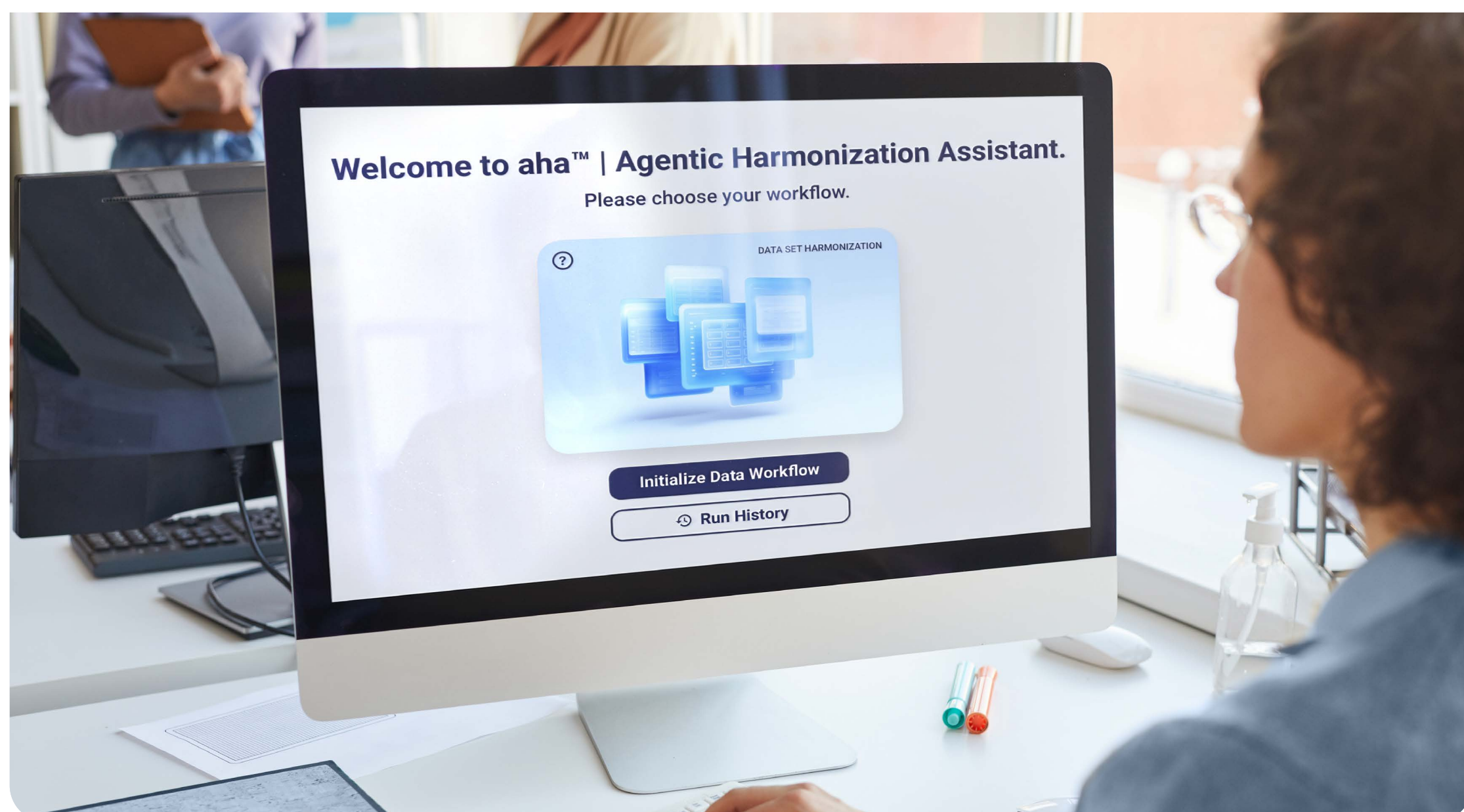


BACKGROUND

The Alzheimer's Disease (AD) research ecosystem contains a rapidly growing collection of clinical, imaging, sociodemographic, and molecular datasets. Yet inconsistencies in schemas, variable encodings, and batch effects require extensive preparation for cross-cohort analyses and constrain their utility for building large-scale trustworthy AI models¹. To address this bottleneck, we introduce the **Agentic Harmonization Assistant (aha™)**, a multi-agent system that reasons, plans, executes, and validates end-to-end dataset harmonization.



METHODS

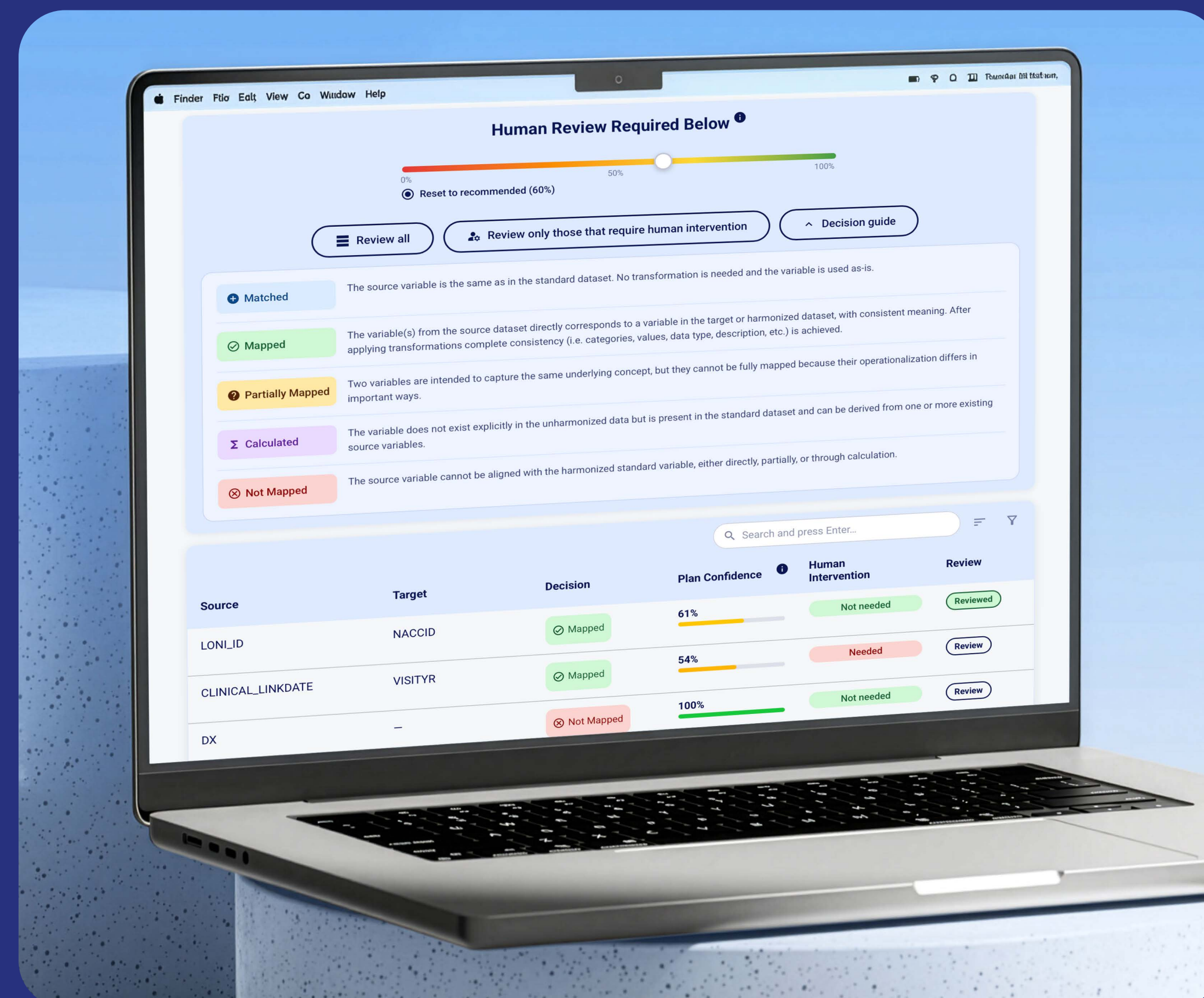
aha™ is powered by multiple large language models operating through in-context learning. It is a four-stage system (**Figure 1**) that begins with a profiling module, which characterizes both the standard and unharmonized datasets using their data dictionaries to generate machine-readable variable profiles that combine descriptive statistics with semantic information. These profiles are then used by the strategy and planning module to create a harmonization plan for each variable, including confidence scores and required transformations, which are refined through an iterative process. The execution module implements the plan. Finally, the validation module evaluates the harmonized output against the standard dataset and generates a consistency analysis.



Figure 1. Multi-agent system workflow for data harmonization.

RESULTS

aha™ reduces harmonization time by 93%, decreasing the required effort from 28 hours of manual work to 2.5 hours for 100 variables. We evaluated **aha™** performance per module.



PROFILING:

We assessed the accuracy of the data extraction performed by the profiling.

We evaluate the correctness of the data extraction on three different domains:

- Data Type: Accuracy, Precision, Recall, and F1-Score measuring the correct selection of the data type between the possible options (float, int, binary, categorical_nominal, categorical_ordinal, text, or date).
- Variable description: Cosine Similarity between the generated description and the ground-truth on the data dictionary.
- Ranges or categories: IoU between the values extracted by the system and the ground-truth on the data dictionary

STRATEGY & PLANNING:

We evaluated the correctness of the selected harmonization plan on 4 items.

- Variable Alignment: F1 score measuring whether predicted variables were correctly matched to ground-truth variables.
- Decision Match: F1 score for selecting the correct harmonization decision (matched, mapped, partially mapped, calculated, or not mapped).
- Transformation Selection: F1 score evaluating whether the correct set of transformation operations (e.g., map, rename, case_when) was selected for each variable.
- Transformation Correctness: Mean accuracy of each operation's parameters (e.g., correct value mappings, rename targets, conditions), scored in [0, 1] per operation across aligned pairs.

EXECUTION:

We examined the proper execution of each transformation plan using RMSE for numerical variables and accuracy for categorical variables.

VALIDATION:

We analyzed the reliability of the reasoning used to approve or reject a harmonized variable. The module evaluates the correctness of the harmonization over five items:

- Concept alignment
- Data type compatibility
- Value/category coverage
- Missing and placeholders
- Stats sanity check.

Additionally, we created a subset of curated correct and incorrect harmonizations, along with manually generated rationales explaining why each variable in this subset was harmonized correctly or not. With this subset and the Prometheus 2 evaluator², we measured the performance of our model in producing accurate reasoning on the five-item rubric and alignment between the rubric and the final decision for each variable (harmonized/not-harmonized). The final score used a penalized factor for lower scores.



