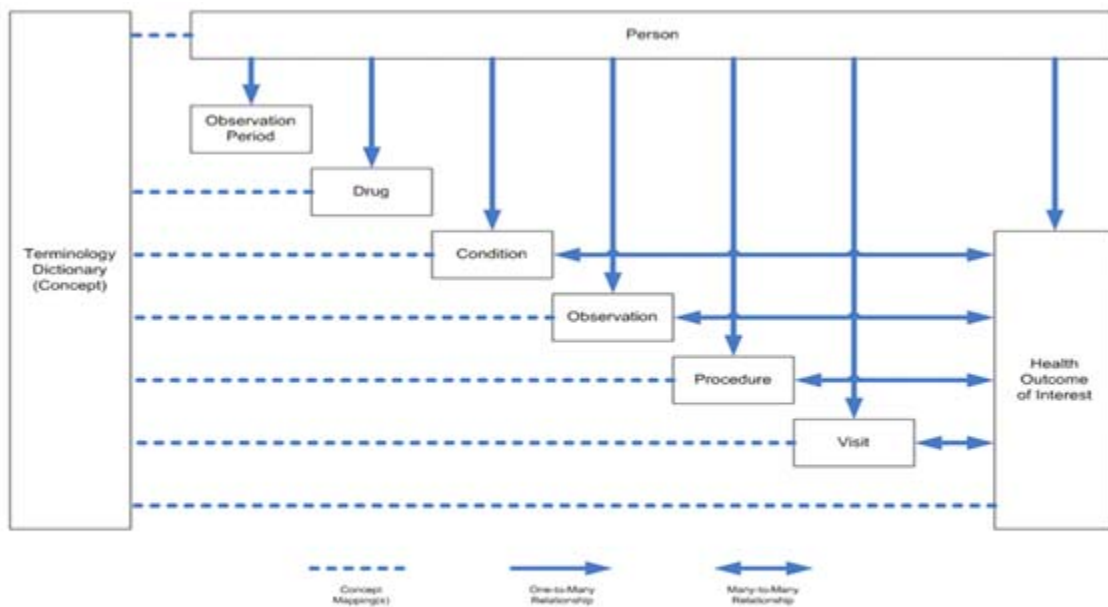


# i3 Drug Safety **FINAL** OMOP Common Data Model (CDM) ETL Mapping Specification Template

23 September 2009

Revised by i3: **31 December 2009**

*i3 modifications in blue*



## Table of Contents

<b>1.0 Introduction</b>	<b>4</b>
<b>2.0 Source Data Mapping Approach</b>	<b>4</b>
<b>3.0 Source Data Mapping</b>	<b>5</b>
3.1 Data Mapping	5
3.1.1 Table Name: PERSON	5
3.1.2 Table Name: DRUG_EXPOSURE	7
3.1.3 Table Name: CONDITION_OCCURRENCE	11
3.1.4 Table Name: VISIT_OCCURRENCE	14
3.1.5 Table Name: PROCEDURE_OCCURRENCE	16
3.1.6 Table Name: OBSERVATION	19
3.1.7 Table Name: OBSERVATION_PERIOD	21
3.2 Source Independent Data Mapping	23
3.2.1 Table Name: DRUG_ERA	23
3.2.2 Table Name: CONDITION_ERA	26
3.3 Reference Tables	28
3.3.1 Table Name: DRUG_EXPOSURE_REF	28
3.3.2 Table Name: CONDITION_OCCURRENCE_REF	28
3.3.3 Table Name: PROC_OCCURRENCE_REF	32
3.3.4 Table Name: OBSERVATION_TYPE_REF	34
3.3.5 Table Name: VOCABULARY_REF	34
3.3.6 Table Name: RELATIONSHIP_TYPE	35

## Document Control

### Change Record

Date	Author	Version	Change Reference
21-Sep-09	OMOP	1.0	New document, describes OMOP ETL for Distributed Partners
02-Oct-09	Florence Wang	1.1	Specific instructions for i3
09-Oct-09	Florence Wang	1.12	Additional details/clarification
16-Oct-09	Florence Wang	1.13	Additional details/clarification
23-Oct-09	Florence Wang	1.14	Additional details/clarification
30-Oct-09	Florence Wang	2	Copy-editing
11-Nov-09	Florence Wang	2.2	Changed to format of new ETL template, provided 06NOV2009
25-Nov-09	Florence Wang	2.3	Validation checks, copy-editing
09-Dec-09	Florence Wang	2.6	Validation checks

### Contributors

Name	Organization	Title
Florence Wang	i3	Epidemiologist

### Reviewers

Name	Role	Title	Date Reviewed
Mark Khayter	OMOP	Technical Consultant	23-Sep-2009
Li Zhou	i3	Analyst	15-Dec-2009
Mike Doherty	i3	Analyst	15-Dec-2009
K. Arnold Chan	i3	Senior Scientist	17-Dec-2009 31-Dec-2009 (public document)

### Document References

Document Title	Type of Reference	Document Location
OMOP ETL Mapping Specification		OMOP Basecamp

## 1.0 Introduction

This document reflects the requirements, assumptions, business rules and transformations for the implementation of the Common Data Model (CDM) as implemented by i3 Drug Safety. The initial ETL process was built using data and transformations as applicable to GE and Thomson.

The purpose of this document is to describe the ETL mapping of the proprietary or licensed data from i3 Drug Safety into the OMOP Common Data Model.

It is based on the OMOP ETL Specifications. General information that is covered by the OMOP ETL Specification will not be covered in this document, but a detailed discussion of the i3 Drug Safety -specific aspects of mapping and converting data to the standard CDM is provided.

The document is composed of three main sections:

- Source Data Mapping. Describes major tables of the CDM schema and special data handling required for each table.
- Source Independent Data Mapping. Describes mapping process of the Drug and Condition Era's.
- Data Mapping Reference tables.

In each section, the tables and their mapping are individually reviewed along with any source specific rules and exceptions.

The intended audience for this document will include both researchers that want to use the experience and learning in order to incorporate them into their own CDM construction.

