Linking Analysis Ready Multi-modal Clinical data

Priya Desai, Somalee Datta

Stanford School of Medicine and Stanford Health Care

Background

STAnford medicine Research data Repository or STARR, is a research ecosystem that contains a collection of linked research ready data warehouses from disparate clinical ancillary systems and a secure data science facility. The ecosystem is designed on the principles of Data Commons and contains reusable data processing pipelines, cohort and analysis tools, training, user support and much more.

STARR data currently includes electronic medical records data, clinical images (radiology, cardiology) and text, bedside monitoring data, and near real time HL7 messages. Processed, "analysis ready" linked data is available for to all Stanford researchers in a "self-service" mode and currently consists of:

- De-identified Electronic Health Records (EHR) from the two Stanford hospitals and clinics in the OMOP Common Data Model (CDM).
- De-identified bedside Monitoring (Waveform) data from Stanford Children's Hospital

Linked patient data in the ecosystem are primarily anchored using person_id, the auto generated identifier for the patient in the CDM from the OHDSI community. When the data is refreshed, the person_id stays stable.

Other data such as imaging metadata from radiology (including MRI's, X Rays, ultrasounds and CT scans), and cardiology are coming soon. These analysis-ready datasets reside in BigQuery, a cloud based data warehouse that leverages the infrastructure of the Google Cloud Platform and offers rapid SQL queries and interactive analysis of massive datasets.

Motivation

As we have brought in the new data types, and their associated metadata we found that extending the OMOP CDM to capture all the additional metadata is a herculean task for multiple reasons. First, only a small number of hospital devices produce data in standard formats. For emerging data types such as bedside monitoring or pathology whole slide imaging, there is simply no accepted standard. Even DICOM, a widely accepted industry standard used in imaging, is not standard across different modalities. There are vendor specific tags within each modality (like radiology, cardiology, radiation oncology, and retinopathy), as well as workflow specific tags within a clinical application. For patients who are consulting at Stanford, we see an even wider range in the choice of tags. Furthermore, the number of populated DICOM tags in the header is also variable, it can be as few as a few hundred or as many as a few thousand.

Second, in the OMOP CDM, any patient data that cannot be represented by any other domains, such as social and lifestyle facts, family history or inpatient flow sheet data can and should be recorded in the Observation table. The Observation table is meant to be the "catch-all" table for any clinical data that cannot be housed in the other OMOP tables. When we first launched OMOP, we decided to pull in all the flowsheet data from approximately 1000 types of flowsheets into the observation table as JSON formatted strings. Our goal was to allow our researchers access to the flowsheet data to run NLP or other sophisticated text processing algorithms. This resulted in an increase of the observation table by

3.5 billion rows and impacted the cost-utility metrics negatively since very few researchers are interested in processing raw flowsheets data.

Third, it is difficult to choose a subset of the metadata that supports the majority of novel research use cases, and standardization within the CDM is a process that requires consensus and time.

We have, therefore, implemented a novel mechanism of keeping all the rich metadata from these ancillary sources, in their own BigQuery datasets while making these data linkable to each other. While BigQuery provides analytical convenience, the approach we present is usable for other databases. Our approach is also aligned with OMOP CDM evolution as we are well poised to bring in elements from these ancillary metadata in the CDM, as the CDM evolves. It also allows us to learn from the researchers' experience and bring that learning to the OHDSI community.

Methods

We present the methodology we implemented for the bedside monitoring data, but the general approach is extensible to any other data type including radiology, pathology, genomics and others.

Our pediatric hospital has 350-450 monitors available to be hooked up to ~500 beds on any given day. This typically results in collection of data from ~280 patients daily and addition of ~180,000 rows in the alerts table, ~10 million rows in wave sample table, 60 million rows in enumeration value table, and 120 million in numeric value table[2]. For the waveform, we have ~40 TB of data across ~50 modalities presenting ~350,000 non-unique patients and ~2.2 million non-unique studies. Instead of bringing the bedside data into the observation table, we have brought the bedside data into a separate BigQuery dataset. This dataset contains the metadata, as well as links to the actual waveform data stored in cloud storage objects.

Since there are currently no recognized standard schemas to store bedside monitoring data in the CDM, we worked with our researchers to identify the most useful parameters for cohort generation, and generated metadata tables that can be linked to the OMOP data via the person_id. See Figure 1 for generalization of our approach.

