MIMIC in the OMOP Common Data Model

N. PARIS and Adrien PARROT, WIND-DSI, AP-HP, Paris, France
nicolas.paris@aphp.fr, adrien.parrot@caramail.fr
T.POLLARD, A.E.W JOHNSON, Laboratory for Computational Physiology, MIT Institute for Medical Engineering and Science

Objectives : In the age of big data, the intensive care unit (ICU) is very likely to benefit from real-time computer analysis and modeling based on close patient monitoring and Electronic Health Record monitoring. MIMIC is still the first open access database in the ICU. Many studies have shown that common data models (CDMs) improve database searching by allowing code, tools and experience to be shared. OMOP CDM is spreading all over the world. We have transformed MIMIC into OMOP (MIMIC-OMOP) and assess the quality of the transformation, the benefits for analysts and the gains and potential for community contributions.
Material & Method: A documented, tested, versioned, exemplified and open repository has been put in place to support the transformation and enhancement of community source code. The resulting data set was evaluated over a 48-hours datathon with 160 participants.
Result: Most of the data correspond to the model and much of the terminology has been standardized with an investment of 2 people for 500 hours. The model demonstrated its ability to support community contributions and was well received during the datathon with 15,000 requests executed with a maximum duration of one minute. Conclusion: The resulting MIMIC-OMOP data set is ready for replicable research and as this is the first freely available OMOP data set with actual data depersonalized expectations are for increased and improved collaborations for ICUs, generalizable to other services.

INTRODUCTION

Intensive care units (ICUs) are care units where the demand for care increases[1] while mortality reaches up to 30%, which is a major health problem[2]. Studies have shown that intensivists use a limited level of evidence to guide decision making[3] and that medical practices are sparse and variable. Knowing that the ICU patient health record is very detailed and that there is a high density environment for data production is a paradox. The increasing adoption of electronic health record (EHR) systems around the world is capturing large amounts of clinical data[4] and data mining has the potential to play an important role in clinical medicine[5]. Indeed, based on important medical informations, expectations are to improve clinical outcomes and practices, enable personalized medicine and guide early warning systems, and also easily enroll a large, multi-center cohort while minimizing costs.

MIMIC (Medical Information Mart for Intensive Care) is a 10 year semi-automatic dataset of over 60,000 intensive care stays with very high granularity (including EKG) from two successive intensive care information systems at the Beth Israel Deaconess Medical Center in Boston. It is the first ICU database available free of charge and has been the subject of more than 300 international publications. However, its monocentric nature makes it difficult to generalize findings to other ICUs. The MIMIC relational data model reflects the original intensive care information system, as evidenced by the two separate inputevent_mv and ouputevent_cv [6] or the two separate terminologies for physiological data. This leads analysts (datascientists, statisticians, etc.) to reconcile this heterogeneity when pre-processing each study.

Some studies have shown that using a common data model (CDM) by generalizing the structural (data model) and conceptual (terminological model) design database allows for multicentre research, exploitation of rare diseases and catalyzes research by sharing practices, source code and tools [7, 8]. As Kahn and Al said [9], ”databases modelling is the process of determining how data are to be stored in a database”. It specifies data types, constraints, relationship and metadata definitions and provides a standardized way to represent resources/data and their relationships. However, some studies have shown that the results are not fully reproducible from one CDM to another [10] or from one centre to another [11]. Some approaches argue that keeping the local conceptual model [12] and the local structural model [13] leads to better results.

On one hand, keeping MIMIC on its specific form will not solve the limitation for multicentric research and, on the other hand, a fully standardized form would introduce other disadvantages, such as loss of datas, lower computational performances. The ideal solution is probably in between to allow local or standardized analysis depending on the research question.

OMOP (Observational Medical Outcomes Partnership Common Data Model) is a CDM originally designed for multi-centre drug-related adverse events and now extends to medical, clinical and genomic cases. OMOP provides structural and conceptual models such as SNOMED for diagnostics, RxNORM for drugs and LOINC for laboratory results. Several examples of database transformation to OMOP have been published [14, 15] and OMOP stores 682 million patient records from around the world [16]. Each clinical area is stored in different dedicated tables. The OMOP conceptual model is based on a closure table pattern [17] capable of ingesting any simple, hierarchical and also graph terminologies such as SNOMED-CT. In addition to local terminologies, OMOP specifies and maintains a set of standard terminologies to be mapped unidirectionally (local to standard) by implementers. Although OMOP has proven its reliability [18], the concept mapping process is known to have an impact on results [19] and the application of the same protocol on different data sources leads to different results [11]. This shows the importance of keeping local codes so that local analysis is always possible. Previous preliminary work has been done on the translation of MIMIC into OMOP [20]. This work remains to be refined and updated to be evaluated.