## 2.0 Source Data Mapping Approach

*In the OMOP ETL Specifications, this section covers the high-level assumptions and approach to extraction, transformation and loading (ETL) of raw source data into the Common Data Model (CDM). This high-level approach should be equivalent between the data sources obtained by OMOP and i3 Drug Safety. However, if a significant divergence becomes necessary and meaningful, it should be discussed here.*

i3 utilized a proprietary research database, the Ingenix Normative Health Informatics (NHI) database, containing longitudinal health care claims and health plan enrollment data dating back to 1993 with the opportunity to link patient and physician survey data to pharmacy and medical claims, socioeconomic measures, and clinical laboratory results. Data included in the database have previously undergone an adjudication process by the health plan. With approval from appropriate institutional review and privacy boards, linkage to medical record data is available.

From this underlying data source, i3 identified all enrolled individuals from 01 January 2002 through 31 December 2008 in the NHI membership. We then sampled 1,000,000 subjects randomly (without replacement) from those with medical and pharmacy coverage, where each person had equal chance of being sampled, regardless of duration of enrollment in the health plan. Among this sample of subjects, we extracted all available health plan enrollment, pharmacy dispensing, and medical data from 01 January 1993 (the earliest available data within the NHI) through 31 August 2009 (the latest available data within the NHI).

For the CDM transformation, we chose a sampling scheme that would allow the sampled population to be reflective of the study population commonly used in epidemiologic studies (i.e. requiring both medical and pharmacy coverage). We also extracted all available longitudinal data for the population to allow for greater flexibility in study design and statistical analysis.

We employed a standard data cleaning program which deleted dispensing claims where both the quantity and days supply dispensed were set to zero and duplicate medical and pharmacy dispensing claims.

## 3.0 Source Data Mapping

*This section will describe mapping process and ETL conversions of data received from your data into Common Data Model.*

### 3.1 Data Mapping

*Describe here how your data are provided, and in what technology (relational database system, SAS files etc.) the CDM will be represented.*

Datasets drawn from the NHI are in SAS file format. The CDM is presented in SAS files. As such, NULL was defined as "." for integers, and defined as missing for character variables.

#### 3.1.1 TABLE NAME: PERSON

*Describe how the Person mapping and transformations are designed.*

Data used to populate the PERSON table were drawn from the member coverage and geographic tables. Variables available in the member coverage table include unique de-identified ID for each individual, sex, year of birth, start and end dates of medical or prescription drug coverage. Variables available in the geographic table include unique de-identified ID for each individual and geographic region.

The field mapping is performed as follows:

1. Missing attributes for concept identifiers were populated with value zero (0) in CDM.
2. Other missing attributes were populated with NULL.

3. A single record was created for each individual. Multiple geographic region records may exist per individual. If so, we chose the geographic region linked with the most recent enrolled period.
4. For areas defined by three-digit zip codes with a population of less than 20,000, three-digit zip code was recoded to 000.
5. Year of birth of those 89 years of age or older was recoded so that the age would be 89.
6. Race is available for a subset of the NHI population, imputed/derived from consumer data. As such, race was not included in the CDM.

Destination Field	Source Field (generic description)	Applied Rule	Comment
PERSON_ID	ID	de-identified unique id	
YEAR_OF_BIRTH	Year of birth		
GENDER_CONCEPT_ID	Sex	8507 for males, 8532 for females, 8551 for unknown  where Sex=first letter of source_to_concept_map.source_code and source_to_concept_map.mapping_type="GENDER" and source_to_concept_map.source_vocabulary code="GE"	Since there were no specific observations for sex specific to NHI, used the codes from GE (CR 07OCT2009).
RACE_CONCEPT_ID		0	
LOCATION_CONCEPT_ID	Zipcode	where first three digit of Zipcode= source_to_concept_map.source_code and source_to_concept_map.mapping_type="ZIPCODE" and source_to_concept_map.source_vocabulary code="GE" or "THOMSON"	Since there were no specific observations for zip code specific to NHI, used the codes from GE/Thomson (CR 07OCT2009).
SOURCE_PERSON_KEY	ID	de-identified unique id	
SOURCE_GENDER_CODE	Sex	M=males  F=females	

Destination Field	Source Field (generic description)	Applied Rule	Comment
SOURCE_LOCATION_CODE	Zipcode		
SOURCE_RACE_CODE		NULL	

### Selected validation checks

1. We compared frequency of sex in NHI with resulting gender variables (source\_gender\_code and gender\_concept\_id) in the CDM.

#### Frequency of sex in NHI

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	502888	50.29	502888	50.29
M	496729	49.67	999617	99.96
U	383	0.04	1000000	100.00

#### Frequency of sex in CDM person table

SOURCE_GENDER_CODE	GENDER_CONCEPT_ID	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	8532	502888	50.29	502888	50.29
M	8507	496729	49.67	999617	99.96
U	8551	383	0.04	1000000	100.00

2. We compared the frequency of the year of birth in NHI sample with the YEAR\_OF\_BIRTH in the CDM. (Data not shown)
3. We compared the total N from NHI with the final CDM sample.

The total N from NHI = 1,000,000 people

The total N from final CDM sample = 1,000,000 people

### 3.1.2 TABLE NAME: DRUG\_EXPOSURE

*Describe how the Drug\_Exposure mapping and transformation are designed.*

Drug exposure data were drawn from the pharmacy dispensing, medical, and facility claims tables. Included were claims for:

1. Filled prescriptions (pharmacy dispensing table)
  - a. Data had separate entries for each filled prescription and every subsequent refill.

- b. Data available on claims included unique de-identified ID, date of dispensing, National Drug Codes (NDC), and quantity/supply dispensed.
  - c. Drugs dispensed during inpatient hospitalizations were not captured.
2. Procedures related to drug administration (medical and facility tables)
- a. Data available on claims included unique de-identified ID, date of procedure, corresponding procedure codes (International Classification of Diseases, Ninth Revision (ICD-9), Current Procedural Terminology (CPT), or Health Care Financing Agency (HCFA) Common Procedure Coding System (HCPCS)) and corresponding diagnoses.
  - b. ICD-9, HCPCS, and CPT codes may be listed in any of the procedure fields.

