ED-STIC - UP-SUD
CNRS - LIMSI (Pierre Zweigenbaum)
INSERM - LIMICS (Xavier Tannier)
AP-HP (Christel Daniel)

Merging Intensive Care Unit Databases

Nicolas Paris

May 9, 2018

Content

MIMIC–OMOP Context

ETL

NLP Pipelines

Datathon

Perspectives

Conclusion

MIMIC–OMOP Context

Context

- hospitals produce clinical databases (DBs) useful for research
- patients demographics, observation metrics, medications, free text reports and more
- merging DBs together would increase power of scientific conclusions

→ US, France, Brazil are federating around MIMIC project

Challenges

DBs are heterogeneous in two ways:

1. data (schema matching):
   - data models, schemas (different practices)
   - languages (free text reports, documentations)
   - measure units (mg, kg . . .)
2. metadata (ontology matching):
   - terminologies, coding systems (sparse use of standards)
   - languages (coding systems, labels, descriptions)

MIMIC OMOP TRANSLATION

Motivations

- highly granular dataset
- demographics, billing codes, labs, notes, medication, charts, waveforms . . .
- easy access: supports many research
- a specific data/terminology model: makes data federation complex

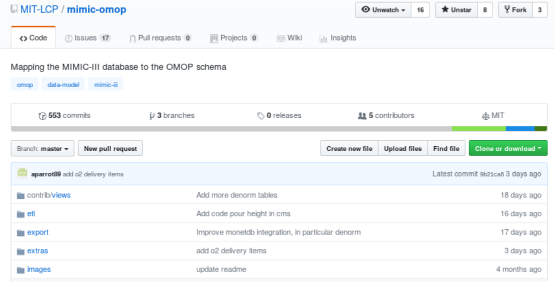
→ time for MIMIC to be translated into CDM

ETL

MIMIC OMOP TRANSLATION

Methods

- collaborative ETL (github.com/MIT-LCP/mimic-omop)
- only SQL code to transform MIMICiii → OMOP5.3
- improve concept mapping by editing csv
- UIMA pipelines to feed omop.NOTE_NLP table



MIT-LCP / **mimic-omop**

Mapping the MIMIC-III database to the OMOP schema

omop data-model mimic-iii

553 commits · **3** branches · **0** releases · **5** contributors · MIT

aparrot89 add o2 delivery items — Latest commit 9b21ca8 3 days ago

contrib/views	Add more denorm tables	18 days ago
etl	Add code pour height in cms	16 days ago
export	Improve monetdb integration, in particular denorm	17 days ago
extras	add o2 delivery items	3 days ago
images	update readme	4 months ago

MIMIC OMOP TRANSLATION
Methods

```
INSERT INTO omop.measurement
SELECT transform(...)
FROM mimiciii.chartevents
JOIN mimiciii.transfers ON (...)
JOIN mimiciii.d_items ON (...)
WHERE filter(...)
```

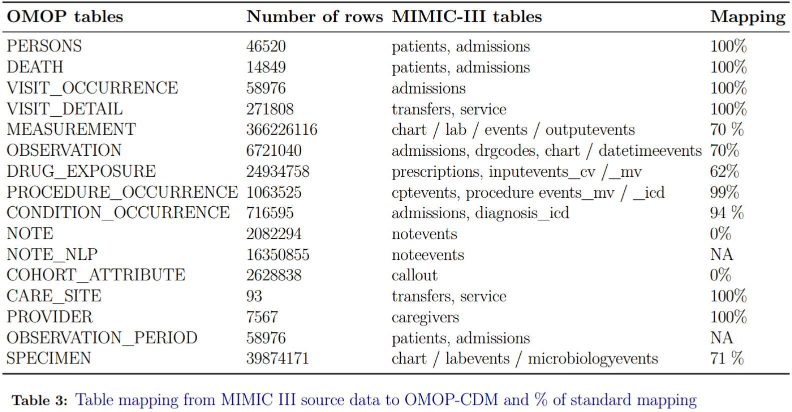
SQL ETL has several advantages:

- easy to review
- easy to maintain
- unit testing (compare counts from both places. . .)
- transforms data in place
- effective (3 hours of computation)

MIMIC OMOP TRANSLATION

Mapping Result

Mapping coverage:



OMOP tables	Number of rows	MIMIC-III tables	Mapping
PERSONS	46520	patients, admissions	100%
DEATH	14849	patients, admissions	100%
VISIT_OCCURRENCE	58976	admissions	100%
VISIT_DETAIL	271808	transfers, service	100%
MEASUREMENT	366226116	chart / lab / events / outputevents	70 %
OBSERVATION	6721040	admissions, drgcodes, chart / datetimeevents	70%
DRUG_EXPOSURE	24934758	prescriptions, inputevents_cv /_mv	62%
PROCEDURE_OCCURRENCE	1063525	cptevents, procedure events_mv / _icd	99%
CONDITION_OCCURRENCE	716595	admissions, diagnosis_icd	94 %
NOTE	2082294	notevents	0%
NOTE_NLP	16350855	noteevents	NA
COHORT_ATTRIBUTE	2628838	callout	0%
CARE_SITE	93	transfers, service	100%
PROVIDER	7567	caregivers	100%
OBSERVATION_PERIOD	58976	patients, admissions	NA
SPECIMEN	39874171	chart / labevents / microbiologyevents	71 %

Table 3: Table mapping from MIMIC III source data to OMOP-CDM and % of standard mapping

MIMIC OMOP TRANSLATION

Feedback

SQL ETL has several advantages:

- 2 people, 2 months
- Intensivist + Data Ingineer

NLP Pipelines

NLP Pipelines

UIMA Section Extractor

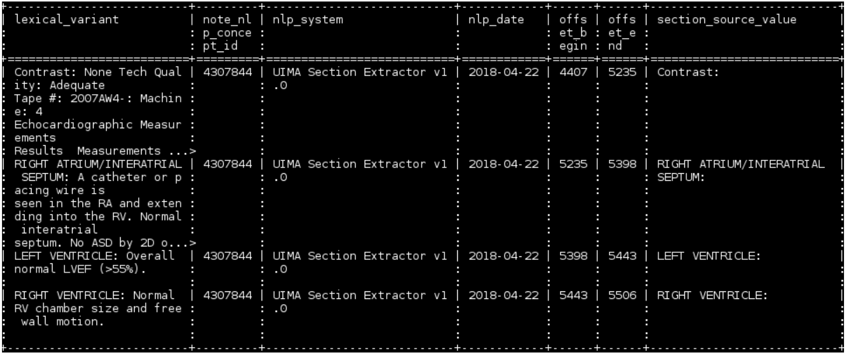
Goal: Split MIMIC notes into sections & feed omop.NOTE_NLP table

1. Automatically extract section patterns (1500)
2. Keep only > 1% used
3. map them together into 400 distinct sections

NLP Pipelines

UIMA Section Extractor

The section mapping csv file:



```
+--------------------------+---------+---------------------------+------------+------+------+--------------------------+
| lexical_variant          | note_nl | nlp_system                | nlp_date   | offs | offs | section_source_value     |
:                          : p_conce :                           :            : et_b : et_e :                          :
:                          : pt_id   :                           :            : egin : nd   :                          :
+==========================+=========+===========================+============+======+======+==========================+
| Contrast: None Tech Qual | 4307844 | UIMA Section Extractor v1 | 2018-04-22 | 4407 | 5235 | Contrast:                |
: ity: Adequate            :         : .0                        :            :      :      :                          :
: Tape #: 2007AW4-: Machin :         :                           :            :      :      :                          :
: e: 4                     :         :                           :            :      :      :                          :
: Echocardiographic Measur :         :                           :            :      :      :                          :
: ements                   :         :                           :            :      :      :                          :
: Results  Measurements ...>         :                           :            :      :      :                          :
| RIGHT ATRIUM/INTERATRIAL | 4307844 | UIMA Section Extractor v1 | 2018-04-22 | 5235 | 5398 | RIGHT ATRIUM/INTERATRIAL |
:  SEPTUM: A catheter or p :         : .0                        :            :      :      : SEPTUM:                  :
: acing wire is            :         :                           :            :      :      :                          :
: seen in the RA and exten :         :                           :            :      :      :                          :
: ding into the RV. Normal :         :                           :            :      :      :                          :
:  interatrial             :         :                           :            :      :      :                          :
: septum. No ASD by 2D o...>         :                           :            :      :      :                          :
| LEFT VENTRICLE: Overall  | 4307844 | UIMA Section Extractor v1 | 2018-04-22 | 5398 | 5443 | LEFT VENTRICLE:          |
: normal LVEF (>55%).      :         : .0                        :            :      :      :                          :
:                          :         :                           :            :      :      :                          :
| RIGHT VENTRICLE: Normal  | 4307844 | UIMA Section Extractor v1 | 2018-04-22 | 5443 | 5506 | RIGHT VENTRICLE:         |
: RV chamber size and free :         : .0                        :            :      :      :                          :
:  wall motion.            :         :                           :            :      :      :                          :
:                          :         :                           :            :      :      :                          :
+--------------------------+---------+---------------------------+------------+------+------+--------------------------+
```