In a recent CDM comparative study [7, 21] OMOP obtained better results in database's evaluation in completeness, integrity, flexibility, simplicity of integration and implementability, adapting to the wider coverage of standard terminologies, a more systematic analysis thanks to an analytical library and to visualization tools, easier SQL models to use. In terms of conceptual approch, OMOP offers a broader set of standardised concepts. In terms of structural CDM, OMOP is very rigorous in how data should be loaded into a particular table when i2b2 for example is very flexible with a general table that solves all data domains. This rigorous approach is necessary for standardization. Previous work has loaded i2b2 with MIMIC-III [22] - however, the concept mapping step has limited the results since i2b2 design does not store local ontologies or informations where OMOP design keeps concept mapping unfinished. OMOP has the advantage of not making the terminology mapping step mandatory by keeping local codes in a usable format. Compared to the Fast Healthcare Interoperability Resources (FHIR) [23], OMOP performs better as a conceptual CDM because the FHIR does not specify the terminology to be used. OMOP relational model can be materialized in csv format and stored in any relational database when FHIR uses json files and needs some processing and more skills to exploit. We believe OMOP shares the advantages of all the above models.

In order to evaluate the transformation of MIMIC into OMOP, we propose to answer the following questions, such as the difficulty of transforming/maintaining an OMOP dataset from an local database, how the initial data is integrated and how much data is lost in the process, how the model should be queried simply and efficiently by analysts, how the design should be enriched by collaborative work, and finally to what extent OMOP can integrate and feed back to intensivists in a real-time context. This work is then evaluated according to 3 axes: Transformations, Contributions and Analyses. The first major contribution of this study is to evaluate OMOP in a freely accessible and well known database. The second major contribution is to provide a freely accessible dataset in OMOP format that could be useful to researchers. The third major contribution is to provide the OMOP community with useful transformations dedicated to intensive care that can be reused on any OMOP data set.

1 MATERIAL & METHOD

1.1 Data Transformation

All transformation processes are freely accessible to the public via the github website [24] maintained by MIT-LCP [6]. The repository is based on git and is designed for sharing, improvement, collaboration and reproducible work. Indeed, github is archived on a universal and durable software archive solution [25]. The github repository centralizes the various resources of this work such as documentation, source code, unit tests, as well as questioning examples, discussions and problem issues. It also indicates web resources such as the physical data model for MIMIC [26] and OMOP[27] datasets and the Achilles' web client[?]. The vast majority of source code is implemented in PostgreSQL 9.6.9 (PgSQL) because it is the primary support for the MIMIC database and allows the community to run our work on limited resources without needing a license. Finally, PgSQL has recently made huge efforts to better manage data processing. Some elaborate data transformations have been implemented as PgSQL functions. MIMIC-III version 1.4.21 (MIMIC) has been loaded into PgSQL with the provided scripts. The OMOP CDM version 5.3.3.1 (OMOP) target tables were created from the provided scripts with some small changes stored in the change script. OMOP which defines 15 standardized clinical data tables, 3 health system data tables, 2 health economics data tables, 5 tables for derived elements and 12 tables for standardized vocabulary. From MIMIC, we only used the clinical and derived tables.

The structural mapping aims at moving the MIMIC data into the right place in OMOP with some data transformations. It consists of three phases: conception, implementation and evaluation.

The conception phase has been looping over each MIMIC table and choosing an equivalent location in OMOP for each column. In general, the MIMIC documentation and the OMOP documentation were sufficiently informative. In several cases, we needed clarification from MIMIC contributors on the dedicated github repository, or from the OMOP community forum. Some complicated choices have been discussed in the repository [28] and can be tracked in the commit log.

The implementation is generally done by an Extract-Transformation-Load (ELT) process which is a methodology for migrating data from a source to a target location. ELT first extracts the data from the source location, then applies the transformations to a dedicated computer and finally loads the resulting data into the target location. Extract-Load-Transform (ELT) is a slightly different methodologie that does not use a dedicated server transformation. The data is extracted and loaded directly into the target location and subsequently transformed into the location. The ELT is composed of PgSQL scripts, each extracting information from the source or concept mapping tables, then transforming and loading an OMOP target table. The order of these scripts is important and is done sequentially through a main script. Measures have been taken to allow fast development cycles. A subset of 100 patients over the 46K total MIMIC patients was selected based on their broad representativeness of the database and cloned into a second instance to serve as a light and representative development set. OMOP indexes that would have slowed data migration with unnecessary calculations have been removed. Finally in last resort some modification of the structural model of OMOP have been made. A dedicated script recaps all of them. It contains columns name modifications, new columns, columns type modifications or database indexing modification.

