MIMIC in the OMOP Common Data Model*

A1FIRSTNAME A1LASTNAME, *AES Member* **AND A2FIRSTNAME A2LASTNAME,** *AES Fellow*
(abc@abc.com) (xyz@abc.com)

Universityxyz, City, Country

Objectives: In the era of big data, Intensive Care Unit (ICU) is highly susceptible to gain from computer realtime analytics and modeling based on patient close monitoring and Electronic Health Records. MIMIC is yet the first free-access ICU database. Numerous studies have shown that Common Data Models (CDM) improve database research by allowing code, tools, experience sharing. OMOP CDM is spreading worldwide. We transformed MIMIC into OMOP and evaluate the quality of the transformation, the benefits for analysts and the gains and potentials for community contributions.
Material & Method: A documented, tested, versioned, exemplified and free-access repository have been setup to support the transformation and community improvement source code. The resulting dataset has been evaluated thought a 48h datathon with 150 participants.
Result: Most of the data fit in the model and a large part of the terminologies have been standardised with an investment of 2 people during 500h. The model demonstrated its capacity to support community contributions and was well-recieved during the datathon with 15K queries have been run with a maximum of 1 minute long.
Conclusion: The resulting MIMIC-OMOP dataset is ready for reproducible research and as the first open-access OMOP dataset with real de-identified patient data available the expectations are an increase and improvement of collaborations for ICUs. Generalizable to other service...

0 INTRODUCTION

Intensive Care Units (ICUs) are particularly sensitive units where demand of care is rising[1] and mortality is up to 30% which is a major health care problem [2]. Studies have shown that intensivists use a limited number of clinical information concepts to guide the decision[3] and that the medical practices are sparse and variable. Knowing that ICU patients' health record are highly detailed including connected devices this is a paradox. The increasing adoption of Electronic Health Records (EHR) systems worldwide makes it possible to capture large amounts of clinical data [4] and big data mining has the potential to play an important role in clinical medicine [5]. Indeed on the basis of broad patient medical informations expectations are to improve clinical outcomes and practices, allow personalized medicine and guide decision thought early warning systems, and also enrol large and multicentric cohort easily while minimizing costs.

By now several commercial or noncommercial, opensource or nonopensource ICU databases have been developed. *MIMIC* (Medical Information Mart for Intensive Care) is a semi-automatic collection of 10 years and over 60k ICU patients stays with very high granularity (including EKG) from two successive critical care information systems (CCI) of the Beth Israel Deaconess Medical Center in Boston. It is the first freely-available ICU database and yet it is subject of 300 international publication. However its monocentric nature makes difficult to generalize the conclusion on other ICUs. The MIMIC relational data model reflects the originals CCI as evidenced by the two distinct inputs tables *inputevent_mv* and *ouputevent_cv* [6] or the two distinct terminologies for the physiological data. This leads analysts (data-scientitsts, statisticians…) to reconcile this heterogeneity during the preprocessing for each studies.

Some studies have shown that using a common data model (CDM) by generalizing the database on structural (data model) and conceptual (terminology model) design enables multicentric research, mining rare disease and catalyses research by allowing sharing practices, source code and tools[7]. Some studies have shown that results are not fully replicable from one to the other CDM [8] or from one to another center [9]. Lighter approaches argue that keeping the local conceptual model [10] or both original conceptual and structural [11] of the database for research leads to better outcomes. On the first hand keeping MIMIC

[*] To whom correspondence should be addressed Tel: +1-240-381-2383; Fax: +1-202-508-3799; e-mail: info@schtm.org

on its specific form won't solve the limitation of the previous paragraph and on the other hand a fully standardized form would introduce other drawbacks. The ideal solution is likely in-between to allow both local or standardized analysis depending on the research question.

Several studies have previously applied CDM to MIMIC. Chronaki et al. [12] have evaluated i2b2 to federate MIMIC and eICU. i2b2 is a medical cohort discovery tool used in more than 200 hospitals over the world. The i2b2' highly flexible structural model is able to ingest various domains and data types in a unique star-schema composed of 5 tables. While i2b2 is highly efficient for cohort discovery, it's model wasn't designed for ad-hoc analysis. The i2b2' conceptual model is a path based table allowing to store any simple or hierarchical local terminology (graph are not supported). The SHRINE project goes a step further by proposing a set of standard terminology to federate multiple i2b2 instances. Each instance needs to map all its local concepts to the standard's one and vice-versa [13] (bi-directional mapping).

More recently Rajkomar et al. have been using *FHIR* (Fast Healthcare Interoperability Resources) as a CDM to apply deep learning on medical records. FHIR is a medical data exchange API specification. FHIR provides a structural CDM that can be materialized as JSON, XML or RDF format. FHIR is flexible and does not specify a standard conceptual model so that each hospital can add extension to implement specific data or share within it's local terminology making each FHIR implementation sensibly divergent. While some research show it as a promising CDM for ad-hoc analysis [14] or cohort discovery [15], its graph nature adds a layer of transformation making usage complicated for analysts.

OMOP structural model consists of several groups of thematic tables (clinical, costs, derived…). Each clinical domains are stored in various dedicated tables. OMOP conceptual model is based on a closure table pattern [16] able to ingest any simple, hierarchical and also graph terminologies such SNOMED-CT used locally. In addition to the local terminologies OMOP specifies and maintains a set of standard terminologies to be mapped uni-directionally (local to standard) by implementers.

Previous preliminary work have been made on translating MIMIC into OMOP [**?**] and still work remains to be affined and upgraded to be furthermore evaluated.

