ETL mapping between MIMIC3_demo and OMOP condtition_occurence table

The goal of this project is to populate the OMOP CONDITION_OCCURENCE table with MIMIC data

To successfully implement the ETL process, I will use the following steps:

1. Understand the source and the target data models
2. Profile the source data to ensure that the data quality is high enough to be worth using
3. Establish the ETL mapping both at the level of tables (structural) and values (semantic)
4. Write the ETL code
5. Execute the ETL code
6. Assess the quality of the resulting tables using data quality measures
7. Finalize documentation

1.Understanding the target CONDITION_OCCURENCE table

Information about the three fields that need to be mapped can be found in the OMOP wiki https://github.com/OHDSI/CommonDataModel/wiki/CONDITION_OCCURRENCE :

- person_id: A foreign key identifier to the Person who is experiencing the condition. The demographic details of that Person are stored in the PERSON table.
- visit_occurence_id: A foreign key to the visit in the VISIT_OCCURRENCE table during which the Condition was determined (diagnosed).
- Condition_source_value: The source code for the Condition as it appears in the source data. This code is mapped to a Standard Condition Concept in the Standardized Vocabularies and the original code is stored here for reference.
- BONUS: condition_start_date: The start of the Condition, date type.

The diagnoses_icd table

Table source: Hospital database.

Table purpose: Contains ICD diagnoses for patients, most notably ICD-9 diagnoses.

Number of rows: 651,047

Links to:

- PATIENTS on `SUBJECT_ID`
- ADMISSIONS on `HADM_ID`
- D_ICD_DIAGNOSES on `ICD9_CODE`

Important considerations

- The ICD codes are generated for billing purposes at the end of the hospital stay.
- All ICD codes in MIMIC-III are ICD-9 based. The Beth Israel Deaconess Medical Center will begin using ICD-10 codes in 2015.
- The code field for the ICD-9-CM Principal and Other Diagnosis Codes is six characters in length, with the decimal point implied between the third and fourth digit for all diagnosis codes other than the V codes. The decimal is implied for V codes between the second and third digit.

Table columns

Name	PostgreSQL data type	Modifiers
ROW_ID	INT	not null
SUBJECT_ID	INT	not null
HADM_ID	INT	not null
SEQ_NUM	INT	
ICD9_CODE	VARCHAR(10)	

1. Understanding the Source MIMIC tables

Information about the MIMIC tables can be found at : https://mimic.physionet.org/mimictables/diagnoses_icd/

- SUBJECT_ID and HADM_ID: Identifiers which specify the patient. SUBJECT_ID is unique to a patient and HADM_ID is unique to a patient hospital stay
- ICD9_CODE: contains the actual code corresponding to the diagnosis assigned to the patient for the given row. Note that all codes, as of MIMIC-III v1.0, are ICD-9 codes.

2. White Rabbit Scan Report

White Rabbit data profiling of the Diagnoses table shows that DIAGNOSES_ICD table for the full MIMIC dataset has 651047 rows, of which the profiler checked 100000. The only column with empty values is the ICD9_CODE column which has 8 rows with missing values from the rows checked.

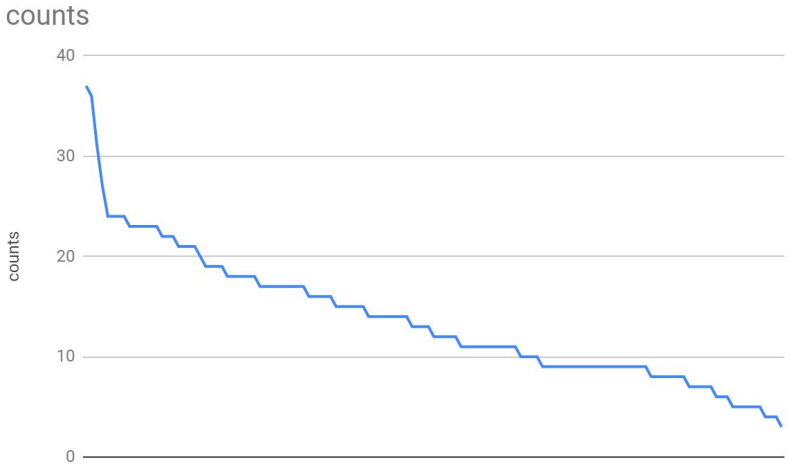
In conclusion, there are not a lot of missing values and it’s worth moving on to the next step. However, the ETL code should include the case of missing values in the ICD9_CODE column.

