Streamlining Research Data Standardization: AI-READI Survey Instrument Data Elements and MoCA Measurement Data Elements are curated and mapped utilizing a Standardized Value Set Mapping Table for transformation into the OMOP Common Data Model

Stephanie S. Hong SB, FAMIA[1], James Cavallon BS[1], Yi-Ju Chen, MS[1], Monique Bangudi, MPH[2], Jessica Mitchell, MS[1], Dawn Matthies PhD[3], Steven Chamberlin, Aaron Cohen MD MS [4], Julie Owens PhD[6], Abigail Lucero[4], Sally Baxter MD, MSc[5], Christopher G Chute, MD DrPH[1], Cecilia S. Lee MD MS[6,7], Aaron Lee, MD MSCI[6,7], and on behalf of the AI-READI consortium

[1] Johns Hopkins University School of Medicine, Baltimore, MD, [2] University of Maryland, College Park, MD, [3] University of Alabama, Birmingham, AL [4] Oregon Health & Science University, Portland, OR, [5] University of California San Diego, San Diego, CA, [6] University of Washington, Seattle, WA, [7] Roger and Angie Karalis Johnson Retina Center, Seattle, WA

Introduction

The Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READi) project is one of the four Data Generation Projects funded by Bridge2AI, an NIH Common Fund Program aimed at setting the stage for widespread adoption of AI in health research.

The goal of AI-READI is to develop a multi-modal atlas of type 2 diabetes mellitus (T2DM) by collecting data from a diverse population while simultaneously creating a roadmap for ethical and equitable research focusing on participant diversity. It aims to create an ethically sourced dataset to advance artificial intelligence and machine learning research for T2DM. The study includes medical data from 4,000 participants from diverse backgrounds with varying levels of T2DM severity. To obtain extensive clinical data from participants, several survey instruments were created and used to collect data.

Data standardization and harmonization is critical in enabling efficient data sharing that meet the FAIR (Findable, Accessible, Interoperable, Reusable) Principles. Here we present challenges and our approaches in standardizing and harmonizing newly collected survey instrument data for meaningful use and how we integrated survey data into the OMOP Common Data Model (CDM).

Method

We enhanced data utility by standardizing each survey question and answer option to fit standardized terminology. Subsequently, we transformed the data into the OMOP CDM, ensuring consistency in format and terminology concepts. Here, we detail the standardization and harmonization process. The data element mapping process can involve a time intensive

semantic review and mapping process with existing standard terminology. The following is a list of data mapping standardization guidelines we followed.

- **Address lack of standardized terminology:** Different survey forms might use different terms to refer to the same concept, requiring careful semantic mapping to ensure accuracy. Custom codes are generated when there is no existing standard terminology code that can be used to capture the survey data element.
- **Curate the response concept in context of the survey question, heterogeneity of survey instruments**: Different survey forms may ask similar questions in varied ways and different answer responses can cause the question concepts to be mapped differently. We reviewed the answers in the context of the survey question for mapping the response data elements.
- **Carefully review mapping abstract concepts**: Some survey questions address abstract concepts or subjective experiences that are difficult to map to standardized data elements. The data element mapping process included review of the existing terminology for comparability, searchability and meaningful use. We followed the FAIR data principles with consideration of aggregate analysis. The data are organized following the OMOP concept domain guidelines.
- **Custom question extension concepts created:** The OMOP CDM primarily focuses on capturing and standardizing positive clinical data, such as visits, diagnoses, procedures, and medication use. This focus can present challenges when trying to record negative responses to a survey question to record the fact that the response was captured both negatively and positively (e.g., absence of a condition or lack of a specific symptom). Custom extension concepts were created to capture required survey instrument concepts including both negative and positive responses.
- **Simple Standard for Sharing Ontological Mappings (SSSOM) Metadata**: The SSSOM metadata in the mapping table includes mapping confidence and predicate ID, crucial for identifying relationships between source data elements and mapped target concept data elements, thereby enhancing mapping quality and transparency. Modifiers were employed to address post-coordination issues; for example, specifying left or right laterality for Photopic and Mesopic contrast sensitivity values was achieved using the modifier column.
- **Survey instrument and mapping table version control**: Even with the best effort to stabilize the survey forms prior to data mapping process, it can change over time, with new versions introducing changes in questions and structure, necessitating ongoing updates to the mapping process.
- **AI-READi Custom Concepts are candidates for OMOP vocabulary expansions.** And the survey concepts are considered for the Survey_conduct domain. The negative survey responses are also captured, which diverges from the typical storage conventions in the OMOP CDM. Researchers should be mindful that the presence of a survey item in the dataset does not necessarily indicate a positive result, as it may represent a negative response of a survey instrument.

Despite challenges from survey instruments, we adopted a standard mapping table format to map AI-READI data elements. This table adheres to the FAIR principles. It utilizes existing terminology codes from the OHDSI vocabulary concept table and locally generated custom codes, organizing data according to OMOP domain conventions for consistent searchability.