CONCLUSIONS

Our interim findings confirm that **aha™** is a powerful and scalable solution for harmonizing AD data across global cohorts. We continue to enhance model performance, enabling accurate outputs that support research and clinical applications. **aha™** will continue to improve through ongoing fine-tuning and integration of additional modalities such as neuroimaging and omics harmonization, positioning it as a foundational platform for AI-driven AD research and precision prevention strategies.

DISCLOSURE OF FUNDING

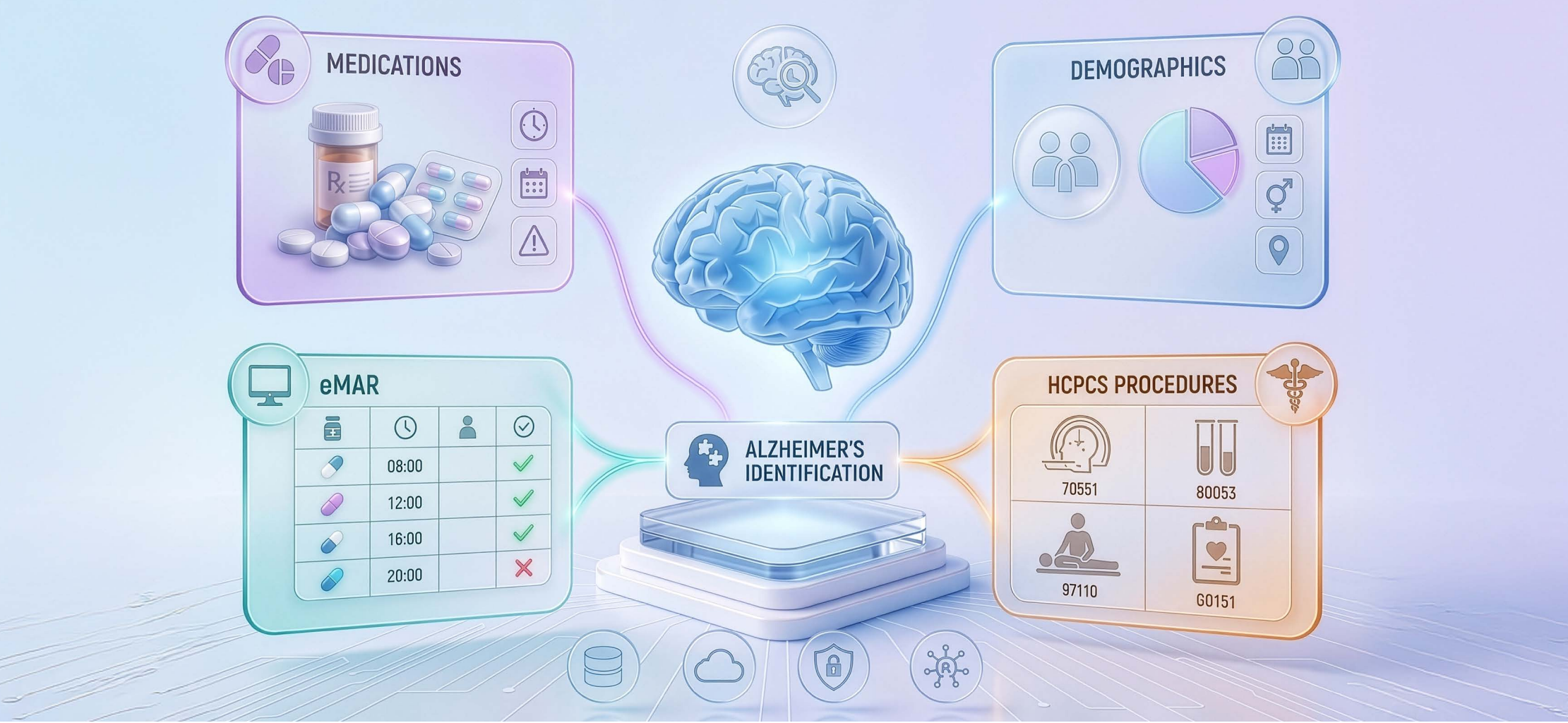
This research was funded by IGC Pharma, LLC.

REFERENCES

1. Hao, X., Li, X., Zhang, GQ. et al. An ontology-based approach for harmonization and cross-cohort query of Alzheimer's disease data resources. BMC Med Inform Decis Mak 23 (Suppl 1), 151 (2023). <https://doi.org/10.1186/s12911-023-02250-z>
2. S. Kim et al. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. Accessed: Jun. 22, 2026. [Online]. Available: <https://arxiv.org/abs/2405.01535>



BACKGROUND



Identifying Alzheimer’s disease and related dementias (AD/ADRD) in real-world settings is often limited by the lack of formal neurocognitive assessments^{1, 2}. Inpatient electronic health records (EHRs) provide a scalable yet underutilized resource for detecting cognitive impairment through hospitalization patterns, medication use, and procedural history^{3, 4}.

EHR Modalities Used:
Patient-level features were extracted from inpatient hospitalizations in the MIMIC-IV v3.1 database⁵.

Medications

Demographics

eMAR

HCPCS Procedures

RESULTS

Integrated models combining all EHR modalities outperformed single-modality models and improved discrimination across cognitive status and dementia subtype classification. The most salient features and their relative importance are summarized in **Figure 1**.

1. **Demographics:** Demographic features, particularly age, helped distinguish patients with cognitive impairment from controls¹¹ but showed limited ability to differentiate dementia subtypes.
2. **eMAR:** Electronic medication administration records captured dynamic inpatient care patterns, including temporal irregularities in medication administration, medication burden, and pharmacy-related features that helped distinguish MCI from dementia¹².
3. **Medications:** Prescription medication features showed a distributed importance profile across therapeutic class, route of administration, dosing frequency, and medication status, with notable relevance among neurological pharmacy categories^{13, 14}.
4. **HCPCS Procedures:** HCPCS procedural codes reflected care intensity and system-level involvement, with cardiovascular, digestive, and hospital observation services serving as meaningful markers across all classification levels¹⁵.
5. **Multimodal Integration:** Integrated multimodal approaches consistently outperformed classifiers based on any single modality, demonstrating superior predictive performance across all levels of the hierarchical classification framework^{16, 17}.

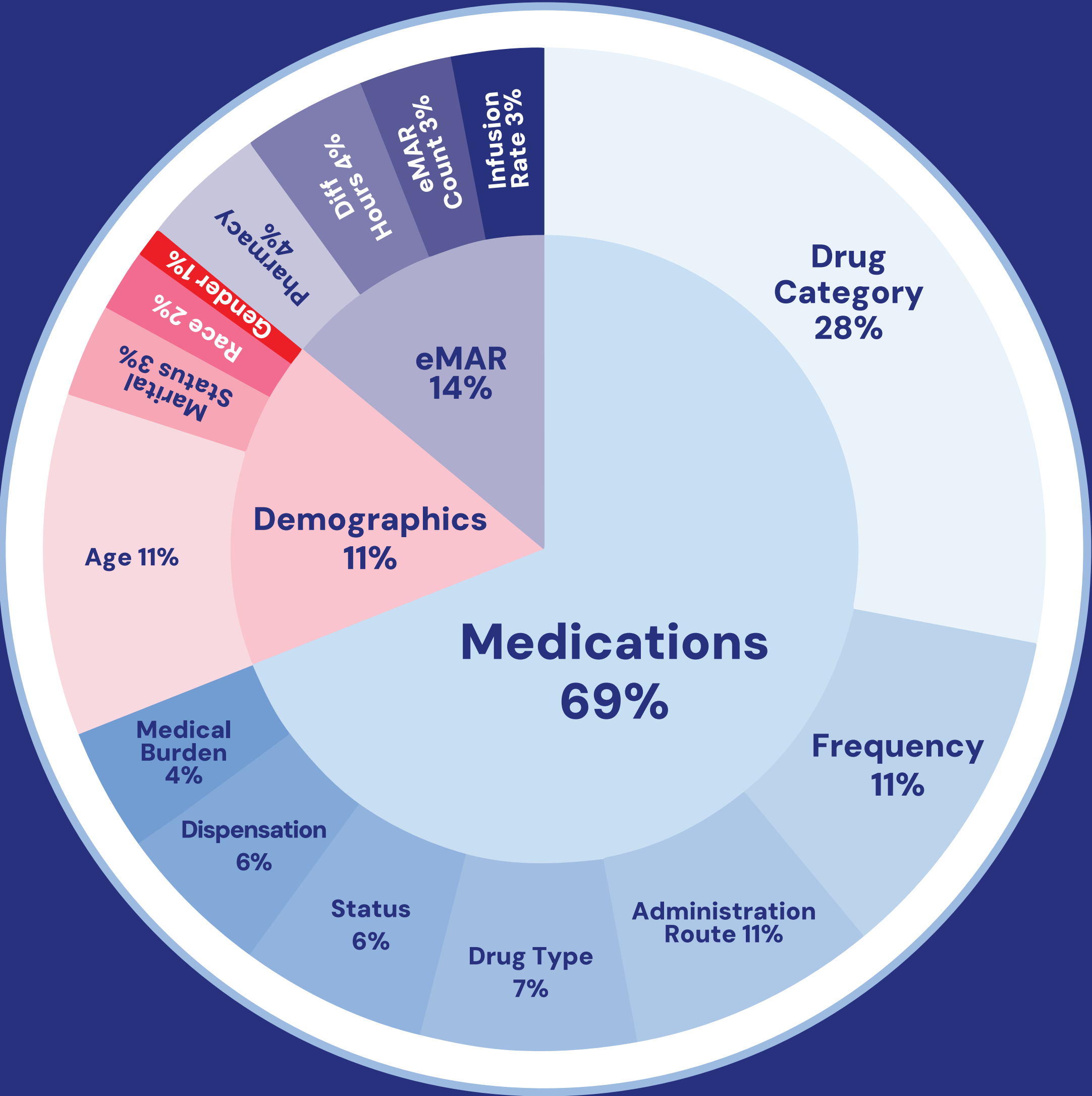


Figure 1. Feature importance per modality for the dementia subtypes classification

METHODS

1. **Feature extraction** from four EHR modalities at the patient level using ICD hospitalization codes^{6, 7}.
2. **Hierarchical classification framework** (see Table 1): controls versus cognitively impaired patients, followed by MCI versus dementia, and finally classification of dementia subtypes^{8, 9}.
3. **Random forest classifiers** trained per modality and as an integrated multimodal model at each classification level¹⁰.
4. **Feature importance analysis** to identify clinically interpretable signatures across modalities.

