

Exploring the Application of the Observational Medical Outcomes Partnership Common Data Model to Multisite Stroke Rehabilitation Research Data

Katherine J. Loomis, PhD, OTR/L, CHT¹, Amisha Kumar, BS², Octavio Marin-Pardo, PhD¹, Grace C. Bellinger, PhD, MPH³, Margaret A. French, PhD, DPT⁴, Ryan T. Roemmich, PhD^{5,3}, Sook-Lei Liew, PhD, OTR/L, FAOTA^{1,6}

¹Chan Division of Occupational Science and Occupational Therapy, University of Southern California, Los Angeles, CA, USA; ²Dornsife College of Letters, Arts, and Sciences, University of Southern California, Los Angeles, CA, USA; ³Department of Physical Medicine and Rehabilitation, Johns Hopkins University School of Medicine; ⁴Department of Physical Therapy and Athletic Training, University of Utah, Salt Lake City, UT, USA; ⁵Center for Movement Studies, Kennedy Krieger Institute, Baltimore, MD, USA; ⁶Division of Biokinesiology and Physical Therapy, Department of Biomedical Engineering, Department of Neurology, Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA, USA

Background

Emerging innovations in artificial intelligence and machine learning (AI/ML) offer powerful tools that can help generate robust knowledge to support the development of precision rehabilitation approaches for varied patient populations.¹ Following research data sharing mandates² and the widespread adoption of electronic health records,³ there are vast quantities of research and clinical data available that can be aggregated into large-scale datasets to train and be analyzed by complex AI/ML models. However, there is a pronounced lack of interoperability across both research-generated and clinical rehabilitation datasets due to siloed, heterogeneous data collection and storage, which presents a massive barrier to aggregation.⁴ The Observational Medical Outcomes Partnership (OMOP) common data model (CDM) is a widely used, open-source, and well-supported framework with an established history of improving EHR data interoperability across numerous medical fields,⁵ and there is growing advocacy for its use specifically for rehabilitation data.^{6,7} In anticipation of increasing adoption in clinical rehabilitation contexts, it is pertinent to explore the OMOP CDM's ability to act as a bridge for incorporating structured rehabilitation research data into clinical datasets, with the goal of enhancing data diversity, volume, and granularity for AI/ML models. As a foundational step, we evaluated the ability of the OMOP CDM⁸ to represent data in a harmonized, multi-site repository of stroke rehabilitation research data: the Enhancing NeuroImaging Genetics through Meta-Analysis Stroke Recovery (ENIGMA-SR) database.⁹

Methods

We evaluated the breadth and depth of OMOP coverage of rehabilitation-related (non-imaging) variables (n=932) from the ENIGMA-SR database (2,400 participants, 62 research cohorts),⁹ spanning two categories: (1) demographics and medical history (DMH) and (2) rehabilitation assessments. To enable more structured and interpretable analyses, variables were represented by "proxy" data elements (n=246), removing qualifiers of laterality, anatomical location, or timing. DMH variables (n=46) and distinct assessments (n=95) were represented by data elements 1-to-1. For complex assessments (n=73; those having multiple items or datapoints), additional data elements were included for assessment subscales and items (one representing all datapoints at each granularity level per assessment).

Using a similar approach to previous studies,^{7,10} based on OHDSI recommendations,⁸ two raters with extensive knowledge of ENIGMA-SR data independently mapped each element to standard OMOP concepts, facilitated by automated USAGI¹¹ recommendations and supplemental information from Athena.¹² Mapping was completed exhaustively for each data element (using all appropriate concepts). Initial rater agreement was assessed for OMOP inclusion (data element mapped to at least one concept) and for specific OMOP concepts used (primary metric: Gwet's agreement coefficient [AC]).¹³ Mapping differences were reconciled via discussion between raters. Final mappings were descriptively analyzed to examine (1) OMOP inclusion of DMH variables and distinct assessments, (2) inclusion of more granular levels of complex assessments (subscales, items), and (3) characteristics of mapped concepts.

Results

Initial rater agreement was good/very good for OMOP inclusion of DMH and assessment data elements (Gwet's AC: 0.79 and 0.89) and for specific OMOP concepts mapped to assessment- and subscale-level data elements (Gwet's AC: 0.79). Initial concept agreement between raters was more variable for DMH and item-level data elements; however, all mapping differences were successfully reconciled to 100%.

