2. Quick Start Guide

This chapter provides a brief overview of the i2b2 CDM star schema and how to use it in combination with an ontology.

2.1. Star Schema Introduction

In this section we describe, at a high level, the main tables and fields in the i2b2 CDM. A detailed description of the full data model is in the i2b2 CDM spreadsheet.

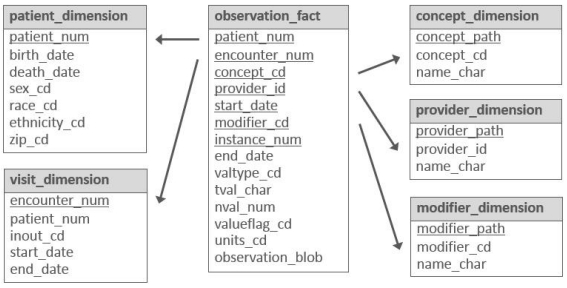

Figure 1. Main tables and fields in the i2b2 CDM star schema. Underlined fields are part of the primary keys of tables. Note that not all tables and fields in the i2b2 CDM are shown here. See the i2b2 CDM spreadsheet for the full data model.

The i2b2 CDM is a data warehouse modeled on the star schema structure first proposed by Ralph Kimball. The database schema looks like a star, with one central fact table surrounded by one or more dimension tables. The most important concept regarding the construction of a star schema is identifying what constitutes a fact.
In healthcare, a logical fact is an observation on a patient. It is important to note that an observation may not represent the onset or date of the condition or event being described, but instead is simply a recording or a notation of something. For example, the observation of 'diabetes' recorded in the database as a 'fact' at a particular time does not mean that the condition of diabetes began exactly at that time, only that a diagnosis was recorded at that time (there may be many diagnoses of diabetes for this patient over time).

The fact table contains the basic attributes about the observation, such as the patient and provider numbers, a concept code for the concept observed, a start and end date, and other parameters described in this document. In the i2b2, the fact table is **OBSERVATION_FACT**. All patient observations are placed in this table, such as diagnoses, procedures, medications, and laboratory test results. A large institution with millions of patients might have billions of rows of observations in this table.

Dimension tables contain further descriptive and analytical information about attributes in the fact table. A dimension table may contain information about how certain data is organized, such as a hierarchy that can be used to categorize or summarize the data. In the i2b2 data mart, there are five dimension tables that provide additional information about fields in the fact table: **PATIENT_DIMENSION**, **CONCEPT_DIMENSION**, **VISIT_DIMENSION**, **PROVIDER_DIMENSION**, **MODIFIER_DIMENSION**.

2.2. Table Descriptions and Examples

In the **OBSERVATION_FACT** table, observation codes, such as ICD-10 diagnosis or NDC medication codes, are placed in the concept_cd field. In addition to the concept_cd, the patient_num, encounter_num, and start_date fields are required for each observation. A provider (observer) ID, modifier code and instance number are used for certain types of observations. An "@" symbol is used as a default value for the provider_id and modifier_cd, and 1 is the default instance_num.

The fact table also contains value objects associated with the observations. A laboratory test result, with a single value, can be stored in one row of the fact table. A complex value, such as blood pressure, for example "120/80-standing", is reduced to three rows with concept "blood pressure" and modifiers "systolic", "diastolic", and "position" (preferably expressed as LOINC or other known standard, but this is not required). The value type of the observation is specified in the valtype_cd field, such as N=number, T=text, D=date, and so forth. Numbers and dates are placed in the nval_char field which is a numeric data type. Text-based values are placed in the tval_char field; and, binary data are stored in the observation_blob field. For numeric values, the tval_char field can be used to indicate an operator, such as "E" (equals), "G" (greater than), or "LE" (less than or equal to). For example, tval_char="L" and nval_num="0.01" means the value is less than 0.01.

patient_num	encounter_num	concept_cd	start_date	modifier_cd	valtype_cd	tval_char	nval_num
1000001	123456	ICD9:462	2007-08-09	@			

