HiRID to OMOP CDM report

By: Julio Portella

Index

- HiRID dataset
- OMOP-CDM
- ETL Process
- Vocabulary translation
- Technical Implementation
- Quality Assurance
- Results
- Observations
- Next Steps

HiRID Database

- Dataset from the Department of Intensive Care Medicine of Bern University Hospital
  - Continuous measurements every 2 minutes
  - Contains Lab values, administered drugs, fluids and nutrition
  - Measurements from bedside monitoring
  - Measurements and settings of medical devices
- Currently available at Physionet (Requires Credentials)
- It has 776’921,131 observations, close to a billion. MIMIC has 312 million of observations
- https://physionet.org/content/hirid/1.0/

HiRID Database

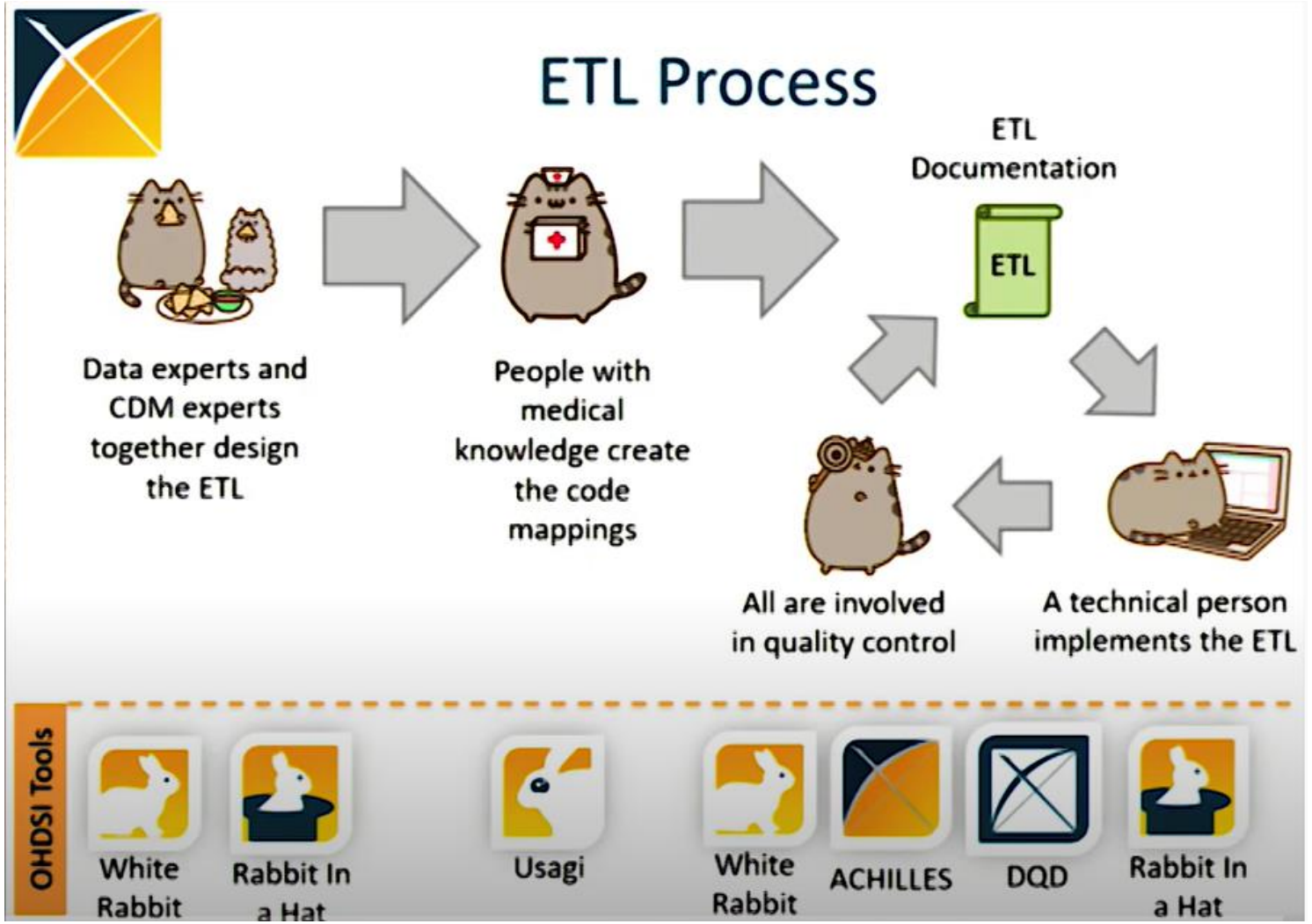
- The dataset contains data of 34 thousand people
- From this group, 64.2% are male and 35.8% are female
- All the patients are over 40 and the median is 65 years old
- The observations and pharmaceutical records are divided in 250 groups, each group has a different amount of random patients.
- There is a summarized version of the HiRID database with less variables and downsampled data used in their published paper. See: https://www.nature.com/articles/s41591-020-0789-4
- The downsampled data has measurements every 5 minutes

