4. Tutorials: Using the i2b2 CDM

This chapter contains a series of tutorials that describe how to model different types of data using the i2b2 CDM. The tutorials start with simpler, more common use cases, and progress to more advanced concepts.

4.1 Diagnoses and Demographics (Basic Facts)

Below are examples of basic facts in the i2b2 CDM.

A diagnosis, such as acute pharyngitis, has an associated code, such as "ICD10:J02.9". The customizable i2b2 ontology defines the concept and codes. Different sites might use their own codes for the acute pharyngitis. The observation of this diagnosis is stored as a record in the OBSERVATION_FACT table. A patient_num (1000001), encounter_num (730868), and start_date (2017-10-22) are required. A provider_id, modifier_cd, and instance_num are also required, but these can be set to their default values, "@", "@" and 1.

The patient_num and encounter_num corresponding to these diagnoses must be listed in the PATIENT_DIMENSION and VISIT_DIMENSION tables. In this example, the patient is female (sex_cd=F) and white (race_cd=W). The encounter is an outpatient visit (inout_cd=O), with start_date and end_date both October 22, 2017.
Note that the date of the observation in the OBSERVATION_FACT table can be different from the dates in the VISIT_DIMENSION. This is illustrated in a second diagnosis of asthma (ICD10:J45.909), which was recorded the day after the patient was discharged from an inpatient visit (encounter_num=798502). This diagnosis also has a provider (the observer). The PROVIDER_DIMENSION lists the provider (provider_id=X1824); it indicates her name is Jane Smith; and, the path can be used to group providers, for example, by department or specialty. In this example, the asthma diagnosis is duplicated. When this occurs, the instance_num must be incremented to ensure that the primary keys of each record are unique.

The i2b2 ontology enables concepts to be modeled in different ways. For example, the ontology can specify that the race_cd field of the PATIENT_DIMENSION be used to store the patient's race. Having all demographics in the PATIENT_DIMENSION table can simplify certain analyses. Alternatively, race can be an observation stored in the OBSERVATION_FACT table (e.g., concept_cd="DEM|Race:W"). This can be useful if demographics are separately recorded for each admission, or if there is a need to store multiple races for a patient.

The CONCEPT_DIMENSION table has vocabulary terms that map to the codes used in the concept_cd field of the OBSERVATION_FACT table. 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. In the examples here, acute pharyngitis and asthma both have paths that are children of the term "Diseases of the Respiratory System". In the i2b2 and tranSMART software, a query for the parent term will match the concept_cd codes corresponding to any of the child nodes.

OBSERVATION_FACT

patient_num	encounter_num	concept_cd	provider_id	start_date	modifier_cd	instance_num
1000001	730868	ICD10:J02.9	@	2017-10-22	@	1
1000001	798502	ICD10:J45.909	X1824	2018-02-12	@	1
1000001	798502	ICD10:J45.909	X1824	2018-02-12	@	2
1000001	798502	DEM|Race:W	@	2018-02-08	@	1

PATIENT_DIMENSION

patient_num	sex_cd	race_cd
1000001	F	W

VISIT_DIMENSION

encounter_num	patient_num	inout_cd	start_date	end_date
730868	1000001	O	2017-10-22	2017-10-22
798502	1000001	I	2018-02-08	2018-02-11

PROVIDER_DIMENSION

provider_path	provider_id	name_char
Medicine\Pulmonary\X1824\	X1824	Jane Smith, MD

CONCEPT_DIMENSION

concept_path	concept_cd	name_char
Dem\Race\White\	DEM|Race:W	White
Diag\ICD10\J00-J99\	ICD10:J	Diseases of the Respiratory System
Diag\ICD10\J00-J99\J02\J02.9\	ICD10:J02.9	Acute pharyngitis, unspecified
Diag\ICD10\J00-J99\J45\J45.909\	ICD10:J45.909	Unspecified asthma, uncomplicated

Figure 2. Example data in the i2b2 CDM.

