Bridging communities: Transforming MIMIC-IV to the OMOP CDM

Michael Kallfelz[1], Andrew Williams[2], Tom Pollard[3], Anna Tsetkova[1], Manlik Kwong[2], Gigi Lipori[4], Jeff Osborn[5], Vojtech Huser[6]

[1] Odysseus Data Services | [2] Tufts University | [3] Massachusetts Institute of Technology | [4] Univ. Florida | [5] Endpoint Health | [6] National Institute of Health

Background

MIMIC-IV[1] is the latest version of a database collected from patients of a tertiary academic medical center in Boston, MA, USA. It contains in particular a large portion of data from ICU patients. It is hosted and maintained by PhysioNet[2], an institution providing access to a large number of resources, managed by the MIT Laboratory for Computational Physiology[3]. However, each of these resources are somewhat siloed and in their own format. Promoting the MIMIC-IV content to be reviewed within the OHDSI community would open up the opportunity to extend this approach to more resources and allow linking multiple ICU datasets. Having the data in OMOP CDM format facilitates the use of OHDSI tools for analysis, while for the PhysioNet datasets, data would have to be explored individually, requiring sound technical and coding knowledge.

The OHDSI community in return is interested in exploring more capabilities by integrating additional objects with clinical data such as biosignal information as represented by waveform data. Having mastered this would also provide a blueprint for OMOP adoption of datasets with ICU data.

Both communities experienced it as beneficial to collaborate and make use of each other's skills and resources when approaching this joint project[4] aiming at a conversion of the MIMIC IV content together with adding waveform information to the process. The OHDSI team contributed strong knowledge of the OMOP ecosystem and best practices while the MIMIC/PhysioNet team contributed clarification of data origins, structure and processes as well as the PhysioNet user community expectations and requirements.

A project[5] converting the previous version, MIMIC III, to OMOP has been used as a template in many aspects.

Methods

The project was approached like a standard ETL project with initial discovery, adoption and subsequent adaptation of previous coding and mappings, regular weekly meetings, continuous improvement of extraction logic and mappings, Unit testing and finally User Acceptance Testing.

Waveform integration was a twofold process, where existing waveform files in wfdb format retrieved from the MIMIC repository were converted and parsed for measurement events so that these can be fed into the ETL, generating additional structured knowledge that was not easily accessible from within the source MIMIC-IV dataset. This exercise was also meant to explore how the somewhat proprietary wfdb format could be converted to a more generally accessible format that can serve as a template for other waveform conversion projects in the OHDSI community. A standardized format that would integrate with the OHDSI toolset could allow an extension of distributed research into secondary analysis of waveforms

based on metadata and measurements extracted from those and made available through classic OMOP structures. For this project a workaround has been chosen to link waveform data to patients as well as measurements derived from them by creating a visit detail entry for every waveform.

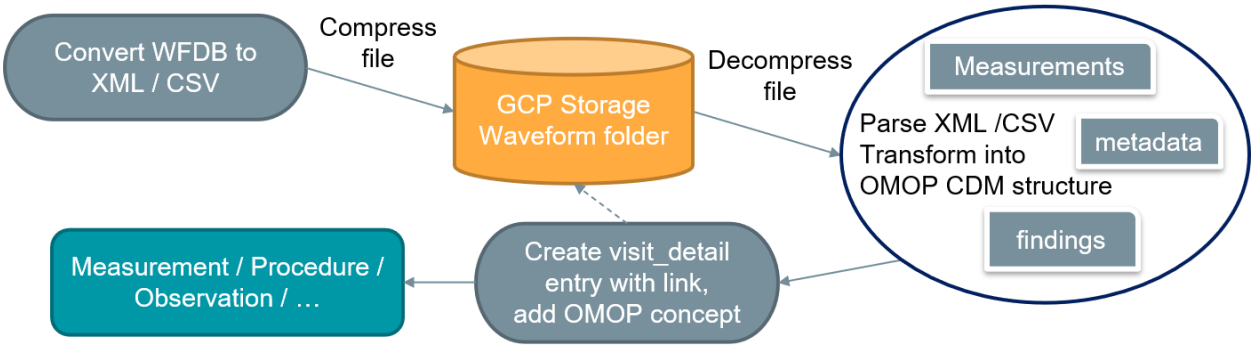

Figure 1. Waveform feature extraction and integration process

PhysioNet has recently adopted hosting on Google Cloud for all datasets, which facilitates simpler operation for PhysioNet users. There is no absolute need to download data, but by using Google credits analyses can be carried out in your own project. The final version of the complete OMOP CDM will be available on Google Cloud as well as a CSV download. The MIMIC IV data was accessed by pulling it from the PhysioNet BigQuery® instance to an ETL development project in a separate BigQuery instance. To allow for easier conversion, the data was transformed into lookup tables before the actual conversion was performed.

Quality checks have been carried out using OHDSI tools such as DQD and Achilles.

As a large number of studies has been published based on the MIMIC database, user acceptance testing was designed by reproducing representative parts of existing studies using OMOP tools such as ATLAS. This serves a number of purposes such as testing the reliability of the transformation and the usefulness of the OHDSI model and ecosystem for MIMIC content. At the time of this submission, the user acceptance testing was still ongoing.

Results

The transformation from the MIMIC IV database structure to the OMOP CDM achieved high coverage but has yet to be completed for selected areas. The code repository and respective documentation have been created in the OHDSI github environment. Over the course of the project, a collaborative approach was taken between the PhysioNet and OMOP teams in mutually understanding limitations and addressing them. In particular the work with the waveform material in wfdb format, its conversion and extraction of additional knowledge and integration with clinical data in the OMOP CDM also reflected back to the PhysioNet team, generating new insights.

The MIMIC team observed that in the Extraction and Transformation process to the OMOP CDM, information loss or loss of detail can sometimes take place because of lack of adequate target structures or as a possible byproduct of converging to a smaller number of standard concepts, which created some concern. Keeping the balance between information loss and usability of the converted data in the OHDSI tools and ecosystem, is subject to individual decisions during the design of the ETL process and a matter of discussion between the communities. The OHDSI team in return noticed that there are potentially

multiple heterogeneous sources for the same data, requiring individual approaches while avoiding duplication of data.

The use of community-built tools and the ecosystem provided by OHDSI is seen favorably by the PhysioNet team, including Data Quality checks and the use of cohort creation tools.

Conclusion

Both research communities welcome the opportunity to collaboratively apply OHDSI technology to the MIMIC space and increase the usage of MIMIC by extending its reach beyond the classic PhysioNet environment. The approach can also be used as a model for conversion of additional PhysioNet datasets, e.g. HiRID (Swiss ICU data) or other datasets covering information captured in an ICU context. The project highlights the need to find a solid approach for integrating objects with the OMOP CDM other than structured data (e.g. imaging or biosignal data) by embedding those with as much structured metadata as possible so that they can be exposed to classic analytical processes.

Remarks

This project received funding from the Bill and Melinda Gates Foundation.

References/Citations

1. https://mimic.mit.edu/iv/
2. https://physionet.org/
3. https://lcp.mit.edu/
4. Kallfelz M, Williams A, Tsvetkova A, Kwong M, Pollard T, Lipori G, Hao S, Osborn J, Huser V. MIMIC data in OMOP Common Data Model. PhysioNet. (in submission).
5. Paris N, Parrot A. https://www.medrxiv.org/content/10.1101/2020.08.14.20175141v1 (and https://github.com/MIT-LCP/mimic-omop).