OMOP CDM

- Managed by the Observational Health Data Sciences and Informatics (OHDSI)
- Most popular common data model
- Created in 2008 and funded by the FDA
- In 2014 moved to OHDSI at Columbia University
- Queries and tools freely available and Open Sourcce
- Patient centered data model
- Currently in version 6.0.0
- Designed for Relational databases but it can be translated to columns databases and Apache Hive

ETL Process

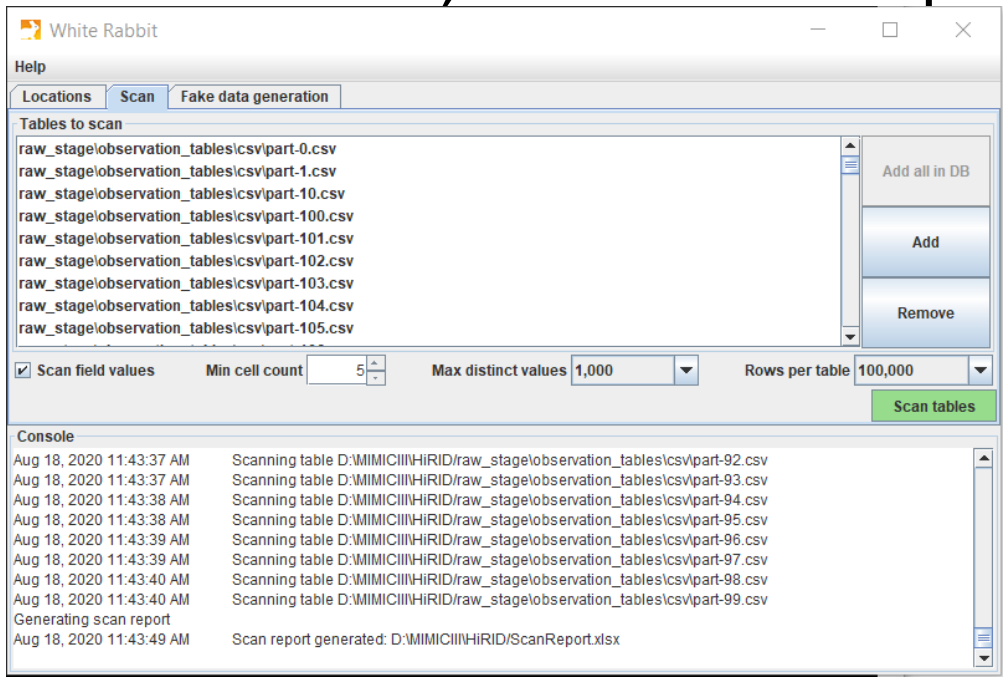
- Cyclical process with 5 parts
  - Design
  - Code mapping
  - Documentation
  - Implementation
  - Validation
- This process was used to translate the database



https://www.youtube.com/watch?v=4vytGDrhHpk

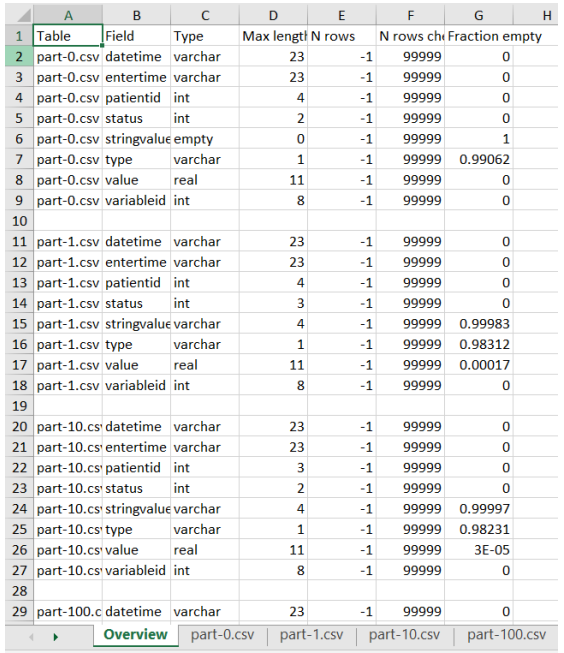
ETL Process

- WhiteRabbit: Tool used to map all the values from a csv file or a database connection
- It scan details can be configured, the more detailed the count the more it will take to scan
- In this case, each csv file represents a table