If the source drug identifier could not be translated into a standard drug concept, only the source drug identifier was stored.

Multiple procedures may be listed on each NHI record. Data were transposed so there would be one record per procedure code.

There may be multiple records for the same drug/procedure per day per individual, reflecting the underlying claims data.

Records of pharmacy dispensing include a flag indicating a reversal of claim. Those records with a flag indicating a reversal were not included in the CDM.

Only procedure codes which mapped to a drug concept were extracted (CR 01OCT2009). Note that the procedures listed in the drug table and the ones listed in the procedure table are not mutually exclusive.

Information on drug exposure end date, reason for stopping drug use, and drug refill were not available on drug dispensing claims.

The field mapping is performed as follows:

Destination Field	Source Field (generic description)	Applied Rule	Comment
DRUG_EXPOSURE_ID	_n_	system generated unique identifier	
DRUG_EXPOSURE_START_DATE	Date of service (dispensing or procedure)	date of dispensing or date of procedure	
DRUG_EXPOSURE_END_DATE		NULL	
PERSON_ID	ID	de-identified unique id	



Destination Field	Source Field (generic description)	Applied Rule	Comment
		mapping_type="PROCEDURE DRUG" and source_to_concept_map.source_vocabulary_code="03"	
DRUG_EXPOSURE_TYPE	Dispensing,  Medical, Facility	if Dispensing then drug_exposure_ref.drug_exposure_type="1"  if Medical or Facility then drug_exposure_ref.drug_exposure_type="4"	Create indicators to distinguish data from dispensing, medical, or facility claims.
STOP_REASON		NULL	
REFILLS		NULL	
DRUG_QUANTITY	Quantity	NDC c: quantity of drug units HCPCS/CPT/ICD-9: NULL	
DAYS_SUPPLY	Days supply	NDC: days supply HCPCS/CPT/ICD-9: NULL	
SOURCE_DRUG_CODE	NDC  Procedure	NDC from dispensing  HCPCS/CPT/ICD-9 codes from medical or facility claims	

### Selected validation checks

Our sampled NHI claims data included 55,247,464 medical and facility claims (829,073 unique individuals) and 24,048,887 dispensing observations (739,662 unique individuals). In all, there were 855,460 unique individuals with claims data for medical visit and/or drug dispensing.

In the transposed claims data, there were 27,193,491 observations (774,123 unique individuals) included in the drug\_exposure table.

We compared the list of drug codes (NDC and procedure codes) in the source\_to\_concept\_map dataset with the set of NDC listed on the claims data. In our claims data, 44,122 unique drug codes (NDC or procedure codes) were found, and 8,133 were not mapped. All of the unmapped codes were NDC (as procedure codes which did not have corresponding concept IDs for mapping in the drug\_exposure table were included only in the procedure\_occurrence table). Therefore, 81.6% of the unique drug codes were mapped.

We printed and checked the contents of the dispensing, medical and facility claims data prior to CDM conversion. The contents of the CDM dataset were printed and compared to the list in specification document. Other checks included comparing printouts of

observations prior to merge procedures with those after merge procedures, examining frequency tables of created variables, and making sure the distribution of the date of service fell within the study date range.

Total number of observations in the CDM

DRUG_ EXPOSURE_ TYPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	24765488	91.07	24765488	91.07
4	2428003	8.93	27193491	100.00

STOP_ REASON	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	27193491	100.00	27193491	100.00

refills	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	27193491	100.00	27193491	100.00

Comparison of date ranges of data in NHI with those in CDM

	NHI Claims Data	CDM Drug_exposure Start
Minimum Date	1-Jan-93	1-Jan-93
Maximum Date	26-Aug-09	21-Aug-09

### 3.1.3 TABLE NAME: CONDITION\_OCCURRENCE

*Describe how the Condition\_Occurrence mapping and transformation are designed.*

Conditions were drawn from the medical claims table and facility claims table, each with up to 9 diagnoses. Discharge status codes were available on inpatient facility claims.

Data available on medical and facility claims included unique de-identified ID, date of diagnoses/procedures, corresponding ICD-9 diagnosis codes (up to 9), corresponding procedures as identified by ICD-9 procedure, CPT, or HCPCS codes, and site of care, as identified by American Medical Association (AMA) site codes.

All claims were flagged as inpatient (IP) claim, emergency department (ED) claim, or outpatient (OP) claim using the following hierarchical logic:

1. ED claims were identified by AMA site code 23 or CPT code 99281-99285.
2. Using a standard program that defined periods of inpatient stay, we identified the IP claims as those remaining claims (not flagged as ED) which fell within periods of inpatient stay.

3. The remaining claims not flagged as an ED claim and fell outside the period of inpatient stay were identified as OP claims.

4. ED, IP and OP flags were created to identify site of service.

Each claim may contain multiple diagnosis codes. Data were transposed so there would be one record per diagnosis code.

There may be multiple records for the same diagnosis per day per individual, reflecting the underlying claims data.

For an inpatient claim with flag of death upon discharge, an additional record was created.

Concept identifiers were populated with value zero (0) if no matching mapping from source codes were available.

Because all diagnosis codes in the NHI were compiled without distinction of header/detail/primary, we created 27 new records (CONDITION\_OCCURRENCE\_TYPE 500-526) to signify diagnosis codes one through 9 in the IP, OP, and ED setting.

Discharge status codes of 20-29, and 40-42 denoted death at discharge.

Note that diagnosis codes missing one or more digits may be included (CR 28OCT2009). For example, the diagnosis code for cholera is 001.x which requires a fourth digit. The diagnosis code 001 (missing a fourth digit) would also be included in the extraction.

Information on condition occurrence end date, stop reason, and diagnosis qualifier were not available on the claims.