A dozen of medical CDM exist and a recent study compared their quality [17] through various characteristics presented in table x. From all the candidates, *OMOP* (Observational Medical Outcomes Partnership Common Data Model) fits better the criteria. OMOP is a CDM originally designed for multicentric Drug adverse Event and now extends to medical, clinical and also genomic use cases. OMOP provides both structural and conceptual model such SNOMED for diagnoses, RxNORM for drug ingredients and LOINC for laboratory results. While OMOP has proven its fiability [18] the fact that concept mapping process is known to have impact on results [19] and that applying the same protocol on different data sources leads to different results [9] reveals the importance of keeping the local codes to allow local analysis. Several example of transforming databases into OMOP have been published [20, 21] and yet OMOP stores 682 milion patients records from all over the world[22]. OMOP had 5 versions, and prones its strong backward compatibility. Among the other CDM, OMOP looks like the best fit as it allows both multicentric standardised analysis as well as monocentric specific modeling and analysis.

In a recent CDM comparison study [7] OMOP performs better in the evaluation database criteria compared with the other models (and PCORnet in particularly) : completeness, integrity, flexibility, simplicity of integration, and implementability, accommodate the broadest coverage of standard terminologies, provides more systematic analysis with analytic library and visualizing tools, provides easier SQL models. Compared to i2b2/SHRINE, the unilateral mapping methodology of OMOP is more effective than the bidirectional [13] SHRINE mapping. Hence OMOP proposes a broader set of standardised concepts. In terms of structural CDM OMOP is highly rigourous in how data shall be loaded in a particular table when i2b2 is highly flexible with a one general table that solution every data domains. This rigourous approach is necessary for standardisation. Previous work have loaded i2b2 with MIMICiii [12] - however, the concept mapping step have limited the results since i2b2 design do not store the local ontology or information where OMOP design allows to keep the concept mapping unfinished. OMOP has this advantage to keep the terminology mapping step not mandatory by keeping the local codes in a usable format. As compared to FHIR, OMOP performs better as a conceptual CDM as FHIR does not specify which terminology to implement. In terms of structural CDM OMOP relational model can be materialized into csv format and stored in any relational database when FHIR materialised as json needs some processing and more skills to be exploited. OMOP shares the advantages of all above models.

In order to evaluate the MIMIC to OMOP transformation we propose to answer to remaining question such the difficulty of transforming/maintaining an OMOP dataset from an existing one, how well the initial dataset is integrated and how much data is lost in the process, how clear and simple the model is to be queried simply and efficiently by analysts, how well design it is to be enriched by collaborative work, and finally in what extend OMOP can integrate and makes feedbacks to intensivists in a realtime context. This work is then evaluated through 3 axes: Transformations, Contributions and Analytics. The *first* major contribution of this study is to evaluate OMOP in a real life and well known freely accessible database. The *second* major contribution is to provide a freely accessible dataset in the OMOP format that might be useful for researchers. The *third* major contribution is to provide the OMOP community some useful transformations dedicated to ICU and that can be reused on any OMOP dataset.

1 MATERIAL & METHOD

1.1 Data Transformation

All the processes are publicly and freely available as a github website [23] included and maintained by the MIT-LCP organization responsive of the MIMIC maintenance and distribution [6]. The repository is based on git and is designed for community sharing, improvement, collaboration and reproducible work. Indeed github is archived on a universal software archive solution [?] and this implies its sustainability. The repository centralizes various resources of this work such documentation, source code, unit tests, as well as query example, discussions and history of issues. It also points to web resources such the physical data model for both MIMIC[?] and OMOP[?] datasets and the Achilles webclient[?].

The vast majority of the source code is implemented in PostgreSQL 9.6.9 (PGSQL)

The *structural transformations* consists on dispatching the MIMIC data into OMOP. The MIMIC source data version version 1.4.21 has been loaded with the provided scripts into a PGSQL instance. A subset of 100 patient chosen upon their broad representativity of the database have been cloned into a second instance to serve as a light, representative development set. Each tables of the source have been added a global sequence incremented from 0 that serves as primary key and link into the OMOP target tables. The OMOP CDM v5.3.1 target tables have been created from the provided scripts with some slight modifications stored in modification script. The indexes that would have slow down the data migration with useless computations. The integrity constraints (primary keys, foreign keys, non nullable columns) have been included to apply sanity checks at runtime.

Extract-Transformation-Load (ETL) processes is a methodology for migrate data from a source to a target location. ETL first extract the data from the source location, then apply transformations on a dedicated computer and finally load the resulting data into the target location. Extract-Load-Transform (ELT) processes is a slightly different methodology that does not use a dedicated transformation server. The data is extracted and directly loaded into the target location and transformed afterwards in place. Vthe programming knowledges needed for code maintenance and to allow end users to participate in this process. PGSQL 9.6 has been chosen as the database support for ELT because it is the primary support of MIMIC database and allows community to run the ELT on limited resources without licensing need. Finally PGSQL have recently made huge effort to handle data-processing better. Some elaborated data transformations have been implemented as PGSQL functions.

The ELT is composed of 22 PGSQL scripts, each extracting information from the source or from concept mapping tables, transforming and loading one OMOP target table. The ordering of those scripts matters and is done sequentially thought a main script.

Each ELT part have been tested through pgTAP, a unit testing framework for PGSQL. This checks for loss of informations, or code regression during the development. The unit tests are composed of 15 scripts, each checking a particular OMOP target table is loaded correctly - most of the tables are covered and tests covers simple counts, aggregated counts or distribution checks.

All character typed columns limited in length have been changed to unlimited since this might cause unpredictable truncation of content, and this has no bad impact on PGSQL's storing size or performance. The *visit_occurrence* and *visit_detail* table have been corrected accordingly to some discussion on the forum. The *note_nlp* table have been extended with some fields accordingly to the documentation online. The character *offset* column have been split into two integers columns because the offset word is a SQL reserved word and it makes sense to fill the resulting *offset_begin* and *offset_end* resulting columns.