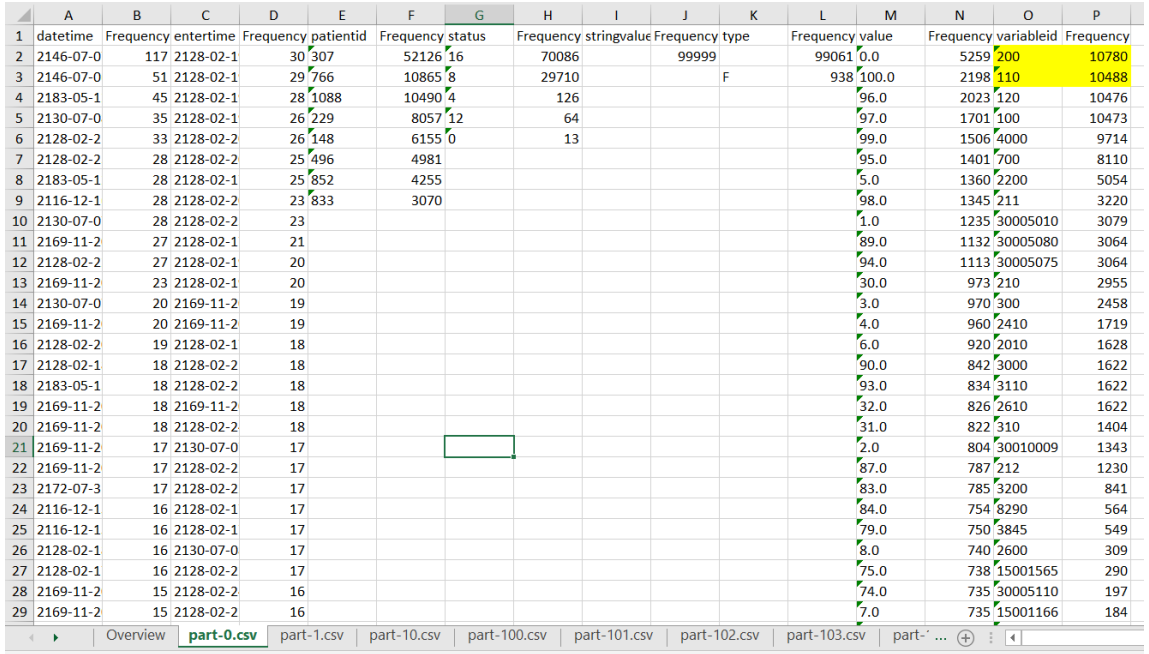
ETL Process

- Once the scan is done it generates an excel spreadsheet named ScanReport.xlsx
- The Scan Report has an overview and an analysis of the most popular values for each table.
- Very useful to know which observations and drugs are the most used



	A	B	C	D	E	F	G	H
1	Table	Field	Type	Max lengt	N rows	N rows ch	Fraction empty	
2	part-0.csv	datetime	varchar	23	-1	99999	0	
3	part-0.csv	entertime	varchar	23	-1	99999	0	
4	part-0.csv	patientid	int	4	-1	99999	0	
5	part-0.csv	status	int	2	-1	99999	0	
6	part-0.csv	stringvalue	empty	0	-1	99999	1	
7	part-0.csv	type	varchar	1	-1	99999	0.99062	
8	part-0.csv	value	real	11	-1	99999	0	
9	part-0.csv	variableid	int	8	-1	99999	0	
10								
11	part-1.csv	datetime	varchar	23	-1	99999	0	
12	part-1.csv	entertime	varchar	23	-1	99999	0	
13	part-1.csv	patientid	int	4	-1	99999	0	
14	part-1.csv	status	int	3	-1	99999	0	
15	part-1.csv	stringvalue	varchar	4	-1	99999	0.99983	
16	part-1.csv	type	varchar	1	-1	99999	0.98312	
17	part-1.csv	value	real	11	-1	99999	0.00017	
18	part-1.csv	variableid	int	8	-1	99999	0	
19								
20	part-10.cs	datetime	varchar	23	-1	99999	0	
21	part-10.cs	entertime	varchar	23	-1	99999	0	
22	part-10.cs	patientid	int	3	-1	99999	0	
23	part-10.cs	status	int	2	-1	99999	0	
24	part-10.cs	stringvalue	varchar	4	-1	99999	0.99997	
25	part-10.cs	type	varchar	1	-1	99999	0.98231	
26	part-10.cs	value	real	11	-1	99999	3E-05	
27	part-10.cs	variableid	int	8	-1	99999	0	
28								
29	part-100.c	datetime	varchar	23	-1	99999	0	

