more Visits for healthcare services for the Person

Figure 1 on the next page illustrates these conceptual entities, and their relationships with the standard health care concepts stored in the Terminology Dictionary. Figure 1 depicts the mappings from each of the CDM conceptual entities to the standard health care concepts in the Terminology Dictionary, and each conceptual entity may have many such mappings. For example, the Visit entity will map the values for Visit Type (e.g., hospital inpatient, hospital outpatient, emergency room, ambulatory/office visit, etc.) to the standard concepts that represent these values. Likewise, the Observation entity will map the values for Observation Type (e.g., laboratory test result) and individual instances of specific Observation Types (e.g., blood glucose test, serum sodium test, etc.) to the standard concepts that represent them.

Generally, any conceptual construct of the CDM that may be assigned a value should have a corresponding standard concept in the Terminology Dictionary. For example, since a blood glucose test may be assigned a value (i.e., the test result, expressed as a number and a unit of measure), there should be in the Terminology a standard concept that represents the blood glucose test. There should also be in the Terminology Dictionary standard concepts that represent all of the possible units of measure that may qualify a blood glucose test result. Some other laboratory tests may actually have a small number of possible discrete values (e.g., positive or negative, present or absent, etc.), and in such cases the result values themselves should also have standard concept codes in the Terminology Dictionary.



**Figure 1:** The conceptual view of the Common Data Model

# 3. Logical Data Model

## Additional Design Principles

The following additional design principles extend the six overall design principles (and particularly Design Principle 1), which were defined in the first chapter of this document, from the abstract conceptual level to the more detailed logical level.

* **Extensibility:** Enable an unlimited number of Persons to be included in any CDM instance. Also, for each Person in the source data set, enable an unlimited number of types of Drug Exposures, Conditions, Observations, Procedures, and Visits to be included in the CDM instance. Accommodate all conceivable values of every conceivable attribute of those Drug Exposures, Conditions, Observations, and Procedures.
* **Flexibility:** Because a CDM instance is populated based on source-to-target data mappings and other metadata rather than “hard-wired” between predefined source and target column pairs, the CDM is flexible on the allowed types and values of Drug Exposures, Conditions, Observations, and Procedures. For example, the Common Data Model described herein can evolve over time to accommodate different types of Observations without necessitating CDM design changes.
* **Reference Concept Codes:** The CDM must be able to reference the standard concepts in the Terminology Dictionary both for types and values of Drug Exposures, Conditions, Observations, and Procedures whenever available. We anticipate that the Terminology Dictionary will contain standard Concept Codes for all drugs, medical conditions, clinical observations, and medical procedures in the source data sets, as well as standard Concept Codes for many of the values of the attributes of these entities.
* **Data Access:** Each database that instantiates the CDM will be deployed as a relational database accessible from any SQL interface. For the purposes of the OMOP Research Lab deployment, access to the CDM instances must be able to be accomplished via the tools listed in Table 1 on the next page.

(continued on the next page)

|  |  |  |
| --- | --- | --- |
| **Tool Name** | **Description** | **Purpose** |
| **SQL\*Plus** | An interactive and batch query tool installed with Oracle’s  Database Server or Client installation. It has a command-line user interface, a  Windows Graphical User Interface (GUI) and the SQL\*Plus web-based user interface. | Through its own command and environment, SQL\*Plus provides access to the Oracle Database.  The tool enables the running of SQL statements interactively or as part of the batch process. |
| **PL/SQL**  (Procedural Language extension of SQL) | PL/SQL is a combination of SQL along with the procedural features of programming languages. It was developed by Oracle Corporation to enhance the capabilities of SQL. | PL/SQL is used to run multi step processes that include logic, decision tress and loops much like a programming language.  While it can be used for running methods, it is currently leveraged for the running of the ETL processes used in the populating of the instances of CDM |
| **SAS Analytics** | SAS provides a range of techniques and processes for the collection, classification, analysis and interpretation of data to reveal patterns, anomalies, key variables and relationships | The OMOP environment has integrated a dedicated SAS server and toolset to seamless retrieve data from CDM in order to create *sas* output files or data sets to be loaded onto dedicated *Results* tables |
| **R** | R is an open-source language for statistical computing, including data manipulation, calculation, and statistical display. It is particularly useful for arrays ( matrices). | The OMOP environment has a dedicated server for R that will allow users to create, edit, and run R as methods, and for special purposes programs such as those that will create the simulated data sets that will be used to validate methods. |

**Table 1:** CDM data access tools for the OMOP Research Lab

## Logical Entity-Relational Diagram

The figure below depicts the entire CDM Logical Model along with the Terminology Dictionary.



Terminology

Dictionary

The next chapter of this document describes the components of this Logical Data Model in detail.

# 4. Logical Entities and Attributes

## PERSON

The Person entity is one of the basic dimensions of analysis and, when combined with the Drug Exposure, Condition, Observation, and Procedure entities, presents the framework for active drug surveillance. The source data for the Person entity comes from patient demographics data that can be de-identified to ensure HIPAA compliance, and the actual extent of these data vary by data source. The Person entity is concept-driven, meaning that the attribute values are stored as standard concept codes rather than original (i.e., “raw”) source values.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| PERSON\_ID | INTEGER | YES | System-generated identifier to uniquely identify each person. |
| YEAR\_OF\_BIRTH | INTEGER | NO | Year of birth of the person. For data sources with date of birth, only the year is extracted. For data sources where the year of birth is not available, the approximate year of birth is derived based on any age group categorization available. |
| GENDER\_CONCEPT\_CODE | VARCHAR(20) | NO | Standard Concept Code for the gender of the person.  The Person Gender is mapped to a standard Gender concept from the Terminology Dictionary and the corresponding concept code is stored here as a reference.  The Concept Code references the CONCEPT entity from the Terminology Dictionary. |
| RACE\_CONCEPT\_CODE | VARCHAR(20) | NO | Standard Concept Code for the race of the Person.  The Person Race is mapped to a standard Race concept from the Terminology Dictionary and the corresponding concept code is stored here as a reference.  The Concept Code references the CONCEPT entity from the Terminology Dictionary. |

(continued on the next page)

|  |  |  |  |
| --- | --- | --- | --- |
| Field | Data Type | Required | Description and Notes |
| LOCATION\_CONCEPT\_CODE | VARCHAR(20) | NO | Standard Concept Code for the location of the person.  The Person Location is mapped to a standard Geographic Location concept from the Terminology Dictionary and the corresponding concept code is stored here as a reference.  The Concept Code references the CONCEPT entity from the Terminology Dictionary. |
| SOURCE\_PERSON\_KEY | VARCHAR(32) | NO | Encrypted key derived from the person identifier from the source data. Necessary when a drug safety issue requires a link back to the person data from the raw source data set.  The Source Person Key always needs to be an encrypted value and no identifier with any medical or demographic significance can be stored.  The OMOP Research environment stores the de-identified unique identifiers for that person from the source data as the Source Person Key. |

### Business Rules

* Person data will remain de-identified as much as possible to comply with Design Principle 5. Accordingly, the CDM will not store the precise date of birth (it will store only the year of birth), nor will it store any identifiers that could be used to re-identify the person data.
* The granularity of the person data from the source system will be maintained. There will be no consolidation or aggregation of individual person records.
* Standard attributes will be stored as concept codes. Original source values will be mapped to the corresponding standard concept codes in the Terminology Dictionary.
* Person source data attributes that are in-scope are limited to race, gender, location, and year of birth. Other person source attributes are out-of-scope.