The structural transformation have been done in few iteration of several phases. The first phase consists of looping over each MIMIC table and choose for each columns an equivalent location in OMOP. In general, the MIMIC documentation and the OMOP documentation were sufficient informative. In several cases, we needed to get clarification from the MIMIC contributors on the dedicated github repository, or from the OMOP community on the dedicated forum. All choices have been discussed in the repository issues [?] and can be tracked into the commit log.

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 multiple exams) we decided to create as many specimen row as laboratory exams because the information is not present in MIMIC. It was the same when date information were not provide (*start/end_datetime* for *drug_exposure*). - chartevents and labevents 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 The MIMIC laboratory results have been restructured to fit in OMOP format. In particular, the numerical value (value_as_number) comes with a mathematical operator (operator_concept_id) and a measurement unit (unit_concept_id). The MIMIC semi-structured raw laboratory data have been structured with a PGSQL function to extract those information.

By design MIMIC aggregates informations from various systems such the emergency specific tool. Thus the transfer information is spread into multiple table, such *admissions*, *transfers* and *icustays*. OMOP centralizes this information in the *visit_detail*. We added emergency stays as a normal location 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 assign a new *visit_detail* pour each stay in ICU (based of the transfers table) whereas mimic preferred to assign new icustay stay if a new admissions occurs ¿ 24h after the end of the previous stay.

The **conceptual transformation** uses the OMOP Vocabulary tables that have been loaded from an export of

Athena [?] of all terminologies without licensing limitations. The MIMIC local codes are also loaded into the concept table with *concept_id* starting from 2.1 billion (below this number is reserved for OMOP terminologies [?]). The MIMIC codes can be distinguished with the *vocabulary_id* equals to "MIMIC code" and a *domain_id* targeting the OMOP table the related data is stored in. Later this domain information is used in the ELT to dispatch the information in the right table. As much as possible information from the MIMIC original tables has been concatenated in the *concept_name* column. Some new concepts have been introduced and were assigned a value starting from 2 billion to distinguish them from MIMIC local concepts.

When comes standardization of the MIMIC local codes into OMOP standards codes there is four distinct cases. In the *first* case MIMIC is by chance already in the OMOP standard terminology (eg: LOINC laboratory results) and consequently both standard and local concepts are the same. In the *second* case the mapping is already provided by OMOP (eg: ICD9/SNOMED-CT) then the domain tables have been loaded accordingly. In the *third* case the mapping is not provided, but is small enough to be done manually in few hours (such as demographic status, signs and symptoms). In the *fourth* case the mapping is not provided and the terminology is huge (such admission diagnoses, drugs). Then only a subset of the most represented code were manually mapped.

When the concept mapping is needed a mapping csv file have been built. This solution can scale for medical users without database engineering background. The spreadsheet has several columns such the local/standard labels, ids and also comments, and evaluation metric and a script load them into the PGSQL once filled. In order to catalyse the mapping process, linguistic based algorithm has proven to be effective [24] and yet OHDSI provides USAGI [?]. We have opted to use simples SQL queries which is flexible enough to be queried on-demand or to generate a pre-filled mapping csv with the best matches. It exploits the PGSQL full-text ranking features and links both local and standard candidates with a scoring function based on their labels. This work was followed by an intensivist check.

While various types of information are stored in the measurement table, the dedicated OMOP concepts types were not enough to distinct them. We added some measurement types. The actual visit_detail doesn't introduce pertinent information and duplicate informations from visit_occurrence table. For admitting_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...)

1.2 Contribution

MIMIC provides a lot of SQL scripts to calculate derived scores and existing cohorts. Some of them have been translated based on the OMOP data and populates the OMOP cohort tables. Unprecedented derived informations have been introduced and loaded such corrected calcemia, kaliemia, P/F ratio, corrected osmolarity

A set of *general denormalized* tables has been built on top of the original OMOP format wich have the *concept_name* from the concept table for both standard and local codes. The concept table is a central piece of the OMOP format and as a result it is involved in many joins to get the concept label. A set of *specialized analytics* tables has been built on top of the original OMOP format. The microbiology events is a reorganization of data from measurements for microorganisms and related antibiograms and is inspired from MIMIC *microbiology_event* table.

The *note_nlp* has been initially designed to store derived final or interim information and its metadata from clinical notes. When final, the extracted informations are intended to be moved to the dedicated domain/table to be then reused as regular structured data. When interim, the information is stored in the table an can serve subsequent analysis. In order to evaluate this table we provided two information extraction pipelines. The *first* pipeline has extracted some numeric values such weight, height, body mass index and cardiac left ventricular ejection fraction within medical notes with a python script. The resulting structured numerical values have been loaded according to its domain in the measurement or the observation tables. The *second* pipeline *section extractor* based on apache UIMA framework splits the notes into sections in order to help analysts to choose or avoid some sections from their analysis. While some methods already exists to extract medical sections [25] the prior work of describing sections was too high, and we went with a nave approach. The sections patterns (such "Illness History") have been automatically extracted from texts from regular expressions, automatically filtered by keeping only one with frequency higher than 1 percent and manually filtered to exclude false positives with a total of 1200 sections. The resulting sections patterns candidate have been then manually regrouped into similar 400 groups. The extracted sections have not been mapped to any standard terminology such LOINC CDO. The reason is the CDO LOINC has decided to stop to maintain and to remove its sections from its standard arguing it is too difficult to maintain, and estimates that this sections are not widely used [26].

