MIMIC in the OMOP Common Data Model*

Nicolas PARIS, WIND-DSI, AP-HP, Paris, France AND Adrien PARROT, WIND-DSI, AP-HP, Paris, France
(nicolas.paris@aphp.fr) (adrien.parrot@caramail.fr)

AP-HP, Paris, France

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-hour 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.

0 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 (CCI) 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 CCI, 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 [?]kahn-data-2012), "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 stan-

*To whom correspondence should be addressed Tel: +33111111111; e-mail: nicolas.paris@aphp.fr

dardized 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 [9] or from one centre to another [10]. The lighter approaches argue that maintaining the local conceptual model [11] or the original conceptual and structural model [12] of the research database leads to better results. On the 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. 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 [13, 14] and OMOP stores 682 million patient records from around the world [15]. Each clinical area is stored in different dedicated tables. The OMOP conceptual model is based on a closure table pattern [16] 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 [17], the concept mapping process is known to have an impact on results [18] and the application of the same protocol on different data sources leads to different results [10]. 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 [19]. This work remains to be refined and updated to be evaluated.

In a recent CDM comparative study [7, 20] OMOP obtained better results in the criteria of the evaluation database compared to the other models: completeness, integrity, flexibility, simplicity of integration and implementability, adapting to the wider coverage of standard terminologies, providing a more systematic analysis with an analytical library and visualization tools, providing SQL models easier to use. That is why 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 [21] - 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 FHIR, OMOP performs better as a conceptual CDM because the FHIR does not specify the terminology to be used. In terms of structural CDM, the 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 processes are freely accessible to the public via the github website [22] 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 [23]. 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[?] and OMOP[?] 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 enormous 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 with the scripts provided in a PgSQL instance. 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. We didn't use the health economics data tables (not provided by MIMIC). Indexes that would have slowed data migration with unnecessary calculations have been removed. Integrity constraints (primary keys, foreign keys, non-nullable columns) have been included to apply integrity checks at runtime. A subset of 100 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. Each source table has been added a global unique sequence incremented from 0 that serves as the primary key and link in the OMOP target tables.

Extract-Transformation-Load (ETL) is a methodology for migrating data from a source to a target location. ETL 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) processes are slightly different methodologies 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.

Each ELT part has been tested using pgTAP, a unit test framework for PgSQL. This allows you to check for loss of information, or code 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.

All character type columns with limited length have been modified as follows unlimited since it could cause unpredictable truncation of content, and it has no negative impact on PgSQL storage size or performance. The visit_occurrence and the visit_detail tables have been corrected accordingly 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 ingeter 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.

The structural transformation took place in several phases. The first phase consists of looping 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 on the dedicated forum. All choices have been discussed in the repository [24] and can be tracked in the commit log.

For the second step, we tried not to infer any results. For laboratory tests when it makes sense to put a specimen (i.e. a body sample) for many laboratory results (because one blood sample can be used for several tests), we decided to create as 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). Chartevents and labevents tables provide many number fields as a string, which is not practical for statistical analysis. We provide a standard and easy enhancement by the community model to extract the numerical value of the string with a PgSQL function. The results of the MIMIC laboratory have been restructured to adapt to OMOP format. In particular, the numerical value (value_as_number) is accompanied by a mathematical operator (concept_operator_id) and a unit of measurement (concept_unit_id). All lines marked in error have not been converted to OMOP format since the MIMIC team plans to delete them at the next release.

By design, MIMIC aggregates information from various systems. 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 [25] and we decided to assign a new visiti_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.

The conceptual transformation uses OMOP vocabulary tables that have been loaded from an Athena export [26] of all terminologies without license limitations.

Local MIMIC codes are also loaded into the concept table with a concept_id identifier from 2.1 billion (below this number is reserved for OMOP terminologies [?]). MIMIC codes can 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 used in the ELT to send the information in the proper table. As the OMOP model did we adopt a "concept-driven methodology", domain of each local concept drive the concept to the right table.

Where possible, relevant information from the original MIMIC tables has been concatenated in the concept_name column. New local MIMIC concepts were introduced and given a value from 2 billion to distinguish them from local MIMIC concepts.

When it came to standardizing local MIMIC codes in OMOP standards codes, there were 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 [27] although OHDSI provides USAGI [?]. 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.

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.

The actual visit_detail table does not introduce relevant information and duplicate informations 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...)

1.2 Contribution

MIMIC provides a large number of SQL scripts to calculate derived scores and define cohorts. Some of them have been implemented in OMOP format and fill OMOP cohort tables. Common derived information was introduced and loaded: corrected serum calcium, corrected serum potassium, P/F ratio, corrected osmolarity, SAPSII.

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. Normalized 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. Microbiologicalevents table is a reorganization of the measurement table datas 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 [?], 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%). 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 [?].

1.3 Data Analytics

A 48-hour open access datathon [?] 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 were been logged.

2 RESULT

2.1 Data Transformation

The MIMIC to OMOP conversion was performed by two developers (a data engineer and an intensivist) for an estimated 500 hours. This includes ELT, git documentation, concept mapping, contributions and unit tests. ELT (with unit tests and generation of ready-to-load archives) on the subset of 100 patients takes five minutes and enables rapid development cycles. On all MIMIC data, the ELT lasts 3 hours. The resulting csv archive is about the same size as the original archive, and it is also the same once instantiated in PGSQL and indexed.

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

The evaluation of a system and a structural model is rather difficult [28] but we have tried to evaluate it through several axes.

The first axe was the unit tests. During the all ETL process we created a lot of unit tests thanks to pgTap library. All are available on our github [22]. All the test passed.