The field mapping is performed as follows:

Destination Field	Source Field (generic description)	Applied Rule	Comment
CONDITION_OCCURRENCE_ID	_n_	system generated unique identifier	
CONDITION_START_DATE	Date of service	Date of service of the claim	
PERSON_ID	ID	de-identified unique id	
CONDITION_END_DATE		NULL	

Destination Field	Source Field (generic description)	Applied Rule	Comment
CONDITION_OCCURRENCE_TYPE	Diagnosis code (up to 9)  Patient status	where condition_occurrence_ref.condition_occurrence_type_desc match claim site of care (IP, ED or OP) and condition_occurrence_ref.condition_occurrence_position match 1-9 in diagnosis code title  or  if Patient status=20-29, or 40-42 then condition_occurrence_ref.condition_occurrence_type=66 (death at discharge).	
CONDITION_CONCEPT_ID	Diagnosis code (up to 9)  Patient status	Diagnosis  where Diagnosis code= source_to_concept_map.source_code and source_to_concept_map. mapping_type="CONDITION" and source_to_concept_map.source_vocabulary_code="02"  Patient status  where Patient status= source_to_concept_map.source_code and source_to_concept_map. mapping_type="DISCHARGE STATUS" and source_to_concept_map.source_vocabulary_code=" THOMSON"	Discharge status 20-29 and 40-42 listed in source_to_concept_map.source_code.  Since there were no observations for MAPPING_TYPE of "DISCHARGE STATUS" specific to NHI, used the codes specific to "THOMSON".
STOP_REASON		NULL	
DX_QUALIFIER		NULL	
SOURCE_CONDITION_CODE	Diagnosis code  Patient status		

### Selected validation checks

We compared the list of ICD-9 diagnosis codes in the source\_to\_concept\_map dataset to the set of ICD-9 diagnosis codes listed on the claims data. In our claims data, there

were 23,006 unique diagnosis codes, 8,572 of which were not mapped. Therefore, 62.7% of the diagnosis codes identified in the claims data were mapped.

We printed and checked the contents of the medical and facility claim datasets prior to CDM conversion. The contents of the CDM dataset were printed and compared with the list in specification document. Other checks included comparing printouts of observations prior to merge procedures with those after merge procedures, examining frequency tables of created variables, and making sure the distribution of the date of service fell within the study date range.

1. Comparison of the number of people having a sample diagnosis code in NHI data with that in the CDM condition\_occurrence table

The number of people having ICD-9 diagnosis code 722.0 in NHI data = 11,273

The number of people having ICD-9 diagnosis code 722.0 in CDM condition\_occurrence table = 11,273

2. Comparison of some sample records in NHI data before the mapping with the corresponding records in CDM condition\_occurrence table after the mapping. (Individual level data not shown)

### **3.1.4 TABLE NAME: VISIT\_OCCURRENCE**

*Describe how the Visit\_Occurrence mapping and transformation are designed.*

Diagnosis data were drawn from the medical and facility tables.

Data available on medical and facility claims included unique de-identified ID, date of diagnoses/procedures, corresponding ICD-9 diagnosis codes (up to 9), corresponding procedures as identified by ICD-9 procedure, CPT, or HCPCS codes (up to 6), and site of care, as identified by AMA site codes.

All claims were flagged as an inpatient (IP) claim, emergency department (ED) claim, or outpatient (OP) claim using the following hierarchical logic:

1. ED claims were identified by AMA site code 23 or CPT code 99281-99285.
2. Using a standard program that defined periods of inpatient stay, we identified the IP claims as those remaining claims (not flagged as ED) which fell within periods of inpatient stay.
3. The remaining claims not flagged as ED claims and fell outside the period of inpatient stay were identified as OP claims.

There may be multiple observations per individual. Each visit to a unique site of service generated one record. Each unique hospitalization stay period generated one record. There may be identical/overlapping dates among the records (e.g., separate records for OP and ED for the same service date).

Concept identifiers were populated with value zero (0) if no matching mapping from source codes were available.

The field mapping is performed as follows:

Destination Field	Source Field (generic description)	Applied Rule	Comment
VISIT_OCCURRENCE_ID	_n_	system generated unique identifier	
VISIT_START_DATE	ED or OP: Date of service		
VISIT_END_DATE	IP: End of hospitalization	ED or OP: NULL IP: End of hospitalization	
PERSON_ID	ID	de-identified unique id	
VISIT_CONCEPT_ID	AMA site code Date of service End of hospitalization	Use AMA site code, Date of service, and End of hospitalization data to classify IP, ED, OP hosp, OP NEC:  If ED visit then VISIT_CONCEPT_ID=8870 (ED)  Else if date of service within IP range then VISIT_CONCEPT_ID=8717 (IP)  Else if OP hospital then VISIT_CONCEPT_ID=8756 (OP hosp)  Else VISIT_CONCEPT_ID=8677 (OP NEC)	Because MAPPING_TYPE = "VISIT" not in source_to_concept_map, used target_concept_id in 23SEP2009 CR email.
SOURCE_VISIT_CODE	AMA site code Date of service End of hospitalization	ED/IP/OP flag called vtype  if ED then vtype = 'ER';  else if IP then vtype = 'IP';  else vtype = 'OP';	

### Selected validation checks

We checked the frequency of visit type against the concept IDs. We also checked the flags for inpatient and ED visits against visit type to confirm visit type was correctly classified. We examined printouts of claims to see if the flags were set correctly. Distribution of service dates was checked to ensure that dates fell within the specified study date range.

Random observations were printed to ensure variables were correctly created. We checked the contents of all incoming and outgoing datasets and compared the number of observations in the claims data with the number of lines in the final dataset.

Tabulation of visit type as defined by the claims data. The ER visit flag takes precedence over the inpatient visit flag.

ER	IP	vtype	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	0	OP	45995126	86.63	45995126	86.63
0	1	IP	4280272	8.06	50275398	94.69
1	0	ER	2699015	5.08	52974413	99.77
1	1	ER	121630	0.23	53096043	100.00

The following table shows visit type as defined by claims against final concept id.

vtype	VISIT_CONCEPT_ID	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ER	8870	2820645	5.31	2820645	5.31
IP	8717	4280272	8.06	7100917	13.37
OP	8677	38090180	71.74	45191097	85.11
OP	8756	7904946	14.89	53096043	100.00

The range of Visit\_Start\_date in the claims data corresponded to the range of Visit\_Start\_date in the CDM: (01JAN1993 – 26AUG2009).

