Data Model Conventions

There are a number of implicit and explicit conventions that have been adopted in the CDM. Developers of methods that run against the CDM need to understand these conventions.

General

The OMOP CDM is platform-independent. Data types are defined generically using ANSI SQL data types (VARCHAR, INTEGER, FLOAT, DATE, DATETIME, CLOB). Precision is provided only for VARCHAR. It reflects the minimal required string length and can be expanded within a CDM instantiation. The CDM does not prescribe the date and datetime format. Standard queries against CDM may vary for local instantiations and date/datetime configurations.

Tables

For the tables of the main domains of the CDM it is imperative that concepts used are strictly limited to the domain. For example, the CONDITION_OCCURRENCE table contains only information about conditions (diagnoses, signs, symptoms), but no information about procedures. Not all source coding schemes adhere to such rules. For example, ICD-9-CM codes, which contain mostly diagnoses of human disease, also contain information about the status of patients having received a procedure. The ICD-9-CM code V20.3 'Newborn health supervision' defines a continuous procedure and is therefore stored in the PROCEDURE_OCCURRENCE table.

Fields

Variable names across all tables follow one convention:

Notation	Description
_SOURCE_VALUE	Verbatim information from the source data, typically used in ETL to map to CONCEPT_ID, and not to be used by any standard analytics. For example, CONDITION_SOURCE_VALUE = '787.02' was the ICD-9 code captured as a diagnosis from the administrative claim.
_ID	Unique identifiers for key entities, which can serve as foreign keys to establish relationships across entities. For example, PERSON_ID uniquely identifies each individual. VISIT_OCCURRENCE_ID uniquely identifies a PERSON encounter at a point of care.
_CONCEPT_ID	Foreign key into the Standardized Vocabularies (i.e. the standard_concept attribute for the corresponding term is true), which serves as the primary basis for all standardized analytics. For example, CONDITION_CONCEPT_ID = 31967 contains the reference value for the SNOMED concept of 'Nausea'
_SOURCE_CONCEPT _ID	Foreign key into the Standardized Vocabularies representing the concept and terminology used in the source data, when applicable. For example, CONDITION_SOURCE_CONCEPT_ID = 45431665 denotes the concept of 'Nausea' in the Read terminology; the analogous CONDITION_CONCEPT_ID might be 31967, since SNOMED-CT is the Standardized Vocabulary for most clinical diagnoses and findings.

_TYPE_CONCEPT_ID	Delineates the origin of the source information, standardized within the Standardized Vocabularies. For example, DRUG_TYPE_CONCEPT_ID can allow analysts to discriminate between 'Pharmacy dispensing' and 'Prescription written'

Vocabulary

Concepts

Concepts in the Common Data Model are derived from a number of public or proprietary terminologies such as SNOMED-CT and RxNorm, or custom generated to standardize aspects of observational data. Both types of Concepts are integrated based on the following rules:

- All Concepts are maintained centrally by the CDM and Vocabularies Working Group. Additional concepts can be added, as needed, upon request by creating a Github issue.
- For all Concepts, whether they are custom generated or adopted from published terminologies, a unique numeric identifier concept_id is assigned and used as the key to link all observational data to the corresponding Concept reference data.
- The concept_id of a Concept is persistent, i.e. stays the same for the same Concept between releases of the Standardized Vocabularies.
- A descriptive name for each Concept is stored as the Concept Name as part of the CONCEPT table. Additional names and descriptions for the Concept are stored as Synonyms in the CONCEPT_SYNONYM table.
- Each Concept is assigned to a Domain. For Standard Concepts, there is always a single Domain. Source Concepts can be composite or coordinated entities, and therefore can belong to more than one Domain. The domain_id field of the record contains the abbreviation of the Domain, or Domain combination. Please refer to the Standardized Vocabularies specification for details of the Domain Assignment.
- Concept Class designations are attributes of Concepts. Each Vocabulary has its own set of permissible Concept Classes, although the same Concept Class can be used by more than one Vocabulary. Depending on the Vocabulary, the Concept Class may categorize Concepts vertically (parallel) or horizontally (hierarchically). See the specification of each vocabulary for details.

