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

Danielle Boyce[1], Colin Bruce Josephson[2], Ray Jiang[3], Samuel Wiebe[2]

[1] Biomedical Informatics and Data Science, Johns Hopkins University, [2]Department of Clinical Neurosciences, University of Calgary, [3]Clinical 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 harmonization[2]. 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) registry[3] 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[7] employing Usagi software[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

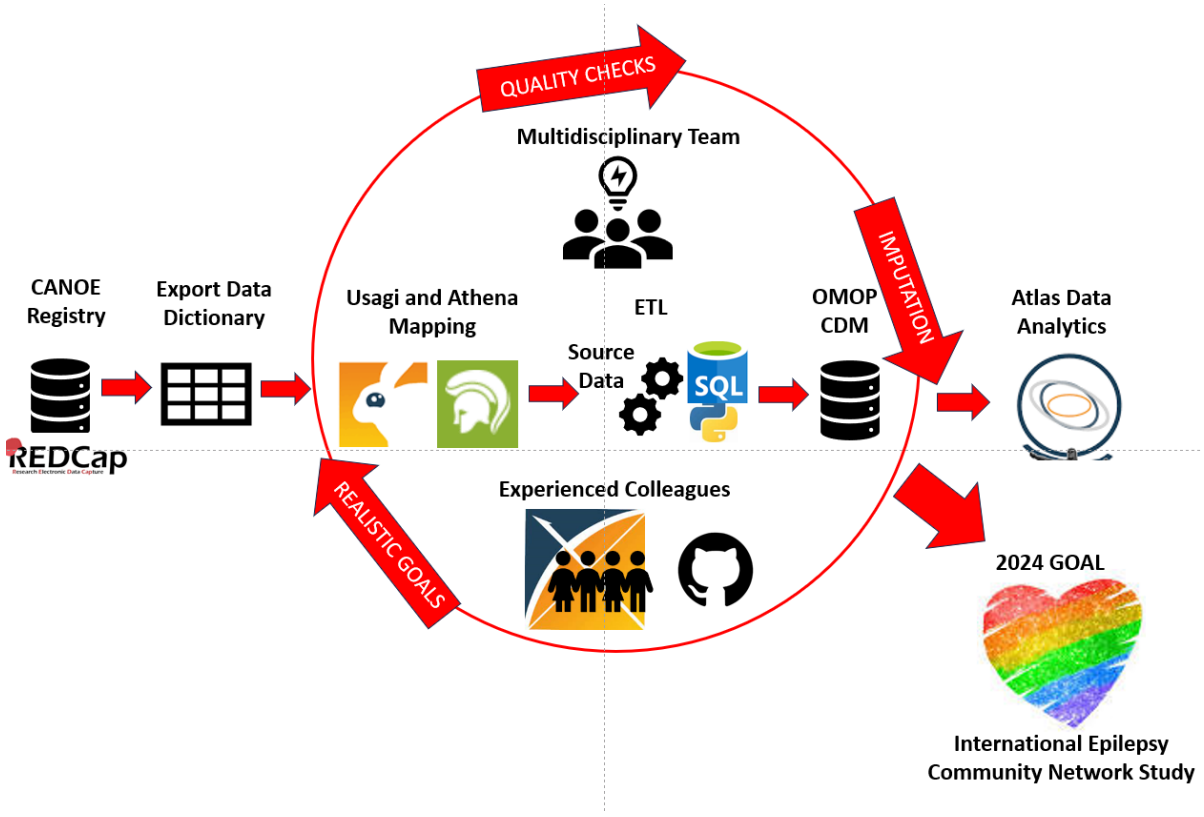
on a small subset of registry participants. However, using Usagi and Athena, the team was able to identify and map most of the concepts, and will implement an updated ETL should the use case arise.

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



References

1. World Health Organization. (2021). Epilepsy. https://www.who.int/news-room/fact-sheets/detail/epilepsy

2. Lhatoo SD, Bernasconi N, Blumcke I, Braun K, Buchhalter J, Denaxas S, Galanopoulou A, Josephson C, Kobow K, Lowenstein D, Ryvlin P, Schulze-Bonhage A, Sahoo SS, Thom M, Thurman D, Worrell G, Zhang GQ, Wiebe S. Big data in epilepsy: Clinical and research considerations. Report from the Epilepsy Big Data Task Force of the International League Against Epilepsy. Epilepsia. 2020 Sep;61(9):1869-1883. doi: 10.1111/epi.16633. Epub 2020 Aug 7. PMID: 32767763.

3. Josephson CB, Engbers JDT, Wang M, Perera K, Roach P, Sajobi TT, Wiebe S; Calgary Comprehensive Epilepsy Program collaborators. Psychosocial profiles and their predictors in epilepsy using patient-reported outcomes and machine learning. Epilepsia. 2020 Jun;61(6):1201-1210. doi: 10.1111/epi.16526. Epub 2020 May 20. PMID: 34080185.

4. PA Harris, R Taylor, R Thielke, J Payne, N Gonzalez, JG. Conde, Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support, J Biomed Inform. 2009 Apr;42(2):377-81.

5. PA Harris, R Taylor, BL Minor, V Elliott, M Fernandez, L O'Neal, L McLeod, G Delacqua, F Delacqua, J Kirby, SN Duda, REDCap Consortium, The REDCap consortium: Building an international community of software partners, J Biomed Inform. 2019 May 9 [doi: 10.1016/j.jbi.2019.103208]

6. Observational Health Data Sciences and Informatics. Common Data Model. OHDSI; 2023. Available from: https://ohdsi.github.io/CommonDataModel/cdm53.html. Accessed 10 Jun 2023.

7. OHDSI. OMOP Common Data Model: OHDSI; 2022 [Available from: https://ohdsi.github.io/CommonDataModel/.

8. OHDSI Community. (2023). Usagi (v.1.4.3) [Computer software]. GitHub. https://github.com/OHDSI/Usagi

9. "ETL--PulmonaryHypertensionRegistries." OHDSI. https://github.com/OHDSI/ETL--PulmonaryHypertensionRegistries/tree/master

10. Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, Wetherill G, Brand M, Didden EM. (2021). Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Medical Research Methodology, 21(1):238. https://doi.org/10.1186/s12874-021-01434-3