1500 heterogeneous sections mapped to 400

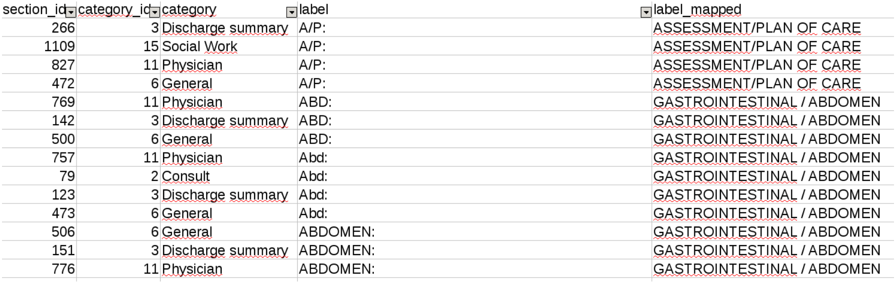
NLP Pipelines
UIMA Section Extractor

UIMA Section Extractor:

- A simple UIMA pipeline
- **INPUT**: section csv + MIMIC NoteEvent csv
- **OUTPUT**: ready to load omop.NOTE_NLP csv
- 2M note into 16M NOTE_NLP rows
- Parallelize the UIMA over apache SPARK (15 minutes job) or on a local computer 2h

NLP Pipelines

UIMA Section Extractor



section_id	category_id	category	label	label_mapped
266	3	Discharge summary	A/P:	ASSESSMENT/PLAN OF CARE
1109	15	Social Work	A/P:	ASSESSMENT/PLAN OF CARE
827	11	Physician	A/P:	ASSESSMENT/PLAN OF CARE
472	6	General	A/P:	ASSESSMENT/PLAN OF CARE
769	11	Physician	ABD:	GASTROINTESTINAL / ABDOMEN
142	3	Discharge summary	ABD:	GASTROINTESTINAL / ABDOMEN
500	6	General	ABD:	GASTROINTESTINAL / ABDOMEN
757	11	Physician	Abd:	GASTROINTESTINAL / ABDOMEN
79	2	Consult	Abd:	GASTROINTESTINAL / ABDOMEN
123	3	Discharge summary	Abd:	GASTROINTESTINAL / ABDOMEN
473	6	General	Abd:	GASTROINTESTINAL / ABDOMEN
506	6	General	ABDOMEN:	GASTROINTESTINAL / ABDOMEN
151	3	Discharge summary	ABDOMEN:	GASTROINTESTINAL / ABDOMEN
776	11	Physician	ABDOMEN:	GASTROINTESTINAL / ABDOMEN

1500 heterogeneous sections mapped to 400

NLP Pieplines

n2c2 challenge

N2C2 NLP Challenge:

- Autodetect inclusion/exlusion criteria
- **INPUT**: section csv + MIMIC NoteEvent csv
- **OUTPUT**: ready to load omop.NOTE_NLP csv
- Heideltime, ctakes, dkpro pipes are used
- Results are pending
- Entity recognition, negation, date can populate the NOTE_NLP table.
- Put results back to measurement table as derived data

⇒ Reuse those pipelines to extend MIMIC-OMOP

NLP Pipelines

UIMA Note Deid

European GDPR regulation are more strict on Personal Health Informations:

- Build-up a tool based on OMOP
- **INPUT**: Known Patient PHI json + Notes csv
- **OUTPUT**: ready to load omop.NOTE csv
- The tool will be used on french dataset

MIMIC OMOP TRANSLATION

NOTE_NLP feedbacks

Some slight modifications:

- **offset** column: this is a reserved SQL keyword → two begin/end integers columns
- **lexical_variant varchar(250)**: why limiting the size of extracted texts ? (ie: sections)
- Added : person_id - visit_occurrence_id - visit_detail_id columns