- Concept Class attributes should not be confused with Classification Concepts. These are separate Concepts that have a hierarchical relationship to Standard Concepts or each other, while Concept Classes are unique Vocabulary-specific attributes for each Concept.
- For Concepts inherited from published terminologies, the source code is retained in the concept_code field and can be used to reference the source vocabulary.
- Standard Concepts (designated as 'S' in the standard_concept field) may appear in CDM tables in all *_concept_id fields, whereas Classification Concepts ('C') should not appear in the CDM data, but participate in the construction of the CONCEPT_ANCESTOR table and can be used to identify Descendants that may appear in the data. See CONCEPT_ANCESTOR table. Non-standard Concepts can only appear in *_source_concept_id fields and are not used in CONCEPT_ANCESTOR table. Please refer to the Standardized Vocabularies specifications for details of the Standard Concept designation.
- The lifespan of a Concept is recorded through its valid_start_date, valid_end_date and the invalid_ reason fields. This allows Concepts to correctly reflect at which point in time were defined. Usually, Concepts get deprecated if their meaning was deemed ambiguous, a duplication of another Concept, or needed revision for scientific reason. For example, drug ingredients get updated when different salt or isomer variants enter the market. Usually, drugs taken off the market do not cause a deprecation by the terminology vendor. Since observational data are valid with respect to the time they are recorded, it is key for the Standardized Vocabularies to provide even obsolete codes and maintain their relationships to other current Concepts .
- Concepts without a known instantiated date are assigned valid_start_date of '1-Jan-1970'.
- Concepts that are not invalid are assigned valid_end_date of '31-Dec-2099'.
- Deprecated Concepts (with a valid_end_date before the release date of the Standardized Vocabularies) will have a value of 'D' (deprecated without successor) or 'U' (updated). The updated Concepts have a record in the CONCEPT_RELATIONSHIP table indicating their active replacement Concept.
- Values for concept_ids generated as part of Standardized Vocabularies will be reserved from 0 to 2,000,000,000. Above this range, concept_ids are available for local use and are guaranteed not to clash with future releases of the Standardized Vocabularies.

Vocabularies

- There is one record for each Vocabulary. One Vocabulary source or vendor can issue several Vocabularies, each of them creating their own record in the VOCABULARY table. However, the choice of whether a Vocabulary contains Concepts of different Concept Classes, or when these different classes constitute separate Vocabularies cannot precisely be decided based on the definition of what constitutes a Vocabulary. For example, the ICD-9 Volume 1 and 2 codes (ICD9CM, containing predominantly conditions and some procedures and observations) and the ICD-9 Volume 3 codes (ICD9Proc, containing predominantly procedures) are realized as two different Vocabularies. On the other hand, SNOMED-CT codes of the class Condition and those of the class Procedure are part of one and the same Vocabulary. Please refer to the Standardized Vocabularies specifications for details of each Vocabulary.
- The vocabulary_id field contains an alphanumerical identifier, that can also be used as the abbreviation of the Vocabulary name.
- The record with vocabulary_id = 'None' is reserved to contain information regarding the current version of the Entire Standardized Vocabularies.
- The vocabulary_name field contains the full official name of the Vocabulary, as well as the source or vendor in parenthesis.
- Each Vocabulary has an entry in the CONCEPT table, which is recorded in the vocabulary_concept_id field. This is for purposes of creating a closed Information Model, where all entities in the OMOP CDM are covered by a unique Concept.

Domains

- There is one record for each Domain. The domains are defined by the tables and fields in the OMOP CDM that can contain Concepts describing all the various aspects of the healthcare experience of a patient.
- The domain_id field contains an alphanumerical identifier, that can also be used as the abbreviation of the Domain.
- The domain_name field contains the unabbreviated names of the Domain.
- Each Domain also has an entry in the Concept table, which is recorded in the domain_concept_id field. This is for purposes of creating a closed Information Model, where all entities in the OMOP CDM are covered by unique Concept.
- Versions prior to v5.0.0 of the OMOP CDM did not support the notion of a Domain.

