1 ICU datas : a paradox

- Reusing medical datas has historically been impossible for a large population and most of datas were simply wasted due data variability and quality challenges
- Intensive care unit ICU are faced to a paradox
  - The level of proof to guide most decisions is low, exacerbated with real-time bedsite decisions and the medical practices are sparse (1).
  - High density environment for data production : prescriptions systems, monitoring (waves), ventilators and large number of exams done in this units
- The practice's variability is due to lack of adherence to best practices, but the vast majority occurs simply because no evidence has been established for the issue in question (2) or because the effects of interventions in the ICU are subject to the exceptional complexity of patient physiology and the variation beetween unique patient and clinical studies
- But the ICU demand of care is rising (3) and the mortality in ICU is up to 30 % which is a major health care problem

2 ICU databases

2.1 aim

The multiple aims were - to create complete and highly detailed patient record - minimize costs while improving the clinical outcomes of individuals and populations thanks to observational clinical research and real time algorithms

2.2 databases (7)

Several commercial or noncommercial, opensource or nonopensource ICU databases have been developed

- Commercial eICU
  - developped in partenariat with Philips
  - available via PhysioNet
  - over 1.5 million ICU stays
  - and is adding 400,000 patient records per year from over 180 subscribing hospitals in the country.
  - patients who were admitted to critical care units in 2014 and 2015.
  - Data are heterogenous and high granularity signal as waveform is not record
- Non commercial CUB-REA database

  - B. Guidet, P. Aegerte
  - http://www.pifo.uvsq.fr/hebergement/cubrea/cr_index.htm
  - collected since 25 years around 300k ICU patients stays low granularity data from 30 distinct ICUs in Paris region.
  - Data are collected semi automatically annually
  - 15 international publications.
- Non commercial OutcomeREA database
  - JF Timsit
  - http://outcomerea.fr/index.php
  - collected since 20 years around 20k ICU patients stays medium granularity data from 20 distinct ICUs in France
  - Data are daily collected manually by senior trained intensivists,
  - This database has been subject of 50 publications.
- Non commercial MIMICIII (Medical Information Mart for Intensive Care) : our case study
  - R. Marc
  - freely-available database via PhysioNet : https://mimic.physionet.org/
  - Data are collected each 5 years, semi automatically.
  - This database is de-identified and open, and one can exploit the data after passing an online exam on clinical ethic.
  - over 300 publications from international researchers independant from the MIT
  - health-related data associated with over forty thousand patients who stayed in critical care units between 2001 and 2012(4).
  - It includes both administrative data (demographic, ICD9, procedures) and clinical data (examination, laboratory results, medication administration and notes)
  - Three types of data are collected :
    - clinical data from hospital information system,
    - death data from the social security database
    - High granulary data as the waveform of EKG, EEG.

2.3 conclusion

The MIMIC-III database is unique in capturing highly granular structured data. But the conception of this database was time consuming and unfortunately only 45,000 unique patients' data from a single center were captured. To produce analyse high number of patient we will have to merge heterogenous databases.

3 Data merging

3.1 aims

Use of EHRs has been increasing world-wide, but most EHRs are different in their structure and not interchangeable.

- more data : may provide better outcomes
- interoperability may provide easy international research and improve reproductibily of it
- decrease costs and investment in developing algorithms and help to performs transferable analyses

3.2 challenges

- but we know that simple merging of databases give poor quality level because of the heterogeneity of datas (9)
- but sharing data creates legal/juridic problems
- but merge may loss datas

3.3 databases modelling and datas exchanges

Common data model (CDM) provides standardized definition of represent resources and their relationships. Many has been developped, certains are open-source: - OMOP model : Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) - incorpore validated standard classification (8) : SNOMED for diagnoses, RxNORM for drug ingredients and LOINC for laboratory results… - provide tables for mapping beetween international classification (ex: ICD9 and SNOMED) - provides more systematic analysis with analytic library from OMOP community : ACHILLES

```
- In this model all the data stay locally at the participant site, the primary analyses are carri

- This model has been already adopted by more than 682 million patient records with databases fro
- Several examples of transforming source databases to CDM already exists (10-11)
```

- I2B2 :
  - good interface for cohort selection

  - i2b2 has been described as being used by more than 200 hospitals6 over the world
  - The central table is called observation_fact table.
  - Compare to OMOP-CDM the hierarchies are organise with a ‘concept path’ column. Two concepts are linked by a single relationship
- PCORnet, the National Patient-Centered Clinical Research Network

