AI-AUGMENTED NARRATIVE PREDICTION FROM CLINICAL NOTES

Raw EHR to OMOP Common Data Model

Project Stakeholder

David Rotenberg
Director, Data Strategy and Business Intelligence
Centre for Addiction and Mental Health (CAMH)

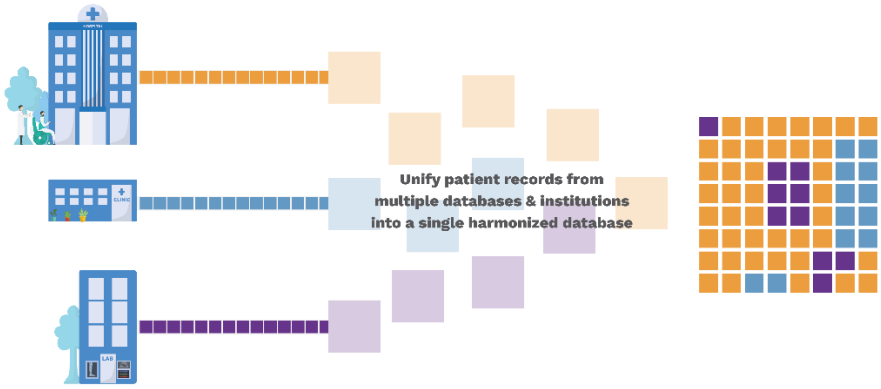
Research Director

Dr. Abhishek Pratap
Group Head of Artificial Intelligence and Digital Health
Krembil Centre for Neuroinformatics