Overview | part-0.csv | part-1.csv | part-10.csv | part-100.csv

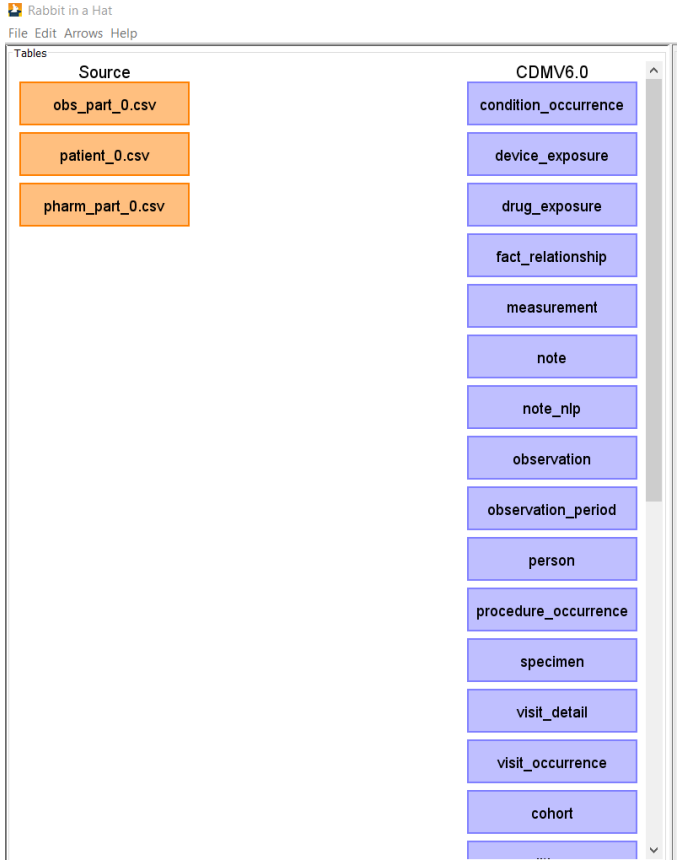


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	datetime	Frequency	entertime	Frequency	patientid	Frequency	status	Frequency	stringvalue	Frequency	type	Frequency	value	Frequency	variableid	Frequency
2	2146-07-0	117	2128-02-1	30	307	52126	16	70086		99999		99061	0.0	5259	200	10780
3	2146-07-0	51	2128-02-1	29	766	10865	8	29710			F	938	100.0	2198	110	10488
4	2183-05-1	45	2128-02-1	28	1088	10490	4	126					96.0	2023	120	10476
5	2130-07-0	35	2128-02-1	26	229	8057	12	64					97.0	1701	100	10473
6	2128-02-2	33	2128-02-2	26	148	6155	0	13					99.0	1506	4000	9714
7	2128-02-2	28	2128-02-2	25	496	4981							95.0	1401	700	8110
8	2183-05-1	28	2128-02-1	25	852	4255							5.0	1360	2200	5054
9	2116-12-1	28	2128-02-2	23	833	3070							98.0	1345	211	3220
10	2130-07-0	28	2128-02-2	23									1.0	1235	30005010	3079
11	2169-11-2	27	2128-02-1	21									89.0	1132	30005080	3064
12	2128-02-2	27	2128-02-1	20									94.0	1113	30005075	3064
13	2169-11-2	23	2128-02-1	20									30.0	973	210	2955
14	2130-07-0	20	2169-11-2	19									3.0	970	300	2458
15	2169-11-2	20	2169-11-2	19									4.0	960	2410	1719
16	2128-02-2	19	2128-02-1	18									6.0	920	2010	1628
17	2128-02-1	18	2128-02-2	18									90.0	842	3000	1622
18	2183-05-1	18	2128-02-2	18									93.0	834	3110	1622
19	2169-11-2	18	2169-11-2	18									32.0	826	2610	1622
20	2169-11-2	18	2128-02-2	18									31.0	822	310	1404
21	2169-11-2	17	2130-07-0	17									2.0	804	30010009	1343
22	2169-11-2	17	2128-02-2	17									87.0	787	212	1230
23	2172-07-3	17	2128-02-2	17									83.0	785	3200	841
24	2116-12-1	16	2128-02-1	17									84.0	754	8290	564
25	2116-12-1	16	2128-02-1	17									79.0	750	3845	549
26	2128-02-1	16	2130-07-0	17									8.0	740	2600	309
27	2128-02-1	16	2128-02-2	17									75.0	738	15001565	290
28	2169-11-2	15	2128-02-2	16									74.0	735	30005110	197
29	2169-11-2	15	2128-02-2	16									7.0	735	15001166	184

Overview | part-0.csv | part-1.csv | part-10.csv | part-100.csv | part-101.csv | part-102.csv | part-103.csv | part-' ...