### Example of Loaded Table

Consider the following example of inbound source data on patients.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SOURCE PATIENT ID** | **PATIENT YEAR of BIRTH** | **PATIENT GENDER** | **PATIENT LOCATION** | **PATIENT RACE** |
| 121107 | 1932 | FEMALE | Phoenix – MSA | Caucasian |
| 127260 | 1933 | FEMALE | Phoenix – MSA | Caucasian |

Sample concept code representation of demographic data from the Terminology Dictionary follows.

|  |  |
| --- | --- |
| **Concept Code** | **Concept Name** |
| C0043157 | Race – Caucasian |
| G9999999 | Gender – Female |
| L8777777 | Location – Phoenix |

Unique system-generated identifiers are used for the PERSON\_ID key: 121107 and 127260 for the two patients.

The one-way hash keys for the source patient identifiers are determined in this example using MD5 hashing.

|  |  |
| --- | --- |
| **Source Person Identifier** | **Source Person Key** |
| 12345 | 827ccb0eea8a706c4c34a16891f84e7b |
| 67890 | 1e01ba3e07ac48cbdab2d3284d1dd0fa |

The above data are represented in the CDM PERSON table as follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PERSON ID** | **YEAR OF BIRTH** | **GENDER CONCEPT CODE** | **RACE CONCEPT CODE** | **LOCATION CONCEPT CODE** | **SOURCE PERSON KEY** |
| 121107 | 1932 | G9999999 | C0043157 | L8777777 | 827ccb0eea8a706c4c34a16891f84e7b |
| 127260 | 1933 | G9999999 | C0043157 | L8777777 | 1e01ba3e07ac48cbdab2d3284d1dd0fa |

## DRUG\_EXPOSURE

Drug Exposure contains all individual records that suggest drug utilization from within the observational source.. Drug Exposure indicators include drug details (captured as standard concept codes in the Terminology Dictionary), drug quantity, number of days supply, period of exposure, and prescription refill data. Drug Exposure is recorded in a variety of ways.

* The “Prescription” section of an EHR captures prescriptions written by physicians.
* Other drugs (both non-prescription products and medications prescribed by other providers) used by a person are recorded in the “Medications” section of the EHR.
* Administrative claim systems capture prescriptions filled at dispensing providers.
* Drug Exposure information as a by-product of certain procedure codes (i.e., procedure codes that refer to professional services related to the administration of certain drugs).

Drug Exposures are indicated in the CDM by standard drug concepts from the Terminology Dictionary. The standard concept code for a drug is stored with the drug reference data, however the concept hierarchy and therapeutic class categorizations from the source data are not stored with the drug reference data (they are stored in the Terminology Dictionary).

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| DRUG\_EXPOSURE\_ID | INTEGER | YES | System-generated identifier to uniquely identify each Drug Exposure. |
| PERSON\_ID | INTEGER | YES | System-generated identifier for the person for who is the subject of the Drug Exposure. Foreign key to the PERSON entity. Demographics for the person are captured in the PERSON entity. |
| DRUG\_EXPOSURE\_START\_DATE | DATE | YES | Start date for the current instance of drug utilization. Valid indicators include a start date of a prescription, the date a prescription was filled, or the date on which a drug administration procedure was recorded. |
| DRUG\_EXPOSURE\_END\_DATE | DATE | NO | End date for the current instance of drug utilization. Not available from all sources. |
| DRUG\_CONCEPT\_CODE | VARCHAR(20) | NO | Standard concept code, from the Terminology Dictionary, related to the Drug concept. Used to map to standard drug information and concept hierarchy in the Terminology Dictionary. |
| DRUG\_EXPOSURE\_TYPE | VARCHAR(3) | YES | Predefined code for the type of Drug Exposure recorded. Defines the indicator from which the Drug Exposure was identified, including medication history, filled prescriptions etc.  Please see the description for the DRUG\_EXPOSURE\_REF entity for more details. |
| SOURCE\_DRUG\_CODE | VARCHAR(20) | YES | Drug identifier as captured in the raw source data. The types of identifiers allowed include National Drug Codes (NDCs), Generic Product Identifier (GPI) codes, etc. |
| STOP\_REASON | VARCHAR(20) | NO | Reason the medication was stopped, where available. Reasons include Regimen Completed, Changed, Removed, etc. |
| REFILLS | INTEGER | NO | Number of refills for the prescription. |
| DRUG\_QUANTITY | INTEGER | NO | Quantity of drug recorded as part of the instance of Drug Exposure. |
| DAYS\_SUPPLY | INTEGER | NO | Number of days of supply of the medication recorded in the Drug Exposure. |

### Business Rules

* Source drug identifiers, including NDC Codes, Generic Product Identifiers, etc. are mapped to standard drug concepts in the Terminology Dictionary. When the Source Drug identifier cannot be translated into Standard Drug Concepts, a Drug exposure entry is stored with only the corresponding Source Drug Code.
* A Drug Exposure Type is assigned to each Drug Exposure, to track the indicator from which the data were drawn or inferred. The Drug Exposure Types are discussed in detail in the DRUG\_EXPOSURE\_REF section of this document.
* Drug Exposures extracted from some of the data sources do not include all required attributes. The following attributes constitute the minimum set required for usable drug exposure data:
  + Patient identifier
  + Source drug identifier/concept
  + Date of Exposure
* Financial details related to the medications are out-of-scope.

### Example of Loaded Table

Consider the following example of inbound source data on medications, from an EHR.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PERSON ID** | **GENERIC PRODUCT IDENTIFIER** | **MEDICATION NAME** | **MEDICATION START DATE** | **MEDICATION END DATE** | **STOP REASON** |
| 121107 | 83200030200313 | Warfarin Sodium Tab 4 MG | 5/9/2003 | 5/9/2003 | Regimen Completed |
| 127260 | 83200030200317 | Warfarin Sodium Tab 6 MG | 4/30/2003 |  |  |
| 127260 | 83200030200317 | Warfarin Sodium Tab 6 MG | 7/27/2003 | 7/27/2003 |  |
| 127260 | 83200030200317 | Warfarin Sodium Tab 6 MG | 8/22/2003 | 8/22/2003 |  |
| 127260 | 83200030200320 | Warfarin Sodium Tab 7.5 MG | 9/7/2003 | 9/7/2003 |  |
| 127260 | 83200030200320 | Warfarin Sodium Tab 7.5 MG | 10/2/2003 | 10/2/2003 | Regimen Completed |

Sample concept code representation of drug data from the Terminology Dictionary follows.

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| 375383004 | Warfarin Sodium Tab 6 MG |
| 375378007 | Warfarin Sodium Tab 7.5 MG |
| 375374009 | Warfarin Sodium Tab 4 MG |

Since the data were drawn from Person Medication lists in the EHR, Drug Exposure type is set to “MEDICAL HISTORY,” which is described in the DRUG\_EXPOSURE\_REF section of this chapter as “Medication History from Electronic Health Records.”