The range of Visit\_End\_date in the claims data corresponded to the range of Visit\_End\_date in the CDM: (04JAN1993 – 31AUG2009).

### 3.1.5 TABLE NAME: PROCEDURE\_OCCURRENCE

*Describe how the Procedure\_Occurrence mapping and transformation are designed.*

Procedure data were drawn from the medical and facility claims table with procedure codes.

Data available on medical and facility claims included unique de-identified ID, date of diagnoses/procedures, corresponding ICD-9 diagnosis codes (up to 9), corresponding procedures as identified by ICD-9 procedure, CPT, or HCPCS codes (up to 6), and site of care, as identified by AMA site codes.

All claims were flagged as inpatient (IP) claim, emergency department (ED) claim, or outpatient (OP) claim using the following hierarchical logic:

1. ED claims were identified by AMA site code 23 or CPT code 99281-99285.
2. Using a standard program that defined periods of inpatient stay, we identified the IP claims as those remaining claims (not flagged as ED) which fell within periods of inpatient stay.
3. The remaining claims not flagged as an ED claim and fell outside the period of inpatient stay were identified as OP claims.
4. ED, IP, and OP flags were created to identify site of service.

Each claim may contain multiple procedure codes. Data were transposed so there would be one record per procedure code.

There may be multiple records for the same procedure per day per individual, reflecting the underlying claims data.

Concept identifiers were populated with value zero (0) if no matching mapping from source codes were available.

Because all procedure codes in the NHI were compiled without distinction of header/detail/primary, we created 21 new records (PROCEDURE\_OCCURRENCE\_TYPE 500-526) to signify procedure codes in the inpatient, outpatient, and ED settings.

The field mapping is performed as follows:

Destination Field	Source Field (generic description)	Applied Rule	Comment
PROCEDURE_OCCURRENCE_ID	_n_	system generated unique identifier	
PROCEDURE_DATE	Date of service	Date of the procedure	
PERSON_ID	ID	de-identified unique id	
PROCEDURE_CONCEPT_ID	Procedure code (up to 6)	<p>HCPCS:</p> <p>where Procedure code= source_to_concept_map.source_code and source_to_concept_map.mapping_type="PROCEDURE" and source_to_concept_map.source_vocabulary_code="05"</p> <p>example:</p> <p>392763 "A4770 "BLOOD COLLECTION TUBE, VACUUM, FOR DIALYSIS, PER 50 "PROCEDURE" 2615003 "05" "05"</p> <p>ICD-9 PROCEDURE:</p> <p>where Procedure code= source_to_concept_map.source_code and source_to_concept_map.mapping_type="PROCEDURE" and source_to_concept_map.source_vocabulary_code="03"</p> <p>example:</p>	

Destination Field	Source Field (generic description)	Applied Rule	Comment
		<p>"75.34 "Other fetal monitoring "PROCEDURE" 2004811 380514 "03 "03"</p> <p>CPT:</p> <p>where Procedure code= source_to_concept_map.source_code and source_to_concept_map.mapping_type="PROCEDURE" and source_to_concept_map.source_vocabulary_code="04"</p> <p>example:</p> <p>382471 "00670 "Anesthesia for extensive spine and spinal cord procedures (eg, spinal instrumentation or vascular procedures)" "PROCEDURE" 2100917 "04 "04"</p> <p>If no matching values found, store value zero (0).</p>	
SOURCE_PROCEDURE_CODE	Procedure code (up to 6)		
PROCEDURE_OCCURRENCE_TYPE	Procedure code (up to 6)	<p>Procedure code</p> <p>where proc_occurrence_ref.proc_occurrence_type_description match claim site of care (IP, ED or OP) and proc_occurrence_ref.proc_occurrence_position match description in procedure code title</p> <p>CPT</p> <p>if CPT code, match proc_occurrence_ref.proc_occurrence_type_description with claim site of care (IP, ED or OP) where proc_occurrence_position=CPT1.</p>	

### Selected validation checks

Among the sampled NHI population, 829,073 unique individuals had claims for medical visits. In the transposed claims data, there were 55,112,064 observations (829,063 unique individuals) included in the procedure\_occurrence table.

We compared the list of ICD-9 procedure codes, CPT codes, and HCPCS codes in the source\_to\_concept\_map dataset (where mapping\_type="procedure") with the set of

ICD-9 procedure codes, CPT codes, and HCPCS codes listed on the claims data (excluding the list of codes mapped to mapping\_type="procedure drug"). In our claims data, 66.1% of the procedure codes were mapped (25,839 unique procedure codes were found, while 8,755 were not mapped).

We ensured that the distribution of service dates fell within range of enrolled dates and compared the contents of incoming datasets (medical and facility claims) against the contents of the outgoing dataset.

Comparison of date ranges of data in NHI with those in CDM

	NHI Claims Data	CDM Procedure_Occurrence
Minimum Date	1-Jan-93	1-Jan-93
Maximum Date	26-Aug-09	26-Aug-09

### 3.1.6 TABLE NAME: OBSERVATION

*Describe how the Observation mapping and transformation are designed.*

The observation table included laboratory data drawn from the NHI Laboratory table. Variables available in this table included unique de-identified ID for each individual, Logical Observation Identifiers Names and Codes (LOINC), LOINC code descriptor, date of laboratory test, numeric or text result of the test, unit of measurement of the test, low and high range of normal for each test, whether the test value was high or low relative to the normal range, and vendor's description of the test.

1. Overall finding (final, positive, negative, trace, etc.) was not available in NHI. Thus, OBS\_VALUE\_AS\_CONCEPT\_ID was coded to zero for all observations.
2. NHI data also contain a text result variable. When available, these results were stored in OBS\_VALUE\_AS\_STRING. When the OBS\_VALUE\_AS\_NUMBER appear questionable, one may utilize the OBS\_VALUE\_AS\_STRING for guidance in research.
3. As we only sought numeric values in OBS\_VALUE\_AS\_NUMBER, all observations were coded to OBS\_TYPE = "LAB".
4. For assigning concept IDs to units of measurement, we first linked all exact upper-case matches with spaces removed to those in the source\_to\_concept\_map. For the remaining unmatched units, we then tried mapping first to the concept\_class='UCUM Standard' units, then to the concept\_class='UCUM Custom' units provided in the concept dataset. Finally, we reviewed and manually assigned unit target concept ID listed in the source\_to\_concept\_map, if possible, to the remaining unmatched units.