Krishna Kothumbaka
20kpk@queenu.ca

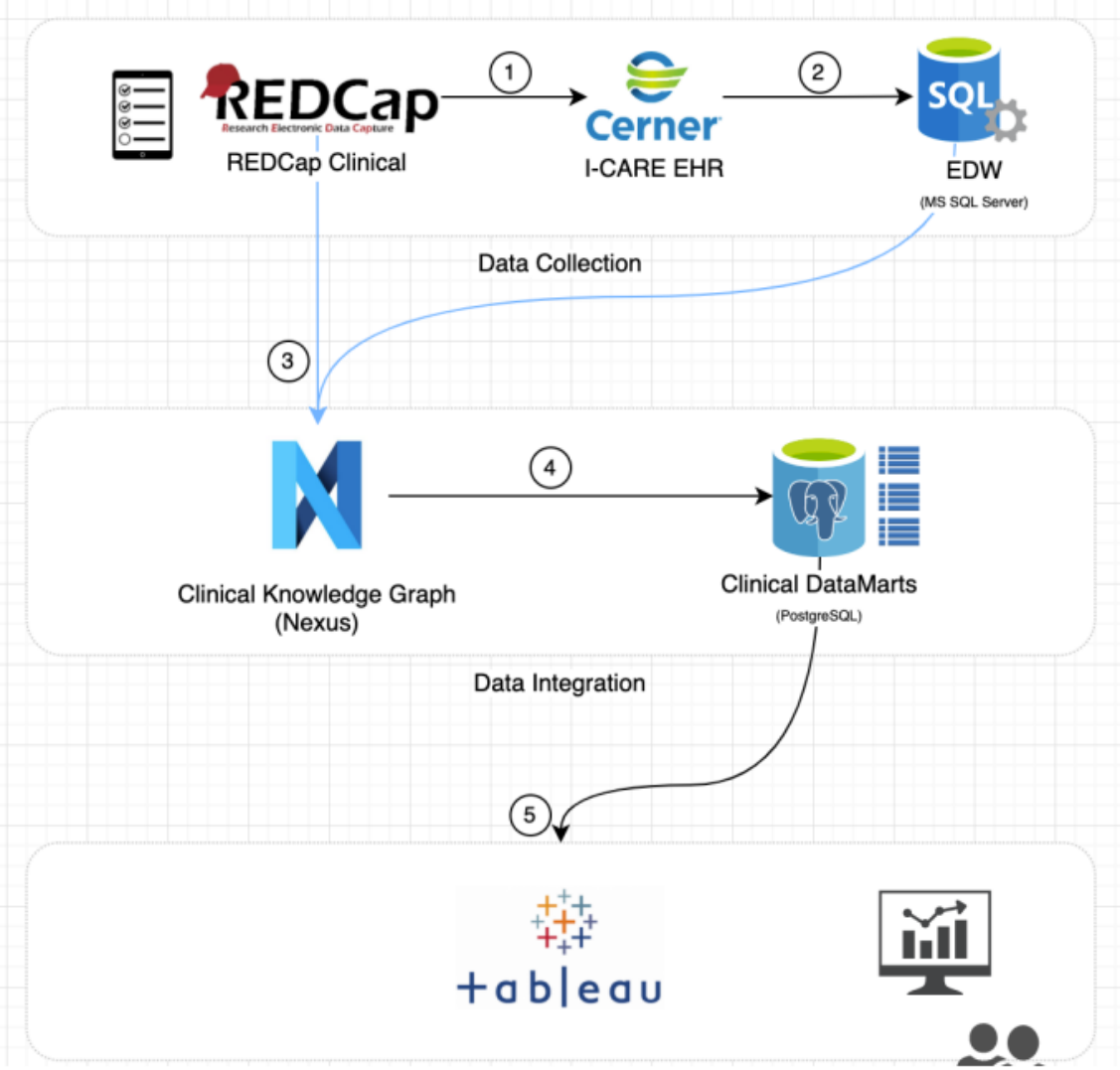
Executive Summary

The Center for Addiction and Mental Health (CAMH) is Canada's largest mental health teaching hospital and one of the world's leading research centers in its field. With a dedicated staff of more than 3,000 physicians, clinicians, researchers, educators, and support staff, CAMH offers outstanding clinical care to more than 34,000 patients each year. One of this project's primary objectives is to facilitate research with evidence-based information to help clinicians, caregivers, and patients make informed healthcare decisions. The research relies on observational data of the patient captured at various touchpoints in the healthcare system as he/she is receiving the care. There are several information systems that capture the EHR and observational data in the clinics, labs, diagnostics, and other medical facilities at CAMH. A typical patient's journey in a hospital setting touches a variety of domain-specific systems (E.g.: Labs, Claims, Pharmacy, Biomedical systems etc.), and most of these systems are proprietary that have their data formats (E.g.: HL7, FHIR etc.) with their underlying data stores, mostly functioning in an isolated manner. An additional data source unique to CAMH is the clinical data - data from patients' mental health-related surveys and assessments. As the number of patients increases, the data creates Big Health Data. Compiling patients' health information from diverse healthcare systems to perform research that is evidence-based at CAMH presents many challenges, both from a technical and managerial point of view. While the people and process challenges are significant on their own, this project primarily addresses the technical challenges to support the end goal of building a platform to perform patient-level prediction analytics.



Introduction

There are two primary healthcare information systems at CAMH. One of them is RedCap which processes and manages the clinical data, and the other is I-Care, from Cerner, which manages the EHR data and processes.



Data from these two transaction-oriented systems is extracted, transformed, and loaded to an EDW (Enterprise Data Warehouse) , backed by SQL Server. EDW in turn, acts as a source to populate the Clinical Knowledge Graph system called Nexus. Nexus is an open-source Swiss brain research initiative. CAMH, Canadian Open Neuroscience Platform (CONP), Human Brain Project are some of the primary users of the Nexus platform. A Knowledge Graph is built from heterogeneous data with a graph-like structure to represent data and their relationships. Nexus is based on the FAIR data principles to provide a flexible data management solution. Nexus exposes data in a format similar to FHIR but represented using an RDF format. Data from Nexus is extracted and loaded into clinical data marts. In the current environment, data flows from multiple systems to multiple systems, each target system picking data relevant for its usage. These systems are a great resource to evaluate and study an individual patient's physical health and mental patterns. But this setup, as is, is not conducive for performing large-scale analytics.