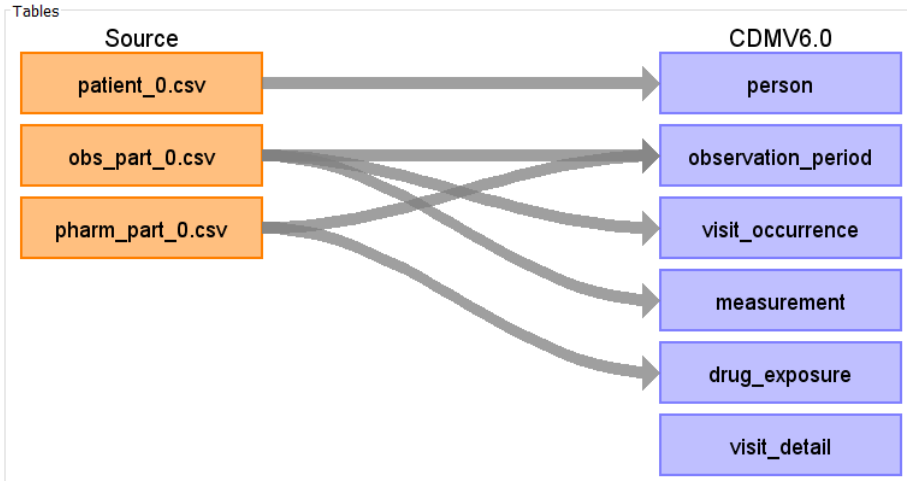
ETL Process

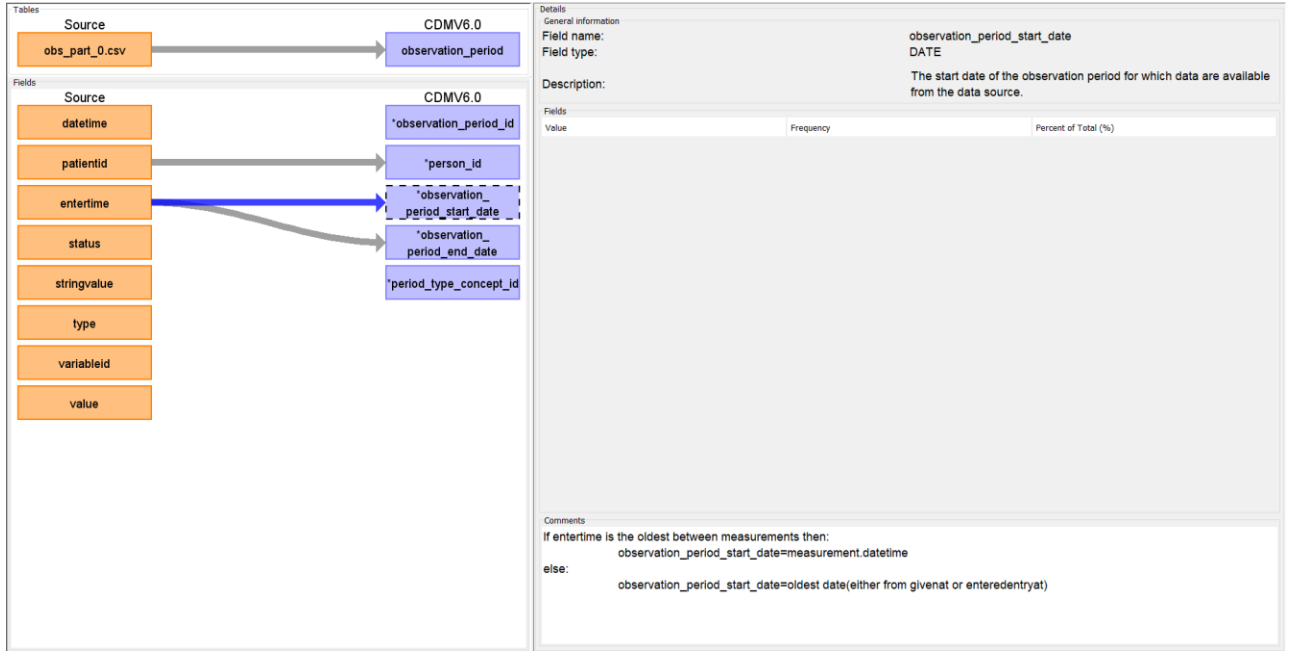
- Rabbit In a Hat: Tool that generates the ETL document.
- Use the ScanReport generated in “White Rabbit” generate the ETL map. Make sure that the ScanReport contains only the required tables



ETL Process

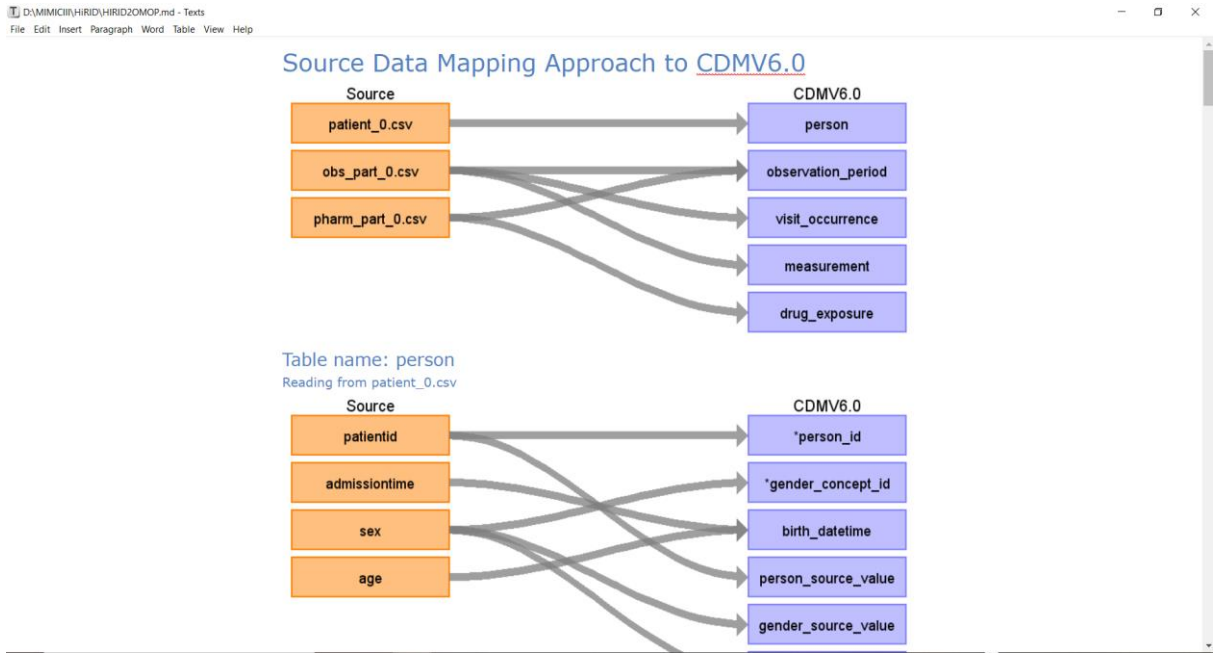
- Given that HiRID has 3 tables, many OMOP tables and fields had to be deduced.
- Tables like measurement, drug_exposure and person can be translated directly from HiRID
- Observation_period and visit_occurrence had to be deduced