1000001	123456	LOINC:1751-7	2007-08-10	@	N	E	4.3
1000001	298765	LOINC:6598-7	2007-09-15	@	N	L	0.01
1000002	890123	LOINC:6598-7	2009-03-20	@	T	NEG	
1000002	543210	VITAL:BP	2010-05-01	systolic	N	120	
1000002	543210	VITAL:BP	2010-05-01	diastolic	N	80	
1000002	543210	VITAL:BP	2010-05-01	position	T	standing	

Table 1. Example observations in the fact table. Shown are a basic diagnosis fact (ICD9 code), laboratory tests (LOINC codes) with numeric and text based results, and a multi-part blood pressure observation stored as three facts (custom VITAL:BP code).

In addition to the fact table, there are five other dimension tables that help express the patient data. The **PATIENT_DIMENSION** table has one row for every patient in the database. The patient_num field is a unique integer for each patient. A separate **PATIENT_MAPPING** table optionally maps the patient_num to a medical record number or other local identifier, which may be non-numeric. The patient_dimension table contains several optional demographic fields, including birth_date, death_date, sex_cd, and race_cd. Demographic concepts can alternatively be placed as observations in the fact table, depending on how an institution choses to model the data.

The **VISIT_DIMENSION** table allows periods to be represented that correspond roughly to patient encounters where observations were recorded. An "encounter" can involve a patient directly, such as a visit to a doctor's office, or it can involve the patient indirectly, such as running several tests tied together by the same tube of the patient's blood. Similar to patient_num, the ecounter_num is a unique integer for each row in the visit_dimension table. A separate **ENCOUNTER_MAPPING** table optionally maps the encounter_num to local encounter billing codes, visit IDs, or other local identifier.

The **CONCEPT_DIMENSION** table has vocabulary terms that map to the codes used in the concept_cd field of the fact table. These terms typically come from standard terminologies, such as International Classification of Diseases (ICD), National Drug Code (NDC), and Logical Observation Identifiers Names and Codes (LOINC). However, the i2b2 and tranSMART software do not recognize a difference between standard and local terminologies. Terms may be grouped into hierarchies. The hierarchical representation used in the concept table is similar to that of a hierarchical file system. The parent term is positioned in the "folder" position of the path, and the child term in the "file" position. For example, in the concept_dimension table (Table 2), the parent "anti-infectives" can have the three children "penicillin", "ampicillin", and "Bactrim". The children map to the NDC codes used in the fact table, but the path shows they are types of anti-infectives.

concept_path	concept_cd	name_char
\Med\		Medications
\Med\anti-infectives\		Anti-Infectives
\Med\anti-infectives\penicillin\	NDC:00002032902	Penicillin
\Med\anti-infectives\ampicillin\	NDC:60429002340	Ampicillin
\Med\anti-infectives\Bactrim\	NDC:00003013850	Bactrim

Table 2. Layout of concept data representations in the concept dimension. Paths group codes into a hierarchy, like a computer file system.

This hierarchical organization allows users to query for a path, such as the general concept of anti-infectives, and the i2b2 or tranSMART software can automatically and efficiently convert this to a set of concept codes to search for in the fact table. For example, the query below finds all patients seen on anti-infectives, we would run the following query:

```
Select distinct(patient_num)
From observation_fact
Where concept_cd in
  (select concept_cd
   from concept_dimension
   where concept_path like
   '\Med\anti-infectives%')
```

The path of the concept is used to find and use all concept_cds that fall into the anti-infectives group. If we only wanted to find patients specifically on Bactrim, we would use the same query with the following concept_path: "\Med \anti-infectives\Bactrim%".

The same approach is used in the **PROVIDER_DIMENSION** and **MODIFIER_DIMENSION** tables. A path can represent a group of providers, such as a hospital department, or the individual clinicians that are part of that department. Examples of modifiers include primary or secondary indicators for diagnoses and dose, route, and frequency for medications.