There may be multiple records for the same lab per day per individual, reflecting the underlying claims data.

The field mapping is performed as follows:

Destination Field	Source Field (generic description)	Applied Rule	Comment
OBS_OCCURRENCE_ID	_n_	system generated unique identifier	
PERSON_ID	ID	de-identified unique id	
SOURCE_OBS_CODE	LOINC		
OBS_CONCEPT_ID	LOINC	where LOINC= source_to_concept_map.source_code and source_to_concept_map.mapping_type="OBSERVATION" and source_to_concept_map.source_vocabulary_code="06" (LOINC)  If no matching values found, store the value zero (0)	
OBS_VALUE_AS_NUMBER	Value (number)		
OBS_DATE	Date of service		test result date
OBS_RANGE_LOW	Low		low range of normal
OBS_RANGE_HIGH	High		high range of normal
OBS_TYPE		LAB	
OBS_VALUE_AS_STRING	Result (text)		Additional field where test result may be recorded.
OBS_VALUE_AS_CONCEPT_ID		Non-existent in NHI, coded as zero.	

Destination Field	Source Field (generic description)	Applied Rule	Comment
OBS_UNITS_CONCEPT_ID	Unit	<p>where Unit = source_to_concept_map.source_code and source_to_concept_map.mapping_type="UOM" and source_to_concept_map.source_vocabulary_code="THOMSON" or "GE".</p> <p>If still unmapped, try mapping to units listed in concept dataset.</p> <p>If still unmapped, manually map to those listed in source_to_concept_map.</p> <p>If no matching values found, store the value zero (0)</p>	Used UOM specific for THOMSON or GE.

### Selected validation checks

In our claims data, over 99.9% of the LOINC codes were mapped (5,291 unique LOINC codes were mapped while one LOINC code was not mapped).

In our claims data, 83.9% of unique units were mapped (663 unique units mapped, 127 not mapped).

The contents of the CDM dataset were printed and compared with the list in the specification document. LOINC codes with unknown descriptions were examined along with invalid results. Each of the codes was examined to make sure they mapped to the dictionary. Duplicate records were examined closely for any issues. Other checks included comparing the printouts of observations prior to merge procedures with those after merge procedures, and examining the frequency of created variables. (Individual level data not shown)

### 3.1.7 TABLE NAME: OBSERVATION\_PERIOD

*Describe how the Observation\_Period mapping and transformation are designed.*

Observation data were drawn from the enrollment table. Enrollment information were consolidated by combining records that indicate continuous enrollment over time with no change in medical or pharmacy coverage, while allowing for a 32-day gap in enrollment. From this consolidation, dates of start and end of enrollment for each period were available. As such, this table may contain multiple records for each individual. The observation periods included were those where both medical and drug coverage existed. During periods where individuals had both medical and drug coverage, we assumed that they were "active". For periods where individuals did not have both

medical and pharmacy coverage, no corresponding observations were recorded for those individuals within those calendar periods.

As confidence in using claims data varies by particular outcomes, we did not specify a level of confidence for our data.

The field mapping is as follows:

Destination Field	Source Field (generic description)	Applied Rule	Comment
OBSERVATION_PERIOD_ID	_n_	system generated unique identifier	
OBSERVATION_PERIOD_START_DATE	Start		
OBSERVATION_PERIOD_END_DATE	End		
PERSON_ID	ID	de-identified unique id	
PERSON_STATUS_CONCEPT_ID		where source_to_concept_map.mapping_type="PATIENT STATUS" and source_to_concept_map.source_vocabulary_code="GE" and SOURCE_CODE="ACTIVE"	As there were no PATIENT STATUS specific to NHI, used codes that were specific to GE.
RX_DATA_AVAILABILITY		Y	
DX_DATA_AVAILABILITY		Y	Variable name confirmed (CR 04NOV2009)
HOSPITAL_DATA_AVAILABILITY		Y	Variable name confirmed (CR 04NOV2009)
CONFIDENCE		NULL	Variable name confirmed (CR 04NOV2009)

### Selected validation checks

1. We compared size of incoming and outgoing datasets to ensure consistency.

The total records in enrollment data from NHI = 1,220,981

The total records in CDM observation\_period table = 1,220,981

## 2. We checked that flags for enrollment were set correctly.

PERSON_STATUS_ CONCEPT_ID	Frequency	Percent	Cumulative Frequency	Cumulative Percent
9181	1220981	100.00	1220981	100.00

RX_DATA_ AVAILABILITY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Y	1220981	100.00	1220981	100.00

DX_DATA_ AVAILABILITY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Y	1220981	100.00	1220981	100.00

HOSPITAL_DATA_ AVAILABILITY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Y	1220981	100.00	1220981	100.00

3. We checked the distribution of dates (enrollment start and end dates) and the distribution of status concept IDs. (Individual level data not shown)

### 3.2 Source Independent Data Mapping

*The following mapping processes ought to work independent of the source feed. Describe here if significant changes have to be made.*

Unless otherwise specified in the sections below, Source Independent Data Mapping will follow specifications as defined in ETL Mapping Specification document.

