Applying the OMOP Common Data Model to Survey Data
Margaret S. Blacketer, MPH1,2, Erica A. Voss, MPH1,2, Patrick B. Ryan, PhD1,2
1Janssen Research & Development, LLC, Raritan, NJ
2OHDSI collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY

ABSTRACT

Background: The Epidemiology Analytics group at Janssen R&D has successfully transformed the NHANES dataset into the OMOP CDM and this study will examine the methods used to achieve this transformation.

Objectives: Evaluate the feasibility of transforming survey data into CDM.

Methods: NHANES questions were loaded into the OHDSI tool Usagi and mappings were created to link the survey questions to standard concepts. These maps are added to the SOURCE_TO_CONCEPT_MAP table which facilitates the transformation.

Results: There was 100% match between the raw data and transformed data break outs.

Conclusions: Survey data is feasible to transform into the CDM as illustrated with NHANES with no information loss from the source and our experience leads us to recommend that a similar process can be followed for other question/response observational data sources, including other surveys and registries.

CONFLICT OF INTEREST STATEMENT

Clair Blacketer, Erica Voss, and Patrick Ryan are full time employees of Janssen Research and Development, a unit of Johnson and Johnson. The work on this study was part of their employment. They also hold pension rights from the company and own stock and stock options.

BACKGROUND

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) has been shown to be an effective way to standardize observational health databases but has not been as commonly applied to survey and registry databases as it has for electronic health records and administrative claims.

The National Health and Nutrition Examination Survey (NHANES) is a program that combines survey information and physical examination results to determine the prevalence of major diseases and risk factors for disease among the U.S. population1.

The NHANES dataset has been successfully transformed into the OMOP CDM and this study will examine the methods used to achieve this transformation and serve as a guide on how to approach the conversion of survey and registry data to the OMOP CDM with little to no loss of data integrity.

OBJECTIVES

Evaluate the feasibility of transforming survey data into CDM.

Assess the completeness of vocabulary mapping for survey questions and responses into standardized vocabularies.

METHODS

We will evaluate the feasibility of survey extract, transform, load (ETL) through examination of our ETL for the NHANES dataset, which is a national open source database comprised largely of data in the form of question/response pairs.

We chose to focus on the 2005-2006 Mental Health Screener since each participant aged 18 years and older was an eligible respondent, allowing for a robust set of questions (survey DPQ020) will be highlighted throughout as an example of how both a question and response are translated into the CDM.

We designed an ETL process for survey questionnaires in columnar format (each question is a column, each response is a record) into the entity, attribute, value (EAV) structure of the OBSERVATION table within CDM (each record has two fields, one for question, one for response); we applied this ETL process to all questions of NHANES to assess if the process was suitable to accommodate all types of questions posed in the survey.

We determined the questions and responses from the source and used Usagi2 to evaluate with a defined threshold of 0.6 and then examined what percent of questions were we able to map to a standard vocabulary model. Table 1 shows how the Mental Health Depression Screener questions were successfully mapped to Logical Observation Identifiers Names and Codes (LOINC) concepts.

RESULTS

From the source data we use 2,939 questions across 71,916 respondents. We were able to transform 100% source data question/answer pairs to the CDM. Of the questions, 366 were previously manually mapped to standard concepts on manual review and 125 were confirmed with Usagi. Using Usagi with an automated threshold of 0.6 achieved another 543 mappings for a total of 688.

There was 100% match between the raw data and transformed data break outs. The NHANES website lists 5,334 cumulative responses to question DPQ020 and after transformation there are 5,334 rows of data in the OBSERVATION table with an OBSERVATION_SOURCE_VALUE.

CONCLUSIONS

Survey data is feasible to transform into the CDM as illustrated with NHANES with no information loss from the source.

Vocabulary mapping is going to be incomplete due to the unstructured nature of the question/response pairs and it will not be solved by manual or automated mapping tools since many of the questions asked in surveys do not have standard concepts.

The structure of the OBSERVATION table can hold question/response pairs but more work needs to be done to determine if a particular question generates data that best belongs in a different domain (e.g. self-reported HbA1c lab values may be better stored in the MEASUREMENT table) but we currently do not have a fully automated process to handle those decisions.

Nonetheless, our experience leads us to recommend that a similar process can be followed for other question/response observational data sources, including other surveys and registries.

REFERENCES