1.3 Data Analytics

A free access 48h datathon [?] was setup in Paris once the MIMIC-OMOP transformation ready for research in order to evaluate OMOP as an alternative data model in a real life event. A total of 150 person and 20 teams from X countries were present for the two days event. 20 projects had been prepared thought a forum. OMOP have been loaded into apache HIVE 1.2.1 into ORC format. Users had access to the ORC dataset from a web interface jupyter notebooks with, python, R or scala. A SQL webclient allowed teams to write SQL from presto on the same dataset. The hadoop cluster was based on 5 computers with 16 cores and 220GO RAM memory. The MIMIC-OMOP dataset has been loaded from a PGSQL instance to

HIVE thought apache SQOOP 1.4.6 directly into the ORC format. Participants had also access to the physical modeling of the database thought Schemaspy to have access to the OMOP physical data model with both table/column comments and primary/foreign key materialising tables relations. All the queries were logged.

2 RESULT

2.1 Transformation

The MIMIC to OMOP conversion was done by two developers (one data engineer and one intensivist) in an estimate amount of 500 hours. This includes the ETL, the git documentation, the concept mapping, the contributions and the unit tests. The ETL (including unit tests, and generation of the ready to load archive) on the 100 patient subset takes five minutes and allows fast cycles of developments. On the full MIMIC dataset the ETL lasts 3 hours. The resulting csv archive is slightly the same size than the original one, and this is also the same once instanced in PGSQL and indexed.

In order to evaluate how well transformed is the data we first adopt the data quality evaluation grid (2.1) developed by Khan and Al [?] which is commonly used as a reference [?].

- size of MIMIC OMOP, row number for the bigest relation (measurement)

For the prescritions MIMIC table 75% (a verifier) of drugs had a gsn code. The conept_relationship table provide mapping between gsn and RxNorm classsifications. To improve the mapping we then proceeded to a manual mapping

The OMOP-CDM contains 37 data tables. We populated 19 tables. From MIMIC we create a standardized model called MIMIC-OMOP.

Quality evaluation criteria 1) Several articles tried to evaluate CDM quality (8, 9). The criterias developped by Khan and Al, which referenced Moody and Shanks (10) metrics were adapted to our study. Our study doesn't want to evaluate the quality of CDM. But we adapt these criterias to assess our ETL work.

Understandability and simplicity will be evaluated in the analytics parts, in real life application.

2) During the all ETL process we created a lot of unit tests thanks to pgTap library. All are available on our github. All the test passed.

3)

Compleness - semantic mapping

The table 2.1 shows where the informations goes and links between MIMIC tables and OMOP tables. Since OMOP is a conceptual model, a same type of data goes in the same table. The best example may be the measurement table which is field by 4 source tables. Is is because of all the numerics datas should go to this table.

Ajouter schema : MIMIC-OMOP_equivalence.png

As shown, all the MIMIC domain are linked to proper OMOP domain. The semantic mapping was not a problem for our work.

Completeness - structural mapping

Baseline characteristics : comparison MIMIC / MIMIC OMOP (basic statistics)

The following table lists the baseline characterization of the population of OMOP compared with MIMIC.

Hopefully most statistics remains similar between two versions. Still some differences exists. Table 2.1 MIMIC contains 61.532 stays in ICU whereas OMOP contains 71.576 ICU stays. That is a 16% increase in stays due to our ETL methodology as explained in the methods.

The table 2.1 shows the number of laboratory measurement per admissions is increased. This is because the laboratory data from MIMIC chartevents has been extracted and treated as laboratory. All the code to created this statistics are provided here (cf extra : basic_statistics.sql).

Loss of datas

We tried to evaluate the percentage of loaded records from the source database to OMOP. We evaluate the percentage of columns and rows lost in the process as done other studies [21]

Depending on tables 40% to 80% of sources columns which do not fit to OMOP where deleted. The exact removed columns are provided in the appendix (cf extras) Almost all the removed columns are redundant with others or provide derived informations. The main concern are some lost timestamps. For example the MIMIC chartevents tables provide the storetime and charttime columns but OMOP only provide one slot for timestamp to stored. Thus storetime was deleted during the ETL. The erroneous rows 2.1 were deleted in the process (marked with a status column in MIMIC tables inputevents_mv, chartevents, procedureevents_mv, note). As MIMIC team told us that they will remove it in the next release because this data are poor quality we decided to do the same. The following table shows the number of rows with errors.

terminology mapping coverage

These results include automatic and manual mapping.

The unmapped concept are the concept_id = 0 (No mapping concept). To improve this mapping we need collaborative work. We provide our csv mapping files on our github. The terminology mapping has been evaluated by a physician. The zero value for concept_id may appear in very different cases. First case the local concept has no equivalent in the standard set of concept. Second case it has not been yet been mapped and may have an standard equivalent. Third case the value is missing and cannot be mapped. In our opinion while all those cases cannot be used for standard queries, they should have a different concept_id in order to be addressed differently.

- % 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_id. This is because MIMIC provide a lots of equivalent free text concepts. For example for the body temperature MIMIC provide 11 distinct concept (Temperature F, Temperature C (calc), Temp Skin [C], Temperature Fahrenheit, Temp Axillary [F], Temperature C, Temperature F (calc), Temperature Celsius, Temp Rectal [F], Temp Rec-

Table 1. Data Transformation Quality Evaluation Metrics

Data Model Dimension	Descriptions
Completeness - semantic mapping	Domain coverage : coverage of sources domains that are accommodated by the standard OMOP model
Completeness - structural mapping	Data coverage : coverage of sources data concepts that mapped to standard OMOP concept
Integrity	”Meaningful data relationships and constraints that uphold th eintent of the data-s original purpose” Khan and Al
Flexibility	The ease to expand the standard model for new datatypes, concepts
Integration	The capacity of the standard model to use multiples terminology and links its to standard one
Implementability	The stability of the models, the community, the cost of adoption
Understandability	The ease of the standard model to be understood
Simplicity	The ease of querying the standard model - the model should contains the minimum of concepts and relationship

tal, Blood Temperature CCO (C)). Our mapping links all of it to one single concept called temperature. All the units have been converted to celcius. TODO: give example of generalisation (admission_location_to_concept)