4.2 Laboratory Tests and Vital Signs (Value Constraints)

Laboratory tests, vital signs, and other data types can have an associated value. The OBSERVATION_FACT table contains six fields to store values: VALTYPE_CD, TVAL_CHAR, NVAL_CHAR, VALUEFLAG_CD, UNITS_CD, and OBSERVATION_BLOB. A description of these fields are at

https://community.i2b2.org/wiki/display/ServerSideDesign/Value+Columns

Note that the i2b2 and tranSMART ontology must be configured properly so that the software uses these value-related fields. A description of how the ontology works with the value fields is at

https://community.i2b2.org/wiki/display/ServerSideDesign
/Example+of+Value+Constraints+Used+in+Queries

Additional information about the i2b2 Ontology Management Cell can be found at

https://community.i2b2.org/wiki/display/ServerSideDesign/Ontology+Management+%28ONT%29+Cell

4.3 Medications, Procedures and Allergies (Modifiers)

A single observation can be represented in the i2b2 CDM as a fact with an unlimited number of modifier codes. A medication, for example, can be modified with dose, route, and frequency. Each of these modifiers is stored as a separate record in the OBSERVATION_FACT table.

For example, a prescription for Aspirin 325 mg QD PO on 4/4/2010 is stored as four records. All four have “med:aspirin” as the CONCEPT_CD and “4/4/2010” as the START_DATE. The “base” record for the observation uses “@” for the MODIFIER_CD. The three modifier records have MODIFIER_CD values of “MED:DOSE”, “MED:FREQ”, and “MED:ROUTE”, with their associated values (325 mg, “QD”, and “PO”) stored in the value fields. In this example, the concept med:aspirin is defined in the CONCEPT_DIMENSION tables; and, the three modifier codes, MED:DOSE, MED:FREQ, and MED:ROUTE, are defined in the MODIFIER_DIMENSION table. The i2b2 and tranSMART software also require these to be included in the ontology. A more detailed description of this example and an example using procedures are available at

https://community.i2b2.org/wiki/display/DevForum/Modifiers+in+i2b2+Data+Model

Modifiers can be used for many other things, such as indicating whether a diagnosis is primary or secondary or adding stage to a cancer diagnosis. Another example using modifiers to indicate patients’ allergies is at

https://community.i2b2.org/wiki/display/DevForum
/Representing+Allergies+in+Star+Schema+with+modifiers

4.4 Clinical Notes, Imaging and Genomics Data (Blob Data)

Clinical Notes

There is a wealth of information within the plain text clinical narrative. Notes can be represented in the i2b2 CDM in a couple ways. First, the entire note can be a single record in the OBSERVATION_FACT table. The concept_cd can be, for example, “NOTE:DischargeSummary”; and, the text of the note is stored in the observation_blob field. This field contains a FULLTEXT index in the Microsoft SQL Server implementation of the i2b2 CDM to enable efficient searches of notes. A description of text search in i2b2 is at

https://community.i2b2.org/wiki/display/DevForum/Text+search+in+i2b2

Alternatively, natural language processing (NLP) software can be used to extract concepts from notes. Each concept can be stored as a separate record in the OBSERVATION_FACT table. An example of this using the NLP CTAKES program is at

https://community.i2b2.org/wiki/display/NLPCTAKES/NLP+cTakes+Home

Imaging Data

The observation_blob field can also be used to store binary data, such as medical images. An example of this was a project called mi2b2 (Medical Imaging Informatics Bench to Bedside), which linked i2b2 to separate PACS (Picture Archiving and Communication System) systems. Although the original images were stored in the PACS systems, a thumbnail version was placed in the i2b2 observation_blob field, so that it could be previewed within the i2b2 user interface. A description of mi2b2 is at

https://community.i2b2.org/wiki/display/mi2b2/mi2b2+User+Documentation

Genomics Data

Like notes, genomics data can be modeled in different ways. A tutorial showing how to load genomic VCF (Variant Call Format) files into the observation_blob field is at

https://community.i2b2.org/wiki/display/IGD/Demo+Data

4.5 Lab Panels and Specimens (Dummy Records)

The encounter_num typically represents a visit. However, it can also be used more generally to group related observations. A "dummy" encounter_num value, for example, can be created that corresponds to a lab panel, such as a complete blood count (CBC). The start_date for this encounter in the VISIT_DIMENSION table could be the specimen date. The results of each test within the panel are stored as separate records in the OBSERVATION_FACT table, using the same encounter_num.

4.6 Clinical Trial Data (Extending the Data Model)

Encounters in clinical trials are often associated with a visit number. The VISIT_DIMENSION table can be extended with a visit_num field to store the visit number. The ontology can include a concept for the visit number that points to this field, so that the user can query for observations that occurred as part of a particular visit number.