DMH data elements generally had far higher OMOP inclusion than rehabilitation assessments (overall, at any granularity level): 84.8% (39/46) vs 58.9% (56/95; **Table 1**). Excluded DMH data elements *all* had stroke-related qualifiers (e.g., 'Handedness after stroke'). Assessments had variable inclusion based on type (i.e., cognitive, global function, psychosocial, or sensorimotor), ranging from 51.5% (17/33) for cognitive to 69.2% (9/13) for psychosocial assessments. OMOP coverage was especially limited for more granular levels of complex assessments, including only 9.4% (3/32) of subscale-level and 19.2% (14/73) of item-level data elements. Comparing the highest level of granularity included for each complex assessment, OMOP had lower granularity (i.e., fewer assessments with item-level coverage) than ENIGMA-SR for 56.2% (41/73) of assessments and higher granularity for only 11% (8/73). Notably, several critical motor rehabilitation assessments frequently collected in stroke research were absent or had very limited coverage: Fugl-Meyer Upper Extremity Motor Assessment, Wolf Motor Function Test, and Action Research Arm Test.

Table 1. OMOP inclusion of ENIGMA-SR rehabilitation data elements by category and granularity

Data category	Data elements	OMOP inclusion*
	<i>n</i>	<i>n</i> (%)
Demographics & medical history	46	39 (84.8%)
Assessments (overall, at any level)	95	56 (58.9%)
Complex assessments, assessment level	73	44 (60.3%)
Complex assessments, subscale level	32	3 (9.4%)
Complex assessments, item level	73	14 (19.2%)

**Inclusion = mapping to ≥ 1 OMOP concept: For distinct assessments, at any or granularity level (assessment-, subscale-, or item-level), and for all other categories, per data element*

Across both data element categories, we observed sizable variability in representation within the OMOP CDM structure, consistent with prior rehabilitation mapping work.⁷ DMH elements were mapped to a median of 6 concepts (range: 1-23), and 58% of these multi-mapped elements had concepts spanning 2-3 different OMOP domains (**Figure 1a**; most often Observation, Condition, and Meas Value). Similarly, assessment-level elements were mapped to a median of 2 concepts (range: 1-8), with 53.6% of multi-mapped elements having concepts from 2 domains (**Figure 1b**; most often Measurement and Observation). Conversely, each subscale- and item-level element was mapped to concepts exclusive to a single domain (**Figure 1c**; either Observation or Measurement). Overall, these differences present major challenges to data harmonization, potentially creating variability in both *where* and *how* the same construct is stored across datasets.

