# Applying the OMOP Common Data Model to Survey Data

Margaret S. Blacketer, MPH[1,2], Erica A. Voss, MPH[1,2], Patrick B. Ryan, PhD[1,2,3]

[1]Janssen Research & Development, LLC, Raritan, NJ
[2]OHDSI collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY
[3]Columbia University, 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 variables 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

Margaret 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[1-3,6].
- 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. population[4].
- 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 Depression Screener because each participant aged 18 years and older was an eligible respondent, allowing for a robust set of answers (question 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 field, 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 in 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 Usagi[5] to evaluate with a defined threshold of 0.6 and then examined what percent of questions we were able to map to a standardized vocabulary. Table 1 shows how the Mental Health Depression Screener questions were successfully mapped to Logical Observation Identifiers Names and Codes (LOINC) concepts.

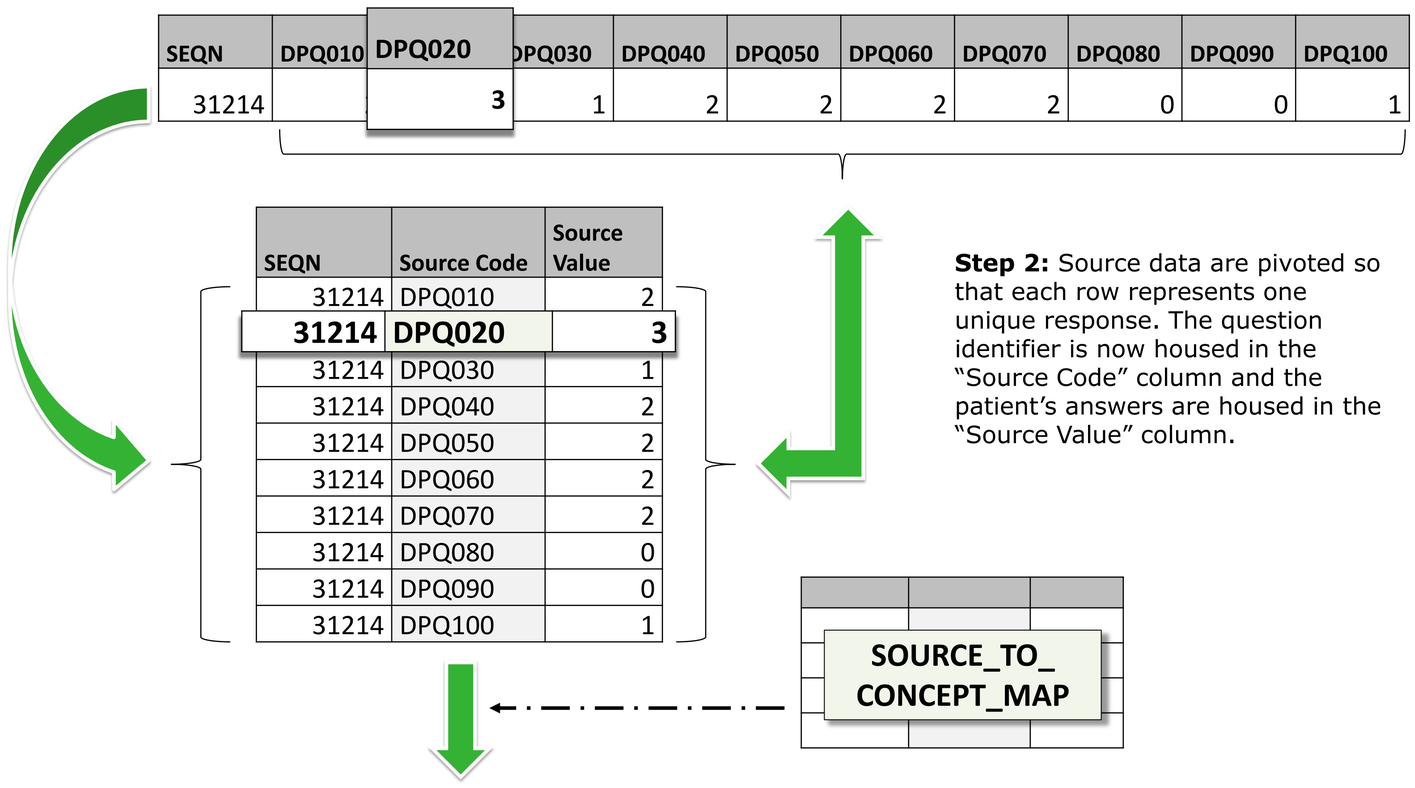
**Table 1:** NHANES Mental Health Depression Screener Questions Mapped to LOINC Concepts using Usagi