  - PCORnet Common Data Model (CDM) hoping to integrate multiple data from different sources and leverages standard terminologies and coding systems for healthcare (including ICD, SNOMED, CPT, HCPSC, and LOINC) to enable interoperability with and responsiveness to evolving data standards.
  - The first version of the CDM was released in 2014
  - Compare to OMOP CDM, PCORNET is less effective for use with a longitudinal community registry (6)
- FHIR, Fast Healthcare Interoperability Resources
  - is a standard for exchanging healthcare information electronically (https://www.hl7.org/fhir/overview.html)
  - Some papers have showed that collaboration between FHIR may provide both applicative software and analytic research and showed great promise(5, 13) nico

4 Our study

The aim of MIT with MIMIC-III is to provide open datas, more collaborative and reproductitible studies with shared codes. In this purpose the transformation from MIMICIII to MIMICIII-OMOP with standardized mapping concept is important and was hightly supported by the MIT. (4)

In this article we provide a example of Extract Transform Load (ELT) implementation of electronic health records (EHR) in intensive care unit by transforming the all MIMIC-III database (expected high frequency datas) to OMOP CDM version 5.3 (last version in date). We'll expose our methodology and we'll discuss about modification we want to propose to the omop community. We'll also discuss about potential loss of information links to this ETL.

1. Vincent JL. Is the current management of severe sepsis and septic shock really evidence based? PLoS Med 2006; 3:e346
2. Vincent JL, Singer M. Critical care: advances and future perspectives. Lancet 2010; 376:1354–1361
3. Angus DC, Kelley MA, Schmitz RJ, White A, Popovich J Jr; Committee on Manpower for Pulmonary and Critical Care Societies (COMPACCS). Caring for the critically ill patient. Current and projected workforce equirements for care of the critically ill and patients with pulmonary disease: can we meet the requirements of an aging population? JAMA 2000;284:2762–2770
4. A.E.W. Johnson, Tom J. Pollard and Al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016-5-24
5. M. Choi and Al. OHDSI on FHIR Platform Development with OMOP CDM mapping to FHIR Resources,Georgia Tech Research Institute, poster
6. M.Garza. Evaluating common data models for use with a longitudinal community registry. Journal of Biomedical Informatics 2016. 333–341

7. Jeff Marshall, Abdullah Chahin and Barret Rush. Chapter 2 Review of Clinical Databases - Springer
8. JM Overhage and Al. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. J Am Med Inform Assoc 2012;19: 54-60
9. G. Hripcsak and Al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers.Stud Health Technol Inform. 2015 ; 216: 574–578
10. F. FitzHenry and Al. Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. Appl Clin Inform 2015; 6: 536–547
11. S. Bayzid and Al. Conversion of MIMIC to OHDSI CDM. National Center for Biomedical Communications, Bethesda, Maryland
12. T. Gruber. Toward principles for the design of ontologies used for knowledge sharing?, International journal of human-computer studies, 1995
13. Nicolas Paris and Al. i2b2 implemented over SMART-on-FHIR # data source

- several other open-source databases
  - eICU (3), freely-available comprising deidentified with more than hundreds of thousands of patients. Data are available to researchers via PhysioNet, similar to the MIMIC database
  - OUTCOMEREA (http://outcomerea.fr/index.php)
  - CUBREA (http://www.pifo.uvsq.fr/hebergement/cubrea/cr_index.htm), with many ICU from APHP with > 2000000 icu stays
- presentation of mimicIII : our case study MIMIC-III (Medical Information Mart for Intensive Care) is freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units between 2001 and 2012(1). It includes both administrative data (demographic, ICD9, procedures) and clinical data (examination, laboratory results, medication administration and notes) Three types of data are collected : clinical data from hospital information system, death data from the social security database and the high granulary data as the waveform of EKG, EEG. In this article we won't speak about high frequency datas.

The aim of MIT with MIMIC-III is to provide open datas, more collaborative and reproductitible studies with shared codes. MIMIC is a large used database with x number of publications. In this purpose the transformation from MIMICIII to MIMICIII-OMOP with standardized mapping concept is important. The mimic documentation is a available online physionet.org/about/mimic/. A public github was created : https://github.com/MIT-LCP/mimic-code with many contributers around the world.