HIERARCHY STAGE	TASK	CLASSES	DEFINITION
General screening	Binary cognitive impairment detection	Normal cognition vs Cognitive impairment	Identifies the presence of any ICD-coded cognitive impairment
Intermediate stratification	Binary MCI vs dementia classification	Mild cognitive impairment vs Dementia	Differentiates mild cognitive impairment from established dementia
Granular classification	Dementia subtype classification	Alzheimer's disease, Lewy body or Parkinson-related dementia, Vascular dementia, Frontotemporal dementia, substance- or exposure-related dementia, other or unspecified dementias	Classifies dementia into different subtypes

Table 1. Hierarchical classification framework

CONCLUSIONS

Machine learning models using routinely collected inpatient EHR data can identify interpretable, modality-specific signatures of AD/ADRD. Integrating complementary EHR modalities offers a scalable approach for Alzheimer’s dementia screening and dementia subtype differentiation in real-world settings without specialized cognitive testing.

DISCLOSURE OF FUNDING

This research was funded by IGC Pharma, LLC.

REFERENCES

1. V. G. Valcour, K. H. Masaki, J. D. Curb, and P. L. Blanchette, "The detection of dementia in the primary care setting," Archives of Internal Medicine, vol. 160, no. 19, pp. 2964–2968, 2000, doi: 10.1001/archinte.160.19.2964.

2. S. Irwin, E. Reddy, M. Doehler, C. C. Lima, and J. H. T. Bates, "Correlates of missed or late versus timely diagnosis of dementia in healthcare settings," Alzheimer's & Dementia, vol. 21, no. 1, e14494, 2025, doi: 10.1002/alz.14494.

3. Q. Li, X. Yang, J. Xu, C. Guo, H. He, J. Bian, Y. Luo, and X. Du, "Early prediction of Alzheimer's disease and related dementias using real-world electronic health records," Alzheimer's & Dementia, vol. 19, no. 8, pp. 3506–3518, Aug. 2023, doi: 10.1002/alz.12967.

4. R. Wei, S. S. Buss, R. Milde, M. Fernandes, D. Sumsion, E. Davis, and M. B. Westover, "Automated phenotyping of mild cognitive impairment and Alzheimer's disease and related dementias using electronic health records," Journal of Biomedical Informatics, vol. 164, p. 104813, Apr. 2025, doi: 10.1016/j.jbi.2025.104813.

5. A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, B. Moody, B. Gow, L.-H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," Scientific Data, vol. 10, no. 1, p. 1, Jan. 2023, doi: 10.1038/s41597-022-01899-x.

6. S. S. Knox, T. Bhatt, T. Bhatt, M. C. Reid, and C. A. Metti, "AI approaches for phenotyping Alzheimer's disease and related dementias using electronic health records," Alzheimer's & Dementia: Translational Research & Clinical Interventions, vol. 11, no. 2, e70089, Apr. 2025, doi: 10.1002/trc2.70089.

7. A. M. Raskin, K. M. Beaulieu-Jones, A. B. Walling, C. Ritchie, L. E. Donahue, and R. A. Lasky-Su, "An algorithm to characterize a dementia population by disease subtype," Alzheimer Disease & Associated Disorders, vol. 33, no. 4, pp. 316–322, 2019, doi: 10.1097/WAD.0000000000000336.

8. R. O. Roberts and D. S. Knopman, "Classification and epidemiology of MCI," Clinics in Geriatric Medicine, vol. 29, no. 4, pp. 753–772, 2013, doi: 10.1016/j.cger.2013.07.003.

9. M. Zhang, K. Lofteröd, M. Boland, and A. Fagerbakke, "Cognitive performance classification of older patients using machine learning and electronic medical records," Scientific Reports, vol. 15, p. 6812, Feb. 2025, doi: 10.1038/s41598-025-90460-y.

10. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

SIGNIFICANCE
AND FUTURE DIRECTIONS.

Routine EHR data provides an accessible and cost-effective pathway for detecting early cognitive impairment at scale across hospital systems. Future work will include additional modalities and validate this framework in external EHR cohorts and assess its integration into real-time clinical decision support for AD/ADRD screening and dementia subtype differentiation.

11. P. Solfrizzi, V. Panza, and F. Frisardi, "Epidemiologic studies of modifiable factors associated with cognition and dementia: systematic review and meta-analysis," BMC Public Health, vol. 14, p. 643, 2014, doi: 10.1186/1471-2458-14-643.

12. J. Stuhler, M. Baum, J. Mayer, and D. Mengel, "Differentiation between early and severe stages of dementia using diagnosis, prescription and utilization patterns," Alzheimer's & Dementia, vol. 21, p. e14524, 2025, doi: 10.1002/alz.14524.

13. J. Farber, D. Levine, and C. Sinsky, "The prescription drug landscape of Alzheimer's disease and related dementias," Innovation in Aging, vol. 6, suppl. 1, p. 388, 2022, doi: 10.1093/geroni/igac059.1543.

14. S. R. Rao, V. Bhatt, and N. Shetty, "Drug utilization pattern in patients with different types of dementia in Western India," International Journal of Alzheimer's Disease, vol. 2014, p. 435202, 2014, doi: 10.1155/2014/435242.

15. M. Zhang, K. Lofteröd, M. Boland, and A. Fagerbakke, "Cognitive performance classification of older patients using machine learning and electronic medical records," Scientific Reports, vol. 15, p. 6812, Feb. 2025, doi: 10.1038/s41598-025-90460-y.

16. S. Alam, G. M. M. Alam, and M. Murad, "Multimodal deep learning models for early detection of Alzheimer's disease stage," Scientific Reports, vol. 11, p. 3254, Feb. 2021, doi: 10.1038/s41598-020-74399-w.

17. A. Kline, M. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo, "Multimodal machine learning in precision health: A scoping review," npj Digital Medicine, vol. 5, p. 171, Nov. 2022, doi: 10.1038/s41746-022-00712-8.

18. A. Z. Kraal, N. Arvanitis, K. A. Hill, and B. Ahluwalia, "Deep learning prediction of mild cognitive impairment using electronic health records," BMC Medical Informatics and Decision Making, vol. 20, suppl. 11, p. 285, Nov. 2020, doi: 10.1186/s12911-020-01297-8.



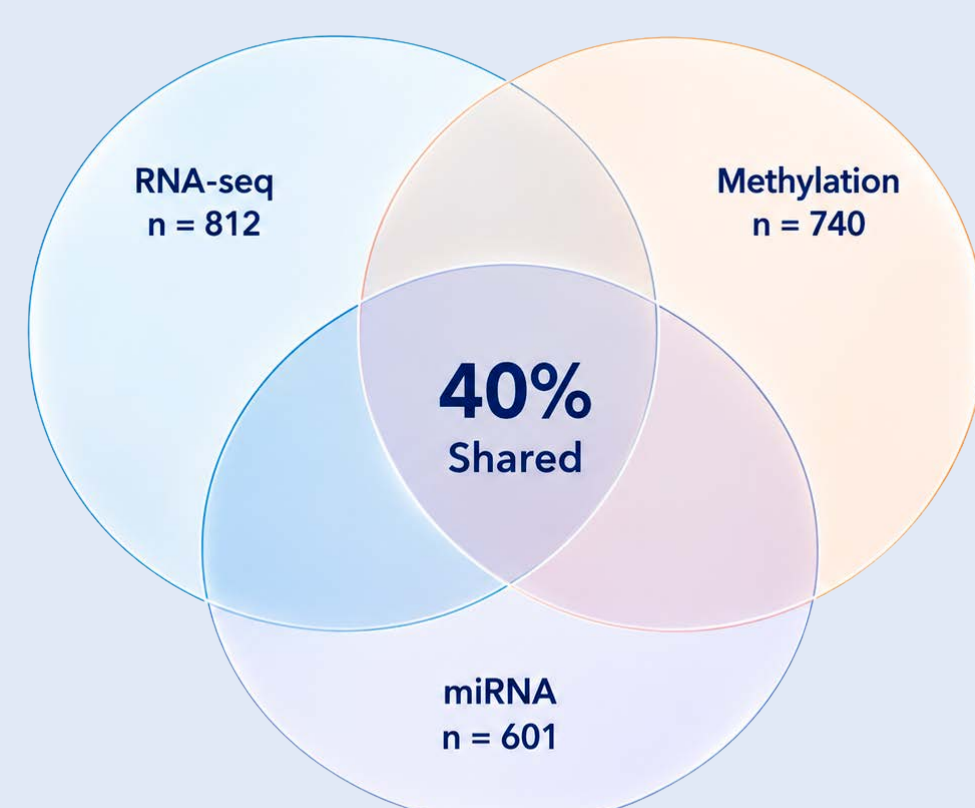
BACKGROUND

Integrative analysis of multi-omics data is critical for understanding the heterogeneous pathophysiology of Alzheimer's Disease (AD). While recent frameworks have addressed inference on incomplete data via contrastive alignment¹, current methods often treat patients as independent instances, neglecting the topological similarity between subjects. Furthermore, reliance on the strict intersection of samples across modalities limits feature learning robustness².

METHODS

We aimed to overcome these limitations by developing an asymmetrical, topology-aware framework that utilizes the "union" of available data to generate hypothesis-driven biomarkers from fragmented clinical cohorts. We utilized the ROSMAP3 aging population post-mortem brain dataset, where the intersection of samples represented only ~40% of the total available data (Figure 1). We designed a two-stage decoupled learning pipeline (Figure 2). First, we pre-trained modality-specific Graph Convolutional Network encoders on the maximal available cohort for each omic type. To capture patient topology, we constructed dynamic K-Nearest Neighbor graphs using cosine similarity for each batch. Second, we froze these encoders and trained a Graphformer4 fusion classifier on the common patient subset. Biological relevance was derived using Integrated Gradients⁵ on attention layers.