#### 3.2.1 TABLE NAME: DRUG\_ERA

1. From the DRUG\_EXPOSURE table, we determined end dates of drug exposure as follows:
  - a. End dates for dispensed drug exposures were assumed by date of dispensing + days supply.
  - b. For non-orally administered drugs, the end date was set to the date of procedure.
2. Drug eras were constructed based on aggregation of sequential drug exposure periods within individuals.
  - a. All drug dispensings with zero or negative days supply were deleted.
  - b. If there were multiple dispensings of the same drug (as defined by drug\_concept\_id) on the same day, we preferentially chose the dispensing with the longest days supply.
  - c. For each person, drug exposures are grouped by the drug concept and sorted on DRUG\_EXPOSURE\_START\_DATE in ascending order.
  - d. Persistence window defined the maximum number of day's gap between periods related to 2 drug exposures in order to aggregate them into the same

- drug era. Options are zero and 30 days persistence window.
- e. Records were combined together by comparing the start and end dates for the drug exposures. There were 2 types of drug exposure that made records valid for the inclusion into era group in order to be considered part of the same era:
    - i. If the gap between the end date for one drug exposure and the start date for the subsequent drug exposure was within the persistence window.
    - ii. If the end date for one drug exposure was after the start date of the subsequent drug exposure occurrence.
  - f. Start date for the drug era was determined as the minimum start date of all the drug exposures that comprise a drug era range.
  - g. End date for the drug era was determined as the maximum end date of all the drug exposures that comprise a drug era range.
  - h. A drug exposure type was applied to all of the drug eras based on the setting used for the persistence window (zero or 30 days); the logic for the drug exposure type was captured in the field mapping below.
    - i. Created drug eras for each group of dispensings within the drug\_exposure table that met the 0-day persistence window requirement
    - ii. Created drug eras for each group of dispensings within the drug\_exposure table that met the 30-day persistence window requirement

All Drug Eras are recorded in the DRUG\_ERA table based on the following field mapping:

Destination Field	Source Field (generic description)	Applied Rule	Comment
DRUG_ERA_ID	_n_	system generated unique identifier	
DRUG_ERA_ST ART_DATE	Date of service (dispensing or procedure)	Minimum Date of service of all drug exposures that make up era  For 0-day persistence window where there was only one dispensing, use Date of service.	
DRUG_ERA_EN D_DATE		maximum End of eligibility for all drug exposure that make up era  For 0-day persistence window where there was only one dispensing:  Dispensing:  drug_era_end_date=Date of service + Days supply  Procedure:	

Destination Field	Source Field (generic description)	Applied Rule	Comment
		drug_era_end_date=Date of service	
PERSON_ID	ID		
DRUG_EXPOSURE_TYPE		For 0-day window, where drug_exposure_ref.persistence_window=0, (drug_exposure_type="6")  For 30-day window, where drug_exposure_ref.persistence_window=30, (drug_exposure_type="7")	
DRUG_CONCEPT_ID	drug_exposure.DRUG_CONCEPT_ID		
DRUG_EXPOSURE_COUNT	count	number of individual drug exposures consolidated into this era,  For 0-day persistence window where there was only one dispensing, drug_exposure_count=1  Duplicate records in dispensing will be counted as one.	

### Selected validation checks

There were 735,573 unique individuals (making up 33,657,731 observations) included in the drug\_era table.

We printed and checked the contents of the drug table prior to CDM conversion. The contents of the CDM dataset were printed and compared with the list in specification document. Other checks included comparing printouts of observations prior to and after merge procedures, examining the frequency of created variables, and making sure the distribution of the date of service fell within the study date range. Careful checks of observations of selected individuals were also made to demonstrate that each step of the conversion occurred as needed.

### Tabulation of Drug\_exposure\_type in the drug era table

The FREQ Procedure

DRUG_EXPOSURE_TYPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
6	21733591	64.57	21733591	64.57
7	11924140	35.43	33657731	100.00

### 3.2.2 TABLE NAME: CONDITION\_ERA

Condition Era table is constructed through an aggregation of individual Condition Occurrences recorded in the CONDITION\_OCCURRENCE table.

1. Condition eras were constructed based on aggregation of sequential condition occurrence periods within individuals.
  - a. For each person, condition occurrences were grouped by the condition concept and sorted on CONDITION\_START\_DATE in ascending order.
  - b. Persistence window defined the maximum number of day's gap between periods related to 2 condition occurrences in order to aggregate them into the same condition era. The options are '0' and '30' day's persistence window.
  - c. Records were aggregated by comparing the start dates for the condition occurrences. Records can be validly included in the era group and considered for the same era if:
    - i. The gap between the start date for one condition occurrence and the start date for the subsequent condition occurrence was within the persistence window.
  - d. Start date for the condition era was determined as the minimum start date of all the condition occurrences that comprise a condition era range.
  - e. End date for the condition era was determined as the maximum start date of all the condition occurrences that comprise a condition era range.
  - f. A condition occurrence type was applied to all of the condition eras based on the setting used for the persistence window. The condition occurrence types are stored in the CONDITION\_OCCURRENCE\_REF table.
    - i. Created condition eras for each group of diagnosis codes within the condition\_occurrence table that met the 0-day persistence window requirement.
    - ii. Created condition eras for each group of diagnosis codes within the condition\_occurrence table that met the 30-day persistence window requirement.

All Condition Eras are recorded in the CONDITION\_ERA table based on the following field mapping:

Destination Field	Source Field (generic description)	Applied Rule	Comment
CONDITION_ERA_ID	_n_	dystem generated unique identifier	
CONDITION_ERA_START_DATE	Date of service	minimum Date of service of all condition occurrences that make up era  For 0-day persistence window where there was only one diagnosis, use Date of service.	

Destination Field	Source Field (generic description)	Applied Rule	Comment
PERSON_ID	ID		
CONFIDENCE		NULL	
CONDITION_ERA_END_DATE		maximum End of eligibility for all condition occurrences that make up era  For 0-day persistence window where there was only one diagnosis:  condition_era_end_date=Date of service	
CONDITION_CONCEPT_ID	condition_occurrence.DRUG_CONCEPT_ID		
CONDITION_OCCURRENCE_TYPE		For 0-day window, where condition_occurrence_ref.persistence_window=0, (condition_occurrence_type="64")  For 30-day window, where condition_occurrence_ref.persistence_window=30, (condition_occurrence_type="65")	
CONDITION_OCCURRENCE_COUNT	count	number of individual condition occurrences consolidated into this era,  For 0-day persistence window where there was only one diagnosis, condition_occurrence_count=1  Duplicate records of condition occurrences will be counted as one.	

### Selected validation checks

We printed and checked the contents of the condition table prior to CDM conversion. The contents of the CDM dataset were printed and compared to the list in specification document. Other checks included comparing the printouts of observations prior to and after merge procedures, examining the frequency of created variables, and making sure the distribution of the date of service fell within the study date range. To ensure the eras were correctly set, select conditions for certain individuals were compared with the resulting condition eras.