Discovery Phase

There is no escaping the fact that CAMH has a heterogeneity of data sources and the challenge this would present to support large-scale analytics with one view of the patient data. Learning from other players in the healthcare industry, the best way to address this challenge is to adopt a Common Data Model that can support research studies involving large-scale analytics. The adoption of a CDM brings consistency to observational research through standardization of its processes. The Common Data Model abstracts away the diverse data sources and their proprietary data formats. The data in the Common Data Model can be refreshed from the upstream systems in near real-time or nightly through a batch ETL process. The real-time aspect of the data is not so relevant for the research use cases; nevertheless, it is possible to do so.

Designing a CDM is a non-trivial undertaking for any organization, even that of CAMH stature. It would consume thousands of man hours and would require input and expertise from across different domains (E.g.: Pharmaceutical, Clinical etc.) in CAMH. The healthcare industry is a bit unique from other industries in that there are different terminologies to represent a single entity (E.g.: Warfarin - The same drug has a

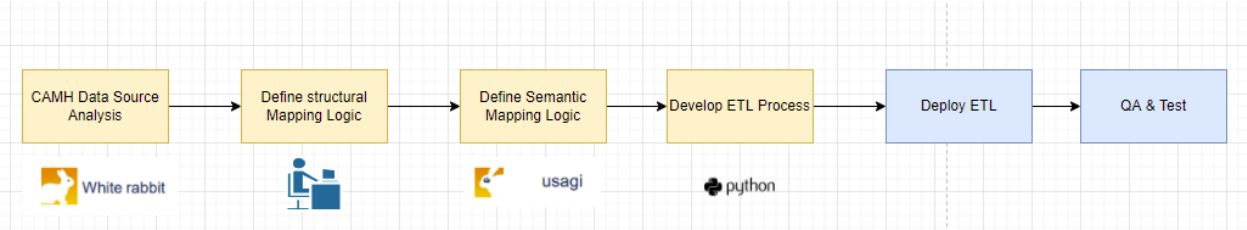
different representation under SNOMED vs RXNorm). The same situation applies conditions (E.g.: Arrhythmia) which tend to be hierarchical in nature with hundreds of representations in different terminologies (SNOMED, LOINC, etc.). The other significant challenge is regional interpretations and modifications to these terminologies. For example, in the mental health domain, there is a Canadian specific version of DSM5 called DSM5CA and this situation proliferates when you take into consideration some European and Southeast Asian countries. While getting the data model right is half the battle, having confidence that the model can capture most of the scenarios and requirements that need to be supported in the long run across research agencies is a daunting undertaking. An alternative to this bespoke CDM design is to adopt a model that has been published and widely adopted in the open domain. One such model is OMOP CDM from the OHDSI community. The broad objectives of OHDSI - openness, community, innovation, and collaboration align with CAMH's. OMOP CDM provides the platform to pursue cross-institutional collaborations. There is a community of researchers across diverse organizations and countries working on OMOP CDM that one can tap into for guidance or validation. One of the side advantages of OMOP CDM is the suite of tools from OHDSI that work with it. These tools add a lot of value for the researchers as they have been developed with the primary purpose of supporting the OHDSI research community, some of whom may not be technologically inclined. Some of the tools from OHDSI are Atlas, Usagi. The Director of Research made a quick and informed decision to adopt OMOP CDM at CAMH.

Implementation

Once the OMOP CDM was decided as the platform for observational research, the first step was to set up a database - both for CAMH source and OMOP CDM. CAMH uses Postgres for their data marts, so it was decided to continue with Postgres for OMOP CDM. OMOP CDM has great support for Postgres. Once the OMOP CDM database was set up, it was populated with data from standard vocabularies from Athena. Athena is a web-based application hosted on the internet by OHDSI. Athena requires registration and once registered, one can download the standardized vocabularies for most common concepts used. Some

vocabularies can be proprietary and would require a license. The other database for data from CAMH was also provisioned. It was populated with anonymized patient data to be used specifically for this project.