COHORT FRAGMENTATION — ROSMAP



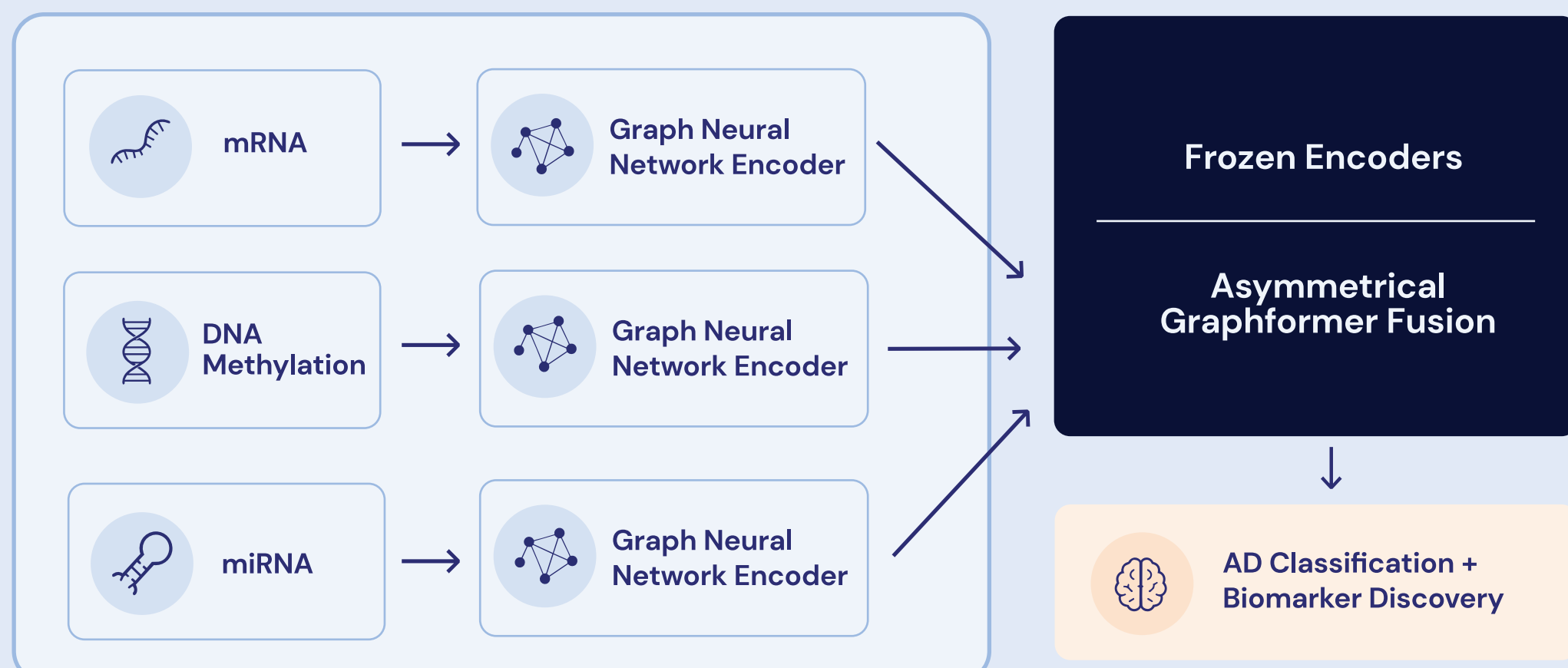
Intersection-based integration discards 60% of samples, severely limiting statistical power for discovery

Figure 1. Structural equation model illustrating multifactorial contributors to cognitive impairment. Psychosis, depression, head trauma, and education were modeled as latent predictors of cognitive impairment using CDR sum of boxes and MoCA as indicators. Standardized coefficients and p-values are shown.

TOPOLOGY-AWARE ASYMMETRICAL GRAPHFORMER FRAMEWORK

Stage 1 – Modality-Specific Graph Pretraining (Union Cohorts)

Stage 2 – Shared Subset Fusion



Interpretability Pipeline:
Graph Attention Weights → Integrated Gradients → Feature Attribution → Pathway Enrichment → Mechanistic Hypotheses

Figure 2. Schematic of the multi-stage graph-based multi-omic fusion model, detailing the pretraining of modality-specific graph neural networks and subsequent joint model fusion for AD classification and biomarker discovery, including the connected mechanistic interpretability pipeline.

RESULTS

The asymmetrical Graphformer achieved an F1 score of 0.922 and accuracy of 0.920 (Figure 3). Interpretation of the top-ranked features revealed a pattern consistent with a "Broken Highway" hypothesis⁶, prioritizing a) the downregulation of kinesin motors (KIF5A, KIF1C)⁷, b) methylation sites linked to axonal regulators (OLFML3/cg17253459), and c) dysregulation of miRNAs (miR-10788, miR-29a9) known to regulate BACE1. These signals were distinct from metabolic stress indicators (e.g. DDIT410), suggesting a multi-stage cascade involving early transport and regulatory failure preceding impaired beta-amyloid clearance (Figure 4).

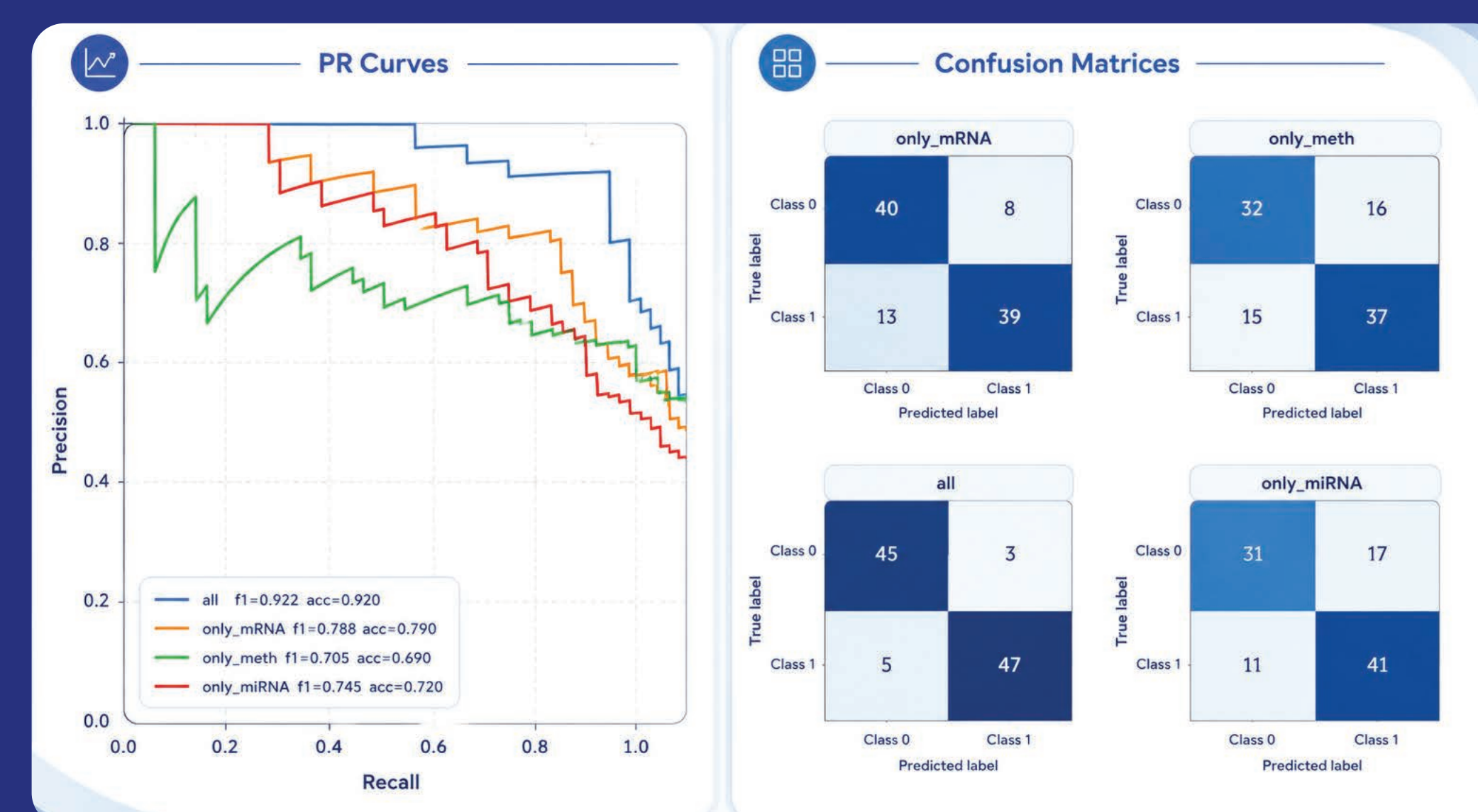


Figure 3. Performance evaluation of single-modality versus multi-omic fusion model for Alzheimer's disease binary classification. (Left) Precision-Recall (PR) curves demonstrate the superior predictive capability of the integrated model with an F1-score of 0.922 and accuracy of 0.920, significantly outperforming isolated modalities (mRNA, DNA methylation, and miRNA). (Right) Corresponding confusion matrices detail the distribution of true and false predictions across classes (0: Normal cognition, 1: Alzheimer's Disease) for each model, highlighting the increased accuracy and reduced misclassification achieved through multi-omic integration.