5 ETL mapping specifications

- The key table for omop is the concept table. The standard vocabulary of OMOP is mainly based on the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)
- A mapping between many classification and the standard omop ones (ICD-9 and snomed-CT for examples) is already provides with concept_relationship.
- Local code for mimiciii such as admission diagnoses, demographic status, drugs, signs and symptoms were manually mapped to OMOP standard models by several participants. For example local drug codes were mapped to the OMOP standardized vocabularies, which use RxNorm. This work was followed and check by a physician. All laboratory exams, exit diagnoses and procedures were already mapped to standard classication. All the csv files used for the mapping are available on github

6 methodology of ETL

All the process is available freely on the github website.

6.1 Preprocessing and modification of mimic

- We added emergency stays as as a normal locations for patients throughout their hospital stay.
- Icustays mimic table was deleted as it is a derived table from transfers table (2) and we decided to assigne a new new visit_detail pour each stay in ICU (based of the transfers table) whereas mimic prefered to assgned new icustay stay if a new admissions occurs > 24h after the end of the previous stay
- We decided to put unique number for each row of mimic database called mimic_id. We think this is very helpful for ETLers

6.2 Technical specifications

- To provide standard and reproductilable precess all the ETL used SQL script.
- subset of 120 patients,
- unit testing during the all process of extraction and SQL script production
- we tried not to infer results. For examens whereas it's logical to put a specimen for many labevents results (as one sample of blood may be used to multilple exams) we decided to create as many specimen row

as laboratory exams because the information is not present. It was the same when date information were not provide (start/end_datetime for drug_exposure)

- concept-driven methodology : as the omop model did we adopt a "concept-driven methodology", domain of each local concept drive the concept to the right table.
- fact_relationship : for drug solution, microbiology / antibiograms, visit_detail and caresite - for example : microorganism are links to their antibiogram thanks to fact_relationship <!– fournir un exemple de SQL pour ca avec un resultat>

6.3 modification of OMOP model

- the less possible
- keep in mind the model of omop as a conceptual model
- constant dialogue with omop community (omop github)
- modifications of OMOP model (few columns)
    - structural (columns type, columns name, new columns)
        - visit_detail : visit_detail table adding of admitting_source_value, admitting_source_concept_id, admitting_concept_id, discharge_to_source_value, discharge_to_source_concept_id, discharge_to_concept_id
    - conceptual (new concepts specific to ICU or general)
        - measurement_type_concept_id
        - the actual visit_detail doesn't introduce pertinent information and duplicate informations from visit_occurrence table. For admitting_from_concept_id and discharge_to_concept_id, we extended the dictionary in order to track bed transfers and ward transfers. For visit_type_concept_id we assigned a new concept for any level of granularity necessary for your use case (ward, bed…)
- modification of MIMIC
    - observation_period provide duplicate information: we fill this table to respect the omop model and tools
    - operators have been extracted to fill operator_concept_id
    - units of measures have been extracted to fill unit_concept_id

1. A.E.W. Johnson, Tom J. Pollard and Al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016-5-24
2. https://mimic.physionet.org/mimictables/icustays/ # table populated with their mimic source table link The OMOP-CDM contains n data

tables. We populated m tables. From MIMICIII we create a standardized model called MIMICIII-OMOP.

Omop tables	Source tables
PERSONS	patients, admissions
DEATH	patients, admissions
VISIT_OCCURRENCE	admissions
VISIT_DETAIL	transfers, service
MEASUREMENT	chartevents, labevents, microbiologyevents, outputevents
OBSERVATION	admissions, chartevents, datetimevvents, drgcodes
DRUG_EXPOSURE	prescriptions, inputevents_cv, inputevents_mv
PROCEDURE_OCCURRENCE	cptevents, procedureevents_mv, procedure_icd
CONDITION_OCCURRENCE	admissions, diagnosis_icd
NOTE	notevents
NOTE_NLP	noteevents
COHORT_ATTRIBUTE	callout
CARE_SITE	trasnfers, service
PROVIDER	caregivers
OBSERVATION_PERIOD	patients, admissions
SPECIMEN	chartevents, labevents, microbiologyevents

- observation_period provide duplicate information: we fill this table to respect the omop model and tools

7 Quality evaluation

7.1 comparaison MIMICIII / MIMIC OMOP (basic statistics)

The table lists the baseline characterization of the population of MIMICIII-OMOP compared with MIMICIII.