The evaluation step is a set of controls to guaranty a correct transformation. Each ELT script has been tested using pgTAP, a unit test framework for PgSQL. This enable checking for loss of information, or regression during development. Each unit test script checks whether a particular OMOP target table is loaded correctly - most tables are covered and tests cover simple counts, aggregate counts or distribution checks. In order to compare overall statistics, some SQL queries have been setup to compare MIMIC and MIMIC/OMOP. The resulting tables are presented in the results section. Integrity constraints (primary keys, foreign keys, non-nullable columns) have been included to apply integrity checks at runtime. In particular, a global unique sequence have been introduced to identify relations between tables. Each source table has been added a global unique sequence incremented from 0 that serves as the primary key and links in the OMOP target tables. Each record in MIMIC is uniquely identified allowing to chain the information with OMOP while simplifying the primary/foreign key maintenance.

The conceptual mapping aims at aligning the MIMIC local terminologies to OMOP standard ones. It consists of three phases: integration, alignment and evaluation. The integration phase is about loading both kind of terminologies into the OMOP vocabulary tables. The OMOP terminologies are provided by the Athena tool [29] and where loaded with the provided programs. We have used an export with all terminologies without licensing limitations. The local terminologies have been extracted from the multiple MIMIC tables and loaded in the unique OMOP concept table. When possible, relevant informations from the original MIMIC tables have been concatenated in the concept_name column. New local MIMIC concepts were introduced in OMOP concept table and a concept_id value starting at 2 billion has been assigned to distinguish them from local MIMIC concepts. MIMIC local codes were also loaded into the vocabulary table with a concept_id identifier started at 2.1 billion (lower numbers are reserved for OMOP terminologies [30]). In OMOP concept table, MIMIC codes can also be distinguished with the vocabulary_id identifier equal to ”MIMIC code” and a domain_id identifier targeting the OMOP table in which the corresponding data is stored. Later, this domain information is sometime used in the ELT to send the information in the proper table with a so called ”concept-driven dispatching”. OMOP documentation explains that conceptual mapping has to be done before the structural mapping because the nature of the OMOP standard concepts guides in which table (domain) the information should be stored. The concept-driven dispatching methodology, enable changing the concept mapping after the transformation without modifying the underlying ELT code because the latter is dynamically based on the concept table content.

The alignment phase to standardizing local MIMIC codes into OMOP standards codes, had four distinct cases. In the first case, MIMIC is by chance already in OMOP standard terminology (e.g. LOINC laboratory results) and, therefore, the standard and local concepts are the same. In the second case, MIMIC is not in the standard OMOP terminology, but the mapping is already provided by OMOP (ex: ICD9/SNOMED-CT), so the domain tables have been loaded accordingly. In the third case, mapping is not provided, but it is small enough to be done manually in a few hours (such as demographic status, signs and symptoms). In the fourth case, mapping is not provided and terminology is enormous (admission diagnosis, drugs). Then, only a subset of the most represented code was manually mapped. When the mapping concept is required manually, a mapping csv file has been built. This solution can be adapted to medical users who do not have training in database engineering. The spreadsheet has several columns such as local/standard labels, ids and also comments, evaluation metrics and a script loads them into the PgSQL when completed. In order to catalyse the mapping process, the language algorithm has proven to be effective [31] although OHDSI provides USAGI [32]. We have chosen to use simple SQL queries that are flexible enough to be queried on demand or to generate a pre-filled csv with the best matches. It uses PGSQL full-text ranking features and links local and standard candidates with a rating function based on their labels. This work was followed by a intensivist check.

The evaluation phase was both qualitative and quantitative. The qualitative evaluation for newly generated mapping has consisted of tagging each mapping with a score between 0 and 1 and eventually write a commentary on each mapped concept. In case the mapping was provided by OMOP the evaluation was made manually by picking some concepts of each terminology. The quantitative evaluation measures the percent of concepts that are mapped to a standard with a SQL query.

1.2 Contribution

MIMIC provides a large number of SQL scripts to calculate derived scores and defined cohorts as known as ”contrib”. Some of them have been implemented on top of the OMOP format to load the OMOP derived tables.

A set of general denormalized tables has been built on top of the original OMOP format that have the concept_name related to the concept_id columns. The concept table is a central element of OMOP and, therefore, it is involved in many joins to obtain the concept label. Denormalized tables accelerate calculation time and provide an easier set of data for analysis.

In addition, a set of specialized materialized analysis views has been built on the original OMOP format. The OMOP microbiologicalevents table is a reorganization of the measurement table data of microorganisms and associated susceptibility testing antibiotics and is based on the MIMIC microbiologicalevents table. The OMOP icustays table allows to quickly obtain the patients admitted in resuscitation and is inspired by the MIMIC icustays tables.

