Conversion of the Canadian Observational Study on Epilepsy (CANOE) REDCap Registry to the OMOP Common Data Model

Danielle Boyce¹, Colin Bruce Josephson², Ray Jiang³, Samuel Wiebe²

¹ Biomedical Informatics and Data Science, Johns Hopkins University, ²Department of Clinical Neurosciences, University of Calgary, ³Clinical Research Unit, Cumming School of Medicine, University of Calgary

Background

Epilepsy is a neurological condition characterized by recurrent seizures that affects people of all ages globally. The World Health Organization (WHO) estimated that approximately 50 million people worldwide have epilepsy, making it one of the most common neurological diseases globally.¹ The International League Against Epilepsy (ILAE) recommends taking a systematic approach to the collection and analysis of “big data” to promote international data harmonization². Our objective was to create a process for conversion of a large epilepsy registry data set into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)⁶. This would allow for participation in the larger Observational Health Data Sciences and Informatics (OHDSI) community and the use of OHDSI software tools.

Methods

The Canadian Observational Study on Epilepsy (CANOE) registry³ is a multi-site study with data on approximately 6,547 people living with disease. The registry collects standardized demographic, clinical, and patient-reported data at every clinical encounter, which are stored and extracted using REDCap (Research Electronic Data Capture)⁴⁻⁵ electronic capture tools hosted at
the University of Calgary's Clinical Research Unit (CRU). The CANOE study is a large, robust
data set that has facilitated several quality improvement and research efforts.

We assembled a multidisciplinary team that included epilepsy specialist physicians,
informaticists, information technology professionals, and a data scientist with OHDSI expertise as
well as lived experience with epilepsy. We implemented an OMOP Extract, Transform, Load
(ETL) process with OMOP CDM version 5.3\textsuperscript{7} employing Usagi software\textsuperscript{8} for vocabulary mappings
and a PostgreSQL database. Programming was performed in Python and SQL. Atlas software was
installed at the University of Calgary CRU. Please see Figure 1 for a detailed diagram of the ETL
process.

**Results**

Records were mapped for all participants in the CANOE registry. The source database
evolved over time and contains measures collected for sub-studies that only apply to a very small
percentage of participants. Therefore, a ‘minimum viable product’ (MVP) data dictionary was
created that included patient and clinician-reported data for baseline and follow-up visits. Members
of the International League Against Epilepsy (ILAE) Big Data Commission, many of whom have
OMOP experience, were consulted to determine the concepts that would be most useful for a “proof
of concept” epilepsy network study. Medication, demographics, epilepsy surgery, hospitalizations,
tests, procedures, and select commonly used validated instruments such as the General Anxiety
Disorder 7 item scale (GAD-7) emerged as the concepts that would contribute to the most
compelling research questions for network studies. Concepts that were de-prioritized included
participant characteristics such as occupation, driving status, and language and items only captured
on a small subset of registry participants. However, using Usagi and Athena, the team was able to identify and map most of the concepts, and will implement an updated ETL should the use case arise.

The structure of the REDCap database and nature of data capture, including absence of some dates required of the OMOP CDM, posed a challenge during the ETL process. For many concepts, dates simply needed to be associated with the visit in question rather than imputed. For instance, a date was captured for each visit, allowing the association of an accurate date with the corresponding concept. Drugs and comorbidities such as psychiatric conditions presented the greatest challenge, as the registry does not capture precise start or end dates for drug and condition-related concepts at baseline. For example, a registry participant is presented with several medications at their baseline visit and asked if they take the medication “at present” or “in the past.” Depending on the research question, it may be important to understand which medications were already prescribed, even in the absence of exact dates. To prevent data loss and ensure consistent mapping of these dates, decision rules were developed with the clinical team, incorporating guidance from OHDSI Registry Workgroup members and published sources.9-10 Fortunately, all registry participants have full outpatient visit dates which are important in establishing a ground for imputation decisions. For example, if a medication was reportedly taken “in the past,” at baseline, the decision was made to code the medication start date as one year prior to the visit, and the stop date three months prior to the visit. These decisions are well-documented to prevent misinterpretation during analysis and will be published on GitHub. Imputation of medication start and end dates was less necessary for follow up visits as medication changes often occur at the follow up visit and all visits have full dates.
Conclusions

Data from a large, multi-center epilepsy registry were transformed into the OMOP CDM with limited data loss and minimal deviation from the source data. Our strategies may be used to encourage more OMOP-friendly registry designs and to efficiently ETL legacy registry data at other centers. Our future plans include publishing our technical documentation on GitHub, continuing to collaborate with the OHDSI Registry Workgroup, expanding our ETL efforts to capture additional registry concepts not included in the current MVP, and to participate in and lead OHDSI network studies.

Figure 1
CANOE Registry ETL Process

International Epilepsy Community Network Study
References