Populating OMOP CDM tables from CAMH's source data must deal with structural and semantic data. Handling structural data is simpler as it is usually straightforward and does not require further analytics and investigation. An example of structural data would be something like a patient's name, address etc. Transforming and loading semantic data involves understanding the source data and its provenance. An example of semantic data would be medication, conditions etc. For example, a drug Warfarin is represented by many vocabularies (E.g.: RxNorm and SNOMED) - some standard and others non-standard. So, it is not a simple mapping based on name. Each source-to-target table had to be evaluated in detail and each column-to-column mapping was performed with a CAMH SME (subject matter expert).



CAMH's current ETL processes employs Python, specifically leverages Pandas and Alchemy libraries for interacting with the Postgres database. An ETL process has been implemented for this project using the same language and libraries, so the work can be integrated into the existing data pipelines and enhanced by CAMH teams in the future. SQL queries from the CAMH source systems were developed in discussion with the CAMH's technical resource. The results of the SQL queries are loaded into Pandas data frames and manipulated to perform any transformations required for the target OMOP CDM schema. In the current setup, the Python ETL process runs as a standalone program with the goal is to integrate this ETL into the existing main batch processes at CAMH.

OHDSI provides a couple of tools to help with the mapping analysis. The WhiteRabbit tool was also employed to perform a scan of the source data. It generated a report that can be used as a guide to design

the ETL; but it only handles all the structural data well. Medications and conditions data are semantic in nature. Their source needs to be mapped to a standard vocabulary. To facilitate this mapping quick lookups, need to be supported. As some of the standard vocabularies' tables are in the order of millions of records, a few ETL support tables just were created with indexes to support fast lookups. We realized that the drug information comes from Cerner and uses Multum vocabulary or terminology. This is a non-standard terminology and must be mapped to the standard one either RxNorm or SNOMED. The current situation is that the data marts at CAMH are missing the source codes for the Multum based drug information in *medication_request* tables and this is a required field for OMOP CDM to support any drug related analytical queries. CAMH is updating its current upstream ETL process to populate the source codes for the medications. Once that is done, the mapping between Multum and one of the standard vocabularies will be done and the *drug_exposure* OMOP CDM table can be populated.

To facilitate the mapping, we use Usagi, a tool from OHDSI. Usagi helps with mappings from source vocabulary to target vocabulary by exploiting textual similarity of code descriptions through a fuzzy logic algorithm. The accuracy measure is also provided for each it finds. It would still require manual touch to confirm or select an alternative mapping or completely ignore the recommendation from the list Usagi provides for each of the source records. While it does not remove the human in the loop, it does make the job much easier for the human, without which it would be a significant and laborious effort to do the mapping by hand one by one.

The conditions data also presented a similar challenge. As CAMH is focused on mental health, a lot of the conditions are defined by the following terminologies - DSMIVTR, DSM5CA, ICD10CA, SNOMEDCT. None of these are standard vocabularies as per OMOP CDM. Some of the vocabularies are very specific to the Canadian region, which makes it a little more challenging to validate the mappings to standard vocabularies. All the conditions defined in the above-mentioned vocabularies are loaded into OMOP CDM as they are. Usagi tool came in handy to create the mappings from these vocabularies to the standard vocabulary concepts. This is a good starting point but as mentioned above they still need to be vetted with

a subject matter expert who is familiar with both the region-specific definitions of these conditions and the generic ones used by OMOP CDM.

As a preliminary test, we attempted to use an open-source tool called PatientExploreR (https://data.ucsf.edu/research/patientexplorer) to test the structural validity of the OMOP CDM. It can generate patient level interactive dynamic reports and visualization of clinical data. It is an R application that uses Shiny framework. Unfortunately, the one of the 3rd party shiny libraries caused issues and required a lot of troubleshooting in the guts of the application. For the time being, this has been put aside to focus more on the ETL process.

An alternative R Package called ROMOP (https://github.com/BenGlicksberg/ROMOP) was used to connect to the OMOP CDM in CAMH to test the basic functionality.

