Linking Analysis Ready Multi-modal Clinical data

Priya Desai, Somalee Datta

Stanford School of Medicine and Stanford Health Care

Stanford MEDICINE

Technology & Digital Solutions

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 including electronic medical records data, clinical images (radiology, cardiology) and text, and bedside monitoring data.

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

Other de-identified 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 .

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.

Motivation

As we have brought in the new data types, we found:

- Very small number of hospital devices produce data in standard formats. Even DICOM is not standard.
- The Observation table is meant to be the "catch-all" table for any clinical data that cannot be housed in the other OMOP tables. Often results in multifold size increase negatively impacting the cost-utility metrics negatively since very few researchers are interested in processing raw flowsheets data.
- 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.

Extending the OMOP CDM to capture all the additional metadata from ancillary clinical datasets is a herculean task!

Our Solution:

1. Keep **all the rich metadata** from these **ancillary sources,** in their **separate** BigQuery datasets while making these **data linkable** to each other.
2. This approach is **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.*
3. While BigQuery **provides analytical convenience**, the approach we present is usable for other databases.

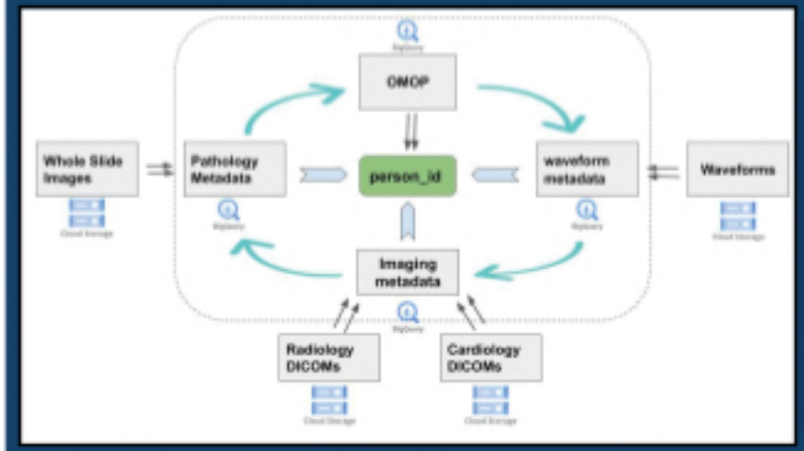

Figure 1: Analysis ready metadata tables from all ancillary clinical datasets (radiology, pathology, genomics), are maintained as separate datasets in BigQuery that can be linked via the person_id. Researchers can use any of the metadata tables to define their cohort, and then refine the cohort by linking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

Methods

1. Currently no recognized standard schemas to store bedside monitoring data in the CDM
2. We worked with our researchers to **identify the most useful parameters for cohort generation**, and **generated de identified metadata tables** that can be **linked to the OMOP** data via the person_id
3. Methodology implemented for the bedside monitoring data. Approach is extensible to any other data type including radiology, pathology, genomics and others.

Data Characteristics:

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

Average daily count of studies	800	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	645	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:

- De-id Patient Study Map table 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).
- 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.

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. It allows us to work with OMOP CDM without losing the granularity that our researchers need, thus assisting the process of adoption and evolution.

References

1. Datta S, Posada J, et. al 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-www/access.html#documentation