

Standardizing Pre-Symptomatic Alzheimer's Trial Data to OMOP for Scalable and Reproducible Research

Haeun Lee¹, MS; Robert Barrett¹; Jen Park¹, PhD; Benjamin Martin¹, PhD; Sarina Azimian Zavareh¹, DDS; Milap Nowrangi^{1,2}, MD; Blake E. Dewey², PhD; Teri Sippel Schmidt¹, MS; Paul Nagy¹, PhD

¹ Biomedical Informatics and Data Science, Johns Hopkins University School of Medicine, Maryland, United States.

² Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Maryland, United States

³ Department of Neurology, School of Medicine, Johns Hopkins University School of Medicine, Maryland, United States

Background

Pre-symptomatic Alzheimer's disease (AD) prevention trials generate longitudinal, multimodal datasets that characterize biomarker-defined disease stages prior to clinical diagnosis. This provides critical evidence for identifying, monitoring and predicting high-risk AD individuals for disease-modifying treatments^{1,2}. The Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease (A4)/ Longitudinal Evaluation of Amyloid Risk and Neurodegeneration (LEARN) initiative has made these data available for secondary research; however, the study-specific data structure and vocabularies limit integration into larger, collaborative research efforts³. Moreover, open-source community-developed tools, such as those developed by Observational Health Data Sciences and Informatics (OHDSI), aim to improve accessibility to data analysis and evidence generation to broader research audiences. This study developed and implemented a data transformation methodology to standardize pre-symptomatic AD trial data for the OMOP CDM, enabling interoperable, scalable, and accessible use within observational research.

Methods

This study used data from the A4 randomized placebo-controlled trial evaluating Solanezumab in cognitively normal adults with preclinical Alzheimer's disease. The dataset includes demographic and clinical information, functional and neuropsychological assessments, multimodal neuroimaging biomarkers, and digital cognitive assessment data. We developed an open-source ETL tool to transform the A4 and LEARN datasets into the OMOP CDM v5.4 with the MI-CDM extension⁴ (Figure 1). The ETL tool was implemented in Python (v3.9) and designed to perform value transformation, rule-based transformation, and standardized concept mapping using OMOP vocabularies with custom concept extensions where necessary. ETL performance was evaluated through field-level mapping quality, information loss, and mapping accuracy, followed by modality-integrated CDM validation via analytic concordance with the original trial findings to assess structural validity and analytic scalability within the OMOP CDM framework⁵.

ETL Pipeline Architecture for Transforming A4/LEARN Preclinical Alzheimer's Disease Trial Data to OMOP CDM v5.4

Standardized data transformation pipeline converting clinical trial data from cognitively normal adults with preclinical AD markers into OHDSI OMOP Common Data Model format for federated research.

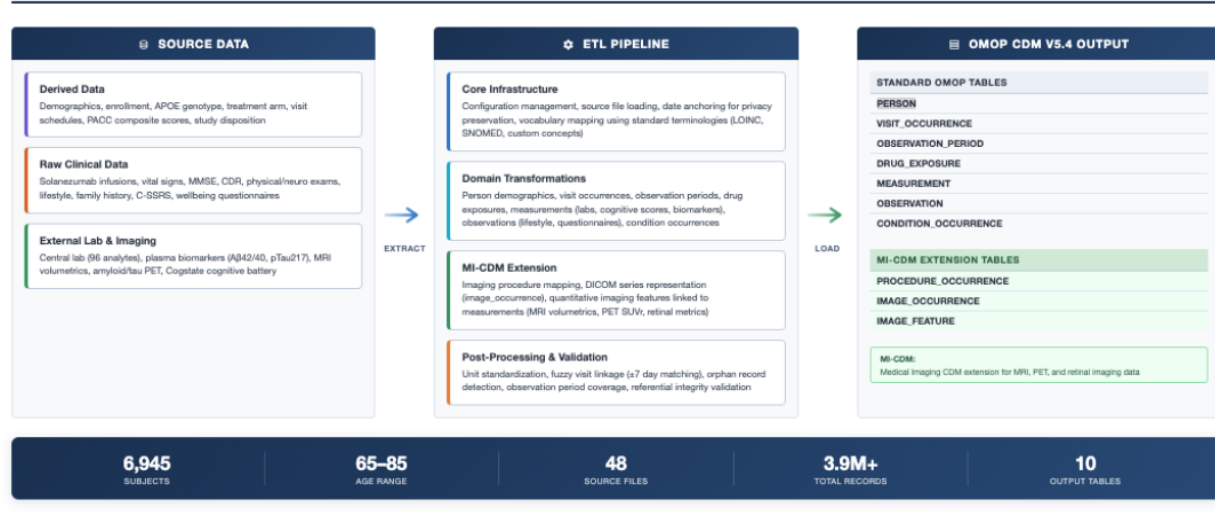


Figure 1. ETL Pipeline Architecture for Transforming Preclinical Alzheimer's Disease Trial Data to OMOP CDM