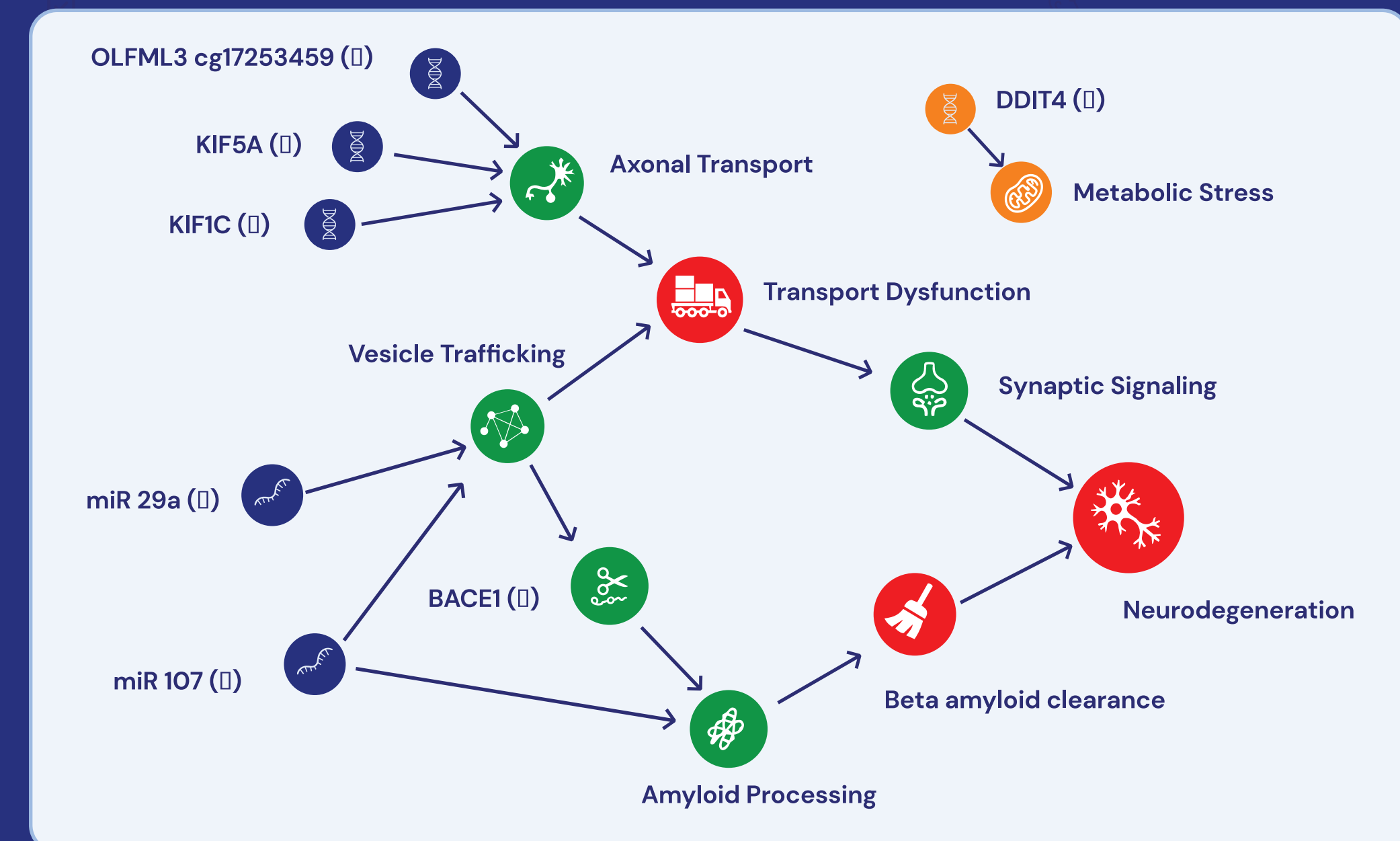
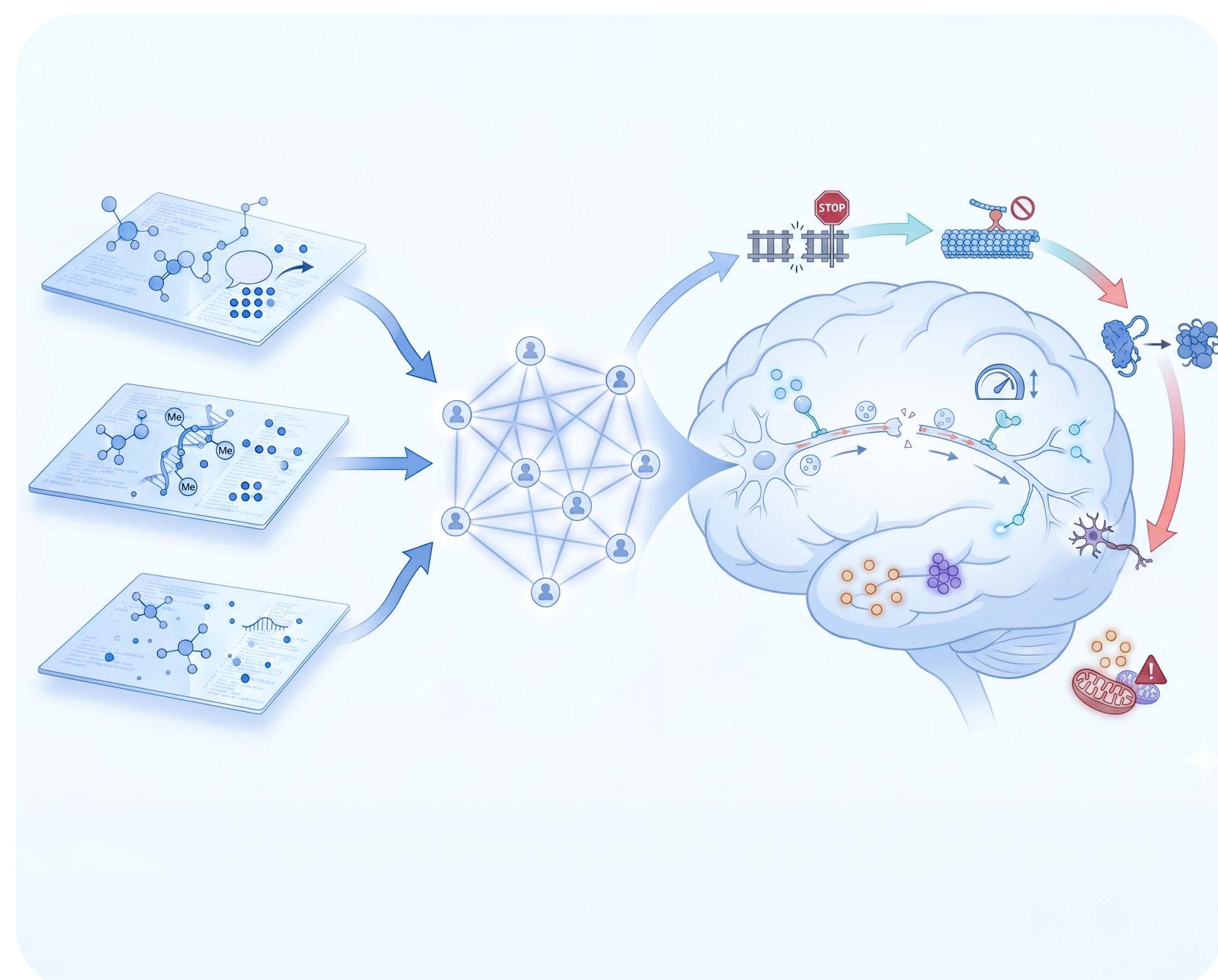


Figure 4. Performance evaluation of single-modality versus multi-omic fusion model for Alzheimer's disease binary classification. (Left) Precision-Recall (PR) curves demonstrate the superior predictive capability of the integrated model with an F1-score of 0.922 and accuracy of 0.920, significantly outperforming isolated modalities (mRNA, DNA methylation, and miRNA). (Right) Corresponding confusion matrices detail the distribution of true and false predictions across classes (0: Normal cognition, 1: Alzheimer's Disease) for each model, highlighting the increased accuracy and reduced misclassification achieved through multi-omic integration.



CONCLUSION

Our approach demonstrates that decoupling patient topology from multimodal fusion enables robust biomarker discovery and improves diagnostic precision in incomplete multi-omics datasets, a common limitation in AD research. By integrating graph-based representations with attention-guided fusion, the model prioritizes biologically plausible markers across molecular layers, including kinesin motors, miRNAs regulating BACE1, and methylation sites, offering a scalable, hypothesis-generating tool to probe mechanistic cascades. Notably, the model suggests a potential sequence in AD pathology, early cytoskeletal transport dysfunction followed by downstream metabolic failure in AD.



DISCLOSURE OF FUNDING

This research was funded by IGC Pharma, LLC.

REFERENCES

- Zhao, C., Liu, A., Zhang, X., Cao, X., Ding, Z., Sha, Q., Shen, H., Deng, H.-W., & Zhou, W. (2023). CLCLSA: Cross-omics Linked embedding with Contrastive Learning and Self Attention for multi-omics integration with incomplete multi-omics data. *arXiv*. <https://doi.org/10.48550/arxiv.2304.05542>
- Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z. and Huang, K. (2021) MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* 12:3445. doi: 10.1038/s41467-021-23774-w
- Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA (2018) Religious Orders Study and Rush Memory and Aging Project. *J Alzheimers Dis.* 64(s1):S161-S189. doi: 10.3233/JAD-179939
- Ying, C., Cai, T., Luo, S., et al. (2021). Do Transformers Really Perform Bad for Graph Representation? *NeurIPS* 2021. <https://doi.org/10.48550/arxiv.2106.05234>
- Sundararajan M, Taly A and Yan Q (2017) Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)* 70:3319–3328. doi: 10.48550/arxiv.1703.01365
- Millecamps S and Julien JP (2013) Axonal transport deficits and neurodegenerative diseases. *Nat. Rev. Neurosci.* 14:161–176. doi: 10.1038/nrn3380
- Wang Q, Li L and Ye Y (2019) Kinesin-1 in neurodegenerative diseases. *Cell. Mol. Life Sci.* 76:2853–2867. doi: 10.1007/s00018-019-03146-5
- Wang WX, Rajeev BW, Stromberg AJ, Ren N, Tang G, Huang Q, Rigoutsos I and Nelson PT (2008) Decreased expression of microRNAs in human Alzheimer's disease. *PLoS One.* 3:e1466. doi: 10.1371/journal.pone.0001466
- Hébert SS, Horré K, Nicolai L, Papadopoulou AS, Mandemakers W, Silahatoglu AN, Kauppinen S, Delacourte A and De Strooper B (2008) Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression. *Proc. Natl. Acad. Sci. U.S.A.* 105:6415–6420. doi: 10.1073/pnas.071026310
- Pérez-Ortiz JM, García-Gutiérrez MS, Navarrete M, Ginés S and Manzanares J (2024) Astrocyte protein RTP801 may contribute to cognitive decline in Alzheimer's disease. *Glia.* 72:1234–1245. doi: 10.1002/glia.24500