Concept Classes

- There is one record for each Concept Class. Concept Classes are used to create additional structure to the Concepts within each Vocabulary. Some Concept Classes are unique to a Vocabulary (for example "Clinical Finding" in SNOMED), but others can be used across different Vocabularies. The separation of Concepts through Concept Classes can be semantically horizontal (each Class subsumes Concepts of the same hierarchical level, akin to sub-Vocabularies within a Vocabulary) or vertical (each Class subsumes Concepts of a certain kind, going across hierarchical levels). For example, Concept Classes in SNOMED are vertical: The classes "Procedure" and "Clinical Finding" define very granular to very generic Concepts. On the other hand, "Clinical Drug" and "Ingredient" Concept Classes define horizontal layers or strata in the RxNorm vocabulary, which all belong to the same concept of a Drug.
- The concept_class_id field contains an alphanumerical identifier, that can also be used as the abbreviation of the Concept Class.
- The concept_class_name field contains the unabbreviated names of the Concept Class.
- Each Concept Class also has an entry in the Concept table, which is recorded in the concept_ class_concept_id field. This is for purposes of creating a closed Information Model, where all entities in the OMOP CDM are covered by unique Concepts.

Concept Relationships

- Relationships can generally be classified as hierarchical (parent-child) or non-hierarchical (lateral).
- All Relationships are directional, and each Concept Relationship is represented twice symmetrically within the CONCEPT_RELATIONSHIP table. For example, the two SNOMED concepts of 'Acute myocardial infarction of the anterior wall' and 'Acute myocardial infarction' have two Concept Relationships:
  - 'Acute myocardial infarction of the anterior wall' 'Is a' 'Acute myocardial infarction', and
  - 'Acute myocardial infarction' 'Subsumes' 'Acute myocardial infarction of the anterior wall'.
- There is one record for each Concept Relationship connecting the same Concepts with the same relationship_id.
- Since all Concept Relationships exist with their mirror image (concept_id_1 and concept_id_2 swapped, and the relationship_id replaced by the reverse_relationship_id from the RELATIONSHIP table), it is not necessary to query for the existence of a relationship both in the concept_id_1 and concept_id_2 fields.

- Concept Relationships define direct relationships between Concepts. Indirect relationships through 3rd Concepts are not captured in this table. However, the CONCEPT_ANCESTOR table does this for hierarchical relationships over several "generations" of direct relationships.
- In previous versions of the CDM, the relationship_id used to be a numerical identifier. See the RELATIONSHIP table.

Relationship Table

- There is one record for each Relationship.
- Relationships are classified as hierarchical (parent-child) or non-hierarchical (lateral)
- They are used to determine which concept relationship records should be included in the computation of the CONCEPT_ANCESTOR table.
- The relationship_id field contains an alphanumerical identifier, that can also be used as the abbreviation of the Relationship.
- The relationship_name field contains the unabbreviated names of the Relationship.
- Relationships all exist symmetrically, i.e. in both direction. The relationship_id of the opposite Relationship is provided in the reverse_relationship_id field.
- Each Relationship also has an equivalent entry in the Concept table, which is recorded in the relationship_ concept_id field. This is for purposes of creating a closed Information Model, where all entities in the OMOP CDM are covered by unique Concepts.
- Hierarchical Relationships are used to build a hierarchical tree out of the Concepts, which is recorded in the CONCEPT_ANCESTOR table. For example, "has_ingredient" is a Relationship between Concept of the Concept Class 'Clinical Drug' and those of 'Ingredient', and all Ingredients can be classified as the "parental" hierarchical Concepts for the drug products they are part of. All 'Is a' Relationships are hierarchical.
- Relationships, also hierarchical, can be between Concepts within the same Vocabulary or those adopted from different Vocabulary sources.