Drug Exposure types are determined, based on the source from which the exposure was recorded, as follows.

|  |  |
| --- | --- |
| **Drug Exposure Type** | **Drug Exposure Type Description** |
| 003 | Medication List |

The above data are represented in the CDM DRUG\_EXPOSURE table as follows.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **DRUG EXPOSURE ID** | **PERSON\_ID** | **DRUG EXPOSURE START DATE** | **DRUG EXPOSURE END DATE** | **DRUG CONCEPT CODE** | **DRUG EXPOSURE TYPE** | **STOP REASON** |
| 1001 | 121107 | 5/9/2003 | 5/9/2003 | 375374009 | 003 | Regimen Completed |
| 1002 | 127260 | 4/30/2003 |  | 375383004 | 003 |  |
| 1003 | 127260 | 7/27/2003 | 7/27/2003 | 375383004 | 003 |  |
| 1004 | 127260 | 8/22/2003 | 8/22/2003 | 375383004 | 003 |  |
| 1005 | 127260 | 9/7/2003 | 9/7/2003 | 375378007 | 003 |  |
| 1006 | 127260 | 10/2/2003 | 10/2/2003 | 375378007 | 003 | Regimen Completed |

## DRUG\_ERA

Drug Era is defined as a span of time when the Person is assumed to be exposed to a particular drug. A Drug Era is not the same as a Drug Exposure; successive periods of Drug Exposure may, under certain rules, be combined to produce one continuous Drug Era. Each drug product name and strength combination is mapped to a separate concept, and exposure to each is treated as a separate Drug Era.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| DRUG\_ERA\_ID | INTEGER | YES | System-generated identifier to uniquely identify each drug era that was constructed based on drug exposure data. |
| PERSON\_ID | INTEGER | YES | System-generated identifier for the Person who is the subject of the Drug Exposure. Foreign key to the PERSON entity. Every Person with a Drug Era must have corresponding demographics data in the PERSON entity. |
| DRUG\_ERA\_START\_DATE | DATE | YES | Start date for the Drug Era constructed based on the individual instances of the Drug Exposure. Defined as the start date of the very first chronologically recorded instance of utilization of a drug. |
| DRUG\_ERA\_END\_DATE | DATE | YES | End date for the Drug Era constructed based on the individual instance of the Drug Exposure. Defined as the end date of the final continuously recorded instance of utilization of a drug. |
| DRUG\_EXPOSURE\_TYPE | VARCHAR(3) | YES | Predefined code for the type of Drug Exposure recorded, such as Drug Era based on Prescription Written, Drug Era based on Procedure Code, Drug Era based on Persistence Window, etc. Allowed Drug Exposure Types are defined in the Drug Exposure Ref section of this chapter. |
| DRUG\_CONCEPT\_CODE | VARCHAR(20) | YES | Standard concept code related to the Drug that is the subject of the study.  The source drug identifier is mapped to a standard drug concept from the Terminology Dictionary and the corresponding concept code is stored here as a reference.  Used to map to standard drug vocabulary and concept hierarchy in the Terminology Dictionary. |
| DRUG\_EXPOSURE\_COUNT | INTEGER | YES | Number of Drug Exposure occurrences used to construct the Drug Era. |

### Business Rules

* For claims related to pharmacy prescriptions, the dispensed date and days supply are used to extrapolate the end date for the period of Drug Exposure. When a person receives recurring prescriptions for the same product and strength, the multiple prescriptions may need to be treated as a single Drug Era. To determine whether this is indeed the case, we must take into account the drug’s “persistence window,” which is the number of days after the Person stops taking a drug, during which the person is deemed to still be affected by the drug. If the number of days between the end date of the prior Drug Exposure and the start date of the subsequent Drug Exposure falls within the persistence window, then the two exposures are considered to belong to the same Drug Era.
* For EHRs, the medications data include the start and stop dates for the medication. The prescription data track only the date on which the medication was prescribed and the date on which the record was created. While the prescription data often include a pointer to identify the corresponding medication record, this is not always the case.
* For a Drug Exposure indicated by procedure codes, usually only a single date is available (i.e., the administration date), which poses a challenge in determining the Drug Era duration.

For example, consider a person who is taking two drugs: Drug A and Drug B. The person has had four prescriptions for Drug A (A1, A2, A3, A4), each with sixty days supply. The person has also had two prescriptions for Drug B (B1, B2). The figure below illustrates the scenario.

Person timeline

Drug A

Drug B

DrugEra1

DrugEra2

DrugEra3

Persistence

window

Persistence

window

A1

A2

A3

A4

B1

B2

To define the Drug Era for Drug A, we consider the timing, duration, overlap, and persistence of the person’s prescriptions for Drug A. A2 was filled before the expected completion of A1. Similarly, A3 was filled before the expected completion of A2. A4 was filled after A3 was completed, but within the persistence window for Drug A. Therefore, we collapse the four prescriptions for Drug A into a single Drug Era (DrugEra1), with the start and end dates equal to the start date for prescription A1 and the end date for prescription A4, respectively.

A significant amount of time elapsed between filling the two prescriptions for Drug B. Because this time exceeded the persistence window for Drug B, we define two distinct Drug Eras for Drug B. The start and end dates for DrugEra2 and DrugEra3 are the start and end dates for prescriptions B1 and B2, respectively.

For the OMOP Research Lab, persistence windows of zero days and thirty days will be used.

### Example of Loaded Table

Consider the following example excerpt from the CDM DRUG\_EXPOSURE table.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **DRUG EXPOSURE ID** | **PERSON\_ID** | **DRUG EXPOSURE START DATE** | **DRUG EXPOSURE END DATE** | **DRUG CONCEPT CODE** | **DRUG EXPOSURE TYPE** | **STOP REASON** |
| 1001 | 121107 | 5/9/2003 | 5/9/2003 | 375374009 | MED HISTORY | Regimen Completed |
| 1002 | 127260 | 4/30/2003 |  | 375383004 | MED HISTORY |  |
| 1003 | 127260 | 7/27/2003 | 7/27/2003 | 375383004 | MED HISTORY |  |
| 1004 | 127260 | 8/22/2003 | 8/22/2003 | 375383004 | MED HISTORY |  |
| 1005 | 127260 | 9/7/2003 | 9/7/2003 | 375378007 | MED HISTORY |  |
| 1006 | 127260 | 10/2/2003 | 10/2/2003 | 375378007 | MED HISTORY | Regimen Completed |

The above example uses the following drug concept codes from a hypothetical Terminology Dictionary.

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| 375383004 | Warfarin Sodium Tab 6 MG |
| 375378007 | Warfarin Sodium Tab 7.5 MG |
| 375374009 | Warfarin Sodium Tab 4 MG |

The drug hierarchy in the Terminology Dictionary indicates that all of the above drug concepts are children of the following high-level drug class concept.

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| 429307006 | Oral form Warfarin (product) |

Drug Exposure types are determined for Drug Eras with a 30 day persistence window as follows.

|  |  |
| --- | --- |
| **Drug Exposure Type** | **Drug Exposure Type Description** |
| 007 | Drug Era, 30 day persistence window |

The Drug Eras constructed from the above data, based on the higher-level drug class concept and using a thirty-day persistence window, would be as follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DRUG ERA ID** | **PERSON ID** | **DRUG CONCEPT CODE** | **DRUG EXPOSURE TYPE** | **DRUG EXPOSURE START DATE** | **DRUG EXPOSURE END DATE** |
| 20001 | 121107 | 429307006 | 007 | 5/9/2003 | 5/9/2003 |
| 20002 | 127260 | 429307006 | 007 | 4/30/2003 | 4/30/2003 |
| 20003 | 127260 | 429307006 | 007 | 7/27/2003 | 10/2/2003 |

## DRUG\_EXPOSURE\_REF

This is a reference listing of various types of Drug Exposures recorded for analysis. The Drug Exposure Type conveys the indicator(s) from which the Drug Exposure was captured, and defines the characteristic of the exposure and the level of aggregation.

The Drug Exposure Types follow.

* Prescription Written (from Electronic Health Records)
* Medication History (from Electronic Health Records)
* Filled Prescription (from Pharmacy Claims)
* Drug from Procedure Code (from Medical Claims)
* Drug Era using a persistence window

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| DRUG\_EXPOSURE\_TYPE | VARCHAR(3) | YES | Predefined code for the type of Drug Exposure recorded. Drug Exposure type is used to define the data source and the type of representation of the Drug utilization recorded. |
| DRUG\_EXPOSURE\_TYPE\_DESC | VARCHAR(120) | YES | Detailed description for the type of drug exposure recorded. |
| PERSISTENCE\_WINDOW | INTEGER | NO | Persistence window used to build the drug era based on drug exposure data.  Persistence window defines the longest time period between two instances of drug utilization for them to be considered as part of the same continuous exposure  Applicable only to Drug Eras. |

### Example of Loaded Table

Reference data for Drug Exposure types are stored in the DRUG\_EXPOSURE\_REF table, which includes the type codes identified and their detailed descriptions. The allowed Drug Exposure types and their descriptions are listed in Appendix A.