Results

Once the decision was made to use OMOP CDM, we started the project with a discovery process to identify the right source to populate the OMOP CDM database. The process took a few months and engaged CAMH subject matter experts and technical resources to provide inputs and reach a final consensus. Before this project, there was no comprehensive data source for patient-level data to support analytical capabilities. There were data marts to support reporting systems. Putting the OMOP CDM in the CAMH network was one of the first achievements. It paves the way for all future analytics and research projects that focus on patient-level analytics. Some of the domain-specific data at CAMH is very specific to Cerner. This project allows for an additional mapping of this data to OMOP CDM. While this may not be useful immediately, it could be something significant if CAMH decides to move off the Cerner platform.

From our research on OHDSI forums, it is highly possible that there is no Canadian healthcare company or institution that has mapped the Canadian specific versions of the DSM5, ICD10 vocabularies. The support for mental health conditions is minimal in the OMOP CDM standards as of today. Based on our discussion on the OHDSI forums, it is possible that CAMH may pave the way forward to introduce some of these

vocabularies, if not standard, at least as a standalone vocabulary, into the OMOP CDM. While we are far off from that point, this project would have contributed to that future state.

This project is not an end. It is one part of an overarching project initiated by David Rottenberg and Dr. Abhishek Pratap at CAMH to build a platform to support patient-level analytics. This project forms the foundation framework for other analytics projects to make use of in the near future. This project has kickstarted the discovery work and created the scaffolding necessary for exciting future work.

Conclusion

Once the medication (*drug_exposure* table in OMOP CDM) data is mapped to one of the standard vocabularies, most of the patient-level data required for performing analytics is available in CAMH's OMOP CDM. We plan to install Atlas inside the CAMH network and point it to the OMOP CDM as a foundational step. Atlas is a web-based tool from OHDSI that enables the design and execution of analyses on patient-level data in the OMOP CDM.

The current setup uses the Python-based ETL process as a standalone step. CAMH employs Apache Nifi, an open-source ETL software. The Python ETL process would have to be incorporated into an Apache Nifi workflow for the nightly batch process; define and provide the configuration capability in Apache Nifi to invoke the ETL process with the relevant configuration values. Most of the data involved in the ETL process is likely to be additive in nature, but where identified, the ETL process would have to be enhanced to detect delta changes to the source tables data.

The current focus is on data from EHR/EMR systems. CAMH, being an institution focused on mental health, captures clinical data in the form of surveys and questionnaires. This data is managed by the Redcap system and is already available in the data marts. There is a plan to extend the ETL to capture this data in OMOP CDM. OMOP CDM has rudimentary support for capturing textual data in notes for NLP purposes. But there are adjunct projects in the open-source community that extend the OMOP CDM schema to perform elaborate NLP processing. With the NLP enhanced OMOP CDM schema, the clinical notes data

can be potentially used to perform AI-Augmented Narrative Prediction. A first step towards integrating OMOP CDM and the NLP domains would be to explore the Apache cTAKES project and find the integration touchpoints in OMOP CDM to facilitate the flow of clinical data free text information into the NLP domain.

Glossary

CDM	Common Data Model
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
DSM5CA	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (Canadian)
DSMIVTR	Diagnostic and Statistical Manual of Mental Disorders, Fourth Ed, Text version
EHR	Electronic Health Record
ETL	Extract Transform Load
FAIR	Findability, Accessibility, Interoperability and Reusability
FHIR	Fast Healthcare Interoperability Resources
ICD10CA	International Classification of Diseases, 10th Edition (Canadian)
ICD9	International Classification of Diseases, 9th Edition
LOINC	Logical Observation Identifiers Names and Codes
NLP	Natural Language Processing
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
RDF	Resource Description Framework
RxNorm	US Specific Terminology in Medicine
SNOMED	Systematized Nomenclature of Medicine
SNOMED-CT	Systematized Nomenclature of Medicine-Clinical Terms