Concept Synonyms

- The concept_synonym_name field contains a valid Synonym of a concept, including the description in the concept_name itself. I.e. each Concept has at least one Synonym in the CONCEPT_SYNONYM table. As an example, for a SNOMED-CT Concept, if the fully specified name is stored as the concept_name of the CONCEPT

table, then the Preferred Term and Synonyms associated with the Concept are stored in the CONCEPT_SYNONYM table.

- Only Synonyms that are active and current are stored in the CONCEPT_SYNONYM table. Tracking synonym/description history and mapping of obsolete synonyms to current Concepts/Synonyms is out of scope for the Standard Vocabularies.
- Currently, only English Synonyms are included.

Concept Ancestor

- Each concept is also recorded as an ancestor of itself.
- Only valid and Standard Concepts participate in the CONCEPT_ANCESTOR table. It is not possible to find ancestors or descendants of deprecated or Source Concepts.
- Usually, only Concepts of the same Domain are connected through records of the CONCEPT_ANCESTOR table, but there might be exceptions.

Source to Concept Map

- This table is no longer used to distribute mapping information between source codes and Standard Concepts for the Standard Vocabularies. Instead, the CONCEPT_RELATIONSHIP table is used for this purpose, using the relationship_id='Maps to'.
- However, this table can still be used for the translation of local source codes into Standard Concepts.
- **Note:** This table should not be used to translate source codes to Source Concepts. The source code of a Source Concept is captured in its concept_code field. If the source codes used in a given database do not follow correct formatting the ETL will have to perform this translation. For example, if ICD-9-CM codes are recorded without a dot the ETL will have to perform a lookup function that allows identifying the correct ICD-9-CM Source Concept (with the dot in the concept_code field).
- The source_concept_id, or the combination of the fields source_code and the source_vocabulary_id uniquely identifies the source information. It is the equivalent to the concept_id_1 field in the CONCEPT_RELATIONSHIP table.
- If there is no source_concept_id available because the source codes are local and not supported by the Standard Vocabulary, the content of the field is 0 (zero, not null) encoding an undefined concept. However, local Source Concepts are established (concept_id values above 2,000,000,000).

- The source_code_description contains an optional description of the source code.
- The target_concept_id contains the Concept the source code is mapped to. It is equivalent to the concept_id_2 in the CONCEPT_RELATIONSHIP table
- The target_vocabulary_id field contains the vocabulary_id of the target concept. It is a duplication of the same information in the CONCEPT record of the Target Concept.
- The fields valid_start_date, valid_end_date and invalid_reason are used to define the life cycle of the mapping information. Invalid mapping records should not be used for mapping information.

Drug Strength

- The DRUG_STRENGTH table contains information for each active (non-deprecated) Standard Drug Concept.
- A drug which contains multiple active Ingredients will result in multiple DRUG_STRENGTH records, one for each active ingredient.
- Ingredient strength information is provided either as absolute amount (usually for solid formulations) or as concentration (usually for liquid formulations).
- If the absolute amount is provided (for example, 'Acetaminophen 5 MG Tablet') the amount_value and amount_unit_concept_id are used to define this content (in this case 5 and 'MG').
- If the concentration is provided (for example 'Acetaminophen 48 MG/ML Oral Solution') the numerator_ value in combination with the numerator_unit_concept_id and denominator_unit_concept_id are used to define this content (in this case 48, 'MG' and 'ML').
- In case of Quantified Clinical or Branded Drugs the denominator_value contains the total amount of the solution (not the amount of the ingredient). In all other drug concept classes the denominator amount is NULL because the concentration is always normalized to the unit of the denominator. So, a product containing 960 mg in 20 mL is provided as 48 mg/mL in the Clinical Drug and Clinical Drug Component, while as a Quantified Clinical Drug it is written as 960 mg/20 mL.
- If the strength is provided in % (volume or mass-percent are not distinguished) it is stored in the numerator_value/numerator_unit_concept_id field combination, with both the denominator_value and denominator_unit_concept_id set to NULL. If it is a Quantified Drug the total amount of drug is provided in the denominator_value/denominator_unit_concept_id pair. E.g., the 30 G Isoconazole 2% Topical Cream is provided as 2% / in Clinical Drug and Clinical Drug Component, and as 2% /30 G.
- Sometimes, one Ingredient is listed with different units within the same drug. This is very rare, and usually this happens if there are more than one Precise Ingredient.