## CONDITION\_OCCURRENCE

Condition Occurrences record individual instances of person conditions extracted from source data. Conditions are recorded in various data sources in different forms with varying levels of standardization. For example:

* Medical claims data include ICD-9 diagnosis codes that are part of a claim for health services and procedures.
* EHRs capture person conditions in the form of diagnosis codes and symptoms as part of problem list, but may not have a way to capture out-of-system conditions.
* Death when observed in person status codes (such as discharge status)

Condition Occurrences are analyzed based on standard condition concepts in the Terminology Dictionary, which also maintains the standard condition hierarchy for use in class-based CDM queries.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| CONDITION\_OCCURRENCE\_ID | INTEGER | YES | System-generated identifier to uniquely identify each person condition.  Every Person with a Condition Occurrence, must have corresponding demographics data in the PERSON entity. |
| PERSON\_ID | INTEGER | YES | Unique identifier for the Person for whom the Condition was recorded. Foreign key to the PERSON entity. Every Person with a recorded Condition Occurrence must have a valid entry for Person demographics in the PERSON table. |
| CONDITION\_CONCEPT\_CODE | VARCHAR(20) | NO | Standard concept code related to the Condition.  The condition code and/or description from the source data set is mapped to a standard Condition concept from the Terminology Dictionary and the corresponding concept code is stored here as a reference.  Used to map to standard conditions vocabulary and concept hierarchy in the Terminology Dictionary. |

(continued on the next page)

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| CONDITION\_OCCUR\_TYPE | VARCHAR(3) | YES | Code indicating the type of Condition occurrence.  Condition Occurrence Type is used to define the source data from which the Condition was recorded, the level of standardization, and the type of occurrence. Conditions are defined as Primary/Secondary Diagnosis, Problem Report and Person Status (including Mortality when available in the source data set).  Foreign key to the CONDITION\_OCCURRENCE\_REF table. |
| SOURCE\_CONDITION\_CODE | VARCHAR(20) | YES | Condition Code as captured in the Source data. Values include ICD-9 diagnosis codes from medical claims and EHRs, and Discharge Status/Disposition codes from medical claims. |
| CONDITION\_START\_DATE | DATE | YES | Date when the instance of the Condition was first recorded. |
| CONDITION\_END\_DATE | DATE | NO | Date when the instance of the Condition was last recorded. |
| STOP\_REASON | VARCHAR(20) | NO | Reason, if known, that the condition was no longer recorded, as indicated in the source data. Valid values include reasons such as Discharged, Resolved, etc.. |
| CONDITION\_QUALIFIER | VARCHAR(20) | NO | Indicator Code for the category of Diagnosis as recorded in the source data. Valid values include qualifiers for a Condition, such as Major Diagnosis, Family History of, History of, Hospitalization, Recurrence, Risk of, Rule-Out, etc. |

### Business Rules

The approach to extraction of Condition Occurrence data is based on the individual data source, but the following guidelines are common to all data sources.

* Source attributes mapped to conditions are checked for standardization. If the source attributes are available as standard diagnosis codes (e.g., ICD-9 Diagnosis Codes) or discharge status codes, then they are mapped to standard concepts in the Terminology Dictionary.
* If the source data are not coded to a national or international standard, then a finite listing of attribute values is created and mapped to standard condition concepts in the Terminology Dictionary.
* A Condition Occurrence Type is assigned based on the data source and type of condition attribute, including:
  + ICD9 Primary Diagnosis from Medical Claims
  + ICD9 Secondary Diagnosis from Medical Claims
  + Person Status from Medical Claims
  + Problem concept from EHRs

More details regarding the Condition Occurrence Types appear in the CONDITION\_OCCURRENCE\_REF section of this chapter.

* Each Condition for every Person, along with its matching standard concept code from the Terminology Dictionary, is extracted from the source data along with the person identifier, start/onset date of the condition, end date for condition (where available), and diagnosis qualifier (DX\_QUALIFIER) attributes.
* Special handling is necessary for source data in which person condition entries are updated by expiration of the current entry and addition of an updated entry. In such cases, only the final version of the record is extracted for inclusion in the CDM.

### Example of Loaded Table

Consider the following example of inbound source data on medications, from GE EHRs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PERSON ID** | **ICD-9 DIAGNOSIS CODE** | **PROBLEM DESCRIPTION** | **PROBLEM START DATE** | **PROBLEM END DATE** | **DX**  **QUALIFIER** |
| 127260 | 787.02 | Nausea | 5/3/2003 | 5/3/2003 | Diagnosis Of |
| 127260 | 787.02 | Nausea | 7/29/2003 | 7/29/2003 | Diagnosis Of |
| 127260 | 531.01 | Acute gastric ulcer without hemorrhage or perforation without obstruction | 8/23/2003 | 8/23/2003 | Diagnosis Of |

The following concept codes correspond to the meanings of the problems captured in the source data:

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| GDS999y | Gastric Ulcers |
| GDS999x | Nausea |

Condition Occurrence types are determined for Conditions extracted from EHR Problem list as follows.

|  |  |
| --- | --- |
| **Condition Occurrence Type** | **Condition Occurrence Type Description** |
| 063 | EHR - Problem List |

The CONDITION\_OCCURRENCE, loaded with the above data, would appear as follows.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CONDITION OCCUR-RENCE ID** | **PERSON ID** | **CONDITION CONCEPT CODE** | **SOURCE CONDITION CODE** | **CONDITION OCCUR TYPE** | **CONDITION START DATE** | **CONDITION END DATE** | **DX QUALIFIER** |
| 3003 | 127260 | GDS999x | 787.02 | 063 | 5/3/2003 | 5/3/2003 | Diagnosis Of |
| 3004 | 127260 | GDS999x | 787.02 | 063 | 7/29/2003 | 7/29/2003 | Diagnosis Of |
| 3005 | 127260 | GDS999y | 531.01 | 063 | 8/23/2003 | 8/23/2003 | Diagnosis Of |

## CONDITION\_ERA

Similar to Drug Eras, Condition Eras are chronological periods of Condition Occurrence. Combining individual Condition Occurrences into a single Condition Era serves at least two purposes.

* It allows us to aggregate chronic conditions that require frequent ongoing care, instead of treating each Condition Occurrence as an independent event.
* It allows us to aggregate multiple, closely-timed doctor visits for the same condition to avoid double-counting the Condition Occurrences.

For example, consider a Person who visits the Primary Care Physician (PCP), who diagnoses the Person with a specific condition and refers the Person to a Specialist. One week later, the Person visits the Specialist, who confirms the PCP’s diagnosis and provides the appropriate treatment to resolve the condition with no further care required. These two independent doctor visits should be aggregated into one Condition Era. Just as with Drug Eras, the persistence windows to be used in determining condition eras are also zero days and thirty days.