Result

After the ETL process, data from 6,945 patients across 48 source files were successfully transformed into 10 OMOP CDM tables, generating 3.9 million records (Table 1). The MEASUREMENT table contained 2,413,202 records covering key domains including vitals, laboratory tests, cognitive assessments, Alzheimer's disease biomarkers, neuroimaging measures, and computerized cognitive testing. The Medical Imaging (MI)-CDM extension integrated imaging acquisition metadata and derived imaging features. Unit concepts were mapped for 54.8% of measurements, with unmapped units primarily reflecting dimensionless scores and imaging-derived metrics. Analytic concordance with the original trial findings was observed, with P-tau217 discrimination (AUROC 0.88 vs. 0.87) and longitudinal progression estimates within $\pm 5\%$ of published values. Table-level record concordance and field-level mapping evaluation demonstrated full person-level capture and structural consistency with the source data.

Table 1. Summary of the Pre-Symptomatic Alzheimer's Trial Data in the OMOP CDM

OMOP Table	Records	Description	Key Data Domains
PERSON	6,945	Subject demographics	Gender, birth year, race, ethnicity
VISIT_OCCURRENCE	99,795	Clinical site visits	Screening, baseline, treatment, follow-up
OBSERVATION_PERIOD	6,945	Per-subject enrollment windows	Start date, end date (longitudinal span)
DRUG_EXPOSURE	74,777	Solanezumab infusion records	Dose levels, IV route
MEASUREMENT	2,413,202	Clinical measurements	Labs, vitals, ECG, cognitive tests, biomarkers
IMAGING FEATURE	675,690	Imaging findings with links to source images and related clinical records	Imaging finding, quantitative measurement, imaging biomarker
IMAGING OCCURRENCE	23,898	Imaging study/series metadata with links to patient and source	DICOM study/series, modality, anatomic site

		DICOM data	
PROCEDURE OCCURRENCE	20,307	Imaging procedures performed during the study	MRI scans, amyloid PET scans, tau PET scans, imaging procedure type
OBSERVATION	580,076	Non-numeric clinical data	Lifestyle, family history, C-SSRS, questionnaires
CONDITION OCCURRENCE	5,262	Abnormal exam findings	Physical/neurological abnormalities

Conclusion

Standardizing pre-symptomatic Alzheimer’s disease prevention trial data to the OMOP CDM demonstrated that complex multimodal clinical trial datasets, including cognitive assessments, plasma biomarkers, and neuroimaging, can be harmonized within a common observational data framework. The transformation achieved high structural completeness and analytic concordance with the original trial analyses, demonstrating that key biomarker discrimination and longitudinal disease progression signals were preserved after standardization. Integration of imaging metadata and derived quantitative imaging features through the MI-CDM extension provided a structured representation of MRI and PET biomarkers alongside clinical and laboratory measurements within the OMOP framework. The transformation process also identified mapping gaps for preclinical disease indicators, including dimensionless cognitive scores, biomarker thresholds, and imaging-derived quantitative metrics used to characterize risk status prior to clinical diagnosis. These findings underscore the importance of continued vocabulary development and validation within OMOP-standardized environments to ensure accurate representation of preclinical disease states and reproducible analyses. Such standardization may facilitate interoperable and scalable Alzheimer’s disease research across institutions, supporting broader evaluation of early risk and disease progression in real-world populations.

References

1. Sperling RA, Donohue MC, Rissman RA, et al. Amyloid and tau prediction of cognitive and functional decline in unimpaired older individuals: longitudinal data from the A4 and LEARN studies. *J Prev Alzheimers Dis.* 2024;11(4):802-813. doi:10.14283/jpad.2024.122.
2. López-Martos D, Suárez-Calvet M, Milà-Alomà M, et al. Longitudinal associations of multimodal core 1 Alzheimer disease biomarkers with cognition in aging and preclinical Alzheimer disease. *Neurology.* 2025;105(9):e214308. doi:10.1212/WNL.0000000000214308.
3. 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.
4. Park WY, Jeon K, Schmidt TS, Kondylakis H, Alkasab T, Dewey BE, You SC, Nagy P. Development of medical imaging data standardization for imaging-based observational research: OMOP common data model extension. *J Imaging Inform Med.* 2024;37(2):899-908. doi:10.1007/s10278-024-00982-6
5. Rissman RA, Donohue MC, Langford O, Raman R, Abdel-Latif S, Yaari R, Holdridge KC, Sims JR, Molina-Henry D, Jimenez-Maggiore G, Johnson KA, Aisen PS, Sperling RA. Longitudinal phospho-tau217 predicts amyloid positron emission tomography in asymptomatic Alzheimer's disease. *J Prev Alzheimers Dis.* 2024;11(4):823-830. doi:10.14283/jpad.2024.134.

Acknowledgments:

The authors gratefully acknowledge the Alzheimer’s Clinical Trial Consortium A4/LEARN as the source of the Data Package and the A4/LEARN funding sources, including National Institute on Aging/NIH, Eli

Lilly and Co., Alzheimer's Association, and others. The authors gratefully acknowledge the support of Gates Ventures for funding and guidance related to this project.