The note_nlp table was originally designed to store final or intermediate derived information and metadata from clinical notes. When definitive, the extracted information is intended to be moved to the dedicated domain or table and then reused as regular structured data. When the information is still intermediate, it is stored in the note_nlp table and can be used for later analysis. To assess this table, we provided two information extraction pipelines. The first pipeline extracted numerical values such as weight, height, body mass index and left ventricular cardiac ejection fraction from medical notes with a SQL script. The resulting structured numerical values were loaded into the measurement or observation tables according to its domain. The second pipeline section extractor based on the apache UIMA framework divides notes into sections to help analysts choose or avoid certain sections of their analysis. While some methods already exist to extract medical sections [33], the prior work of describing sections was too high, and we opted for a naive approach. Section templates (such as ”Illness History”) have been automatically extracted from text with regular expressions, then filtered to keep only the most frequent (frequency > to 1%).

1.3 Data Analytics

A 48-hour open access datathon [34] was set up in Paris AP-HP (Assistance Publique des Hopitaux de Paris) once the MIMIC-OMOP transformation was ready for research to evaluate OMOP as an alternative data model in a real event. This datathon was organised in collaboration with the MIT. Scientific questions had been prepared in an online forum. Participants could introduce themselves and propose a topic or choose an existing one. OMOP has been loaded into apache HIVE 1.2.1 in ORC format. Users had access to the ORC dataset from a web interface jupyter notebooks with python, R or scala. A SQL web client allowed teams to write SQL from presto to the same dataset. The hadoop cluster was based on 5 computers with 16 cores and 220GB of RAM memory. The MIMIC-OMOP dataset has been loaded from a PgSQL instance to HIVE thought apache SQOOP 1.4.6 directly in ORC format. Participants also had access to the Schemaspy database physical model to access the OMOP physical data model with both table/-column comments and key primary/foreign relationships materializing the relationships between the tables. All queries have been logged.

2 RESULT

The evaluation of a system and a structural model is rather difficult [35]. Several articles attempted to assess the quality of the CDM [9, 21]. The criteria developed by Khan and Al[36], which refer to the metrics Moody and Shanks [35], have been adapted to assess the quality of the data transformation (table 1). Those metrics are mentionned along the results to bring some additional comparison metrics.

2.1 Data Transformation

The MIMIC to OMOP conversion was performed by two developers (a data engineer and an intensivist) for 500 hours. This includes ELT, git documentation, concept mapping, contributions and unit tests. ELT (with unit tests and generation of ready-to-load archive) on a subset of 100 patients takes five minutes and enables fast development cycles. The ELT lasts 3 hours to process the whole MIMIC database. The

Table 1. Transformation Quality Evaluation Metrics

Data Model Dimension	Descriptions
Completeness - structural mapping	Domain coverage : coverage of sources domains that are accommodated by the standard OMOP model
Completeness - conceptual mapping	Data coverage : coverage of sources data concepts that mapped to standard OMOP concept
Integrity	”Meaningful data relationships and constraints that uphold the intent of the data’s original purpose” [36]
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

resulting csv archive is about the same size as the original archive, and MIMIC-OMOP is also the same size as MIMIC once loaded in PgSQL and indexed.

The Structural Mapping results presented in the table 2 show the structural mapping, i.e. where the information goes from the MIMIC to the MIMIC-OMOP tables. Among of the 37 OMOP tables, the one related to hospital costs were not applicable, some related to derived data were not populated and some tables related to vocabulary are pre-loaded with terminology informations. The 26 tables of MIMIC have been dispatched into 19 OMOP tables. The reduced number of tables results from the differences in design of both models. OMOP stores all the terminologies into one table whereas MIMIC has one table for each typology and the same applies for facts data that are grouped by nature in OMOP while MIMIC tables are more specialized and respects the source EHR’s design. For example the measurement gather measured information and combines 4 source tables resulting in 366272371 rows which is 20% more than the largest MIMIC table. To some extends this is a regression in terms of performances. Two important tables are provided by OMOP to represent the relationship between the data : concept_relationship and fact_relationship. We used them to bind the drugs into a solution, for microbiology / antibiograms and for visit_detail and caresite links. The following SQL query (listing 1) shows how a microorganism is linked to its susceptibility test by a fact_relationship. This results are in favor with a good flexibility. However this flexibility affects the simplicity and the performances of the model by increasing the number of joins within SQL queries.