This model generally fits well for acute conditions, but may be less robust for chronic conditions. For example, chronic conditions that do not require regular follow-up may be recorded as multiple Condition Eras because the absence of data in the periods between visits does not justify the aggregation of the eras. Because our persistence window is small, we are likely to aggregate multiple visits in rapid succession for the same condition, but unlikely to combine infrequent visits for chronic conditions (e.g. a person with Rheumatoid Arthritis who visits his rheumatologist every three months). However, the small window also reduces the likelihood that we will falsely classify independent events into the same Condition Era.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| CONDITION\_ERA\_ID | INTEGER | YES | System-generated identifier to uniquely identify each Condition Era. |
| PERSON\_ID | INTEGER | YES | Unique identifier for the Person for whom the Condition was recorded. Foreign key to the PERSON entity. Every Person with a recorded Condition Era must have a valid entry for Person demographics in the PERSON table. |

(continued on the next page)

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| CONDITION\_CONCEPT\_CODE | VARCHAR(20) | YES | Standard concept code related to the Condition.  The Condition code and/or the description from the source data set is mapped to a standard Condition concept from the Terminology Dictionary and the corresponding concept code is stored here as a reference.  Used to map to standard conditions vocabulary and concept hierarchy in the Terminology Dictionary. |
| CONDITION\_ERA\_START\_DATE | DATE | YES | Staring date for Condition Era. This is the first date on which the earliest constituent Condition occurrence was first recorded. |
| CONDITION\_ERA\_END\_DATE | DATE | YES | Ending date for Condition Era. This is the last date on which the final constituent Condition occurrence was last recorded. |
| CONDITION\_OCCUR\_TYPE | VARCHAR(3) | YES | Code indicating the type of Condition Occurrence.  Condition Occurrence Type is used to define the source data from which the Condition was recorded, the level of standardization, and the type of occurrence. Conditions are defined as Primary/Secondary Diagnosis, Problem Report, and Person Status (including Mortality when available in the source data).  Foreign key to the CONDITION\_OCCURRENCE\_REF table. |
| CONDITION\_OCCURRENCE\_COUNT | INTEGER | YES | Number of Constituent Condition Occurrences in the Condition Era. |
| CONFIDENCE | NUMBER | NO | Degree of confidence based on the source data for condition and the type of condition recorded. |

### Business Rules

A Condition Era represents the span of time for which a person can be considered to have a given condition. An example is illustrated graphically below. Imagine a person who has been diagnosed with two conditions during his insurance coverage period: Condition A and Condition B. The person has been diagnosed with Condition A has four times (A1, A2, A3, A4), and has been diagnosed with Condition B twice (B1, B2).

Person timeline

Condition A

Condition B

ConditionEra1

ConditionEra2

ConditionEra3

Persistence

window

A1

A2

A3

A4

B1

B2

To define condition persistence for Condition A, we look at the timing of successive diagnoses. Here, A2 is within the persistence window of A1. Similarly, A3 is within the persistence window of A2, and A4 is within the persistence window of A3. Thus, the four diagnoses of Condition A should be collapsed into ConditionEra1, with the start date equal to the diagnosis date for A1, and the end date equal to the diagnosis date for A4.

Regarding Condition B, there has elapsed significant time between diagnoses B1 and B2. Therefore, we cannot assume dependence between the diagnoses. Because this time exceeded the persistence window for B1, we define two distinct Condition Eras, one that corresponds to B1, and another that corresponds to B2.

### Example of Loaded Table

Consider the following example excerpt from the CDM’s CONDITON\_OCCURRENCE table.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CONDITION OCCUR-RENCE ID** | **PERSON ID** | **CONDITION CONCEPT CODE** | **SOURCE CONDITION CODE** | **CONDITION OCCUR TYPE** | **CONDITION START DATE** | **CONDITION END DATE** | **DX QUALIFIER** |
| 3003 | 127260 | GDS999x | 787.02 | DIAGNOSIS FROM PROBLEM LIST | 5/3/2003 | 5/3/2003 | Diagnosis Of |
| 3004 | 127260 | GDS999x | 787.02 | DIAGNOSIS… | 7/29/2003 | 7/29/2003 | Diagnosis Of |
| 3005 | 127260 | GDS999y | 787.02 | DIAGNOSIS… | 8/23/2003 | 8/23/2003 | Diagnosis Of |

The above example uses the following condition concept codes from a hypothetical Terminology Dictionary.

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| GDS999y | Gastric Ulcers |
| GDS999x | Nausea |

The ontology within the Terminology Dictionary indicates that the above two concepts are children of the following higher-level condition class concept.

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| GDS9999 | Digestive system disorders |

Condition Occurrence types are determined for Condition Eras, based on a thirty-day persistence window, as follows:

|  |  |
| --- | --- |
| **Condition Occurrence Type** | **Condition Occurrence Type Description** |
| 063 | Condition Era, 30 day persistence window |

The sample Condition Era is constructed from the condition data based on a thirty-day persistence window, indicated by a CONDITION\_OCCURRENCE\_TYPE of “Era, 30-Day Persistence.” The sample representation of the above data in the CONDITION\_ERA table follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CONDITION ERA ID** | **PERSON ID** | **CONDITION CONCEPT CODE** | **COND OCCURRENCE TYPE** | **CONDITION START DATE** | **CONDITION END DATE** |
| 40001 | 121107 | GDS9999 | 063 | 5/3/2003 | 5/3/2003 |
| 40002 | 127260 | GDS9999 | 063 | 7/29/2003 | 8/23/2003 |

## CONDITION\_OCCURRENCE\_REF

The Condition Occurrence Type indicates the indicator(s) from which the Condition Occurrence was drawn or inferred, and indicates whether a condition (diagnosis) was primary or secondary and their relative positioning within a person’s condition record.

A detailed listing of the Condition Occurrence types is included in Appendix B.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| CONDITION\_OCCUR\_TYPE | VARCHAR(3) | YES | Code indicating the type of Condition occurrence.  Condition Occurrence Type is used to define the source data indicators from which the Condition was identified, the level of standardization, and the type of occurrence. Conditions are defined as Primary/Secondary Diagnosis, Problem Report, and Person Status (including Mortality when available in the source data).  Foreign key to the CONDITION\_OCCURRENCE\_REF table. |
| CONDITION\_OCCUR\_TYPE\_DESC | VARCHAR(120) | YES | Detailed description of the Condition Occurrence Type. Valid values include Diagnosis from Inpatient Claims, Diagnosis from Outpatient visits, Diagnosis from ER visits, etc. |
| PERSISTENCE\_WINDOW | INTEGER | NO | Persistence window used to construct the Condition ERA. Persistence windows of 0 days and 30 days are used. |

## VISIT\_OCCURRENCE

The VISIT\_OCCURRENCE entity aggregates all person visits to health care providers, including inpatient, outpatient, and ER visits. Visits are recorded in various data sources in different forms with varying levels of standardization. For example:

* Medical Claims include Inpatient Admissions, Outpatient Services, and Emergency Room visits.
* Electronic Health Records may capture person visits as part of the activities recorded.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| VISIT\_OCCURRENCE\_ID | INTEGER | YES | System-generated identifier to uniquely identify each occurrence of a person’s visit to a health care provider. |
| PERSON\_ID | INTEGER | YES | Unique identifier for the Person for whom the Condition was recorded. Foreign key to the PERSON entity. Every Person with a recorded Condition Era must have a valid entry for Person demographics in the PERSON table. |
| VISIT\_CONCEPT\_CODE | VARCHAR(20) | YES | Standard concept code related to the Visit.  The type of Visit code and/or description from source data is mapped to a standard concept from the Terminology Dictionary and the corresponding concept code is stored here as a reference.  Used to map to standard visit vocabulary and concept hierarchy in the Terminology Dictionary. |
| VISIT\_START\_DATE | DATE | YES | Visit start date. |
| VISIT\_END\_DATE | DATE | NO | Visit end date. |
| SOURCE\_VISIT\_CODE | VARCHAR(20) | YES | Type/source of the visit data. Valid entries include office visits, hospital admissions, etc.  For the OMOP research environment, type-of-service codes and activity type codes can be used as source visit codes. |

### Business Rules

A Visit Occurrence is recorded for each visit to a health care facility. Each visit is standardized by assigning a corresponding concept code based on the type of facility visited and the type of services rendered.