items	OMOP-MIMIC	MIMICIII
Persons (Number)	46.520	46.520
Admissions (Number)	58.976	58.976
Icustays (Number)	61.532	71.576
Age (Mean)	64 ans, 4 months	64 years, 4 monts
Gender, Female (Number, %)	20.399 (43 %)	20.399

items	OMOP-MIMIC	MIMICIII
Length of stay, hospital (median)	6.59 (Q1-Q3 : 3.84 - 11.88)	6.46 (Q1-Q3 : 3.74 -11.79)
Length of stay, ICU (median)	1.87 (Q1-Q3 : 0.95 - 3.87)	2.09 (Q1-Q3 : 1.10 - 4.48)
Mortality, ICU (Number, %)	5815 (9%)	5814 (9%)
Mortality, hospital (Number, %)	4559 (6%)	4511 (7%)
Lab measurement per admissions (mean)		

papier + test

cf extra : basic_statistics.sql

7.2 loss of data (try to quantify it)

- percent of records loaded from the source database to the CDM
  - percent of columns
  - percent of rows as have done other studies (1)
- Row

items	rows per persons
Nb patients	100 %
Nb admissions	100 %
Procedures	%
Admissions diagnosis	%
Exit diagnosis	%
Laboratory exams	%
Physical exams	%
Drugs	%
Notes	%

remark all the error lign are deleted (prescriptions, inputevents_mv, chartevents, procedureevents_mv, note)

- Columns % of sources columns which doesn't fits to CDM storetime!!

7.3 terminology mapping coverage

- ICD-9-CM A part of source data for condition_occurrence was ICD-9 codes. The OMOP common standard vocabulary, SNOMED-CT, did not cover all ICD-9-CM codes (95%) Moreover, not all ICD-9-CM codes can have one-to-one mapping to SNOMED, some are one-to-many (28%)(2)
- LOINC
- RxNorm
- % of standard_concept_id = 0 (No mapping concept) per table Need colaborative work
- % of domain_id not in adequation with table name
  - some are logical because observation domain may be measurement table and vice verca
- we have mapped many source concept to one standard concept is it the same meaning? distribution of values sometimes very different

7.4 ACHILLES evaluation

ACHILLES is open-source software application developped by OHDSI and Achilles Heel provided data quality checker Other team used this tool to practice data quality assess(4). Our result …

8 Community sharing

We provided many derived values. Community is welcome to improve it - F/P, corrected Ca / K, BMI - Note_NLP with section splitting. The algorythm is freely accessible here - SOFA, SAPSII

9 Feedbacks of real MIMICIII-OMOP testing

- this work has been done with APHP to test OMOP model in real statistical condition. A datathon was organised in collaboration with the MIT.(3) We also test the big data APHP platforms.
- most of queries under 30 second ; simplified model VS MIMIC ; to much normalized for data scientist)

10 others

- estimation of number of work hours
- ethnicity_concept_id : only two strange concept_name hispanic or non_hispanic
- size of MIMIC OMOP, row number for the bigest relation (measurement)
- chartevents and lavents provide many number field as a string which is not handy for statistical analyse. We provide a standard and easy improval by the community model to extract numerical value from string
  - operators have been extracted to fill operator_concept_id column
  - numeric value has been extracted to fill value_as_number column
  - units of measures have been extracted to fill unit_concept_id column

1. F. FitzHenry Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. Appl Clin Inform 2015; 6: 536–547
2. https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html
3. http://blogs.aphp.fr/dat-icu/
4. Y.Dukyong and Al.Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research.Healthcare Informatics Research 2016; 54

11 Discussion

database modelling

- advantages of normalized
- disadvantages of normalized
- advantages of packed tables
- disadvantages of packed tables
- need to propose an optimized version of OMOP -> XT: is that the idea of flattenning the tables to improve query-ability? OMOP Analytics? technical performances -> NP: indeed, precalcul most of joins so that queries are focused on one or two table instead of 5 and more ; in this case, replicating information is not a problem because it is a frozen dataset and replicated field all come from one unique field and this is then no error prone by design XT: What are the pros and cons of this solution against keeping the structure but producing some ”query plans” for the frequent accesses? Several participants at dat-icu told me that at the end of the end, most of the teams struggled to finally produce queries that were very similar to their neighbors’ queries

terminology mapping

- athena existing standard mapping
- athena missing concepts

What else could have been done instead?

- use of automatic concept mapping processes: but their performances are not yet as good as human manual mapping.
- large community mapping: this was too early to open the work to public so that effort and direction could be kept

OMOP model - enables observational studies to be conducted using multiple data sources - while confidential personal health data remain with the original data holders

12 Conclusion

- time and ressources consuming
- allow worldwide shared algorithm