Listing 1. Original table microbiology SQL query

```
SELECT measurement_source_value
 , value_as_concept_id
 , concept_name
FROM measurement
JOIN concept resistance
      ON value_as_concept_id = concept_id
JOIN fact_relationship
      ON measurement_id =  fact_id_2
JOIN
(
    SELECT measurement_id AS id_is_staph
    FROM measurement m
    WHERE TRUE
    AND measurement_type_concept_id = 2000000007
        -- 'Labs - Culture Organisms'
    AND value_as_concept_id = 4149419
        -- 'Staph aureus coag +'
    AND measurement_concept_id = 46235217
        -- 'Bacteria identified in Blood product
           unit.autologous by Culture'
)␣staph␣ON␣id_is_staph␣=␣fact_id_1
WHERE␣TRUE
AND␣measurement_type_concept_id␣=␣2000000008
␣␣--␣'Labs - Culture Sensitivity'
```

The table 3 presents the basic characterization of the MIMIC-OMOP population in relation to the MIMIC and assesses the overall quality of structural mapping. Fortunately most statistics remain similar between the two versions with still few differences. The table 3 MIMIC contains 61,532 intensive care stays while OMOP contains 71,576 intensive care stays. This represents a 16% increase in stays due to our ELT methodology as explained in the methods. This table shows that the number of laboratory measurements per admission is increased. This is because MIMIC-OMOP gathers laboratory data from both the MIMIC dedicated laboratory table and the chartervents table which is usually not considered for this purpose. By design, MIMIC aggregates information from various systems[28]. Thus, the transfer information is divided into several tables, such as admissions, transfers and icustays. OMOP centralizes this information in the detail of the visit_detail. We added emergency stays as a normal location for patients throughout their hospital stay (unlike what had been done by MIMIC). Icustays raw mimic table has been removed because it is a table derived from the transfer table [37] and we decided to assign a new visit_detail for each ICU stay (based on the transfer table) while mimic preferred to assign a new icustay stay if a new admission occurs > 24h after the end of the previous stay. For laboratory tests when it makes sense to put a specimen (i.e. a blood sample) for many laboratory results (because one blood sample can be used for several tests), we decided to create as

Table 2. MIMIC to OMOP data flows

OMOP tables	Number of rows	MIMIC tables
CONCEPT	46520	d_cpt, d_icd_procedures, d_items, d_labitems
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	noteevents
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

many rows of samples as laboratory tests because the information is not present in MIMIC. The same was true when date information was not provided (start /end_datetime for drug_exposure).

We estimated the loss of information during the ELT process by measuring the percentage of both columns and rows lost in the process as other previous studies have done [15]. From 40% to 80% of the columns in the sources have been deleted. Almost all the deleted columns were redundant with others or provided derived information. The main concern is the loss of some timestamps. For example, the MIMIC chartevents tables provides the storetime and charttime columns, but OMOP only provides one location to store timestamp. Thus, MIMIC storetime column was eliminated during ELT. As mentioned in the methods the error lines have been deleted in the process (marked with a status column in the MIMIC tables inputevents_mv, chartevents, procedureevents_mv, note). According to the tables 4, four MIMIC tables have lost rows in the process. All of them were tagged in MIMIC as erroneous or cancelled informations since OMOP does not consider those information to be loaded. In the other hand some data has been added which is discussed later in the contribution section.

A set of minor modifications of the OMOP structure was made in order to feet the data. All character typed columns with limited length have been modified to unlimited length since it could cause unpredictable truncation of content, while having no negative impact on PgSQL storage size or performance. The visit_occurrence and the visit_detail tables have been corrected accordingly to some discussions on the OHDSI forum. The nlp_note table has been completed by fields corresponding to the online documentation. The character offset column has been divided into two integer type 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.

All the code to create these statistics is provided on the github repository [24]. During the ELT process were created a lot of unit tests thanks to the pgTap library. All are available on our github [24]. All the tests passed. Moreover OMOP had a 100% match of the integrity constraints and the relationships of the data models. The second axe was Achilles evaluation. Like many previous authors, we used the Achille software to assess data quality [38]. It is an open-source analysis software produced by OHDSI [39]. This tool is used for data characterization, data quality assessment (Achilles' heel) and health observation data visualization [39]. It has been common practice to perform Achilles tests and use it as a quality assessment in many works. After 18 hours of computations Achilles Heel issued 12 errors and y warnings. This result is correct compared to other studies [38] We believe that this tool has several limitations. It does not evaluate the structural change, it is difficult to understand some error messages and we decide to process more evaluation tests.

The Conceptual Mapping results are presented in the table 5.