### Example of Loaded Table

Consider the following example visit data extracted from Medical Claims.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PERSON ID** | **TYPE OF SERVICE CODE (BETOS CODES)** | **TYPE OF VISIT** | **VISIT START DATE** | **VISIT END DATE** |
| 127260 | M2A | Hospital Admission | 5/3/2003 | 5/4/2003 |
| 127260 | P9A | Outpatient Dialysis | 7/29/2003 | 7/29/2003 |
| 127260 | M1A | Physician Office Visit | 8/23/2003 | 8/23/2003 |

The following concept codes correspond to the meanings of the types of visits that were indicated in the source data.

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| V-M2A | Hospital Admission |
| V-P9A | Outpatient Dialysis |
| V-M1A | Physician Office Visit |

The above data, represented in the CDM’s VISIT\_OCCURRENCE table, follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **VISIT OCCURRENCE ID** | **PERSON ID** | **VISIT CONCEPT CODE** | **SOURCE VISIT CODE** | **VISIT START DATE** | **VISIT END DATE** |
| 5003 | 127260 | V-M2A | M2A | 5/3/2003 | 5/4/2003 |
| 5004 | 127260 | V-P9A | P9A | 7/29/2003 | 7/29/2003 |
| 5005 | 127260 | V-M1A | M1A | 8/23/2003 | 8/23/2003 |

## PROCEDURE\_OCCURRENCE

Procedure occurrences record individual instances of person procedures extracted from source data. Procedures are recorded in various data sources in different forms with varying levels of standardization. For example:

* Medical Claims include CPT4, ICD-9-CM, and HCPCS procedure codes that are submitted as part of a claim for procedures performed.
* Electronic Health Records that capture CPT, HCPCS procedures as orders.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| PROCEDURE\_OCCURRENCE\_ID | INTEGER | YES | System-generated identifier to uniquely identify each procedure occurrence. |
| PROCEDURE\_CONCEPT\_CODE | VARCHAR(20) | YES | Standard concept code related to the procedure.  The Procedure code and/or description from the source data is mapped to a standard Observation concept from the Terminology Dictionary and the corresponding concept code is stored here as a reference.  Used to map to standard procedures vocabulary and concept hierarchy in the Terminology Dictionary. |
| PERSON\_ID | INTEGER | YES | Unique identifier for the Person for whom the Condition was recorded. Foreign key to the PERSON entity. Every Person with a recorded Condition Era must have a valid entry for Person demographics in the PERSON table. |
| PROCEDURE\_DATE | DATE | YES | Date on which the procedure was performed |

(continued on the next page)

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| PROC\_OCCUR\_TYPE | VARCHAR(3) | YES | Code indicating the type of Procedure Occurrence.  Procedure Occurrence Type is used to define the source data from which the Procedure was recorded, and the type of occurrence. Please see appendix for all Procedure occurrence types that are recorded.  Foreign key to the PROC\_OCCURRENCE\_REF table. |
| SOURCE\_PROCEDURE\_CODE | VARCHAR(6) | YES | Procedure Code as captured in the Source data. Values include CPT4, ICD-9-CM, HCPCS, and other procedure codes. |

### Business Rules

Procedure Occurrences are recorded for each procedure performed on a person. Each procedure is standardized by assigning a concept code corresponding to the meaning of the procedure code and code type used.

### Example of Loaded Table

Consider the following example procedure data extracted from Electronic Health Records.

|  |  |  |  |
| --- | --- | --- | --- |
| **PERSON ID** | **CPT Code** | **PROBLEM DESCRIPTION** | **PROCEDURE DATE** |
| 127260 | 71020 | Chest X-Ray | 5/3/2003 |
| 127260 | 93925 | Lower Extremity Arterial Duplex, Bilateral | 7/29/2003 |
| 127260 | 72110 | X-ray exam of lower spine | 8/23/2003 |

The following concept codes correspond to the meanings of the procedures that were captured in the source data:

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| ABC456 | Chest X-Ray |
| DEF457 | Arterial Duplex Ultrasound Study |
| IJK458 | X-ray exam of lower spine |

Procedure Occurrence types are determined for Procedures extracted from the EHR Order list as follows.

|  |  |
| --- | --- |
| **Procedure Occurrence Type** | **Procedure Occurrence Type Description** |
| 027 | EHR - Order List |

The above data, represented in the CDM’s PROCEDURE\_OCCURRENCE table, follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PROCEDURE OCCURRENCE ID** | **PERSON ID** | **PROCEDURE OCCURRENCE TYPE** | **PROCEDURE CONCEPT CODE** | **SOURCE PROCEDURE CODE** | **PROCEDURE DATE** |
| 5003 | 127260 | 027 | ABC456 | 71020 | 5/3/2003 |
| 5004 | 127260 | 027 | DEF457 | 93925 | 7/29/2003 |
| 5005 | 127260 | 027 | IJK458 | 72110 | 8/23/2003 |

## PROCEDURE\_OCCURRENCE\_REF

The Procedure Occurrence Type defines the indicators from which the Procedure Occurrence was drawn or inferred, and indicates whether a Procedure was primary or secondary and their relative positioning within a Patient Procedure record.

A detailed listing of the Procedure Occurrence type codes and their associated descriptions are listed in Appendix B.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| PROC\_OCCUR\_TYPE | VARCHAR(3) | YES | Code indicating the type of Procedure occurrence.  Procedure Occurrence Type is used to define the source data indicators from which the Procedure was identified, the level of standardization, and the type of occurrence.  A detailed listing of the Procedure Occurrence Types is recorded in Appendix C.  Foreign key to the PROC\_OCCURRENCE\_REF table. |
| PROC\_OCCUR\_TYPE\_DESC | VARCHAR(120) | YES | Detailed description of the Procedure Occurrence Type. Valid values include Procedures from Inpatient Claims, Procedures from Outpatient visits etc. |

## OBSERVATION

The Observation entity aggregates all general observations from the following categories:

* Lab observations (i.e., test results) from Medical Claims
* Lab and other observations from Electronic Health Records
* Function tests performed on patients from Electronic Health Records
* A person’s chief complaint as captured in Electronic Health Records
* Other observations from various data sources that cannot be otherwise categorized within the domains provided (Drug, Condition, Procedure, Visit)

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| OBSERVATION\_OCCURRENCE\_ID | INTEGER | YES | System-generated identifier to uniquely identify each observation occurrence. |
| PERSON\_ID | INTEGER | YES | Unique identifier for the Person for whom the Condition was recorded. Foreign key to the PERSON entity. Every Person with a recorded Condition Era must have a valid entry for Person demographics in the PERSON table. |
| SOURCE\_OBSERVATION\_CODE | VARCHAR(20) | YES | Observation code as it appears in the source data. |
| OBSERVATION\_CONCEPT\_CODE | VARCHAR(20) | YES | Standard concept code related to the Observation.  The type of Observation from the source data set is mapped to a standard Observation concept from the Terminology Dictionary and the corresponding concept code is stored here as a reference.  Used to map to standard observations vocabulary and concept hierarchy in the Terminology Dictionary. |
| OBSERVATION\_TYPE | VARCHAR(3) | YES | Code for the type of observation that was recorded. |
| OBS\_VALUE\_AS\_NUMBER | NUMBER(11,3) | NO | Observation result stored as a numeric value. Applicable to observations where the result is expressed as a numeric value. |
| OBS\_VALUE\_AS\_STRING | VARCHAR(60) | NO | Observation result stored as character string. Applicable to observations where the result is expressed as a character string. |

(continued on the next page)

|  |  |  |  |
| --- | --- | --- | --- |
| Field | Data Type | Required | Description and Notes |
| OBS\_VALUE\_AS\_CONCEPT\_CODE | VARCHAR(20) | NO | Observation result stored as Concept Code.  Applicable to observations where the result can be expressed as a standard concept from the Terminology Dictionary. Usually these are the results of observations that have a relatively small number of discrete allowed values (e.g., positive/negative, present/absent, low/high, etc.) |
| OBSERVATION\_DATE | DATE | YES | Date of the Observation |
| OBS\_UNITS\_CONCEPT\_CODE | VARCHAR(20) | NO | Unit of measure used for observation result when measured as a numeric value. The units are stored as a concept code from the Terminology Dictionary.. |
| OBS\_RANGE\_LOW | NUMBER(11,3) | NO | Lower limit of the numeric range of the observation value. Not applicable if the observation results are non-numeric and categorical. Must be in the same units of measure as the observation value. |
| OBS\_RANGE\_HIGH | NUMBER(11,3) | NO | High limit of the numeric range of the observation value. Not applicable if the observation results are non-numeric and categorical. Must be in the same units of measure as the observation value. |

### Business Rules

The approach to extraction and representation of Observation data are based on the individual data source, but the following guidelines are common to all data sources.