Flexibility =========== OMOP had a 100% match of the data models constraints and relationship. Two important tables are provided with the OMOP models to match relationship : concept_relationship and fact_relationship. The fact_relationship table which is a important part of the OMOP CDM. It is used to represent the relationship between data. We used to link drugs in a solution, for microbiology / antibiograms and for visit_detail and caresite. The SQL following query shows how a microorganism is linked to its antibiogram thanks to fact_relationship

Integration ===========

OMOP Terminology coverage has been previously evaluated as excellent [17]. We used the OMOP mapping for NDC-RxNorm, ICD9-SNOMED, CPT4-SNOMED. It was really helpful because MIMIC provides lots of non standard terminology already mapped by OMOP community We tried to evaluated this OMOP mapping. We check 100 items for each mapping used (NDC, ICD9 and CPT4). ICD9 and CPT4 are correcly mapped to SNOMED. But only 85% of NDC are linked to a correct RxNorm code. In part due to incorrect NDC code (from MIMIC), in part because only 78% of NDC codes are mapped to Rxnorm.

But 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%)[27]

Implementability =============== - OMOP available since 9 years - this models and its provided concepts are licence free - the community is large and was very helpful. - Full versions (V6, 7 etc.) are usually released each year (1-Jan) and are not backwards compatible. Minor versions (V5.1, 5.2 etc.) are not guaranteed to be backwards compatible though an effort is made to make sure that current queries will not break. Micro versions (V5.1.1, V5.1.2 etc.) are released irregularly and often, and contain small hot fixes or backward compatible changes to the last minor version. (7) TODO: forum centric + github + themis -¿ confusing

Finally OHDSI provides ACHILLES as data quality tool for OMOP. It has been a common practice to run the tests and use this as a quality evaluation in numerous work [**?**]

Achilles for quality assessment ============================ A many previous authors, we used Achilles software to evaluate the data quality(4). It is open-source analytics software produced by OHDSI (6). This tool is used for data characterization, data quality assessment (Achilles Heel), and the visualization of observational health data (6). ACHILLES calculates summary statistics and includes a unique function for checking data quality, named Achilles Heel. Other team used this tool to practice data quality assess(4). Achilles Heel issued x errors and y warnings.

- 18h 50k patients: this testifies the model needs structural optimisations - difficult pour ajoute fr. - extension achilles how to ? - comparison with other paper about error/warnings. (Korean Yoon 36-¿28 errors) TODO: minimize achille errors (12-¿? errors)

2.2 Contribution

The *denormalized derived* tables appears to improves the computations costs and the verbosity of the SQL queries. In addition the resulting tables are much more human readily with the label of the concepts directly in the fact table. Hence a bit of denormalization improves greatly the experience of the data scientist by adding some redundancy in the data while not breaking existing SQL queries. The drawback is it would make it more tricky to update and keep the dataset consistent - which is non applicable because OMOP is generally a static dataset. The *microbiology derived* table simplifies the experience for data-scientists.

We provided many derived values. Community is welcome to improve it - From noteevents : weight, height, LVEF - From measurement : SOFA, IGSII, F/P, corrected Ca / K, BMI, corrected osmolarity

We also provided materialized views with denormalized structures : microbiology tables This is a more ICU centric data structures that may help for analytics.

- ethnicity_concept_id : only two strange concept_name Hispanic or non_Hispanic

Table 2. MIMIC to OMOP data flows

OMOP tables	Number of rows	MIMIC tables
PERSONS	46520	patients, admissions
DEATH	14849	patients, admissions
VISIT_OCCURRENCE	58976	admissions
VISIT_DETAIL	271808	transfers, service
MEASUREMENT	366272371	chartevents, labevents, microbiologyevents, outputevents
OBSERVATION	6721040	admissions, chartevents, datetimevvents, drgcodes
DRUG_EXPOSURE	24934758	prescriptions, inputevents_cv, inputevents_mv
PROCEDURE_OCCURRENCE	1063525	cptevents, procedureevents_mv, procedure_icd
CONDITION_OCCURRENCE	716595	admissions, diagnosis_icd
NOTE	2082294	notevents
NOTE_NLP	16350855	noteevents
COHORT_ATTRIBUTE	2628838	callout
CARE_SITE	93	transfers, service
PROVIDER	7567	caregivers
OBSERVATION_PERIOD	58976	patients, admissions
SPECIMEN	39874171	chartevents, labevents, microbiologyevents

2.3 Analytics

- French Paris hospital organized a datathon with MIMIC-OMOP 25 teams, 160 participants had 48 hours to undertake a clinical project using the OMOP MIMIC database thought 15000 requests. They had the opportunity to create mixed teams : clinicians brought the questions which need data mining, along with their expertise of the data ; data scientists judged the technical feasibility and eventually implement the various analysis needed This datahon had tested OMOP model in real statistical condition. A datathon was organised in collaboration with the MIT.(3)
- AP-HP calculation clusters, able to access to the data pre-loaded in Jupyter environments, where will be installed the most popular tools and libraries in R and Python, with Hadoop Spark

schema big data platform

MIMIC-OMOP dataset was 10 GO sized as ORC format. On the other side the same dataset is sized 400 GO on a PGSQL instance. Generating the indexes ORC automatically generates indexing. However these are much light than the PGSQL btree indexes specified by the OMOP DDL.

Writing standard queries (ie: with standard concepts) need to be familiar with the concept_relationship design and also master graph nature of some terminology such SNOMED-CT in order to grab all potential codes that might be related to the one analysts think of primarily. This complexity is inherent of terminology complexity and the closure table **??** design handle perfectly.