Often we have mapped many source concepts to a unique standard concept_id because MIMIC provides a large number of equivalent concepts. For example, for body temperature, MIMIC provides 11 distinct concepts (Temperature F, Temperature C (calc), Temp Skin [C], Temperature Fahrenheit, Temp Axillary [F], Temperature C, Temperature F (calc), Temperature Celsius, Temp Rectal [F],

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

Table 4. Row level Data lost

Relations	Rows lost
inputevents_mv	10,00%
chartevents	0.04%
procedureevents_mv	3,00%
Note	0.04%

Terminology mapping was evaluated by a physician. We tried to evaluate this automatic OMOP mapping. We check 100 elements for each mapping used (NDC, ICD9 and CPT4). CIM9 and CPT4 are correctly mapped to SNOMED (100%). But only 85% of NDCs are linked to a correct RxNorm code. Partly because of an incorrect NDC drug code (from MIMIC), partly because only 78% of NDC codes are mapped to Rxnorm. Moreover, even if this does not seem to have affected our ELT we know that not all ICD-9-CM codes can have a one-to-one match with SNOMED, some are one to several (28%) [40]. This results are in favor with a good integration of the model. OMOP's terminology coverage has already been rated as excellent [21]. We used the OMOP provided mapping to standardize a consequent set of MIMIC non-standard terminologies (NDC-RxNorm, ICD9-SNOMED, CPT4-SNOMED).

In several cases, OMOP had not sufficient concepts adapted to ICU specific cases. The actual visit_detail table does not introduce relevant information and duplicate information from visit_occurrence table. For admitting_concept_id and discharge_to_concept_id columns, we extended the dictionary to track bed transfers and room transfers. For visit_type_concept_id column we assigned a new concept for any level of granularity necessary for your use case (ward, bed…) These added concepts are susceptible to be replaced by new OMOP standard concepts in the future and have been introduced with concept_ids between 2 billion and 2.1 billion to distinguish them with OMOP concepts (0 to 2B) and MIMIC locals (>2.2B).

The unmapped concepts are the concept id = 0 (no mapping concept). The value zero for concept_id can appear in different cases. In the first case, the local concept has no equivalent in the standard concept set. In the second case, it has not yet been mapped and may have a standard equivalent. In the third case, the value is missing and cannot be mapped. In our opinion, although not all of these cases can be used for standard queries, they should have a different concept identifier in order to be treated differently (not only concept_id = 0). Some of the domains_id do not match the table name, it makes sense because the observation domain can be measurement table and vice versa. Although various types of information are stored in the measurement table, the dedicated OMOP concepts for the measurement_type_concept_id column were not sufficient to distinguish them. We have added some.

2.2 Contribution

Some MIMIC raw information have been transformed and added to meet the structural model. The laboratory values have been splitted into operators,

Table 5. Terminology Mapping coverage

Omop tables (domain)	Records	% Mapped records	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%

values, and units when needed with a dedicated stored procedure. The free text conditions have been normalized and mapped to OMOP standard codes to meet the conceptual model. The note section extraction pipeline, 1200 sections were collected and then manually filtered to exclude false positives. 400 similar groups were highlighted. The extracted sections have not been mapped to standard terminology such as LOINC CDO. The reason for this is that the CDO LOINC decided to delete its sections from its standard, considering that these sections were not widely used [41]. Common derived information were introduced and loaded: corrected serum calcium, corrected serum potassium, P/F ratio, corrected osmolarity, SAPSII.

Denormalized derived tables improve calculation costs and SQL query verbosity. In addition, the resulting tables are much more human readable with the concept label directly in table and greatly reduces joins. Therefore, a little denormalization greatly improves the analysts experience of the data model and the simplicity by adding some redundancy in the data while not interrupting existing SQL queries. Moreover, these normalized views are backward compatible and remain standardized allowing the creation of multicentric algorithms.

Materialized derived views from microbiologyevents and icustays simplify the experience for scientists (listing 2). This results reflect the lack of simplicity of the model in its original form but this can be easily overcome.

Listing 2. Optimized and denormalized table microbiology SQL query

```
SELECT
 , o.organism_source_value
 , o.organism_concept_id
 , o.antibiotic_concept_id
 , o.antibiotic_concept_name
 , o.antibiotic_source_value
 , o.antibiotic_interpretation_concept_id
 , o.antibiotic_interpretation_concept_name
 , o.MIC_operator_concept_id
 , o.MIC_operator_concept_name
 , o.MIC_value_as_number
 , o.MIC_source_value
FROM microbiology
WHERE TRUE
AND value_as_concept_id = 4149419
   -- 'Staph aureus coag +'
AND measurement_concept_id = 46235217
   -- 'Bacteria identified in Blood product
      unit.autologous by Culture';
```

As indicated in the methods section, we have provided many derived values. Again, the community is welcome to evaluate and improve them.

This results are in favor with a a good flexibility of the model allowing to store derived data.