* Source attribute values mapped to Observations are checked for standardization. If the source attribute values are available as national or international standard codes (e.g., LOINC codes), then they are mapped to standard concepts in the Terminology Dictionary.
* If the source data are not coded to a national or international standard, then a finite listing of attribute values is created and mapped to standard Observation concepts in the Terminology Dictionary.
* The type of result recorded for the Observation is important for further processing of the Observation data. Knowledge of the whether an Observation result is captured as a numeric value (with the range of values considered normal), standard concept code, or non-standard text will inform the handling of the Observation data.
* An Observation Type is assigned based on the type of source data from which the Observation was extracted and type of result expected. More details regarding the Observation types appear in the OBSERVATION\_TYPE\_REF section of this chapter.
* Each Observation for every Person, along with its matching standard concept code from the Terminology Dictionary, is extracted from the source data along with the person identifier. Also extracted are related attributes including date of the Observation, type of observation, type of result, result as a number/text/concept code and reference range for numeric results.

Special handling is necessary for source data in which person condition entries are updated by expiration of the current entry and addition of an updated entry. In such cases, only the final version of the record is extracted for transformation and loading into the CDM.

### Example of Loaded Table

Consider the following example Observation data extracted from Lab Claims and Observations.

|  |  |  |  |
| --- | --- | --- | --- |
| **PERSON ID** | **OBSERVATION CODE** | **OBSERVATION DESCRIPTION** | **OBSERVATION DATE** |
| 127260 | 13457-7 | LDL CHOLESTEROL | 5/3/2003 |
| 127260 | 6690-2 | White Blood Count | 7/29/2003 |
| 127260 | 7332 | Smoking Status | 8/23/2003 |

The following concept codes correspond to the observations that were captured in the source data.

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| LRS101 | Lipid Panel – LDL Check |
| LRS201 | White Blood Count Check |
| LRS301 | Smoking Status |

The following concept codes correspond to the meanings of the units of measure associated with the lab observations in the source data:

|  |  |
| --- | --- |
| **Concept Code** | **Concept Description** |
| LRU101 | Lipid Panel – LDL Check |
| LRU201 | White Blood Count Check |

The above data, represented in the CDM’s OBSERVATION table, follows.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBS  OCCUR-RENCE  ID** | **PERSON ID** | **OBS CONCEPT CODE** | **OBS SOURCE ID** | **OBS DATE** | **OBS TYPE** | **OBS VALUE AS NUMBER** | **OBS VALUE AS STRING** | **OBS UNITS** | **OBS RANGE HIGH** | **OBS RANGE LOW** |
| 5003 | 127260 | LRS101 | 13457-7 | 5/3/2003 | LON | 124 |  | LRU101 | 130 | 0 |
| 5004 | 127260 | LRS201 | 6690-2 | 7/29/2003 | LON | 6000 |  | LRU201 | 10000 | 4500 |
| 5005 | 127260 | LRS301 | 7332 | 8/23/2003 | LOT |  | PASSIVE SMOKER |  |  |  |

## OBSERVATION\_TYPE

Assignment of an Observation type is essential to determine the type of source data, level of standardization, and coding, as well as the type of result recorded for the observation. The Observation Types include the following.

* Lab Observation Numeric Result
* Lab Observation Text
* Lab Observation Concept Code Result
* Observations from EHRs. These are tracked separately and not rolled into other Lab Observation categories
* EHR observations with text results
* Chief Complaint

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| OBSERVATION\_TYPE | VARCHAR(3) | YES | Code representing the type of observation that was recorded. |
| OBSERVATION\_TYPE\_DESC | VARCHAR(255) | YES | Detailed description of the type of observation recorded. |

### Example of Loaded Table

The Observation Types identified so far follow.

|  |  |
| --- | --- |
| **OBSERVATION\_TYPE** | **OBSERVATION\_TYPE\_DESC** |
| LON | Lab Observation Numeric Result |
| LOT | Lab Observation Text |
| LOC | Lab Observation Concept Code Result |
| EHR | Observation recorded from Electronic Health Records |
| TEM | Observation recorded from Electronic Health Records with text results |
| CHC | Chief Complaint |

## OBSERVATION\_PERIOD

The Observation Period entity is designed to track person status over time. Understanding of the availability of Person data during the Observation Period helps isolate time periods when reliable data for a Person was not available in the data source being used for the analysis. Person Status (Characterized as Covered/Not Covered /Unknown) is mapped to the corresponding concept from the Terminology Dictionary and the corresponding concept code is stored for each status entry. Observation Period entity is also used to track Prescription/Medication coverage availability for the person during the period of study.

Knowing the Person Status during a period of time under study would help refine active drug surveillance by accounting for changes in the person’s medical coverage data availability. However, not all Person Status details are available from all data sources.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Required** | **Description and Notes** |
| OBSERVATION\_PERIOD\_ID | INTEGER | YES | System-generated identifier to uniquely identify each Observation Period. |
| PERSON\_ID | INTEGER | YES | Unique identifier for the Person for whom the Condition was recorded. Foreign key to the PERSON entity. Every Person with a recorded Condition Era must have a valid entry for Person demographics in the PERSON table. |
| OBSERVATION\_START\_DATE | DATE | YES | Start date of the Observation Period for which person history data is available from the data provider. |
| OBSERVATION\_END\_DATE | DATE | YES | End date of the Observation Period for which person history data is available from the data provider. |
| PERSON\_STATUS\_CONCEPT\_CODE | VARCHAR(20) | NO | Clinical status of the person. Valid Values include Active, Diseased, Unknown.  The status is stored as a concept code from the Terminology Dictionary. |
| DRUG\_COVERAGE\_INDICATOR | VARCHAR(1) | NO | Flag to indicate whether medication/prescription coverage was available for the patient during the time period covered by the study.  The valid values are as follows:  Y: Drug Coverage available  N: Drug Coverage  U: Unknown  These data are not available from all sources and is recorded where available. |

### Business Rules

Tracking Person Status during an Observation Period requires unique handling for each raw data source from which Person data are extracted.

* The status of a Person determines whether the health-related data of a person was recorded during the Observation Period.
* Each patient can have many Observation Periods.
* For data sources in which the status of a Person for each calendar month or year is recorded as a separate entry, even if there are no changes, a single consolidated Person Status entry is recorded in the Common Data Model.
* Medication/prescription coverage is tracked for the period of study for data sources in which it is applicable.