We manually curated AI-READI survey instrument questions, response data elements, and MoCA score data elements, ensuring consistency. Using the standardized value set mapping table, we transformed the data into OMOP CDM format. This mapping table format supports straightforward data transformation, indicating mapping confidence levels for reference.

The data element mapping process involves connecting sources and documenting using codes/code systems, facilitating data transformation into the OMOP CDM schema format. The adopted standardized value set mapping table simplifies transformation into OMOP Common Data Model through straightforward joins on source data elements for semantic concept harmonization.

Standardized Value Set Mapping Table Columns Described:

Mapping Table Column	***Description***
FORM_NAME	*Survey Instrument Name*
FIELD_TYPE	*Data Element Field ID type*
FIELD_ID	*Data Element field ID*
SRC_CODE	*Permissible Value*
SRC_CODE_ID	*Permissible Data Element Value ID*
SRC_CD_DESCRIPTION	*Permissible Value description*
TARGET_CONCEPT_ID	*OMOP concept id*
LOCAL_CONCEPT_ID	*Custom OMOP concept id generated, when needed*
TARGET_CONCEPT_NAME	*Concept name*

TARGET_DOMAIN_ID	*Domain id*
TARGET_VOCABULARY_ID	*Vocabulary id*
TARGET_CONCEPT_CLASS_ID	*OMOP concept class*
TARGET_STANDARD_CONCEPT	*Standard concept*
TARGET_CONCEPT_CODE	Target concept code
PREDICATE_ID	The ID of the predicate level or relation that relates the subject and target of the concept
CONFIDENCE	Mapping confidence
MODIFIER	Modifier to aid post coordination mapping

Figure 1

The survey data output is saved in the csv format. The survey source data with the data element id and data element source value of each field are joined using the mapping table described above to produce the harmonized dataset. The table is designed such that you can join on the data element field id and the value in order to retrieve the corresponding target_concept_id. Once retrieved, each field's corresponding target_concept_id can be used to insert the concept into the respective OMOP domain according to the target_domain_id specified in the mapping table.

Results

Summary of Data Standardization and Mapping Efforts from 47 Survey Instruments Across 4000 Participants:

	Mapped with existing Standard OMOP Concepts	Mapped with AI-READI Custom Extension concepts
1821 AI-READI Survey Data Elements mapped	1337 (349 distinct OMOP concepts)	484 (344 distinct concepts)

Figure 2

- Of the 2704 curated data elements, not all concepts required transformation into the OMOP CDM.
- 1821 data elements are mapped from 47 survey instruments
- 1337 data elements are mapped to 349 existing terminology codes
- 344 AI-READi custom codes are created to support new survey concepts
- 1784 AI-READi data elements are transformed to OMOP CDM
- 37 MoCA data elements are mapped
    - 27 Mapped with AI-READI Custom Extension concepts
    - 10 Mapped with existing Standard OMOP Concepts

Target Domain ID	Distinct Non-Custom Code Count	Distinct Custom Code Count
Observation	105	197
Device	1	1
Measurement	22	58
Meas Value	182	85
Procedure	1	5
Route	1	0
Condition	24	0
Unit	13	0
Total	**349**	**344**

Figure 3

Conclusion

Across the 47 AI-READI survey instruments, 2704 data elements were curated and mapped using a standardized value-set mapping table. To support concepts where existing terminology was unavailable, custom AI-READI codes were created for harmonization. While not all data elements required transformation into the OMOP CDM, the mapping table structure was designed to facilitate this process. Currently, 168 custom concepts are stored in the Observation domain, though they are candidates for the Survey_Conduct domain. Moving these concepts to the Survey_Conduct domain would allow for a more consistent approach to mapping survey data in the OMOP CDM and enhance integration with standard real-world data.

It is also important to note that negative survey responses are captured, which diverges from typical OMOP CDM storage conventions. Researchers should be aware that the presence of a survey item in the dataset does not necessarily indicate a positive response, as it may represent a negative result.

Reference

AI-READi Data Element Mapping Table:
https://github.com/AI-READI/DataElementMaps/blob/main/mappings-csv/AIREADi_Pilot_Redcap_Data_Dictionary_Mappings/Redcap_Data_Dictionary_Mappings.csv

MoCA Data Element Mapping table:
https://github.com/AI-READI/DataElementMaps/blob/main/mappings-csv/MOCA_Data_Dictionary_Mappings/MOCA_Data_Dictionary_Mappings.csv

OMOP CDM Documentation - OMOP Common Data Model Documentation

OHDSI Book of OHDSI- This book provides detailed insights into the methodology, tools, and best practices for working with OMOP CDM.

OHDSI Forums - The OHDSI forums are a valuable resource for community discussions.

A Simple Standard for Sharing Ontology Mappings (SSSOM) Specification - URI: https://w3id.org/sssom/schema