

Standardizing the A4 Alzheimer's Disease Research Dataset into the OMOP Common Data Model to Enable Accessible Precision Medicine

Sarina Azimian Zavareh¹,DDS; Teri Sippel Schmidt¹,MS, Benjamin Martin¹,PhD, Milap Nowrangi^{1,2},MD ; Paul Nagy¹,PhD; Blake E. Dewey^{1,3},PhD

¹Department of Biomedical Informatics and Data Science, Johns Hopkins University School of Medicine, Baltimore, Maryland ²Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland ³Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, Maryland

Background

Clinical trials provide high-quality longitudinal data with carefully collected imaging, biomarkers, treatments, and outcomes; however, study-specific structures limit integration with real-world clinical populations. Transforming clinical trial datasets into the OMOP Common Data Model (CDM) enables comparison with observational health data to evaluate generalizability, validate findings, and support precision medicine research¹. The Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease (A4) Study is a deeply phenotyped clinical trial Alzheimer's disease that includes cognitive assessments, amyloid and tau PET imaging, MRI measurements, biomarkers, and treatment information^{2,3,4}. This work standardized A4 terminology into the OMOP CDM by mapping research variables to standardized vocabularies (including imaging outcomes in the OMOP Medical Imaging CDM). Creating an OMOP-aligned resource allows researchers and clinicians to use shared OHDSI tools for cohort discovery, phenotype development, and clinical analytics, without needing to understand study-specific data structures. Standardizing A4 into OMOP CDM enables longitudinal clinical, imaging, biomarker, treatment, and outcome data to be analyzed within a shared framework and used alongside external EHR-based Alzheimer's cohorts. This supports precision medicine applications, including identification of candidates for anti-amyloid therapy, prediction of treatment response³, and amyloid-related imaging abnormalities (ARIA)⁵, and evaluation of relationships between imaging biomarkers, disease progression, and patient outcomes⁶.

Methods

Variables from four A4 study dictionaries were reviewed and standardized using the OMOP vocabulary framework⁷: clinical variables (demographics, cognitive assessments, medical history), external variables (imaging, biomarkers, and laboratory measurements), derived variables (calculated scores and research-generated measures), and visit variables (study timeline and protocol information). Each variable name, description, and metadata definition was evaluated to identify semantically equivalent standardized concepts. For each mapped variable, the OMOP concept ID, concept name, source vocabulary term, vocabulary type (including SNOMED, LOINC, UK Biobank, and other vocabularies), and concept code were documented to create a reusable mapping resource. Variables without equivalent concepts were assigned custom concepts to preserve research-specific meaning while maintaining OMOP compatibility. Initial LLM-assisted approaches were explored; however, expert validation remained necessary due to the complexity of research-specific variables. Mapping followed a staged review process: Initial concept

selection and vocabulary mapping were performed by one reviewer, followed by weekly review sessions with an OMOP vocabulary expert and informatics faculty member. Final validation was completed by two clinical and informatics experts to resolve ambiguous mappings and ensure semantic accuracy, clinical relevance, and OMOP compatibility.

Results

The standardization process included 1,813 variables and required approximately 120 hours across 9 weeks. As shown in Figure 1, among all variables, 752 (41.5%) were represented using existing standardized vocabulary concepts, while 1,061 (58.5%) required custom concepts to maintain detailed research-specific information. Each study database contributed different numbers of custom concepts. Three databases were mapped to around 50% custom concepts: clinical (359/619, 58.0%), external (443/835, 53.1%), and derived (107/207, 51.7%), but visit-related variables all required custom concepts (152/152, 100%). Many custom concepts represented complex research-derived variables, in which study-specific definitions, calculations, or multiple clinical concepts were combined into a single variable. These findings demonstrate the feasibility of transforming complex public research datasets into OMOP MI-CDM, reducing technical barriers, and allowing broader use of high-value datasets for reproducible research and precision medicine applications⁵.

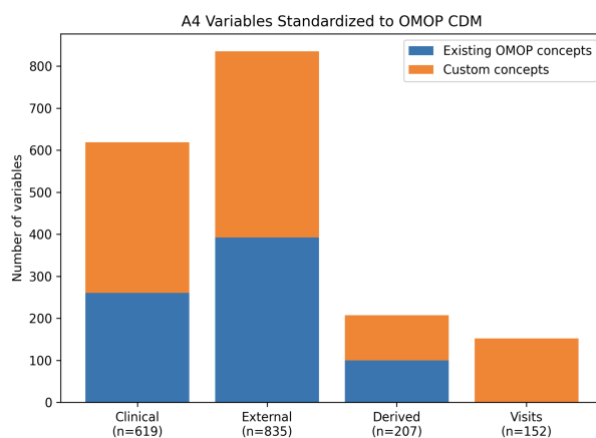


Figure 1. Distribution of A4 variables mapped to existing OMOP vocabulary concepts and custom concepts across four study dictionaries. Among 1,813 reviewed variables, 752 (41.5%) were represented using existing concepts and 1,061 (58.5%) required custom concepts to preserve research-specific information.

Conclusion

Standardizing the A4 dataset into OMOP MI-CDM provides a scalable pathway to transform complex public research datasets into reusable resources that integrate with real-world clinical data^{1,7}. This work enables researchers to move beyond study-specific data structures by using standardized vocabularies and OHDSI tools for cohort discovery, reproducible analyses, and cross-dataset validation⁷. Mapping the A4 dataset establishes a framework that can be extended to additional public datasets, increasing

accessibility and accelerating precision medicine research. While this harmonization process required significant manual effort and expert validation, human-in-the-loop LLM workflows may improve scalability by assisting with concept identification, semantic matching, and mapping review for future datasets.

We acknowledge the A4 Study investigators, USC Alzheimer's Therapeutic Research Institute (ATRI), participants, and collaborators. A4 data sharing was supported by the Epstein Family Alzheimer's Research Collaboration and Gates Ventures.

References

1. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clin Pharmacol Ther.* 2017;102(6):924-933. doi:10.1002/cpt.857
2. Sperling RA, Rentz DM, Johnson KA, Karlawish J, Donohue M, Salmon DP, et al. The A4 Study: stopping AD before symptoms begin? *Sci Transl Med.* 2014;6(228):228fs13. doi:10.1126/scitranslmed.3007941
3. Sperling RA, Donohue MC, Raman R, Rafii MS, Johnson K, Masters CL, et al. Trial of solanezumab in preclinical Alzheimer's disease. *N Engl J Med.* 2023;389(12):1096-1107. doi:10.1056/NEJMoa2305032
4. Jimenez-Maggiore GA, Schulz AP, Donohue MC, et al. Maximizing the utility of Alzheimer's disease trial data: sharing of baseline A4 and LEARN data. *J Prev Alzheimers Dis.* 2024;11(4):889-894. doi:10.14283/jpad.2024.120
5. Sperling RA, Jack CR Jr, Black SE, Frosch MP, Greenberg SM, Hyman BT, et al. Amyloid-related imaging abnormalities in amyloid-modifying therapeutic trials: recommendations from the Alzheimer's Association Research Roundtable Workgroup. *Alzheimers Dement.* 2011;7(4):367-385. doi:10.1016/j.jalz.2011.05.2351
6. Jack CR Jr, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA Research Framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* 2018;14(4):535-562. doi:10.1016/j.jalz.2018.02.018
7. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574-578.