### Example of Loaded Table

Consider the following example data extracted from Medical Claims.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PERSON ID** | **PERSON STATUS** | **RX COVERAGE INDICATOR** | **START DATE** | **END DATE** |
| 127260 | Active | Available | 1/1/2003 | 9/30/2003 |
| 127260 | Active | Available | 10/1/2003 | 12/31/2003 |
| 127260 | Not-Eligible | Not Available | 1/1/2004 | 12/31/2004 |

The following concept codes correspond to the meanings of the Person Status values that were present in the source data.

|  |  |
| --- | --- |
| **Person Status Concept Code** | **Person Status Concept Code Description** |
| PS101 | Person Active |
| PS102 | Person Not-Eligible |

The above data are consolidated and represented in the OBSERVATION\_PERIOD table as follows.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **OBSERVATION PERIOD ID** | **PERSON ID** | **PERSON STATUS CONCEPT CODE** | **OBSERVATION START DATE** | **OBSERVATION END DATE** |
| 80001 | 127260 | PS101 | 1/1/2003 | 12/31/2003 |
| 80003 | 127260 | PS102 | 1/1/2004 | 12/31/2004 |

# Appendix A: Drug Exposure Type Codes

Drug Exposure types are used to define the indicators from which exposures have been extracted. They also define the characteristics of the exposure and the level of aggregation. The following Drug Exposure Types are allowed.

|  |  |  |
| --- | --- | --- |
| **DRUG EXPOSURE TYPE** | **DRUG EXPOSURE TYPE DESCRIPTION** | **PERSISTENCE WINDOW**  **(in days)** |
| 1 | Prescription Dispensed |  |
| 2 | Prescription Written |  |
| 3 | Medication List |  |
| 4 | Physician Administered Drug (Identified as Procedure) |  |
| 5 | Inpatient Administration |  |
| 6 | Drug Era – 0 day window | 0 |
| 7 | Drug Era – 30 days window | 30 |

# Appendix B: Condition Occurrence Type Codes

The Condition Occurrence Type indicates the indicator(s) from which the Condition Occurrence was drawn or inferred, and indicates whether a condition (diagnosis) was primary or secondary and their relative positioning within a patient condition record.

The following Condition Occurrence Types are allowed.

|  |  |  |  |
| --- | --- | --- | --- |
| **CONDITION OCCURRENCE TYPE** | **CONDITION OCCURRENCE TYPE DESCRIPTION** | **CONDITION OCCURRENCE POSITION** | **PERSISTENCE WINDOW**  **(in days)** |
| 1 | Inpatient Detail | Primary |  |
| 2 | Inpatient Detail | 1 |  |
| 3 | Inpatient Detail | 2 |  |
| 4 | Inpatient Detail | 3 |  |
| 5 | Inpatient Detail | 4 |  |
| 6 | Inpatient Detail | 5 |  |
| 7 | Inpatient Detail | 6 |  |
| 8 | Inpatient Detail | 7 |  |
| 9 | Inpatient Detail | 8 |  |
| 10 | Inpatient Detail | 9 |  |
| 11 | Inpatient Detail | 10 |  |
| 12 | Inpatient Detail | 11 |  |
| 13 | Inpatient Detail | 12 |  |
| 14 | Inpatient Detail | 13 |  |
| 15 | Inpatient Detail | 14 |  |
| 16 | Inpatient Detail | 15 |  |
| 17 | Inpatient Header | Primary |  |
| 18 | Inpatient Header | 1 |  |
| 19 | Inpatient Header | 2 |  |
| 20 | Inpatient Header | 3 |  |
| 21 | Inpatient Header | 4 |  |
| 22 | Inpatient Header | 5 |  |
| 23 | Inpatient Header | 6 |  |
| 24 | Inpatient Header | 7 |  |
| 25 | Inpatient Header | 8 |  |
| 26 | Inpatient Header | 9 |  |
| 27 | Inpatient Header | 10 |  |
| 28 | Inpatient Header | 11 |  |
| 29 | Inpatient Header | 12 |  |
| 30 | Inpatient Header | 13 |  |

(continued on the next page)

|  |  |  |  |
| --- | --- | --- | --- |
| **CONDITION OCCURRENCE TYPE** | **CONDITION OCCURRENCE TYPE DESCRIPTION** | **CONDITION OCCURRENCE POSITION** | **PERSISTENCE WINDOW**  **(in days)** |
| 31 | Inpatient Header | 14 |  |
| 32 | Inpatient Header | 15 |  |
| 33 | Outpatient Detail | 1 |  |
| 34 | Outpatient Detail | 2 |  |
| 35 | Outpatient Detail | 3 |  |
| 36 | Outpatient Detail | 4 |  |
| 37 | Outpatient Detail | 5 |  |
| 38 | Outpatient Detail | 6 |  |
| 39 | Outpatient Detail | 7 |  |
| 40 | Outpatient Detail | 8 |  |
| 41 | Outpatient Detail | 9 |  |
| 42 | Outpatient Detail | 10 |  |
| 43 | Outpatient Detail | 11 |  |
| 44 | Outpatient Detail | 12 |  |
| 45 | Outpatient Detail | 13 |  |
| 46 | Outpatient Detail | 14 |  |
| 47 | Outpatient Detail | 15 |  |
| 48 | Outpatient Header | 1 |  |
| 49 | Outpatient Header | 2 |  |
| 50 | Outpatient Header | 3 |  |
| 51 | Outpatient Header | 4 |  |
| 52 | Outpatient Header | 5 |  |
| 53 | Outpatient Header | 6 |  |
| 54 | Outpatient Header | 7 |  |
| 55 | Outpatient Header | 8 |  |
| 56 | Outpatient Header | 9 |  |
| 57 | Outpatient Header | 10 |  |
| 58 | Outpatient Header | 11 |  |
| 59 | Outpatient Header | 12 |  |
| 60 | Outpatient Header | 13 |  |
| 61 | Outpatient Header | 14 |  |
| 62 | Outpatient Header | 15 |  |
| 63 | Problem List |  |  |
| 64 | Condition Era |  | 0 |
| 65 | Condition Era |  | 30 |
| 66 | Death at Discharge |  |  |

# Appendix C: Procedure Occurrence Type Codes

The Procedure Occurrence Type defines the indicators from which the Procedure Occurrence was drawn or inferred, and indicates whether a Procedure was primary or secondary and their relative positioning within a Patient Procedure record.

The following Procedure Occurrence Types are allowed.

|  |  |  |
| --- | --- | --- |
| **PROCEDURE OCCURRENCE TYPE** | **PROCEDURE OCCURRENCE TYPE DESCRIPTION** | **PROCEDURE OCCURRENCE POSITION** |
| 1 | Inpatient Detail | Primary |
| 2 | Inpatient Detail | 1 |
| 3 | Inpatient Header | Primary |
| 4 | Inpatient Header | 1 |
| 5 | Inpatient Header | 2 |
| 6 | Inpatient Header | 3 |
| 7 | Inpatient Header | 4 |
| 8 | Inpatient Header | 5 |
| 9 | Inpatient Header | 6 |
| 10 | Inpatient Header | 7 |
| 11 | Inpatient Header | 8 |
| 12 | Inpatient Header | 9 |
| 13 | Inpatient Header | 10 |
| 14 | Inpatient Header | 11 |
| 15 | Inpatient Header | 12 |
| 16 | Inpatient Header | 13 |
| 17 | Inpatient Header | 14 |
| 18 | Inpatient Header | 15 |
| 19 | Outpatient Detail | Primary |
| 20 | Outpatient Detail | 1 |
| 21 | Outpatient Header | Primary |
| 22 | Outpatient Header | 1 |
| 23 | Outpatient Header | 2 |
| 24 | Outpatient Header | 3 |
| 25 | Outpatient Header | 4 |
| 26 | Outpatient Header | 5 |
| 27 | Outpatient Header | 6 |
| 28 | EHR Order |  |