diagnoses_icd	row_id	integer	6	651047	100000	0
diagnoses_icd	subject_id	integer	5	651047	100000	0
diagnoses_icd	hadm_id	integer	6	651047	100000	0
diagnoses_icd	seq_num	integer	2	651047	100000	0.00008
diagnoses_icd	icd9_code	character varying	5	651047	100000	0.00008

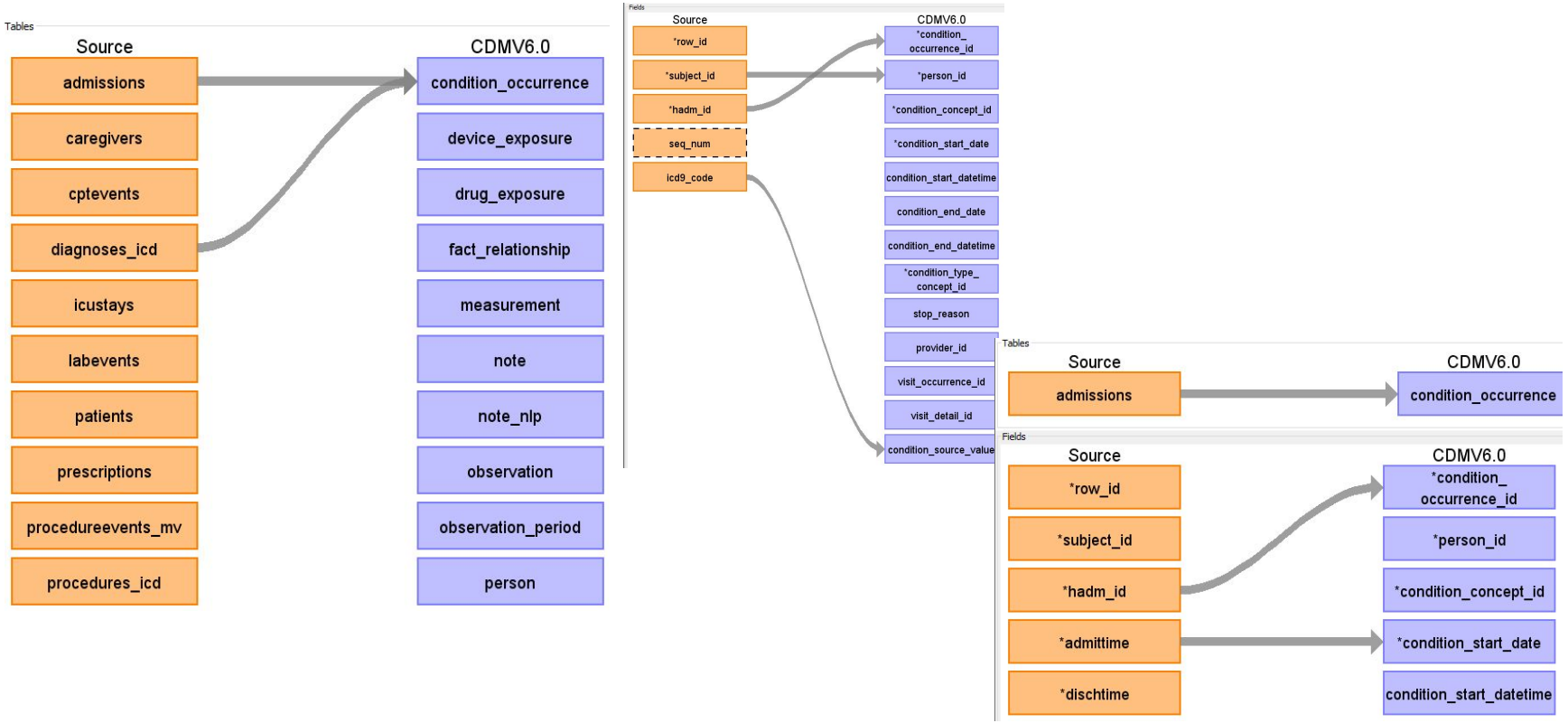
Profiling of SUBJECT_ID

In the Diagnosis_ICD table, the White Rabbit scan report shows a SUBJECT_ID, 109 with 103 Diagnoses. This is due to multiple hospital admissions.

The number of hospital admissions follows a long tail distribution with a sharp peak due to a few hospital admissions of very sick patients with a lot of different diagnoses.



3. ETL mapping between MIMIC3 and OMOP: Rabbit in a Hat



3. ETL mappings OMOP: MIMIC3

- person_id: DIAGNOSES_ICD.SUBJECT_ID
- visit_occurence_id: DIAGNOSES_ICD.HADM_ID
- condition_source_value: DIAGNOSES_ICD.ICD9_CODE
- BONUS: condition_start_date: ADMISSIONS.ADMITTIME

4. The ETL code

```
/* ETL mapping of MIMIC3 to OMOP CONDITION_OCCURENCE */

WITH

    cond_occurence1 AS (SELECT diag.SUBJECT_ID AS person_id, diag.HADM_ID AS
condition_occurence_id, diag.ICD9_CODE as condition_source_value, adm.ADMITTIME AS
condition_start_date

                    FROM mimic3_demo.ADMISSIONS adm

                    JOIN mimic3_demo.DIAGNOSES_ICD diag USING (HADM_ID))

SELECT * FROM cond_occurence1
```

5. Results: OMOP table with MIMIC data