ETL Process

- Once the connections and mapping logic was done, the ETL documentation can be generated
- White Rabbit also generates the ETL test framework
- The generated ETL document is a reference, it is expected to change it during the whole process



ETL Process

- Many deductions were done using “Rabbit in a Hat”
- Patient’s birthday were deduced from the admission datetime minus their age at the moment
- The observational period start and end date are from the oldest and newest date from the observations and drugs reports
- The table visit_occurrence has a similar logic but it only uses the dates from the observations
- Drug_exposure’s refills column was deduced by counting the number of times that a drug was taken.
- Many terms had to be translated using OMOP CDM vocabularies or were deduced by the context of the database.

Vocabulary translation

- 70% of the vocabulary was translated, 2 missing observations and 195 drugs
- Missing data by vocabulary items

Vocabulary Type	% of the total data
Observations	0.025% of all observations
Pharmaceuticals	13% of all medications

- Most of the observations were translated into LOINC, the rest used SNOMED

Technical Implementation

- Before translating the tables, the database was created using OMOP CDM queries
- Most relevant tools are DQD and Achilles
- Given that DQD is not available for OMOP v6, Achilles is going to be used
- Given that

Technical Implementation

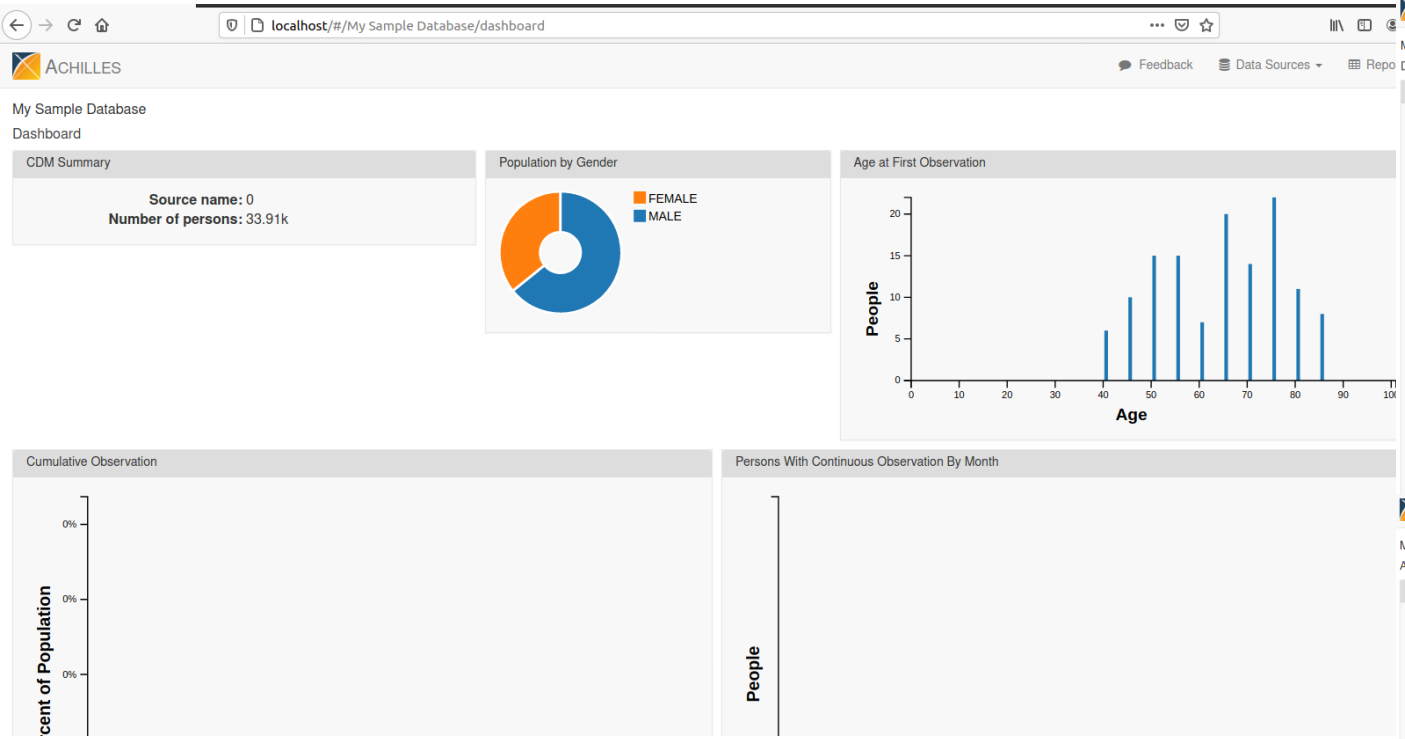
- After translating the dataset and adding the constraints into the database, quality assurance tools were used
- The technical implementation was done using Python
- The data was uploaded and processed using Panda’s dataframe
- Rows with unmappable elements and errors were removed
- After creating the tables, they were bulk uploaded into the databse
- Once the data was uploaded the foreign keys and restrictions were added into the databse

