Themis Part 2 - The ETL data quality conversion initiative

Melanie Philofsky, RN, MS 1; Meghan Pettine, MS 2; Karthik Natarajan, PhD 3
1 Odysseus Data Services, Cambridge, MA, USA; 2 IQVIA, Plymouth Meeting, PA, USA; 3 Columbia University, New York, NY, USA

Background

The Observational Health Data Sciences and Informatics partnership (OHDSI) is an open source community and network of federated sites that have converted their observational healthcare data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). The key objectives of the CDM include standardization of semantic representation, format, and analysis of the data. Once the data are in the CDM, a researcher from the network can write one query to be run by all sites ensuring the query, definitions, and analytic methods are consistent. However, because each data partner's source system is different, there is no one tool or query that can be used for transforming source data into the CDM structure and codes. The OHDSI community has developed a growing body of highly detailed specifications and conventions for the extract, transform, and load (ETL) process of creating a CDM. Currently, there is a gap in the community to ensure the ETL conversion into CDM follows the conventions and model specifications. To address this gap, the Themis hack-a-thon focused on providing ways to verify that the source data to CDM data transformations conform to Themis specifications.

Methods

The two-day Themis hack-a-thon was to focus on creating methods and measures for assessing the quality of source-to-OMOP ETL conversions. We reviewed each existing OMOP CDM convention for testability, translated it into a verifiable SQL test, and classified it into a widely used data quality framework [1]. In evaluating each convention, it was determined some conventions were more ETL guidelines as opposed to a testable rule that could translate into SQL. The Themis hack-a-thon group also discussed the difference between the current Achilles tests and the proposed convention tests and started a discussion on certifying OMOP instances using both the Themis and Achilles derived data quality and conformance tests. Ultimately, this information will allow researchers to make an informed decision if an OMOP data resource meets a minimal requirement to be used in a network study.

Results

Using the published data quality framework, assessments are divided into two "contexts": Verification and Validation. Verification are data quality features that do not require access to external data sources, only requiring features that are intrinsic to the data source. Validation requires access to external comparisons, which may be known gold standards or other high quality OMOP data sets. Within both Verification and Validation contexts, three categories (with subcategories) were defined: Conformance (Value, Relational or Computational) , Completeness or Plausibility (Uniqueness, Atemporal or Temporal). As has been observed in other national data networks data quality programs [2], most current Themis topics and conventions discussed fell into the Verification context instead of the Validation context. Conformance was the most popular category. Below is a table stating examples of checks discussed and how they are categorized:

	Verification	**Validation**
Conformance	"The Visit during which the Device was first used is recorded through a reference to the VISIT OCCURRENCE table" (Relational) "Valid Device Concepts belong to 'Device' domain. The concepts of this domain are derived from the DI portion of a UDI or based on other source vocabularies, like HCPCS" (Value) "The drug_concept_id field only contains concepts that have the concept_class 'Ingredient'. The Ingredient is derived from the Drug Concepts in the Drug Exposure table that are aggregated into the Drug Era Record" (Computational)	Distribution of Lab Values Valid Zip Code Checks National NPI Code Checks *All considered General Conformance
Completeness	For fields that are optional, percentage are populated versus mapped to 0	Distribution of records by Domain across multiple CDMs

Plausibility	In the Drug Exposure Table, all start dates should be less than the end date (Temporal) Visit end dates are mandatory and some times may not be available in the source. We can then derive them For example an Outpatient Visit would have the same end date as start date. (Atemporal) Care Site is a unique combination of location_id and place_of_service_source_value. (Uniqueness)	% of Inpatient, Outpatient, Emergency Room Visits compared to other CDMs Distribution of Concept Prevalence compared to other CDMs or comparable data sets *All considered General Plausibility

It was found there is a low concordance between what Achilles Heel currently tests and the conventions and checks discussed during the Themis Hack-a-thon. Each table convention was discussed along with additional checks. It was determined there will be a second phase to this project. The second phase will incorporate queries written to check for THEMIS conventions, table conventions and additional conventions discussed by the group.

Discussion

Based on the conclusions of the above work, THEMIS will be putting a certification of the OMOP CDM into place. This certification will come in the form of queries that will be run on the data after it is converted into the CDM. The queries will test for the individual conventions discussed and decided by THEMIS, which will utilize a Data Quality Framework that has been discussed by the larger OHDSI community. The resulting product will be the development of an ETL Conformance Tool to measure conformance to the CDM conventions.

Next Steps

The conventions in place need to be revised in order to provide a description of greater clarity. Secondly, the Data Quality Framework will be incorporated into the ETL Conformance Tool which will be independent yet congruent with ACHILLES. It will differ from ACHILLES by testing for conventions that could take multiple factors into consideration rather than just a pass/fail of one field. To support rules in the Validation Context, future work will also incorporate summary statistics from the data for comparison across the network. The community is encouraged to share certain items that this tool will bring forth in the data. This will allow for comparison across networks to understand the quality of their data.

References

1 Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw S-T, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs (Generating Evidence & Methods to improve patient outcomes) [Internet]. 2016 Sep 11 [cited 2016 Sep 12];4(1). Available from: http://repository.edm-forum.org/egems/vol4/iss1/18

2 Callahan TJ, Bauck AE, Bertoch D, Brown J, Khare R, Ryan PB, Staab J, Zozus MN, Kahn MG. A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. EGEMS (Wash DC). 2017 Jun 12;5(1):8.