3 DISCUSSION

The choice of the ELT has several advantages over using a dedicated ETL software. It factorizes both people knowledges and computer resources allowing analysts to become implementers and review the code or contribute to the transformation with SQL as single language and technology.

By choosing a public git repository for both documentation and source code support this allows analysts to learn about the project and also learn how to contribute [28].

Any data transformation is susceptible to comes with bugs that can have huge impact in medical research. The RDBMS basements such transactions, normalisation and integrity constraints are built-in safeguards that have been useful along the process. Besides the implemented unit-tests ensures pasts or futures bugs are behind us. A ideal but complex [29] validation method would be to reproduce existing studies on OMOP and make sure results are consistent but yet only data distribution concordance have been verified.

The computation time of the ETL on the PGSQL instance on a modest personal computer is compatible with a community work where collaborator can clone the source code and setup a development instance to reproduce or improve the work. The choice of the ELT based mostly in SQL code lets end users with SQL background only to review and enhance the work. As a result, the targeted community is as broad as possible and we expect translational profiles to get involved.

The datathon has shown that distributed platforms with commodity hardware provides SQL tools allowing OLAP analysis with great performances that overcome OLTP

Table 3. Baseline characteristics MIMIC versus OMOP

items	MIMIC	OMOP
Persons (Number)	46.520	46.520
Admissions (Number)	58.976	58.976
Icustays (Number)	71.576	61.532
Gender, Female (Number, %)	20.399	20.399 (43 %)
Age (Mean)	64 years, 4 months	64 ans, 4 months
0-5	8110	8110
6-15	1	1
16-25	1434	1434
26-45	5962	5962
46-65	17375	17375
66-80	15793	15793
>80	10301	10301
Emergency	42071	42071
Elective	7706	7706
Surgical patients	19246	19246
Length of stay, hospital (median)	6.46 (Q1-Q3 : 3.74 -11.79)	6.59 (Q1-Q3 : 3.84 - 11.88)
Length of stay, ICU (median)	2.09 (Q1-Q3 : 1.10 - 4.48)	1.87 (Q1-Q3 : 0.95 - 3.87)
Mortality, ICU (Number, %)	5814 (9%)	5815 (9%)
Mortality, hospital (Number, %)	4511 (7%)	4559 (6%)
Lab measurements per admissions (mean)	478	678
Procedures per admissions (mean)	4.6	4.6
Drugs per admissions (mean)	82.8	82.8
Exit dignosis per admissions (mean)	11.0	11.0

RDBMS weaknesses. Hence it takes advantage of SQL language analytics features such grouping, windowing, joining and mathematic functions that often lack in NOSQL databases.

It is important that OMOP keeps a level of normalisation in order to simplify the ETL and make it consistent. However once done, it is judicious to give access to data-scientist to more denormalized tables and more specialized tables. Multiple concerns exists about OMOP performances and optimization. However there will never be a perfect multi-use case table, and this is the reponsibility of the data scientist to build his own tables, simplified, specialized for his research and answer efficiently and clearly his needs.

Derived data integrates quite well in OMOP. We made use of note_nlp to store information derived from notes, measurement to store numerical information and cohort_attributes to store scores. However it is still unclear if derived data should be stored per domain or if it should be stored in dedicated derived tables. We found out that there is a lack of tables to track provenance and description of such data. In addition we have been confused wether the derived information from notes should be directed into the dedicated domain table. However, notes may contain information from family members and they should'nt be lost or go into patient centered tables. For this reason, we think

An other missing aspect is some quality tables to assess and measure the quality of data. MIMIC had some column to keep track of corrupted information. It would be of interest to be able to keep the messy data and allow research on data cleaning and data quality and avoid removing information.

Last but not least, as stated in the introduction a good CDM for ICU would allow near realtime early warning systems and model inference on fresh data. OMOP is clearly designed to provide static dataset and does not have mecanisms for realtime ingestion, and data version control - it is not a datawarehouse. Making analysis on statics datasets is essential in order to have reproducible results. However when the algorithm needs to be moved at bed side, there is a need for a freshness of data, and a way to identify the patient that OMOP won't easily provide. That being said a solution such FHIR is a great way to implement realtime inference from EHR data and that's how FHIR and OMOP are complementary that yet have been investigated [**?**] but needs further optimizations.

Table 4. Row level Data lost

Relations	Error Percentage
inputevents_mv	10,00%
chartevents	0.04%
procedureevents_mv	3,00%
Note	0.04%

During this work the OMOP forum was very active. Working groups. It is a challenge to manage such large community from all moderator, contributors and from a user perspective. It appears it is not doable for most of people to get involved. The forum is full of details and information. It contrasts with the implementation guide that suffer from not being as well detailed. We think the OMOP community would greatly benefit from systematic and synthetic synchronisation between forum, mailing lists, github and end user documentation.

The datathon real life test has revealed the strong need to make the physical data model including comments on columns and table accessible and we found out that the open-source tool schema spy was of good help. In addition limit the misunderstanding by delivering SQL examples and we found out that the git repository is the best place for documenting and interact with the community.

4 CONCLUSION

OMOP model is very powerful as it allows a large spectrum of analysis from the black-boxes specialized local models to evidence based statistical analysis thanks to a easy to learn and accessible format. As seen OMOP model efficiency has some weaknesses as it looks to put the cursor more on consistency than on performances. However we have shown that it's easy to overcome the issues and enhance OMOP with a set of design or technology optimization and dedicated structure that in the end remains standards and shareable because derive from the original model.