The second axe was Achilles evaluation. Like many previous authors, we used the Achille software to assess data quality [29]. It is an open-source analysis software produced by OHDSI [30]. This tool is used for data characterization, data quality assessment (Achilles' heel) and health observation data visualization [30]. It has been common practice to perform Achilles tests and use it as a quality assessment in many works. Achilles Heel issued 12 errors and y warnings. This result is correct compared to other studies [29] 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.

Several articles have attempted to assess the quality of the CDM (8, 9). The criteria developed by Khan and Al[?], which refer to the metrics Moody and Shanks [28], have been adapted for our study. Our study does not want to assess the quality of the CDM. But we adapt these criteria to assess the quality of the data transformation (table 1). Comprehensiveness, integrity, flexibility, integration, implementability will be discussed in this part (data transformation). Understandability and simplicity will be assessed in the analytical parts, in the actual application. All the code to create these statistics is provided on the github repository [22].

The table 2 shows the structural mapping, i.e. where the information goes and links between the MIMIC tables and the MIMICIII-OMOP tables go. The largest relationship is the measuring table with 366272371 rows. Since OMOP is a conceptual model, the same type of data goes into the same table. The best example can be the measurement table which is field by 4 source tables. That's because all numerical data should go to that table. All MIMIC domains are linked to the OMOP domain. Structural mapping was not a problem for our work.

The table 3 presents the basic characterization of the MIMIC-OMOP population in relation to the MIMIC and assesses the overall quality of our semantic mapping.

Fortunately most statistics remain similar between the two versions. There are still some 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 the laboratory data from MIMIC chartervents have been extracted and treated as a laboratory.

We tried to estimate the percentage of records loaded from the source database at MIMIC-OMOP. We estimate the percentage of columns and rows lost in the process as other studies have done [14].

According to the tables, 40% to 80% of the columns in the sources that do not correspond to OMOP have been deleted. The exact columns removed are provided on the article github. 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 provide the storetime and charttime columns, but OMOP provides only 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). The following table 4 shows the number of lines with deleted errors.

This table ?? shows the results of automatic and manual mapping. The unmapped concepts are the concept id = 0 (no mapping concept). To improve this mapping, we need collaborative work. Terminology mapping was evaluated by a physician. The value zero for concept_id can appear in very 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

Table 1. Data Transformation Quality Evaluation Metrics

Data Model Dimension	Descriptions
Completeness - structural mapping	Domain coverage : coverage of sources domains that are accommodated by the standa
Completeness - semantic mapping	Data coverage : coverage of sources data concepts that mapped to standard OMOP c
Integrity	"Meaningful data relationships and constraints that uphold th eintent of the data-s o
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 stan
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 o

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. Often we have mapped many source concepts to a standard concept_id. This is 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 Fahren- heit, Temp Axillary [F], Temperature C, Temperature F (calc), Temperature Celsius, Temp Rectal [F], Temp Rectal, Blood Temperature CCO (C)). Our mapping links all this to a single concept called temperature. All units have been converted to Celcius.

OMOP had a 100% match of the constraints and relationships of the data models and is a flexibility is high. Two important tables are provided with OMOP models to match the relationships : concept_relationship and fact_relationship. It is used to represent the relationship between the data. We used it to bind the drugs into a solution, for microbiology / antibiograms and for visit_detail and caresite links. The following SQL query shows how a microorganism is linked to its susceptibility test by a fact_relationship.

```
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 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␣measurement_type_concept_id␣=␣2000000008␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
␣␣--␣concept.concept_name␣=␣'Labs - Culture Sensitivity'
```

OMOP's terminology coverage has already been rated as excellent [20]. We used OMOP mapping for NDC-RxNorm, ICD9-SNOMED, CPT4-SNOMED. This was really useful to integreate MIMIC database because it provides a lot of non-standard terminology already mapped by the OMOP community. 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 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%) [31]

OMOP has been available for 9 years. Its models and concepts provided are license free, the community is large and has been very helpful. Full versions are generally published annually and are not backwards compatible. Minor versions are not guaranteed to be backwards compatible, although an effort is made to ensure that current requests will not break. Microversions are published irregularly and often, and contain small corrections or changes that are backward compatible with the latest minor version [32]

2.2 Contribution

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 data scientist's experience 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. The community should do more.

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

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

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 OMOP or MIMIC-OMOP. Overall the teams found MIMIC-OMOP easy to learn. All teams managed to produce results at the end of the datathon.

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 [33].

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 [34] 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 notei_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

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

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. In addition, we do not know if the information derived from the notes should be directed to the dedicated domain table. However, the notes may contain information from family members and should not be lost or found in patient-centred tables.

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.

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 [24] 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 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).

[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] 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).

[10] 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).

[11] 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.

[12] 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, “Multicentre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project,” Phar-

macoepidemiol Drug Saf, vol. 25 Suppl 1, pp. 156–165 (2016 Mar).

[13] 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).

[14] 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).

[15] 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).

[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] 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).

[18] 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).

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

[20] 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).

[21] 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).

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

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

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

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

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

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

[28] 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.

[29] 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).

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

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

[32] "OMOP-CD ," https://github.com/OHDSI/CommonDataModel/, accessed: 2018-08-10.

[33] 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).

[34] 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.

[35] 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.).

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

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

[38] 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., [35], [36].

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

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 [37], [38].

NOMENCLATURE

a_c = condensation coefficient condensation coefficient condensation coefficient
TLR = Toll-like receptor
PAMPs = pathogen-associated molecular patterns condensation coefficient condensation

THE AUTHORS