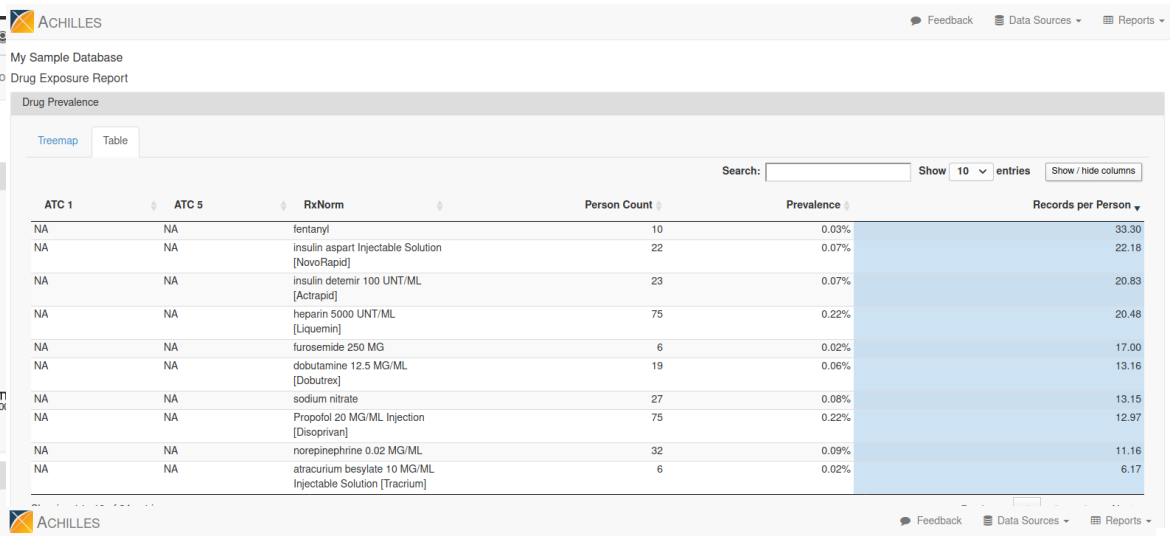
Quality Assurance

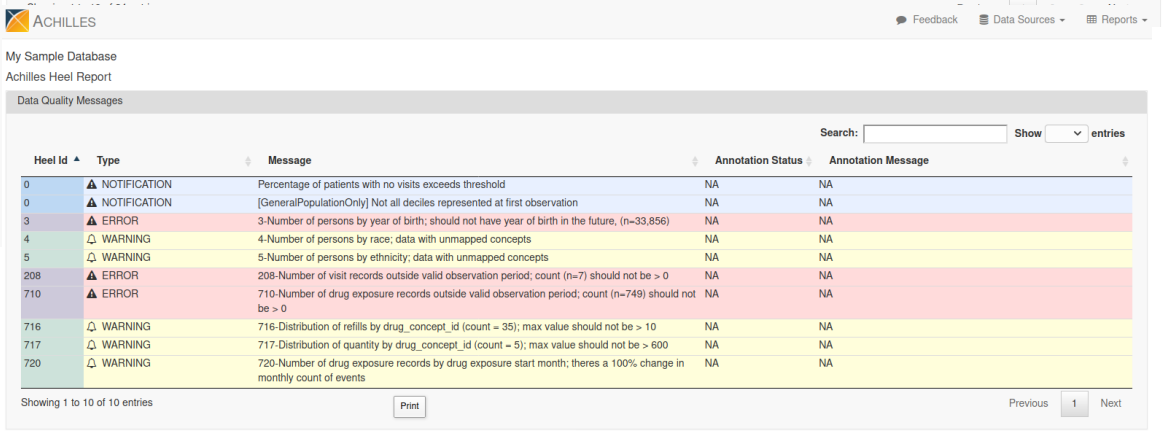
- Before translating the tables, the database was created using OMOP CDM queries
- Most relevant tools are DQD and Achilles
- Given that DQD is not available for OMOP v6, Achilles is going to be used
- Using the R library for Achilles, the analysis was positive
- For Achilles Web, the data was sampled. The algorithm has a mode that allows the user to upload only 10% of the existing data

Results

- Translation successful with minor errors and warnings







Results

- Observations (HiRID)

datetime	entertime	patientid	status	stringvalue	type	value	variableid
00:00.0	20:27.0	148	8			0	30005080
00:00.0	16:11.6	148	8			0	30005110
00:00.0	20:26.9	148	8			0	30005010
00:00.0	20:27.0	148	8			0	30005075
35:00.0	36:37.7	148	12			73	4000
35:00.0	36:37.6	148	12			98	200
36:00.0	37:37.7	148	8			94	4000
36:00.0	37:37.7	148	8			118	200
37:00.0	38:37.2	148	12			-2.5	211

- Observations (OMOP-CDM)