Row	person_id	condition_occurence_id	condition_source_value	condition_start_date
1	10043	168674	486	2185-04-14T00:23:00
2	10043	168674	00845	2185-04-14T00:23:00
3	10043	168674	2875	2185-04-14T00:23:00
4	10043	168674	28529	2185-04-14T00:23:00
5	10043	168674	49121	2185-04-14T00:23:00
6	10043	168674	51881	2185-04-14T00:23:00
7	10043	168674	42831	2185-04-14T00:23:00
8	10043	168674	4280	2185-04-14T00:23:00
9	10094	168074	2554	2180-02-29T18:54:00
10	10094	168074	25000	2180-02-29T18:54:00
11	10094	168074	3970	2180-02-29T18:54:00
12	10094	168074	2273	2180-02-29T18:54:00
13	10094	168074	4280	2180-02-29T18:54:00
14	10094	168074	486	2180-02-29T18:54:00
15	10094	168074	70706	2180-02-29T18:54:00

Table JSON

First < Prev Rows 1 - 15 of 1761 Next >

6. Assessing data quality

- Checking that the same number of rows in the mimic3_demo.DIAGNOSES_ICD (all the data) is present in the new table: 1761 records

```
/* ETL mapping of MIMIC3 to OMOP CONDITION_OCCURENCE */
WITH cond_occurence1 AS (SELECT SUBJECT_ID AS person_id,
                                HADM_ID AS visit_occurence_id,
                                ICD9_CODE AS condition_source_value
                         FROM mimic3_demo.DIAGNOSES_ICD)
SELECT COUNT(*) FROM cond_occurence1
```

Row	f0_
1	1761

6. Assessing data quality

- There are 100 distinct person_id in the new condition_occurence table & there were 100 distinct subjects in the Diagnoses_ICD table