2.3 Analytics

The French Hospital of Paris (AP-HP) organized a datathon with MIMIC-OMOP. 25 teams, 160 participants had 48 hours to undertake a clinical project using the database MIMIC-OMOP through 15000 requests with a maximum duration of one minute. They had the opportunity to create mixed teams: clinicians brought the issues that required data mining, as well as their data expertise; data scientists judged the technical feasibility and finally implemented the various analyses needed. Writing standard queries (i.e. with standard concepts) requires knowing the organization of relational models (SQL) and also mastering the graphical nature of certain terminologies such as SNOMED-CT in order to capture all potential codes that might be related to the one analysts think of first. This complexity is inherent in terminology complexity and the closure table ??. It is therefore not specific to MIMIC or OMOP. Overall the teams found MIMIC-OMOP easy to learn and managed to produce results at the end of the datathon. This results are in favor with a good understandibility and simplicity of the model.

3 DISCUSSION

The choice of ELT has several advantages over the use of dedicated ELT/ETL software. It factors both people's knowledge and computer resources allowing analysts to become implementers and revise code or contribute to transformation with SQL as the unique language and technology.

By choosing a public git repository for documentation and source code support, this allows analysts to learn more about the project and learn how to contribute [42].

Any data transformation is likely to generate bugs that can have a huge impact in medical research. The foundations of the RDBMS, such as transactions, standardization and integrity constraints, are integrated safeguards that have been useful throughout the process. In addition to the implemented, unit tests ensure that past or future bugs are behind us. An ideal but complex validation method [43] would be to replicate existing OMOP studies and ensure that the results are consistent.

The calculation time of the ETL on the PGSQL instance on a modest personal computer is compatible with a community work where the collaborator can clone the source code and configure a development instance to reproduce or improve the work. The choice of ELT based mainly on SQL code allows end users with SQL background only to review and improve the work. As a result, the target community is as broad as possible and we expect translation profiles to be involved.

The datathon showed that platforms distributed with basic hardware provide SQL tools for OLAP analysis with excellent performance that overcome OLTP RDBMS weaknesses. Therefore, it takes advantage of SQL language analysis functions such as grouping, windowing, assembling and mathematical functions that are often missing in NOSQL databases.

It is important that OMOP maintains a level of standardisation in order to simplify ETL and make it consistent. However, once done, it makes sense to give access to scientific data at more denormalized and special tables. There are many concerns about OMOP's performance and optimization. However, there will never be a perfect multi-purpose case table, and it is the responsibility of the data scientist to build his own, simplified, specialized tables for his research and to respond effectively and clearly to his needs.

The derived data integrate quite well into OMOP. We used note_nlp to store information derived from scores, measurement to store numerical information and cohort_attribute to store scores. However, it is not yet clear whether derived data should be stored by domain or whether it should be stored in dedicated derived tables. We found that there are no tables to track the source and description of these data.

Another missing aspect is the quality tables for assessing and measuring data quality. MIMIC had a column to keep track of corrupted information. It would be interesting to be able to keep the disordered data and allow research on data cleaning and quality and avoid deleting data in training.

Last but not least, as noted in the introduction, a good CDM for the ICU would allow for near real-time early warning systems and inference modelling on fresh data. OMOP is clearly designed to provide a static data set and does not have real-time ingestion and data version control mechanisms - it is not a data warehouse. Analysis of static data sets is essential for reproducible results. However, when the algorithm needs to be moved to the bed side, it is necessary to have fresh data and a means of identifying the patient that OMOP will not easily provide. That said, a solution like FHIR is a great way to implement real-time inference from EHR data, and that's how FHIR and OMOP are complementary. This has already been studied [?] but needs further optimisation.

During this work, the OMOP forum was very active. It is a challenge to manage such a large community of all moderators, contributors and from the user's point of view. It seems that it is not possible for most people to get involved. The forum is full of details and in training. It contrasts with the implementation guide, which suffers from not being as detailed. We believe that the OMOP community would greatly benefit from a systematic and synthetic synchronization between the forum, mailing lists, github and end user documentation.

The real life test of the datathon revealed the strong need to make the physical data model accessible, including comments on columns and tables, and we discovered that the open-source tool schema spy tool was a good help. In addition, we found that the git repository is the best place to document and interact with the community.

The conversion of MIMIC to OMOP format is finally the first step in our plan. In France, the exchange of patient data between centres is rightly very regulated. We think that MIMIC-OMOP could be used to learn OMOP and build algorithms. Then it is necessary to send the data we imagine to send the algorithms in the centers having transcribed their database in OMOP format. This would resolve confidentiality issues.

TODO:

4 CONCLUSION

The OMOP model is very powerful because it allows a broad spectrum of analysis from specialized local models to evidence-based statistical analysis in an easy-to-learn and accessible format.