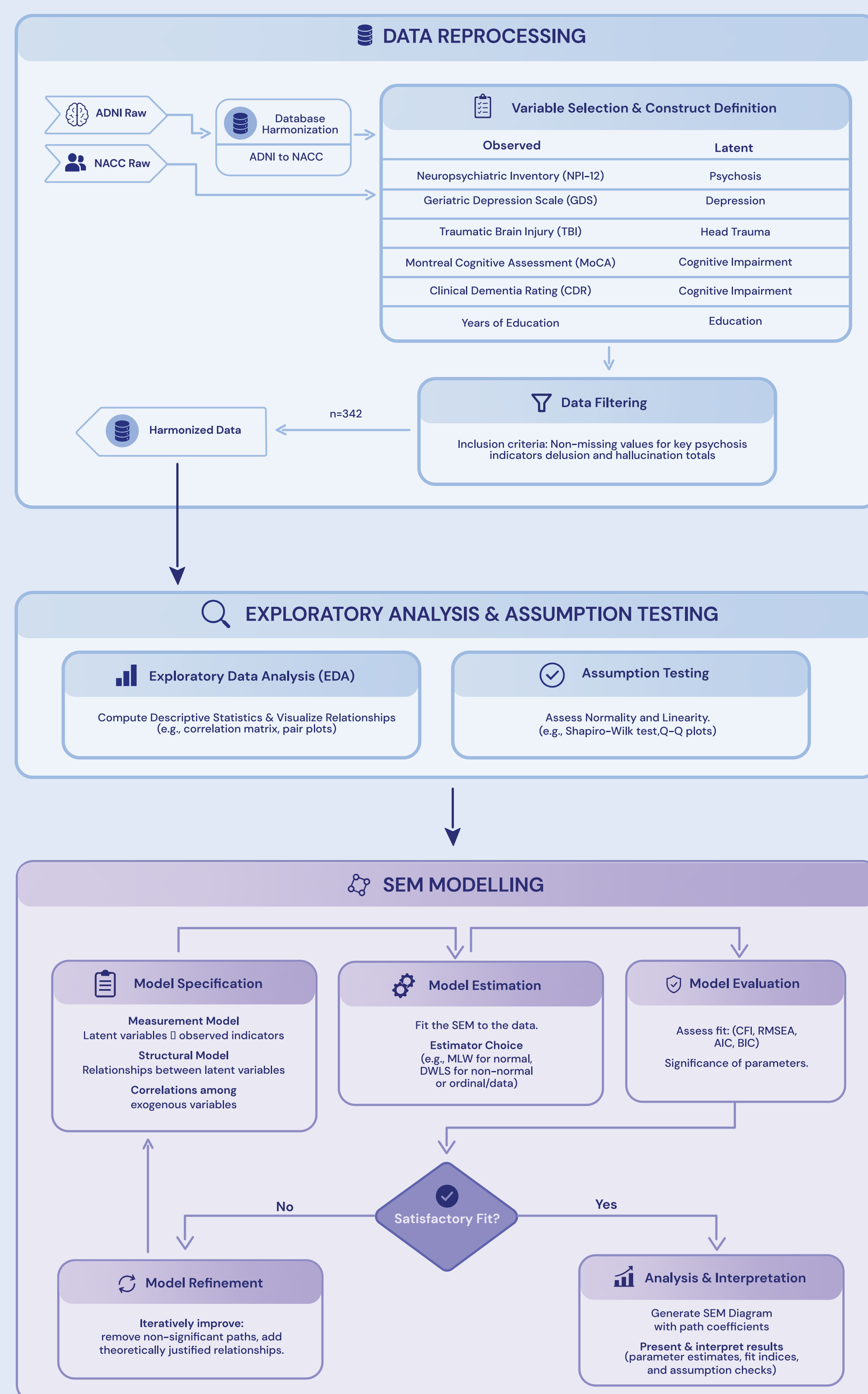


SCAN ME

BACKGROUND

- Cognitive impairment in older adults arises from complex interactions among psychiatric symptoms, life experiences, and sociodemographic factors, yet these domains are often examined in isolation using univariate approaches.
- Structural equation modeling (SEM) has been applied in Alzheimer's disease research mostly to specific domains such as cerebrovascular disease or biomarkers¹, but rarely to model multiple clinical and sociodemographic factors within a single framework.
- This study applies SEM to examine the relative and correlated contributions of psychiatric symptoms, head trauma, and education to cognitive impairment, using data from ADNI and NACC.

METHODS



Participants and Data:

- n=342 participants drawn from clinical-based studies in ADNI and NACC.
- Inclusion criteria required non-missing values for delusion and hallucination totals; other clinical variables allowed missingness.

Constructs:

- Psychosis**: sum of NPI-12 delusion and hallucination items
- Depression**: Geriatric Depression Scale (GDS)
- Head trauma**: standardized NACC TBI indicators
- Education**: years of formal schooling
- Cognitive impairment**: CDR sum of boxes and MoCA

Analysis:

- SEM implemented in Python's Semopy, allowing parameter estimation with partially observed data without listwise deletion.

RESULTS

- Psychosis ($\beta=0.83$, $p<0.001$), depression ($\beta=1.19$, $p<0.001$), and head trauma ($\beta=0.68$, $p<0.001$) showed strong positive associations with cognitive impairment.
- Education showed a significant negative association ($\beta=-0.57$, $p<0.001$).
- Psychosis and depression were positively correlated ($r=0.54$, $p<0.01$)
- All factor loadings were significant ($p<0.01$) except the covariance between psychosis and education ($p=0.307$).

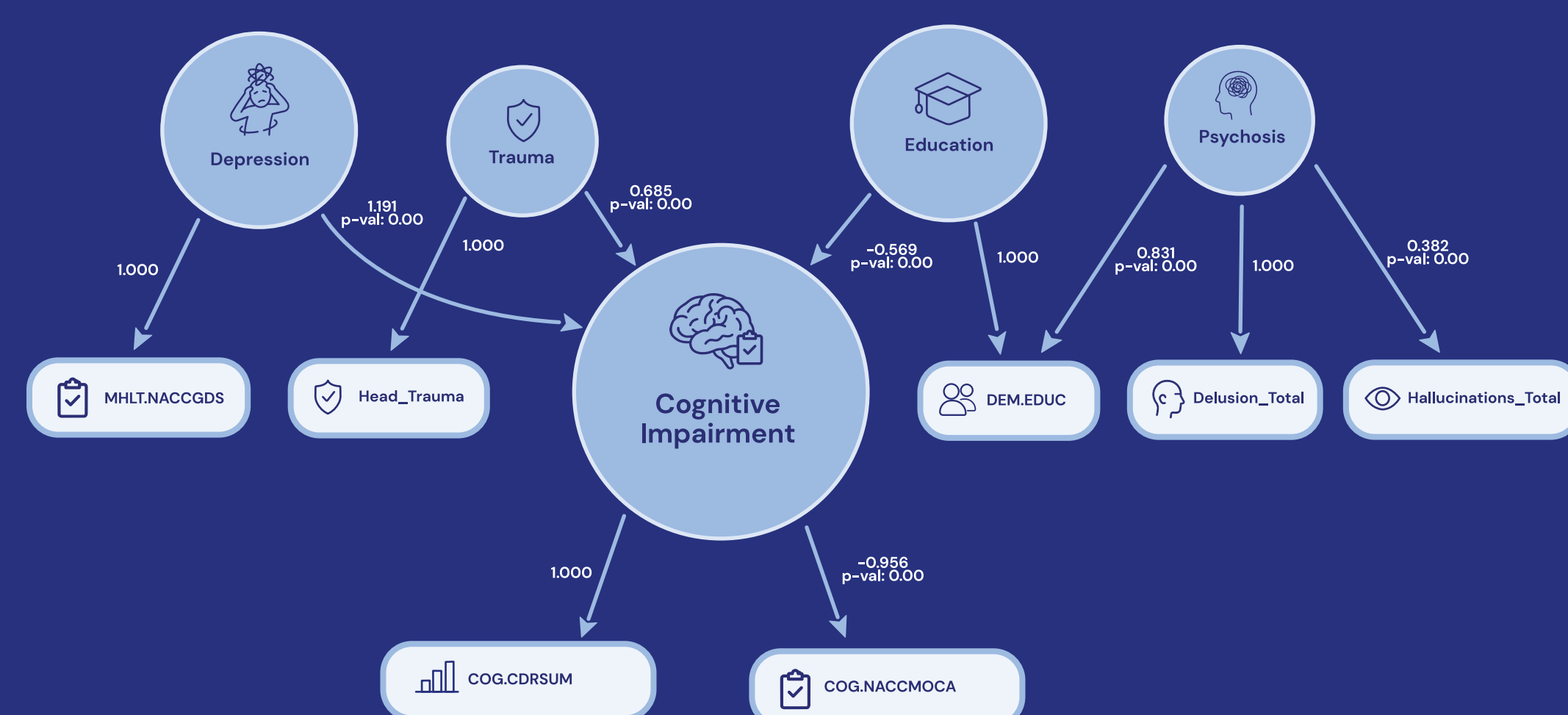


Figure 1. Structural equation model illustrating multifactorial contributors to cognitive impairment. Psychosis, depression, head trauma, and education were modeled as latent predictors of cognitive impairment using CDR sum of boxes and MoCA as indicators. Standardized coefficients and p-values are shown.



DISCUSSION

- The protective effect of education supports the cognitive reserve hypothesis. Higher educational attainment may delay the clinical onset of neurodegeneration without altering the underlying pathology.
- The association between head trauma and cognitive impairment is consistent with evidence that TBI accelerates neurodegenerative processes.
- The moderate psychosis–depression correlation likely reflects shared neurobiological mechanisms and frequent co-occurrence in older adults; a pattern that single-domain analyses tend to overlook.
- Unlike prior SEM studies in AD that focused on cerebrovascular mediators¹ or quality-of-life response shifts², this framework integrates psychiatric, traumatic, and sociodemographic domains within a single analysis.
- The non-significant psychosis–education covariance suggests these factors contribute through largely independent pathways.

CONCLUSION

Overall, this SEM analysis supports a multifactorial model of cognitive impairment in which psychosis, depression, and head trauma act as risk factors, while education shows a protective association. These findings are consistent with a multifactorial etiology of Alzheimer's disease and highlight the utility of SEM for uncovering complex pathways.



DISCLOSURE OF FUNDING

This research was funded by IGC Pharma, LLC.

REFERENCES

- Tang C, Ding Y, Yang J, He D. Cerebrovascular disease in Alzheimer's disease: Brain structure as a critical mediator of cognitive decline. J Prev Alzheimers Dis. 2025 Sep;12(8):100209. doi: 10.1016/j.tpad.2025.100209. Epub 2025 May 29. PMID: 40447488; PMCID: PMC12413723.
- Wang X, Xu X, Han H, et al. Using structural equation modeling to detect response shift in quality of life in patients with Alzheimer's disease. International Psychogeriatrics. 2019;31(1):123–132. doi:10.1017/S1041610218000595



BACKGROUND

Neuropsychiatric symptoms (NPS) affect most individuals with Alzheimer's disease (AD) and are associated with accelerated disease progression, increased caregiver burden, institutionalization, and higher mortality.¹ Despite their clinical importance, NPS are highly heterogeneous and often overlap across symptom domains, limiting the ability of traditional clustering methods to identify meaningful behavioral phenotypes.⁵ Topological Data Analysis (TDA) is a machine learning framework capable of identifying intrinsic structures in complex, nonlinear datasets without assuming predefined geometric relationships.⁴ In this study, we applied TDA to Neuropsychiatric Inventory (NPI)² data from the Alzheimer's Disease Neuroimaging Initiative (ADNI)³ to identify data-driven neuropsychiatric pathways in symptomatic AD patients.

METHODS



ADNI Cohort

Study Population

- Primary etiological AD diagnosis
- NPI data available for ≥ 10 of 12 domains
- Most recent visit included



NPI Assessments
(12 Domains)

Data Preparation

- NPI scores z-score standardized
- Cutoff: total NPI score 24
- 402 symptomatic AD patients
- Missing scores imputed (domain medians)



Topological Data Analysis

- Persistent homology (Ripser)
- Mapper algorithm (KeplerMapper)
- UMAP projection