Having such analytics power has a prior transformation and maintenance cost. The MIMIC to OMOP transformation has needed efforts that remains reasonable. It is and will ever be a work in progress as the standard concept mapping is a quite infinite process with constant improvement. Fortunately the released version is ready for research and already offers the same perimeter of data as the original MIMIC version and even more with the derived data.

As compared to the original MIMIC data model, working on OMOP offers the opportunity to write standard code and analysis that might benefit from or to other users internationally. The MIMIC-OMOP database is available online on physionet as well as the original MIMIC database. All the existing work is publicly available on github [**?**] and has been designed to be easily reviewed, copied or enriched easily accordingly to OMOP or MIMIC free philosophy by any end-user aware of SQL.

Future work on evaluation of the existing concept mapping though practical research studies on both local and standard coding will be made. In addition we expect to enhance the OHDSI USAGI concept mapping tool to allow international concept mapping suggestion to transform other foreign ICU databases. Finally investigation on how to articulate FHIR and OMOP to get best of two worlds (patient level versus multicentric level informations) and improve care practically at bed side will be done.

5 ACKNOWLEDGMENT

This research was conducted in fall 2008 when Vesa Välimäki was a visiting scholar at CCRMA, Stanford University. His visit was financed by the Academy of Finland (project no. 126310). The authors would like to Dr. Henri Penttinen for his comments and for the snare drum sample used in this work.

6 REFERENCES

[1] D. C. Angus, M. A. Kelley, R. J. Schmitz, A. White, J. Popovich, "Caring for the critically ill patient. Current and projected workforce requirements for care of the critically ill and patients with pulmonary disease: can we meet the requirements of an aging population?" *JAMA*, vol. 284, no. 21, pp. 2762–2770 (2000 Dec).

[2] E. Azoulay, C. Alberti, I. Legendre, C. B. Buisson, J. R. Le Gall, "Post-ICU mortality in critically ill infected patients: an international study," *Intensive Care Med*, vol. 31, no. 1, pp. 56–63 (2005 Jan).

[3] J. L. Vincent, "Is the current management of severe sepsis and septic shock really evidence based?" *PLoS Med.*, vol. 3, no. 9, p. e346 (2006 Sep).

[4] M. K. Ross, W. Wei, L. Ohno-Machado, ""Big data" and the electronic health record," *Yearb Med Inform*, vol. 9, pp. 97–104 (2014 Aug).

[5] Y. Zhang, S. L. Guo, L. N. Han, T. L. Li, "Application and Exploration of Big Data Mining in Clinical Medicine," *Chin. Med. J.*, vol. 129, no. 6, pp. 731–738 (2016 Mar).

[6] A. E. Johnson, T. J. Pollard, L. Shen, L. W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci Data*, vol. 3, p. 160035 (2016 May).

Table 5. Terminology Mapping coverage

Omop tables (domain)	Total_records	% Mapped_records	Total_concepts_source	% Mapped_concepts_source
PERSONS	93040	100,00%	43	100,00%
VISIT_OCCURRENCE	58976	100,00%	34	100,00%
VISIT_DETAIL	396930	100,00%	28	100,00%
MEASUREMENT				
OBSERVATION				
DRUG_EXPOSURE	24934751	37,00%	7410	53,00%
PROCEDURE_OCCURRENCE	1063525	99,00%	2218	98,00%
CONDITION_OCCURRENCE	716595	92,00%	6984	95,00%
CARE_SITE	144	100,00%	58	100,00%
SPECIMEN	39874171	70,00%	92	77,00%

[7] J. J. Gagne, "Common Models, Different Approaches," *Drug Saf*, vol. 38, no. 8, pp. 683–686 (2015 Aug).

[8] Y. Xu, X. Zhou, B. T. Suehs, A. G. Hartzema, M. G. Kahn, Y. Moride, B. C. Sauer, Q. Liu, K. Moll, M. K. Pasquale, V. P. Nair, A. Bate, "A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics: Implications for Active Drug Safety Surveillance," *Drug Saf*, vol. 38, no. 8, pp. 749–765 (2015 Aug).

[9] D. Madigan, P. B. Ryan, M. Schuemie, P. E. Stang, J. M. Overhage, A. G. Hartzema, M. A. Suchard, W. DuMouchel, J. A. Berlin, "Evaluating the impact of database heterogeneity on observational study results," *Am. J. Epidemiol.*, vol. 178, no. 4, pp. 645–651 (2013 Aug).

[10] H. Morgenstern, B. Rafaely, "Spatial Reverberation and Dereverberation Using an Acoustic Multiple-Input Multiple-Output System," *J. Audio Eng. Soc*, vol. 65, no. 1/2, pp. 42–55 (2017 Jan.Feb.), doi:https://doi.org/10.17743/jaes.2016.0063.

[11] O. H. Klungel, X. Kurz, M. C. de Groot, R. G. Schlienger, S. Tcherny-Lessenot, L. Grimaldi, L. Ibanez, R. H. Groenwold, R. F. Reynolds, "Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project," *Pharmacoepidemiol Drug Saf*, vol. 25 Suppl 1, pp. 156–165 (2016 Mar).

[12] C. Chronaki, A. Shahin, R. Mark, "Designing Reliable Cohorts of Cardiac Patients across MIMIC and eICU," *Comput Cardiol (2010)*, vol. 42, pp. 189–192 (2015).

[13] A. J. McMurry, S. N. Murphy, D. MacFadden, G. Weber, W. W. Simons, J. Orechia, J. Bickel, N. Wattanasin, C. Gilbert, P. Trevvett, S. Churchill, I. S. Kohane, "SHRINE: enabling nationally scalable multi-site disease studies," *PLoS ONE*, vol. 8, no. 3, p. e55811 (2013).