Datathon

MIMIC-OMOP Datathon

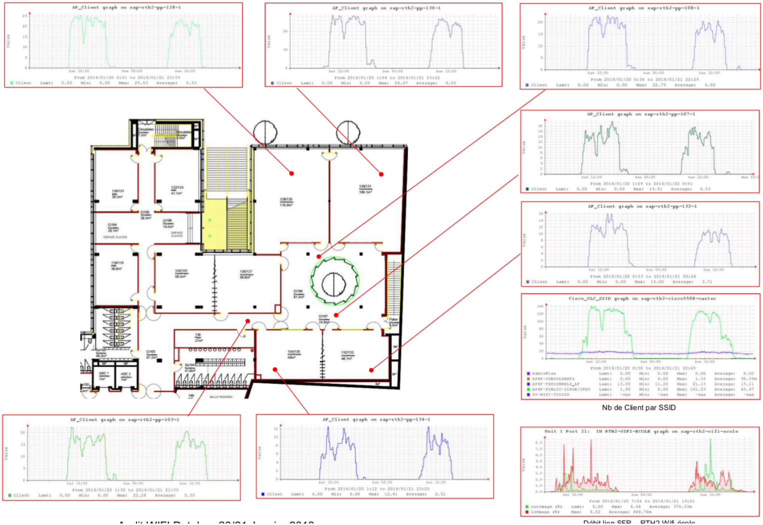
Details

January 2018, Paris datathon:

- 150 participants
- 20 projets
- Fresh MIMIC-OMOP dataset
- Distributed hadoop platform
- 15000 SQL queries during the week-end

MIMIC-OMOP Datathon

Wifi Access



Nb de Client par SSID

Audit WIFI Datathon 20/21 Janvier 2018

Débit lien SFR – RTH2 Wifi école

MIMIC-OMOP Datathon

Feedbacks

Participant feedback:

- New OMOP Model understood in 10 hours
- SQL too verbose (too much join to concept table)
- tables alone do not provide informations (mostly integers columns)
- Useful NOTE_NLP table
- MIMIC team have been interested in OMOP and are know collaborating actively

Perspectives

Next Steps

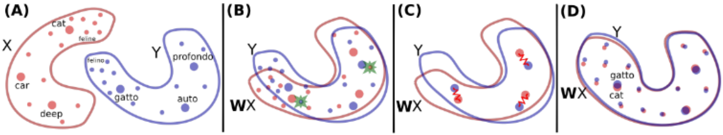
- Feed NOTE_NLP with Ctakes
- Transform French dataset into OMOP

Next Steps

Generating a Bilingual Lexicon

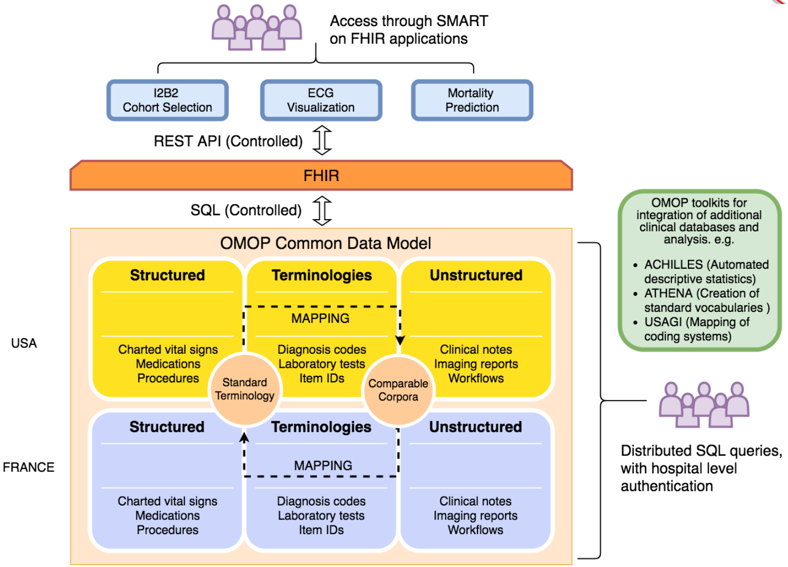
- Exploit the standardized structure / terminologies
- Exploit structure / unstructured data
- Exploit state of the art methods

Figure: WORD TRANSLATION WITHOUT PARALLEL DATA



Conclusion

Conclusion



Access through SMART on FHIR applications

OMOP toolkits for integration of additional clinical databases and analysis. e.g.

- ACHILLES (Automated descriptive statistics)
- ATHENA (Creation of standard vocabularies)
- USAGI (Mapping of coding systems)

Distributed SQL queries, with hospital level authentication

Thank you for your attention