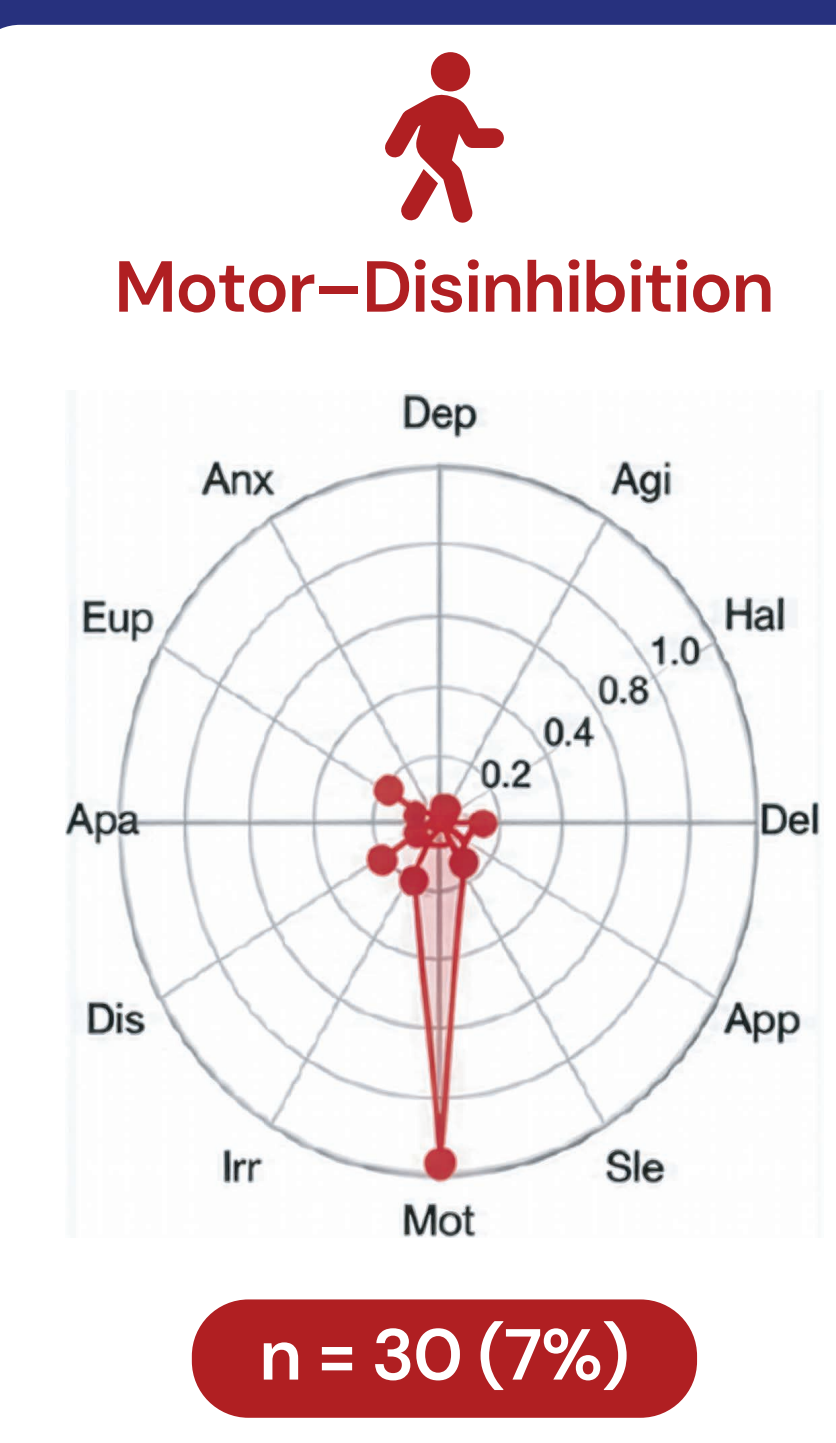
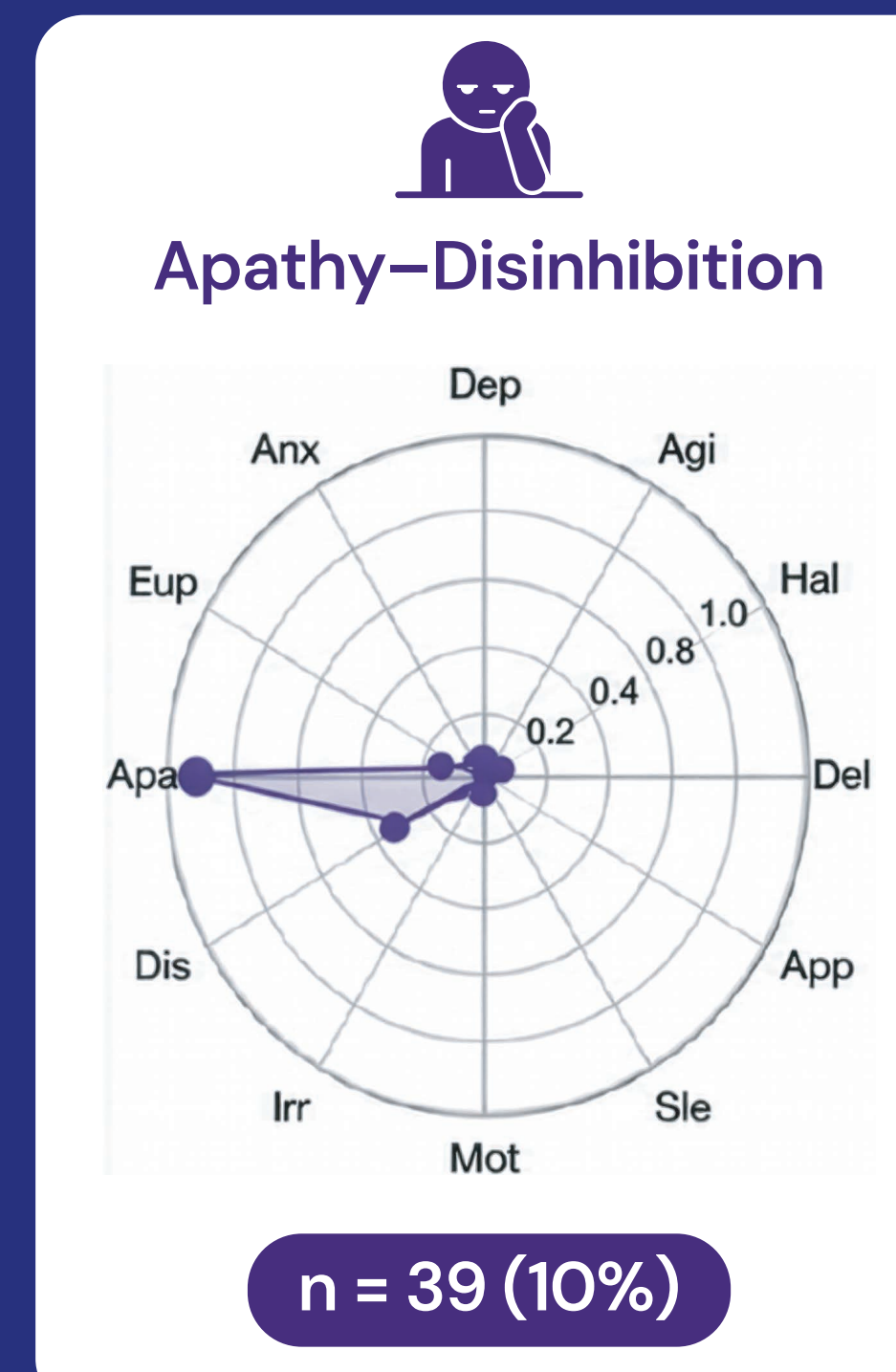
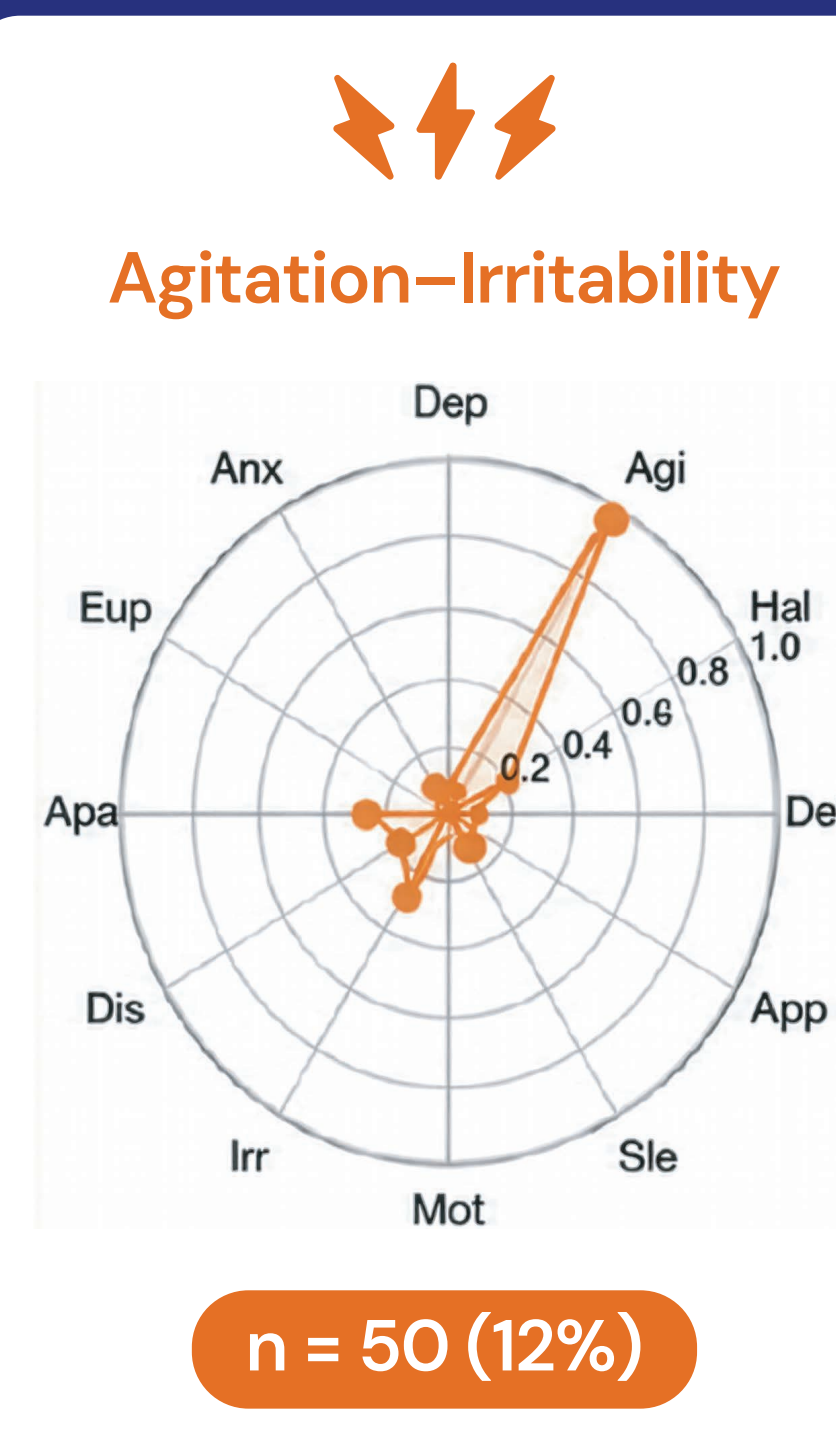
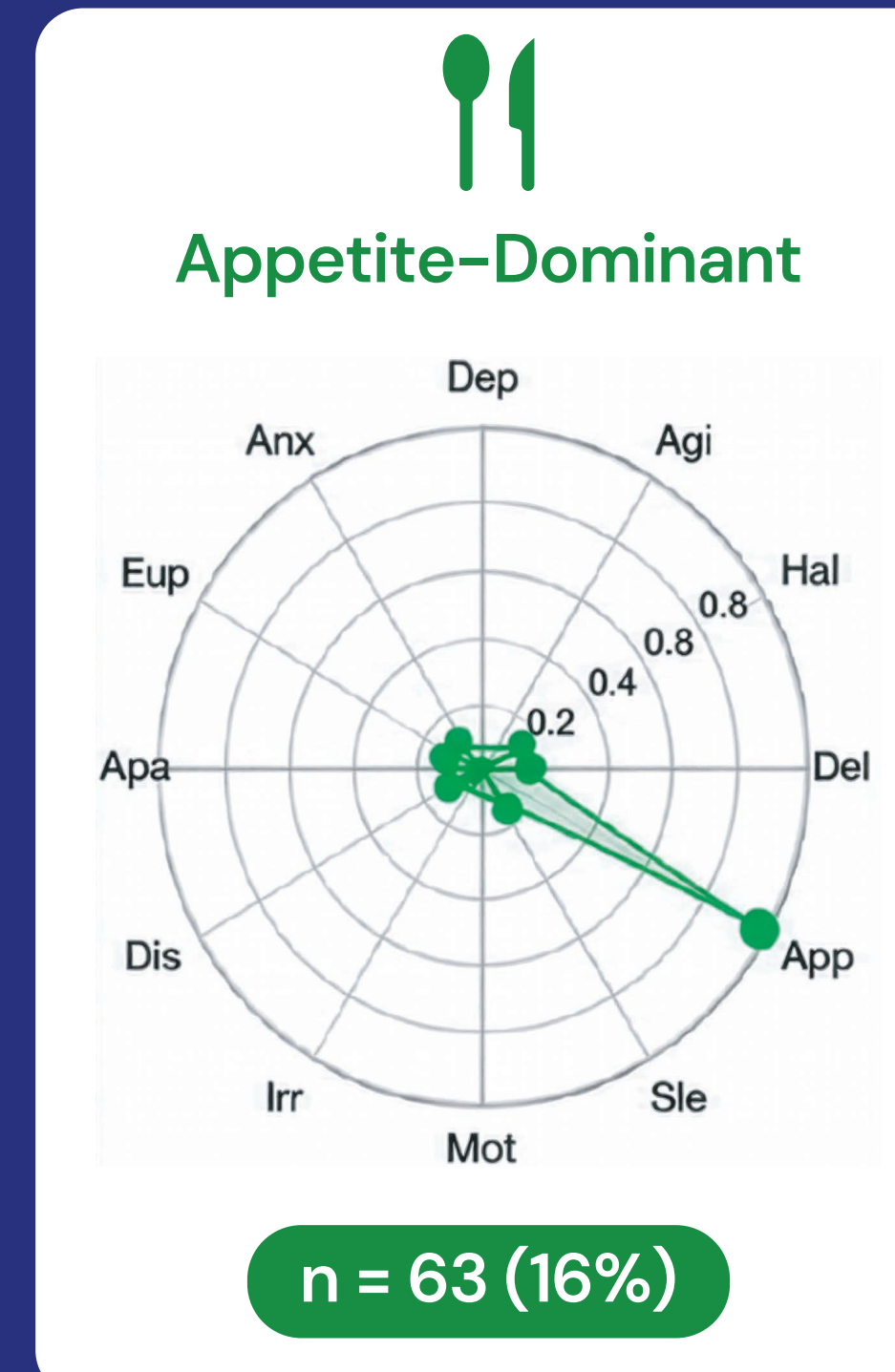
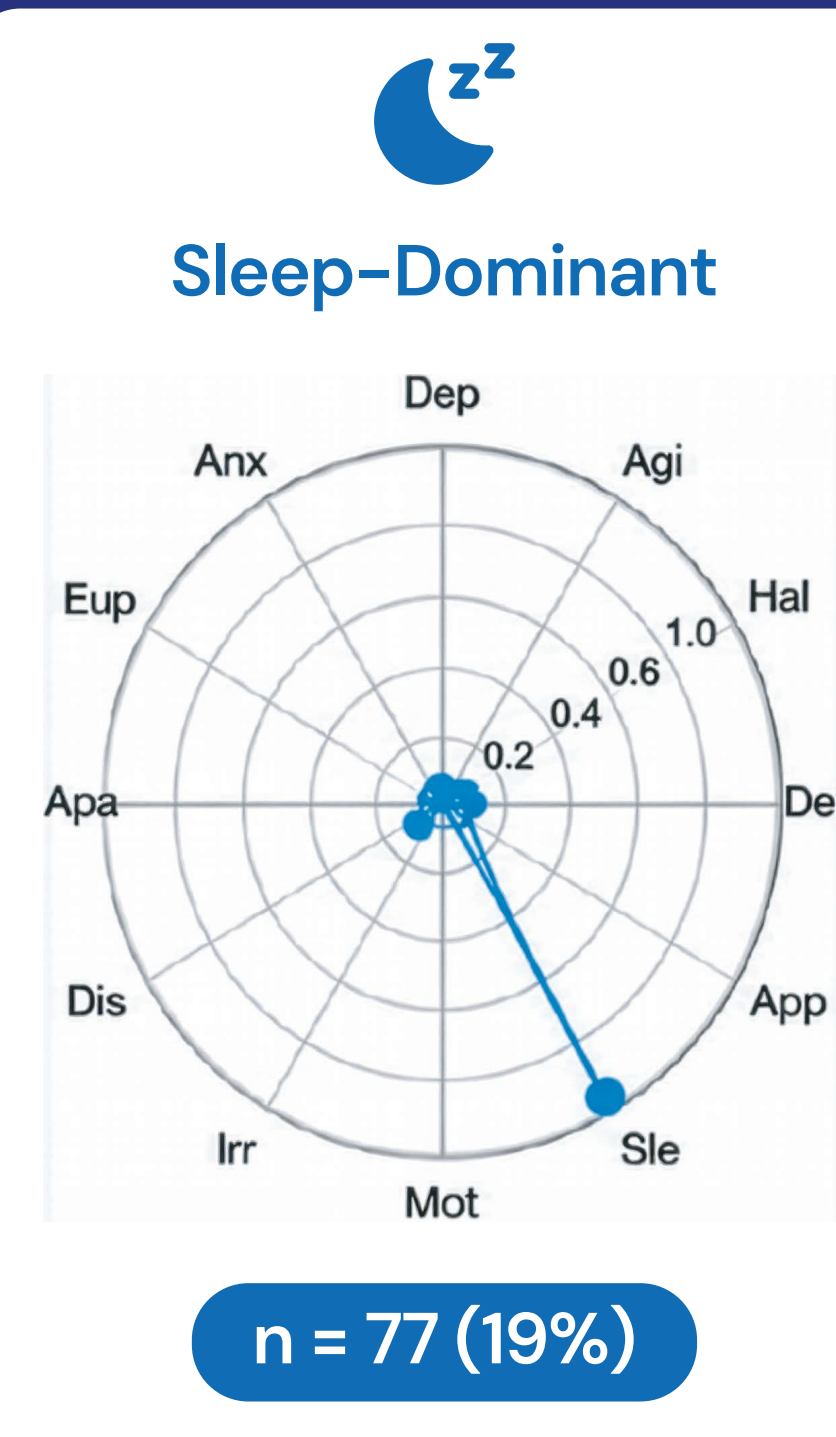