For example, 'Penicillin G, 30 Benzathine 150000 UNT/ML / Penicillin G, Procaine 150000 MEQ/ML Injectable Suspension' contains Penicillin G in two different forms.

- Sometimes, different ingredients in liquid drugs are listed with different units in the denominator_ unit_concept_id. This is usually the case if the ingredients are liquids themselves (concentration provided as mL/mL) or solid substances (mg/mg). In these cases, the general assumptions is made that the density of the drug is that of water, and one can assume 1 g = 1 mL.
- All Drug vocabularies containing Standard Concepts have entries in the DRUG_STRENGTH table.
- There is now a Concept Class for supplier information whose relationships can be found in CONCEPT_ RELATIONSHIP with a relationship_id of 'Has supplier' and 'Supplier of'

Mapping

Representing content as Concepts

In CDM data tables the meaning of the content of each record is represented using Concepts. Concepts are stored with their CONCEPT_ID as foreign keys to the CONCEPT table in the Standardized Vocabularies, which contains Concepts necessary to describe the healthcare experience of a patient. If a Standard Concept does not exist or cannot be identified, the Concept with the CONCEPT_ID 0 is used, representing a non-existing or un-mappable concept.

Records in the CONCEPT table contain all the detailed information about it (name, domain, class etc.). Concepts, Concept Relationships and other information relating to Concepts is contained in the tables of the Standardized Vocabularies.

Concept IDs and Source Values

Many tables contain equivalent information multiple times: As a Source Value, a Source Concept and as a Standard Concept.

- Source Values contain the codes from public code systems such as ICD-9-CM, NDC, CPT-4 etc. or locally controlled vocabularies (such as F for female and M for male) copied from the source data. Source Values are stored in the *_SOURCE_VALUE fields in the data tables.

- Concepts are CDM-specific entities that represent the meaning of a clinical fact. Most concepts are based on code systems used in healthcare (called Source Concepts), while others were created de-novo (CONCEPT_CODE = 'OMOP generated'). Concepts have unique IDs across all domains.
- Source Concepts are the concepts that represent the code used in the source. Source Concepts are only used for common healthcare code systems, not for OMOP-generated Concepts. Source Concepts are stored in the *_SOURCE_CONCEPT_ID field in the data tables.
- Standard Concepts are those concepts that are used to define the unique meaning of a clinical entity. For each entity there is one Standard Concept. Standard Concepts are typically drawn from existing public vocabulary sources. Concepts that have the equivalent meaning to a Standard Concept are mapped to the Standard Concept. Standard Concepts are referred to in the _CONCEPT_ID field of the data tables.

Source Values are only provided for convenience and quality assurance (QA) purposes. Source Values and Source Concepts are optional, while Standard Concepts are mandatory. Source Values may contain information that is only meaningful in the context of a specific data source.

Type Concepts

By Mik Kallfelz and Dmitry Dymshyts

Type Concepts (ending in _TYPE_CONCEPT_ID) are present in many tables. They are special Concepts with the purpose of indicating from where the data are derived in the source. For example, the Type Concept field can be used to distinguish a DRUG_EXPOSURE record that is derived from a pharmacy-dispensing claim from one indicative of a prescription written in an electronic health record (EHR).

- Type concepts help determining the provenance of a record in the OMOP CDM. Many tables hold a specific _type_concept_id field for which valid concepts can be used to indicate a particular source of that record. For a condition it can be helpful to know, if it was derived from an EHR system or insurance claims. For a drug exposure it should be very useful to distinguish between prescriptions and actual administrations.
- In respect to the target table, matching type concepts should be chosen during the ETL step while processing sources. There are various representations in the list of type concepts of which some are quite specific for one table and others can be applied for many tables / domains, as they are quite generic. There is however no

plausibility check or dependency between type concepts and tables which means they have to be chosen correctly.