Data Output | Explain | Messages | Notifications

	measurement_id [PK] bigint	person_id bigint	measurement_concept_id integer	measureme date	measurement_datetime timestamp without time z	measurement_time character varying (1(	measurement_type_ integer	opera integ	value_as_numbe numeric	valu integ	unit_ integ	rang num	rang num	prov bigir	visit_occurr bigint	visit bigir	measurement character vary	measurement_source integer
1	5598	148	36309521	2183-05-13	2183-05-13 08:34:02.35	08:34:02	42530833	[null]	0.0	[null]	[null]	[null]	[null]	[null]	1	[null]	0.0	210
2	4944	148	3002137	2183-05-13	2183-05-13 05:53:58.96	05:53:58	42530833	[null]	1846.8541	[null]	[null]	[null]	[null]	[null]	1	[null]	1846.8541	30005010
3	4681	148	3013502	2183-05-13	2183-05-13 04:51:07.09	04:51:07	42530833	[null]	97.0	[null]	[null]	[null]	[null]	[null]	1	[null]	97.0	4000
4	3887	148	36309521	2183-05-13	2183-05-13 01:46:02.24	01:46:02	42530833	[null]	0.5	[null]	[null]	[null]	[null]	[null]	1	[null]	0.5	210
5	617	148	36309521	2183-05-12	2183-05-12 13:43:00	13:43:00	42530833	[null]	0.6	[null]	[null]	[null]	[null]	[null]	1	[null]	0.6	211
6	2378	148	21499295	2183-05-12	2183-05-12 19:54:20.14	19:54:20	42530833	[null]	0.0	[null]	[null]	[null]	[null]	[null]	1	[null]	0.0	30005080
7	1167	148	3013502	2183-05-12	2183-05-12 14:42:00	14:42:00	42530833	[null]	96.0	[null]	[null]	[null]	[null]	[null]	1	[null]	96.0	4000
8	793	148	36309521	2183-05-12	2183-05-12 14:02:00	14:02:00	42530833	[null]	0.7	[null]	[null]	[null]	[null]	[null]	1	[null]	0.7	211

Results

- Pharma Records (HiRID)

datetime	entertime	patientid	status	stringvalue	type	value	variableid
00:00.0	20:27.0	148	8			0	30005080
00:00.0	16:11.6	148	8			0	30005110
00:00.0	20:26.9	148	8			0	30005010
00:00.0	20:27.0	148	8			0	30005075
35:00.0	36:37.7	148	12			73	4000
35:00.0	36:37.6	148	12			98	200
36:00.0	37:37.7	148	8			94	4000
36:00.0	37:37.7	148	8			118	200
37:00.0	38:37.2	148	12			-2.5	211

- Drug_exposure (OMOP-CDM)

	drug_exposure_id [PK] bigint	person_id bigint	drug_concept_id integer	drug_exposu date	drug_exposure_start_datetime timestamp without time zone	drug_exposu date	drug_exposure_end_datetime timestamp without time zone	verbatim_end_date date	drug_type_concept_id integer	stop_reason character va	refills integer	quantity numeric	days_supply integer	sig tex
1	10559	2470	40953611	2155-04-23	2155-04-23 04:11:11.71	2155-04-23	2155-04-23 04:11:11.71	2155-04-23	38000180	0	483	4.166664	0	0
2	24177	11317	1154029	2158-12-10	2158-12-10 07:27:21.62	2158-12-10	2158-12-10 07:27:21.62	2158-12-10	38000180	0	129	2.430555	0	0
3	65263	29217	43712853	2187-02-11	2187-02-11 21:21:32.48	2187-02-11	2187-02-11 21:21:32.48	2187-02-11	38000180	0	16	2.65	0	0
4	15891	6586	35130768	2133-07-31	2133-07-31 07:21:51.1	2133-07-31	2133-07-31 07:21:51.1	2133-07-31	38000180	0	1	0.0	0	0
5	16356	7638	40953611	2150-09-10	2150-09-10 06:13:46.56	2150-09-10	2150-09-10 06:13:46.56	2150-09-10	38000180	0	67	04.62963	0	0
6	53266	24171	19036476	2113-07-16	2113-07-16 08:00:00	2113-07-16	2113-07-16 08:11:25.683	2113-07-16	38000180	0	0	60.0	0	0
7	50899	21805	40953611	2125-09-04	2125-09-04 10:11:35.43	2125-09-04	2125-09-04 10:11:35.43	2125-09-04	38000180	0	19	0000001	0	0
8	363	307	40953611	2128-02-17	2128-02-17 17:46:17.96	2128-02-17	2128-02-17 17:46:17.96	2128-02-17	38000180	0	14	4.166664	0	0

Observations

- The number of rows for observation dataset is in the 100 thousands, uploading 1 group from jupyter dataset to postgres takes around 3 minutes
- With the current methods applied the total upload time will take 21 hours
- The ETL is a cyclical process, there were many iterations before getting the final version of the algorithm
- The table observational_period was created because there was an observation from Achilles web
- There are fields that are related with procedures like chemotherapy, they have to mapped as procedures, thus creating a new table
- The date of birth transformation is right. However, the Achilles json file only shows patients over 40. More research has to be done,

Next Steps

- Publish the algorithm to the OHDSI community and anyone interested
- Upload the data to the cloud and measure the performance. Currently the work was done locally in a Laptop(Core i7, 32GB of RAM)