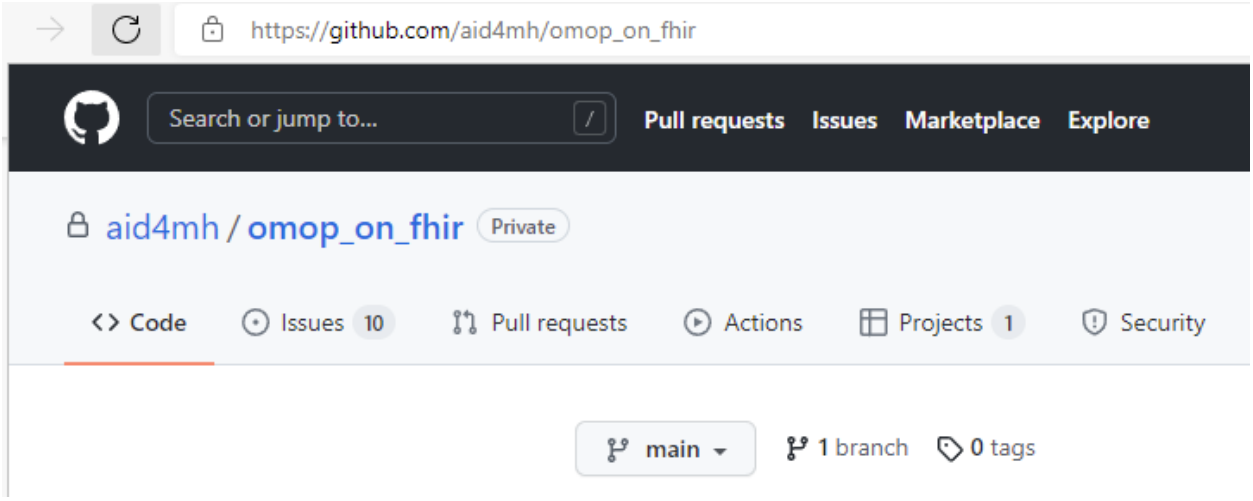
Appendix

Project Website

There is a website for the project at https://aid4mental.health/projects/etoc/ set up by the Research Director's team.

Github Project

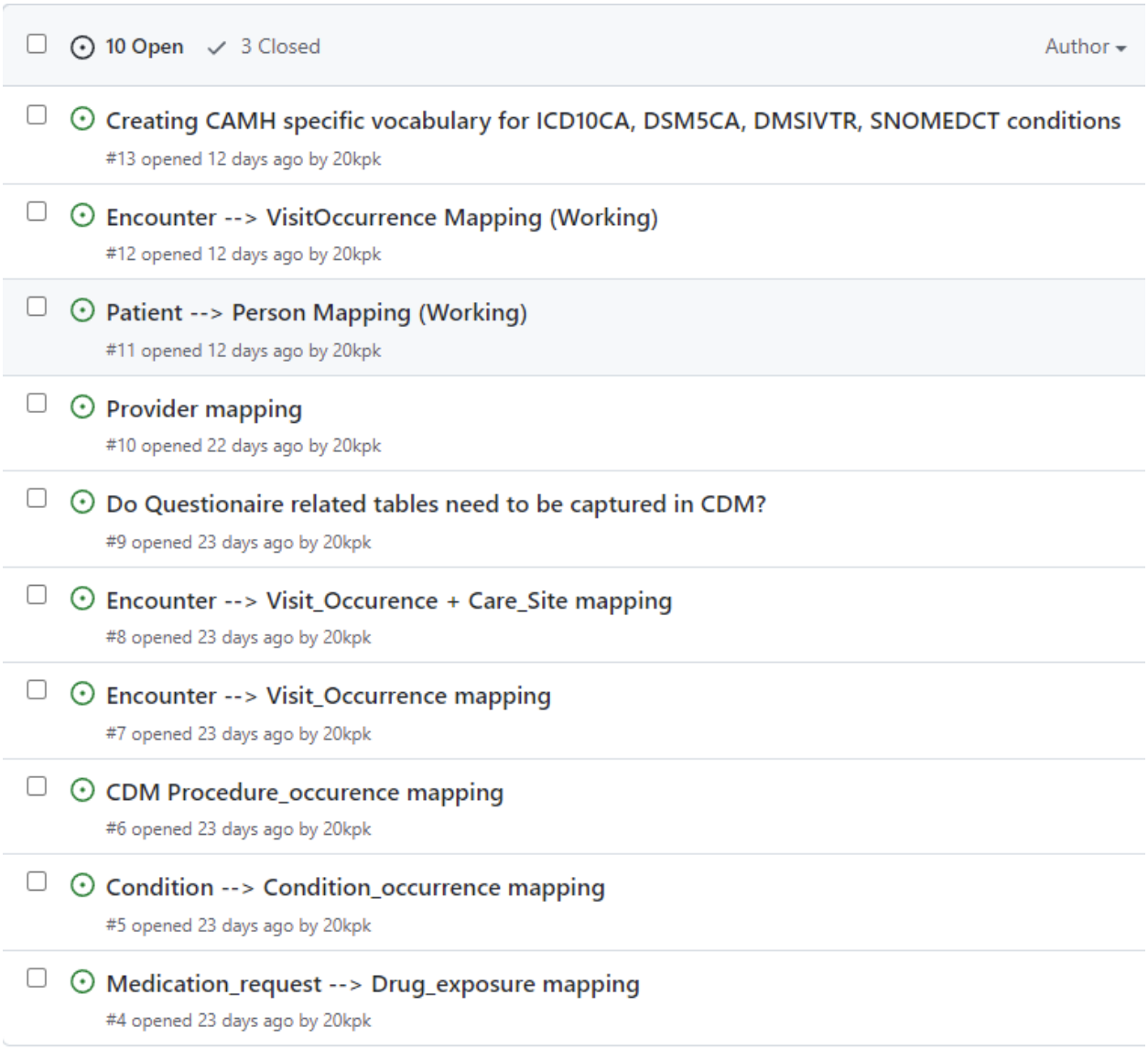
There is a github repository set up for this project. However, it is a private project owned by the Research Director. It is not available publicly. Having a github repository that captures the work allows for the work to be passed over to future research students, who can continue or build on this.