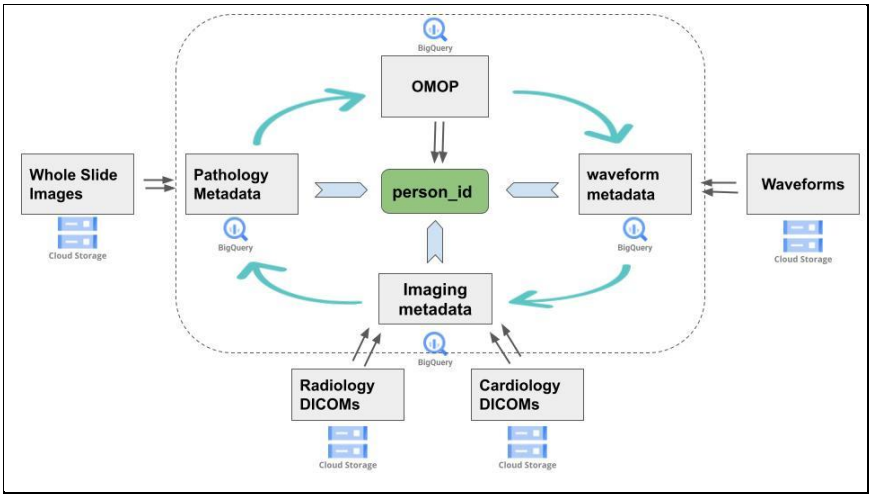

Figure 1: Analysis ready metadata tables from all ancillary clinical datasets (radiology, pathology, genomics), are maintained as separate datasets in BigQuery that can be linked via the person_id. Researchers can use any of the metadata tables to define their cohort, and then refine the cohort by

linking to the other tables.

Data Characteristics:

A. Flowsheet Data (1,883 templates, 46,85 groups and 25,777 distinct types of row measures available at the Stanford Hospitals)

Average Templates per patient	18	Includes templates like Acute Care assessment, PEWS, Custom Formula Data etc. Flowsheet templates are customized to each institution.
Average groups per patient	34	Includes groups like Cardiovascular/Peripheral Vascular, Respiratory etc.
Average measure types per patient	156	Includes measurements like Edema, PEWS Score Previous Value, SpO2, Nutrition, N2O etc.

B. Waveform Data (Feb 2017 to March 2021, ~500 beds)

Average daily count of studies	400	A study corresponds to continuously monitored patient data
Average daily count of patients	280	Patients are from different clinical units.
Average num of rows added to Alarms & Alerts table per patient per day	643	Includes alerts & alarms for measurements like Pressure levels, SpO2 levels etc with severity status. Data relayed in 1 sec intervals.
Average num of rows added to Wave sample table per patient per day	35,715	Includes continuous waveforms of Central Venous Pressure(CVP),Electrocardiograms (ECG), Left/Right Arterial Pressure etc for upto 28 waveforms/patient.
Average num of rows added to Numeric Value table per patient per day	428,571	Includes vitals such as Heart Rate (HR), Pulse Oximetry (SpO2), Partial pressure of carbon dioxide (PaCO2) etc of the patient

Results

The deidentified bedside monitoring metadata dataset[3] contains 2 main tables:

1. The de-id Patient Study Map table has ~ 278,000 rows and contains person_id, study_id, bed labels, and study start and end dates that have been jittered with the unique offset used for all dates for that patient (in the deid OMOP data).
2. The deid Study Details table allows researchers to select studies that only contain waveforms of specific interest e.g. ECG or SpO2, Respiratory rates(RR), alerts and alarm values, and define their cohorts using the study map metadata which can then be linked to the OMOP dataset. The deid Study Details table contains over 3.68 millions rows.

Conclusion

The decision to generate multiple auxiliary datasets containing relevant patient metadata that can be

queried and linked as needed has proved to be very beneficial to the rapidly evolving STARR ecosystem. This allows us to work with OMOP CDM without losing the granularity that our researchers need, thus assisting the process of adoption and evolution.

References/Citations

1. Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D, Mesterhazy J, Pallas P, Desai P, Nigam Shah, A new paradigm for accelerating clinical data science at Stanford Medicine, arXiv:2003.10534, Mar 2020, https://arxiv.org/abs/2003.10534
2. Malunjkar S, Weber S, Datta S, A highly scalable repository of waveform and vital signs data from bedside monitoring devices, arXiv:2106.03965, Jun 2021, https://arxiv.org/abs/2106.03965
3. STARR pediatric Philips PIC iX bedside monitoring metadata dictionary: https://med.stanford.edu/starr-wave/access.html#documentation