| Source Code | Source Description* | Target Concept ID | Concept Name |
|---|---|---|---|
| DPQ010 | Little interest or pleasure in doing things? | 3042924 | Little interest or pleasure in doing things in last 2 weeks [Reported.PHQ] |
| **DPQ020** | **Feeling down, depressed or hopeless?** | **40757792** | **Feeling or appearing down, depressed, or hopeless in last 2 weeks.frequency [Observed PHQ-9 MDSv3]** |
| DPQ030 | Trouble falling or staying asleep, or sleeping too much? | 3045933 | Trouble falling or staying asleep, or sleeping too much in last 2 weeks [Reported.PHQ] |
| DPQ040 | Feeling tired or having little energy? | 42870477 | I feel tired - have no energy in the last 2 weeks or more [M3] |
| DPQ050 | Poor appetite or overeating? | 3044098 | Poor appetite or overeating in last 2 weeks [Reported.PHQ] |
| DPQ060 | Feeling bad about yourself - or that you are a failure or have let yourself or your family down? | 3043801 | Feeling bad about yourself - or that you are a failure or have let yourself or your family down in last 2 weeks [Reported.PHQ] |
| DPQ070 | Trouble concentrating on things, such as reading the newspaper or watching TV? | 3045019 | Trouble concentrating on things, such as reading the newspaper or watching television in last 2 weeks [Reported.PHQ] |
| DPQ080 | Moving or speaking so slowly that other people could have noticed? Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual? | 3043785 | Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual in last 2 weeks [Reported.PHQ] |
| DPQ090 | Thoughts that you would be better off dead or of hurting yourself in some way? | 3043462 | Thoughts that you would be better off dead, or of hurting yourself in some way in last 2 weeks [Reported.PHQ] |
| DPQ100 | How difficult have these problems made it for you to do your work, take care of things at home, or get along with people? | 40772146 | How difficult have these made it for you to do your work, take care of things at home, or get along with other people [Reported.PHQ] |

*Each question was asked in relation to the past 2 weeks.

**Table 2:** As reported by NHANES, the raw data breakout showing the frequency of each response to the 2005-2006 Mental Health Depression Screener question DPQ020[4].

| DPQ020 Response Frequency | | | |
|---|---|---|---|
| Code or Value | Value Description | Count | Cumulative |
| 0 | Not at all | 3,769 | 3,769 |
| 1 | Several days | 769 | 4,538 |
| 2 | More than half the days | 179 | 4,717 |
| 3 | Nearly every day | 114 | 4,831 |
| 9 | Don't know | 5 | 4,836 |
| . | Missing | 498 | 5,334 |




**Step 1:** NHANES Mental Health Depression Screener data loaded. SEQN represents a unique patient identifier and each column represents a survey question. In the below example person 31214 responded to question DPQ020 with a three, which in this instance means "Nearly every day".

**Step 2:** Source data are pivoted so that each row represents one unique response. The question identifier is now housed in the "Source Code" column and the patient's answers are housed in the "Source Value" column.

| PERSON_ID | OBSERVATION_ CONCEPT_ID | ... | VALUE_AS_ NUMBER | VALUE_AS_STRING | VALUE_AS_CONCEPT _ID | OBSERVATION_SOURCE _VALUE |
|---|---|---|---|---|---|---|
| 31214 | 3042924 | | 2 | More than half the days | 45878994 | DPQ010 |
| **31214** | **40757792** | **...** | **3** | **Nearly every day** | **45882010** | **DPQ020** |
| 31214 | 3045933 | ... | 1 | Several days | 45879886 | DPQ030 |
| 31214 | 42870477 | ... | 2 | More than half the days | 45878994 | DPQ040 |
| 31214 | 3044098 | ... | 2 | More than half the days | 45878994 | DPQ050 |
| 31214 | 3043801 | ... | 2 | More than half the days | 45878994 | DPQ060 |
| 31214 | 3045019 | ... | 2 | More than half the days | 45878994 | DPQ070 |
| 31214 | 3043785 | ... | 0 | Not at all | 45883172 | DPQ080 |
| 31214 | 3043462 | ... | 0 | Not at all | 45883172 | DPQ090 |
| 31214 | 40772146 | ... | 1 | Somewhat difficult | 45877108 | DPQ100 |

**Step 3:** The SOURCE_TO_CONCEPT_MAP table is used to map the questions and answers into the OBSERVATION table. In this instance the OBSERVATION table is the destination because the framework of the table allows for the storage of both the question and answer as concepts while also allowing the question identifier (OBSERVATION_SOURCE_VALUE) to be captured.

## 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 668.
- 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 of DPQ020[4].

## 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 all 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

1. OMOP Common Data Model. [Webpage]. 2015; http://www.ohdsi.org/data-standardization/the-common-data-model/, 20 Jul 2015.
2. Voss E, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association.* 2015 May 2015;22(3):553-564.
3. Ogunyemi O, Meeker D, Kim H-E, Ashish N, Farzaneh S, Boxwala A. Identifying Appropriate Reference Data Models for Comparative Effectiveness Research (CER) Studies Based on Data from Clinical Information Systems. *Medical Care* 2013; S45-S52. Available at. Accessed 8, 51.
4. National Health and Nutrition Examination Survey. 2014; http://www.cdc.gov/nchs/nhanes/about_nhanes.htm. Accessed 13Aug2015.
5. Schuemie M. Usagi. 2015; http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:usagi. Accessed 24 Aug 2015, 2015.
6. (2015). "Data Network." 2015, from http://www.ohdsi.org/web/wiki/doku.php?id=resources:data_network.