[14] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, P. J. Liu, X. Liu, M. Sun, P. Sundberg, H. Yee, K. Zhang, G. E. Duggan, G. Flores, M. Hardt, J. Irvine, Q. V. Le, K. Litsch, J. Marcus, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. Howell, C. Cui, G. Corrado, J. Dean, "Scalable and accurate deep learning for electronic health records," *CoRR*, vol. abs/1801.07860 (2018), URL `http://arxiv.org/abs/1801.07860`.

[15] H. Morgenstern, B. Rafaely, "i2b2 implemented over SMART-on-FHIR," *J. Audio Eng. Soc*, vol. 65, no. 1/2, pp. 42–55 (2017 Jan.Feb.), doi: https://doi.org/10.17743/jaes.2016.0063.

[16] "Bill Karwin's blog Rendering Trees with Closure Tables," `https://karwin.blogspot.com/2010/03/rendering-trees-with-closure-tables.html`, accessed: 2010-09-30.

[17] M. Garza, G. Del Fiol, J. Tenenbaum, A. Walden, M. N. Zozus, "Evaluating common data models for use with a longitudinal community registry," *J Biomed Inform*, vol. 64, pp. 333–341 (2016 12).

[18] J. M. Overhage, P. B. Ryan, C. G. Reich, A. G. Hartzema, P. E. Stang, "Validation of a common data model for active safety surveillance research," *J Am Med Inform Assoc*, vol. 19, no. 1, pp. 54–60 (2012).

[19] C. Reich, P. B. Ryan, P. E. Stang, M. Rocca, "Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases," *J Biomed Inform*, vol. 45, no. 4, pp. 689–696 (2012 Aug).

[20] C. Maier, L. Lang, H. Storf, P. Vormstein, R. Bieber, J. Bernarding, T. Herrmann, C. Haverkamp, P. Horki, J. Laufer, F. Berger, G. Honing, H. W. Fritsch, J. Schuttler, T. Ganslandt, H. U. Prokosch, M. Sedlmayr, "Towards Implementation of OMOP in a German University Hospital Consortium," *Appl Clin Inform*, vol. 9, no. 1, pp. 54–61 (2018 01).

[21] F. FitzHenry, F. S. Resnic, S. L. Robbins, J. Denton, L. Nookala, D. Meeker, L. Ohno-Machado, M. E. Matheny, "Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership," *Appl Clin Inform*, vol. 6, no. 3, pp. 536–547 (2015).

[22] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C.

Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Noren, Y. C. Li, P. E. Stang, D. Madigan, P. B. Ryan, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers," *Stud Health Technol Inform*, vol. 216, pp. 574–578 (2015).

[23] "MIMIC-OMOP repository," `https://github.com/MIT-LCP/mimic-omop`, accessed: 2010-09-30.

[24] P. A. Bernstein, J. Madhavan, E. Rahm, "Generic schema matching, ten years later," *PVLDB*, p. 2011.

[25] J. C. Denny, A. Spickard, K. B. Johnson, N. B. Peterson, J. F. Peterson, R. A. Miller, "Evaluation of a method to identify and categorize section headers in clinical documents," *J Am Med Inform Assoc*, vol. 16, no. 6, pp. 806–815 (2009).

[26] "LOINC news," `https://loinc.org/news/loinc-version-2-63-and-relma-version-6-22-are-now-available/`, accessed: 2010-09-30.

[27] "Terminology Mapping ICD9CM to SNOMED-CT," `https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html`, accessed: 2010-09-30.

[28] A. E. Johnson, D. J. Stone, L. A. Celi, T. J. Pollard, "The MIMIC Code Repository: enabling reproducibility in critical care research," *J Am Med Inform Assoc*, vol. 25, no. 1, pp. 32–39 (2018 Jan).

[29] A. E. W. Johnson, T. J. Pollard, R. G. Mark, "Reproducibility in critical care: a mortality prediction case study," presented at the F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, J. Wiens (Eds.), *Proceedings of the 2nd Machine Learning for Healthcare Conference*, vol. 68 of *Proceedings of Machine Learning Research*, pp. 361–376 (2017 18–19 Aug), URL `http://proceedings.mlr.press/v68/johnson17a.html`.

[30] D. Preis, "Phase Distortion and Phase Equalization in Audio Signal Processing—A Tutorial Review," *J. Audio Eng. Soc.*, vol. 30, no. 11, pp. 774–779 (1982 Nov.).

[31] J. S. Abel, D. P. Berners, "MUS424/EE367D: Signal Processing Techniques for Digital Audio Effects," (2005), unpublished Course Notes, CCRMA, Stanford University, Stanford, CA.

[32] C. Roads, "Musical Sound Transformation by Convolution," presented at the *Int. Computer Music Conf.*, pp. 102–109 (1993).

[33] C. Roads, *The Computer Music Tutorial* (MIT Press, Cambridge, MA), 1st ed. (1996).

APPENDIX

Filtering an audio signal with an allpass filter does not usually have a major effect on the signal's timbre. The allpass filter does not change the frequency content of the signal, but only introduces a phase shift or delay. Audibility of the phase distortion caused by an allpass filter in a sound reproduction system has been a topic of many studies, see, e.g., [30], [31].

$$\phi(\omega) = -\omega + 2\arctan\left(\frac{a_1 \sin\omega}{1 + a_1 \cos\omega}\right) \tag{1}$$

In this paper, we investigate audio effects processing using high-order allpass filters that consist of many cascaded low-order allpass filters. These filters have long chirp-like impulse responses. When audio and music signals are processed with such a filter, remarkable changes are obtained that are similar to the spectral delay effect [32], [33].

NOMENCLATURE

a$_c$ = condensation coefficient condensation coefficient condensation coefficient

TLR = Toll-like receptor

PAMPs = pathogen-associated molecular patterns condensation coefficient condensation

THE AUTHORS