As we have seen, the effectiveness of the OMOP model has some weaknesses because it seems to focus on consistency rather than performance. However, we have shown that it is easy to overcome problems and improve OMOP with a set of design or technology optimization and a dedicated structure that ultimately remains a standard and shareable because it derives from the original model.

To have such analysis power has a cost of transformation and prior maintenance. The transformation of MIMIC into OMOP has required efforts that remain reasonable. It is and always will be a work in progress because standard concept mapping is an almost infinite process with constant improvements. Fortunately, the published version is search-ready and already offers the same scope of data as the original MIMIC version and even more with the derived data.

Compared to the original MIMIC data model, working on OMOP offers the ability to write standard code and analyses that could benefit other users internationally. The MIMIC-OMOP database is available online on physionet as well as the original MIMIC database. All existing works are publicly available on github [28] and have been designed to be easily revised, copied or enriched according to the OMOP or MIMIC philosophy by any end user who knows SQL.

Future work on the evaluation of existing concept mapping through practical research studies on local and standard coding will be carried out. In addition, we plan to enhance the USAGI OHDSI concept mapping tool to allow the international concept mapping suggestion to transform other foreign ICU databases. Finally, research on how to articulate FHIR and OMOP to get the best of both worlds (information at the patient level versus information at the multicenter level) and improve near- bedside care will be done.

5 GRANT

This research received non specific grant from any funding agency in the public, commercial, or not-for-profit sectors

6 REPOSITORY WORK

All the latex files, statistics, pdf of the article are provide online :

7 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).

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

[8] P. R, L. T, “Data enclaves for sharing information derived from clinical and administrative data,” JAMA (2018), doi:10.1001/jama.2018.9342, URL +http://dx.doi.org/10.1001/jama.2018.9342.

[9] M. G. Kahn, D. Batson, L. M. Schilling, “Data Model Considerations for Clinical Effectiveness Researchers,” Medical Care, vol. 50, pp. S60–S67 (2012 Jul.), doi:10.1097/MLR.0b013e318259bff4, URL https://insights.ovid.com/crossref?an=00005650-201207001-00013.

[10] 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).

[11] 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).

[12] 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.

[13] 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).

[14] 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).

[15] 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).

[16] 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. Madi-

gan, P. B. Ryan, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers," Stud Health Technol Inform, vol. 216, pp. 574–578 (2015).

[17] "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.

[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] J. G. Md Shamsuzzoha Bayzid, Vojtech Huser, "Conversion of MIMIC to OHDSI CDM," (2016).

[21] 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).

[22] 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).

[23] "fhir-doc," https://www.hl7.org/fhir/, accessed: 2018-08-10.

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

[25] "Universal Archive web solution," https://www.softwareheritage.org/, accessed: 2010-09-30.

[26] "mimic-schemaspy," https://github.com/MIT-LCP/mimic-omop/blob/master/mimic/doc/schemaspy/index.html, accessed: 2018-08-10.

[27] "omop-schemaspy," https://github.com/MIT-LCP/mimic-omop/blob/master/omop/doc/schemaspy/index.html, accessed: 2018-08-10.

[28] "MIMIC-OMOP github," https://github.com/MIT-LCP/mimic-omop/issues/, accessed: 2018-08-10.

[29] "athena standardized vocabularies," https://www.ohdsi.org/analytic-tools\/athena-standardized-vocabularies/, accessed: 2018-08-10.

[30] "OMOP-doc," http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:concept, accessed: 2018-08-10.

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

[32] "usagi," https://github.com/OHDSI/Usagi, accessed: 2018-08-10.

[33] 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).

[34] "mimic omop datathon AP-HP website," http://blogs.aphp.fr/dat-icu/, accessed: 2018-08-10.

[35] D. L. Moody, G. G. Shanks, "Improving the quality of data models: empirical validation of a quality management framework," vol. 28, no. 6, pp. 619–650, doi:10.1016/S0306-4379(02)00043-1, URL http://linkinghub.elsevier.com/retrieve/pii/S0306437902000431.

[36] M. G. Kahn, D. Batson, L. M. Schilling, "Data Model Considerations for Clinical Effectiveness Researchers," vol. 50, pp. S60–S67, doi:10.1097/MLR.0b013e318259bff4, URL https://insights.ovid.com/crossref?an=00005650-201207001-00013.

[37] "icustays documentation," https://mimic.physionet.org/mimictables/icustays/, accessed: 2018-08-10.

[38] D. Yoon, E. Ahn, M. Young Park, S. Yeon Cho, P. Ryan, M. J. Schuemie, D. H. Shin, H. Park, R. W. Park, "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," vol. 22, p. 54 (2016 02).

[39] "athena standardized vocabularies," https://www.ohdsi.org/web/achilles/, accessed: 2018-08-10.

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

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

[42] 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).

[43] 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.

THE AUTHORS