Tracking Tasks/Decisions

There was a lot of discovery work and research that was part of this project. Most of the discussions, decisions and conclusions happened in email threads. To capture the key decision points, the Research Director suggested using Issues List in the above Github project. Shown below is a screenshot of the outstanding decisions/questions.

.



OHDSI Forums

The OHDSI online forums were relied upon heavily to search for similar experiences and questions when a there was a showstopper. There were some questions which required some expert OMOP CDM knowledge from the community (in one example, we needed confirmation on how to we planned to handle custom vocabularies). Shown below is one of my interactions on the forums

https://forums.ohdsi.org/t/mapping-dsm5ca-and-dsmivtr-conditions/16039

OHDSI

Mapping DSM5CA and DSMIVTR Conditions

Vocabulary Users

Krishna Krishna K 18d

Hello,
I am working on an OMOP based project for my Master's. As part of it, I am performing ETL from a FHIR like resource into OMOP CDM. Some of the conditions in the source system are coming from DSMIVTR and DSM5CA vocabulary. I could not find a mapping for these in my database or in Athena. I have searched the forums but could not find any information regarding my issue. Has someone run into this? If so, how did you go about addressing this requirement? As a newbie to the OMOP world, looking for any information/recommendations that could help me get over this challenge. Thank you!

Reply

created	last reply	3	26	2	1
18d	14d	replies	views	users	link

mik Michael Kallfelz 15d

Hi @Krishna , excuse my ignorance, but is DSM5CA the DSM-5 version used in Canada? But DSMIVTR is probably not the Turkish version of DSM-4, is it?
In any case: it seems that you are facing what almost everyone faces when doing an ETL: while the almighty Athena has a very comprehensive set of vocabularies and concepts, it is not covering each and every terminology and concept in the world that is. Athena covers lots of sources and mappings to standard target concepts but for the ones that are not covered, you have to do the heavy lifting yourself. How many concepts are we talking about? Do you have frequencies / counts of occurrence in the source data alongside with? You probably know that already, but the process described in the Book of OHDSI including the use of the Rabbit family of tools should already take you a long way. I am quite confident that you can identify good targets for example in SNOMED. If you however can describe a valid use case, find more support in the OHDSI community and maybe even some funding and hands to help, maybe at least the DSM-5 could be introduced to Athena as a new license restricted vocabulary (given that the APA permits).