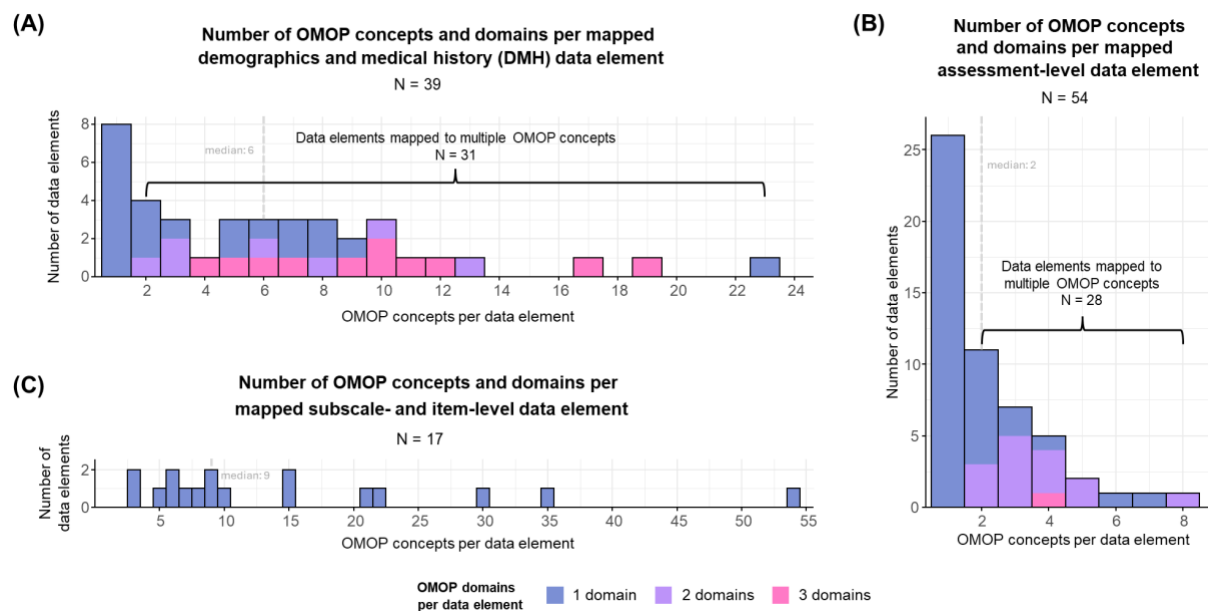


Figure 1. Number of OMOP concepts and domains per mapped (A) demographics and medical history (DMH) data element, (B) assessment-level data element, and (C) subscale- and item-level assessment data elements

Conclusion

The OMOP CDM provides strong coverage of DMH data and moderate top-level coverage of rehabilitation-related assessments housed within the ENIGMA-SR database, though has far more limited coverage of more granular assessment data. Furthermore, substantial variability in concepts and domains may present additional challenges for the specific use case of harmonizing and aggregating research-generated and clinical rehabilitation data into AI/ML-ready datasets. The most immediate steps that can be taken to help realize OMOP's potential as a bridge between these data sources are expanding the vocabulary to include more rehabilitation-related concepts and building cross-walks to research-oriented data standards (e.g., National Institute of Neurological Disorders and Stroke [NINDS] NeuroRehab Common Data Elements).¹⁴ For the OHDSI community, these findings highlight targets for vocabulary development and related opportunities to expand the utility of the OMOP CDM into a novel, but highly relevant and impactful context.

Acknowledgment

This work was supported by NIH/NICHD P50HD118603 and NIH/NICHD P50HD118624.

References

1. French MA, Roemmich RT, Daley K, et al. Precision rehabilitation: optimizing function, adding value to health care. *Arch Phys Med Rehabil.* 2022;103(6):1233-1239.
2. Gonzales S, Carson MB, Holmes K. Ten simple rules for maximizing the recommendations of the NIH data management and sharing plan. *PLoS Comput Biol.* 2022;18(8):e1010397.
3. Modi S, Feldman SS. The value of electronic health records since the Health Information Technology for Economic and Clinical Health Act: systematic review. *JMIR Med Inform.* 2022;10(9):e37283.
4. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. *BMC Med Inform Decis Mak.* 2023;23(1):18.
5. Weeks J, Pardee R. Learning to share health care data: a brief timeline of influential common data models and distributed health data networks in U.S. health care research. *EGEMS (Wash DC).* 2019;7(1):4.
6. Cavanaugh R, Manes MR, Banks CL, et al. Challenges and opportunities of using common data models in rehabilitation. *SportRxiv.* 2026. Preprint.
7. French MA, Hartman P, Hayes HA, Ling L, Magel J, Thackeray A. Coverage of physical therapy assessments in the Observational Medical Outcomes Partnership Common Data Model. *Appl Clin Inform.* 2024;15(5):1003-1012.
8. Observational Health Data Sciences and Informatics. The Book of OHDSI. Accessed February 1, 2026. <https://ohdsi.github.io/TheBookOfOhdsi/>
9. Liew SL, Zavaliangos-Petropulu A, Jahanshad N, et al. The ENIGMA Stroke Recovery Working Group: big data neuroimaging to study brain-behavior relationships after stroke. *Hum Brain Mapp.* 2022;43(1):129-148.
10. Cai CX, Halfpenny W, Boland MV, et al. Advancing toward a common data model in ophthalmology: gap analysis of general eye examination concepts to standard Observational Medical Outcomes Partnership (OMOP) concepts. *Ophthalmol Sci.* 2023;3(4):100391.
11. OHDSI. Usagi. Accessed February 1, 2026. <https://ohdsi.github.io/Usagi/>
12. OHDSI. Athena: automated terminology harmonization, extraction and normalization for analytics. Accessed February 1, 2026. <https://athena.ohdsi.org>
13. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol.* 2008;61(1):29-48.
14. Carlozzi NE, Mendoza-Puccini MC, Marden S, et al. Common data elements for rehabilitation research in neurologic disorders (NeuroRehab CDEs). *Arch Phys Med Rehabil.* 2025;106(7):981-988.