Name:	Md. Shamsuzzoha Bayzid
Affiliation:	University of Texas at Austin
Email:	shams.bayzid@gmail.com
Presentation type (select one):	Poster

Conversion of MIMIC to OHDSI CDM

Md Shamsuzzoha Bayzid, MS[1], Vojtech Huser, MD, PHD[2], Joydeep Ghosh, PhD[1]
[1]University of Texas at Austin, Austin, Texas; [2] Lister Hill National Center for Biomedical Communications, Bethesda, Maryland

Abstract

The Observational Medical Outcome Partnership (OMOP) provides a Common Data Model (CDM) for standardizing the format and content for electronic health records (EHRs) and claims data so that that it can be analyzed by a library of standard methods written for OMOP CDM. MIMIC II and MIMIC III are very popular critical care databases that are freely available and quite comprehensive in terms of EHR recordings in ICU settings. However they do not conform to OMOP CDM. We have designed and developed a PostgreSQL-based extraction, transformation and loading (ETL) that generates CDM v5-compatible CSV files from MIMIC II data files. Availability of CDM-shaped MIMIC dataset allows OHDSI researchers to easily work with MIMIC, and also exposes the large and expanding research community working with MIMIC data to leverage and further develop OHDSI tools.

Introduction

The widespread availability of observational health data, collected throughout the health care spectrum in the form of EHRs, insurance claims data etc., enables meaningful use of multiple disparate health databases. However, the format and schema of various data sources can be quite different, which makes the systematic analysis of disparate databases challenging. OMOP CDM provides a common data structure, and standard definitions and terminologies to facilitate the use of disparate data sources[1]. It allows us to develop standardized tools that can be run on different databases specified according to OMOP CDM. Over the past few years, many databases have been converted to OMOP CDM. The Health Improvement Network (THIN) and the Premier hospital databases are two prominent examples[2]. In this paper, we present a PostgreSQL-based ETL implementation for transforming the MIMIC II demo database[3] to OMOP CDM. We also discuss some challenges that we faced and overcame during the transformation process.

Materials and Methods

MIMIC II clinical demo database contains comprehensive clinical data from 4,000 deceased intensive care units (ICU) patients (of over 32,000 total patients in the full non-demo database). It contains physiologic signals and vital signs from a variety of ICUs (medical, neonatal, surgical etc.) collected between 2001 and 2008. OHDSI CDM v 5.0.1 defines 14 standardized clinical data tables, 5 health system data tables, 4 health economics data tables, 3 tables for derived elements and 8 tables for standardized vocabulary. Although in 2016, an updated version of MIMIC was released (MIMIC III), we used MIMIC II because it provides a demo subset of deceased patients that is not yet available for MIMIC III.

We mapped 10 CDMv5 standardized clinical data tables, 2 health system data tables and the metadata table (CDM_SOURCE). Note that there is no cost related information in the MIMIC database. Therefore we could not generate the cost related tables. Also, the currently implemented ETL does not populate the derived tables. We have done this translation in two phases: In *Phase 1,* we focused on transforming the data into proper CDM tables and columns (using the `source_value` columns extensively). For columns requiring CDM vocabulary concept_ids, in phase 1, we put concept zero in such columns. For the terminology work in *Phase 2,* we mapped the concept ids

using CDM v.5 vocabularies (http://www.ohdsi.org/web/athena/). Table 1 shows the table mapping from MIMIC II native format to the CDM for a subset of the tables that we mapped. The PostgreSQL scripts are publicly available for download at https://github.com/shamsbayzid/mimic-cdm or at https://github.com/OHDSI/sandbox/tree/master/ETL-mimic.

Table 1. Table mapping from MIMIC II source data to OHDSI CDM for a subset of the tables that we mapped.

OHDSI CDM table	MIMIC II source table	Comments
person	d_patients	Gender mapped, MIMIC `subject_id` used was mapped to person_id. Phase 2 was done for this table.
death	d_patients	Only in hospital death were imported, death_type_concept_id cannot be mapped in Phase 2.
condition_occurrence	Icd9	MIMIC provides ICD9 data; phase 2 was done for this table.
visit_occurrence	icustay_days	9203(emergency visit) was used as the `visit_concept_id` in Phase 2.
procedure_occurrence	procedureevents, d_codeditems	MIMIC provides ICD9 CM procedure data. Procedure source_value was in an incompatible format with CDM vocabulary; phase 2 was not done for this table.
drug_exposure	medevents, d_meditems	
measurement	d_labitems, labevents	We ignored visit attribution for labs and used lab result time. MIMIC provides lab results coded in LOINC and in Phase 2 we linked those to concept_ids.
Note	noteevents	Note types (RADIOLOGY_REPORT (65884 entries), DISCHARGE_SUMMARY (3930 entries), MD Notes (179 entries), Nursing/Other (101934 entries)) were mapped to OMOP document concepts; Phase 2 was done for this table.

We were able to successfully map demographic, diagnostic, procedural and medication data to the OHDSI CDM. The resulting MIMIC2 demo database contains 4000 patients with 34,828 data rows in the visit_occurrence table (with 5,844 distinct icustay_ids being recorded), 53,486 data rows in the condition_occurrence table (with 2,719 unique conditions being diagnosed), 25,288 data rows in the procedure_occurrence table (with 943 distinct procedures being reported), 3,740,682 data rows in the measurement table (with 537 distinct lab tests being reported), 1,048,968 data rows in the drug_exposure table (with 59 distinct drugs being prescribed), and 171,927 data rows in the note table.

The data transformation was done by a single developer in less than 40 hours (excluding 20 hours of time for understanding MIMIC documentation and install MIMIC). To test model conformance to the CDM, we executed Achilles Heel and found no errors in the columns that were in the scope of our project. We plan to upload Achilles files to the OHDSI public Achilles instance and extend our work to MIMIC III data (which utilizes similar format to MIMIC II).

Salient issues encountered during the translation include: (1) Multiple ethnicities (CDM does not support multiple). In full mimic2 dataset, one patient can have more than one ethnicity (combined); (2) Condition_occurrence table: not enough data to obtain the `condition_start_date` and `condition_end_date`; (3) Note table: there are some notes (e.g., discharge summary) without any associated `icustay_id`.

Conclusion

Our published ETL script allows other teams to quickly obtain a testing CDM-compliant database. Compared with existing CMS synthetic demo data, it includes realistic patients and includes clinical notes data (NOTE table).

References

1. JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012; 19(1):54-60.
2. Makadia R, Ryan PB. Transforming the Premier Perspective® Hospital Database into the observational medical outcomes partnership (OMOP) common data model. eGEMs 2014; 2(1):1110.
3. https://physionet.org/mimic2/demo/ (visited on June 15, 2016).