● There is now one specific vocabulary for type concepts which replaced a number of previously existing tables. For example, where previously there was a dedicated vocabulary for Drug Type Concepts, now we would choose the respective ones from the overall vocabulary (and ignore some of the old ones):

Drug Type	Type Concept
Inpatient administration	EHR administration record
Physician administered drug (identified from EHR problem list)	EHR administration record
Physician administered drug (identified from referral record)	EHR administration record
Physician administered drug (identified from EHR observation)	EHR administration record
Physician administered drug (identified from EHR order)	EHR administration record
Prescription dispensed in pharmacy	EHR dispensing record
Dispensed in Outpatient office	EHR dispensing record
Medication list entry	EHR medication list
Prescription written	EHR prescription

	EHR prescription issue record
Prescription dispensed through mail order	Mail order record
NLP derived	NLP
Patient Self-Reported Medication	Patient self-report
Physician administered drug (identified as procedure)	Pharmacy claim
Drug era - 0 days persistence window	
Drug era - 30 days persistence window	
Randomized Drug	

Time span of available data

Data tables for clinical data contain a datetime stamp (ending in _DATETIME, _START_DATETIME or _END_DATETIME), indicating when that clinical event occurred. As a rule, no record can be outside of a valid OBSERVATION_PERIOD time period. Clinical information that relates to events that happened prior to the first OBSERVATION_PERIOD will be captured as a record in the OBSERVATION table as 'Medical history' (CONCEPT_ID = 43054928), with the OBSERVATION_DATETIME set to the first OBSERVATION_PERIOD_START_DATE of that patient, and the VALUE_AS_CONCEPT_ID set to the corresponding CONCEPT_ID for the condition/drug/procedure that occurred in the past. No data occurring after the last OBSERVATION_PERIOD_END_DATE can be valid records in the CDM. * When mapping source data to the CDM, if time is unknown the default time of 00:00:00 should be chosen.

Source Values, Source Concept Ids, and Standard Concept Ids

Each table contains fields for Source Values, Source Concept Ids, and Standard Concept Ids.

- Source Values are fields to maintain the verbatim information from the source database, stored as unstructured text, and are generally not to be used by any standardized analytics. There is no standardization for these fields and these columns can be used as the local CDM builders see fit. A typical example would be an ICD-9 code without the decimal from an administrative claim as condition_source_value = '78702' which is how it appeared in the source.
- Source Concept Ids provide a repeatable representation of the source concept, when the source data are drawn from a commonly-used, internationally-recognized vocabulary that has been distributed with the OMOP Common Data Model. Specific use cases where source vocabulary-specific analytics are required can be accommodated by the use of the *_SOURCE_CONCEPT_ID fields, but these are generally not applicable across disparate data sources. The standard *_CONCEPT_ID fields are **strongly suggested** to be used in all standardized analytics, as specific vocabularies have been established within each data domain to facilitate standardization of both structure and content within the OMOP Common Data Model.

The following provide conventions for processing source data using these three fields in each domain:

When processing data where the source value is either free text or a reference to a coding scheme that is not contained within the Standardized Vocabularies:

- Map all Source Values to the corresponding Standard CONCEPT_IDs. Store the CONCEPT_IDs in the TARGET_CONCEPT_ID field of the SOURCE_TO_CONCEPT_MAP table.
    - If a CONCEPT_ID is not available for the source code, the TARGET_CONCEPT_ID field is set to 0.

When processing your data where Source Value is a reference to a coding scheme contained within the Standardized Vocabularies:

- Find all CONCEPT_IDs in the Source Vocabulary that correspond to your Source Values. Store the result in the SOURCE_CONCEPT_ID field.