1. We compared the number of people having a sample diagnosis code in condition\_occurrence with that in the condition\_era tables.

The number of people having a sample diagnosis 722.0 (condition\_concept\_id = 74725 ) in condition\_occurrence = 11,273

The number of people having a sample diagnosis 722.0 (condition\_concept\_id = 74725 ) in condition\_era\_0 = 11,273

The number of people having a sample diagnosis 722.0 (condition\_concept\_id = 74725 ) in condition\_era\_30 = 11,273

- We compared sample records in condition\_occurrence with the corresponding records in condition\_era tables. (Individual level data not shown)

### 3.3 Reference Tables

The following contain reference tables that were derived from the OMOP Thomson and GE source data. They reflect the content of those databases. It is assumed that you will update these tables to describe your data more adequately.

#### 3.3.1 TABLE NAME: DRUG\_EXPOSURE\_REF

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

#### 3.3.2 TABLE NAME: CONDITION\_OCCURRENCE\_REF

Condition Occurrence Reference table serves as the reference listing of various types of Condition Occurrences recorded for analysis. The Condition Occurrence Type conveys the indicator(s) from which the Condition Occurrence was captured and defines the characteristic of the occurrence and the level of aggregation.

This table is loaded based on a reference list of Occurrence types, descriptions and persistence window settings. The current listing is as follows (i3 added in blue):

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	

Condition Occurrence Type	Condition Occurrence Type Description	Condition Occurrence Position	Persistence Window (in days)
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	
31	Inpatient Header	14	
32	Inpatient Header	15	

Condition Occurrence Type	Condition Occurrence Type Description	Condition Occurrence Position	Persistence Window (in days)
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	

Condition Occurrence Type	Condition Occurrence Type Description	Condition Occurrence Position	Persistence Window (in days)
62	Outpatient Header	15	
63	Problem List		
64	Condition Era		0
65	Condition Era		30
66	Death at Discharge		
500	IP	1	
501	IP	2	
502	IP	3	
503	IP	4	
504	IP	5	
505	IP	6	
506	IP	7	
507	IP	8	
508	IP	9	
509	ED	1	
510	ED	2	
511	ED	3	
512	ED	4	
513	ED	5	
514	ED	6	
515	ED	7	
516	ED	8	
517	ED	9	
518	OP	1	
519	OP	2	
520	OP	3	
521	OP	4	
522	OP	5	

Condition Occurrence Type	Condition Occurrence Type Description	Condition Occurrence Position	Persistence Window (in days)
523	OP	6	
524	OP	7	
525	OP	8	
526	OP	9	

### 3.3.3 TABLE NAME: PROC\_OCCURRENCE\_REF

Procedure Occurrence Reference table serves as the reference listing of various types of Procedure Occurrences recorded for analysis. The Procedure Occurrence Type conveys the indicator(s) from which the Procedure Occurrence was captured, and defines the characteristic of the occurrence.

This table is loaded based on a reference list of occurrence types, position and descriptions. The current listing is as follows (i3 added in blue):

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

Procedure Occurrence Type	Procedure Occurrence Type Description	Procedure Occurrence Position
17	Inpatient Header	14
18	Inpatient Header	15
19	Outpatient Detail	Primary
20	Outpatient Detail	1
21	Outpatient Header	Primary
21	Outpatient Header	1
22	Outpatient Header	2
23	Outpatient Header	3
24	Outpatient Header	4
25	Outpatient Header	5
26	Outpatient Header	6
27	EMR Order	
500	IP	1
501	IP	2
502	IP	3
503	IP	4
504	IP	5
505	IP	6
506	IP	CPT1
507	ED	1
508	ED	2
509	ED	3
510	ED	4
511	ED	5
512	ED	6
513	ED	CPT1
514	OP	1
515	OP	2

Procedure Occurrence Type	Procedure Occurrence Type Description	Procedure Occurrence Position
516	OP	3
517	OP	4
518	OP	5
519	OP	6
520	OP	CPT1

### 3.3.4 TABLE NAME: OBSERVATION\_TYPE\_REF

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
- EHR observations with text results (e.g., reason for visit)
- Chief Complaint

Data in the OBSERVATION\_TYPE\_REF table is as follows:

Observation Type	Observation Type Description
CHC	Chief Complaint
EHR	Observation recorded from Electronic Health Records
LAB	Lab Observation
PRL	Problem List from Electronic Health Records
TEM	Observation recorded from Electronic Health Records with text results

### 3.3.5 TABLE NAME: VOCABULARY\_REF

The Vocabulary Reference entity includes a list of all standard terminologies from which Concepts have been extracted for observational analysis using the Common Data Model. The reference table is populated with a single record for each Vocabulary source and includes a descriptive name for the Vocabulary source.

For a complete listing of the VOCABULARY\_REF table see Standard Terminology Specifications Document.

i3 added (in blue):

VOCABULARY_CODE	VOCABULARY_NAME
01	SNOMED
02	ICD9 CM
03	ICD9 Procedure
04	CPT
05	HCPCS
06	LOINC
07	NDFRT
08	RxNorm
09	NDC
52	THOMSON
51	GE
15	MedDRA
10	GPI
54	OMOP Intermediate Concept – Drug
55	OMOP Generic
53	OMOP Intermediate Concept-Condition
11	UCUM
12	HL7 ADMINISTRATIVE SEX
13	CDC RACE/ETHNICITY
14	CMS PLACE of SERVICE
500	i3

### 3.3.6 TABLE NAME: RELATIONSHIP\_TYPE

A Concept Relationship is standardized via the Relationship Type entity. The Relationship Type codes are adopted from SNOMED-CT. Where the relationships are hierarchical, the Relationship Type captures the “IS A” string that identifies it as a Subtype relationship. Where the relationship is an Object Attribute Value relationship, the Relationship Type holds the Concept that defines the Attribute.

For a complete listing of the RELATIONSHIP\_TYPE table see Standard Terminology Specifications Document.