- If the source code follows the same formatting as the distributed vocabulary, the mapping can be directly obtained from the CONCEPT table using the CONCEPT_CODE field.
- If the source code uses alternative formatting (ex. format has removed decimal point from ICD-9 codes), you will need to perform the formatting transformation within the ETL. In this case, you may wish to store the mappings from original codes to SOURCE_CONCEPT_IDs in the SOURCE_TO_CONCEPT_MAP table.
- If the source code is not found in a vocabulary, the SOURCE_CONCEPT_ID field is set to 0
- Use the CONCEPT_RELATIONSHIP table to identify the Standard CONCEPT_ID that corresponds to the SOURCE_CONCEPT_ID in the domain.
  - Each SOURCE_CONCEPT_ID can have 1 or more Standard CONCEPT_IDs mapped to it. Each Standard CONCEPT_ID belongs to only one primary domain but when a source CONCEPT_ID maps to multiple Standard CONCEPT_IDs, it is possible for that SOURCE_CONCEPT_ID to result in records being produced across multiple domains. For example, ICD-10-CM code Z34.00 'Encounter for supervision of normal first pregnancy, unspecified trimester' will be mapped to the SNOMED concept 'Routine antenatal care' in the procedure domain and the concept in the condition domain 'Primagravida'. It is also possible for one SOURCE_CONCEPT_ID to map to multiple Standard CONCEPT_IDs within the same domain. For example, ICD-9-CM code 070.43 'Hepatitis E with hepatic coma' maps to the SNOMED concept for 'Acute hepatitis E' and a second SNOMED concept for 'Hepatic coma', in which case multiple CONDITION_OCCURRENCE records will be generated for the one source value record.
  - If the SOURCE_CONCEPT_ID is not mappable to any Standard CONCEPT_ID, the _CONCEPT_ID field is set to 0.
- Write the data record into the table(s) corresponding to the domain of the Standard CONCEPT_ID(s).
  - If the Source Value has a SOURCE_CONCEPT_ID but the SOURCE_CONCEPT_ID is not mapped to a Standard CONCEPT_ID, then the domain for the data record, and hence it's table location, is determined by the DOMAIN_ID field of the CONCEPT record the SOURCE_CONCEPT_ID refers to. The Standard _CONCEPT_ID is set to 0.
  - If the Source Value cannot be mapped to a SOURCE_CONCEPT_ID or Standard CONCEPT_ID, then direct the data record to the most appropriate CDM domain based on your local knowledge of the intent of the source data and associated value. For example, if the un-mappable Source Value came from a 'diagnosis' table then, in the absence of other information, you may choose to record that fact in the CONDITION_OCCURRENCE table.

Each Standard CONCEPT_ID field has a set of allowable CONCEPT_ID values. The allowable values are defined by the domain of the concepts. For example, there is a domain concept of 'Gender', for which there are only two allowable standard concepts of practical use (8507 - 'Male', 8532- 'Female') and one allowable generic concept to represent a standard notion of 'no information' (concept_id = 0). This 'no information' concept should be used when there is no mapping to a standard concept available or if there is no information available for that field. The exceptions are MEASUREMENT.VALUE_AS_CONCEPT_ID, OBSERVATION.VALUE_AS_CONCEPT_ID, MEASUREMENT.UNIT_CONCEPT_ID, OBSERVATION.UNIT_CONCEPT_ID, MEASUREMENT.OPERATOR_CONCEPT_ID, and OBSERVATION.MODIFIER_CONCEPT_ID, which can be NULL if the data do not contain the information (THEMIS issue #11).

There is no constraint on allowed CONCEPT_IDs within the SOURCE_CONCEPT_ID fields.

Custom SOURCE_TO_CONCEPT_MAP

When the source data uses coding systems that are not currently in the Standardized Vocabularies (e.g. ICPC codes for diagnoses), the convention is to store the mapping of such source codes to Standard Concepts in the SOURCE_TO_CONCEPT_MAP table. The codes used in the data source can be recorded in the SOURCE_VALUE fields, but no SOURCE_CONCEPT_ID will be available.

Custom source codes are not allowed to map to Standard Concepts that are marked as invalid.