TDA Pipeline

Statistical Validation

- ANOVA, permutation testing, effect size estimation

RESULTS



Radar profiles of five neuropsychiatric pathways in Alzheimer's disease (symptomatic AD patients, N=402). Each radar chart shows the normalized NPI profile for one pathway. Axes represent 12 NPI domains. Abbreviations: Del=Delusions, Hal=Hallucinations, Agi=Agitation, Dep=Depression, Anx=Anxiety, Eup=Euphoria, Apa=Apathy, Dis=Disinhibition, Irr=Irritability, Mot=Motor, Sle=Sleep, App=Appetite.

Data-driven TDA identified eight distinct neuropsychiatric pathways, of which five were most clinically relevant. All pathways demonstrated statistically distinct NPI profiles (ANOVA $p < 0.001$ for 8/12 domains; permutation test $p < 0.001$; effect size $\eta^2 = 0.30$). All pairwise comparisons remained significant after Bonferroni correction ($p < 0.001$).



DISCUSSION AND CONCLUSION

- Topological Data Analysis identified five phenotypically distinct and clinically recognizable neuropsychiatric pathways in Alzheimer's disease through an entirely data-driven framework.
- TDA captured nonlinear, multidimensional symptom interactions that conventional clustering approaches miss.
- Identified symptom constellations align with previously reported NPS co-occurrence networks, supporting the phenotypic plausibility of the structures.
- Findings support pathway-based stratification to improve patient characterization, optimize clinical trial design, and guide personalized therapeutic strategies in Alzheimer's disease.

DISCLOSURE OF FUNDING

This research was funded by IGC Pharma, LLC.

REFERENCES

1. Zhao QF, et al. The prevalence of neuropsychiatric symptoms in Alzheimer's disease: systematic review and meta-analysis. *J Affect Disord.* 2016;190:264-271.
2. Cummings JL, et al. The Neuropsychiatric Inventory: comprehensive assessment of psychopathology in dementia. *Neurology.* 1994;44(12):2308-2314.
3. Weiner MW, et al. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. *Alzheimers Dement.* 2013;9(5):e111-e194.
4. Lum PY, et al. Extracting insights from the shape of complex data using topology. *Sci Rep.* 2013;3:1236.
5. Hui SKW, et al. Neuropsychiatric Inventory domains cluster into neuropsychiatric syndromes in Alzheimer's disease: a systematic review and meta-analysis. *Brain Behav.* 2022;12:e2734.



TAKE-HOME MESSAGE

Topological Data Analysis uncovers phenotypically distinct and clinically recognizable neuropsychiatric subtypes in Alzheimer's disease that may improve precision stratification and targeted therapeutic approaches.

BACKGROUND

The lack of non-invasive, accessible, and scalable biomarkers for Alzheimer's disease and related dementias (AD/ADRD) remains a major challenge. While blood-based biomarkers are actively studied, alternative modalities show promise. Audio biomarkers, easily collected via standardized tasks or natural conversation, can capture speech features associated with AD/ADRD pathology¹⁻³.

METHODS

We extracted multilingual audio recordings (English, Spanish, Mandarin, and German) from the DementiaBank subrepository of TalkBank. Figure 1 presents distributions of the data for training and testing, and shows the diagnosis distribution and the type of tasks available in each language. The diagnoses found in this dataset consist of healthy controls (HC), mild cognitive impairment (MCI), Alzheimer's disease (AD), and other dementias (OD). Time-aligned transcripts were used to isolate participant speech. All recordings were processed through a voice enhancement pipeline to improve clarity and reduce acoustic variability, including background noise reduction, dynamic range compression, frequency equalization, and loudness normalization.

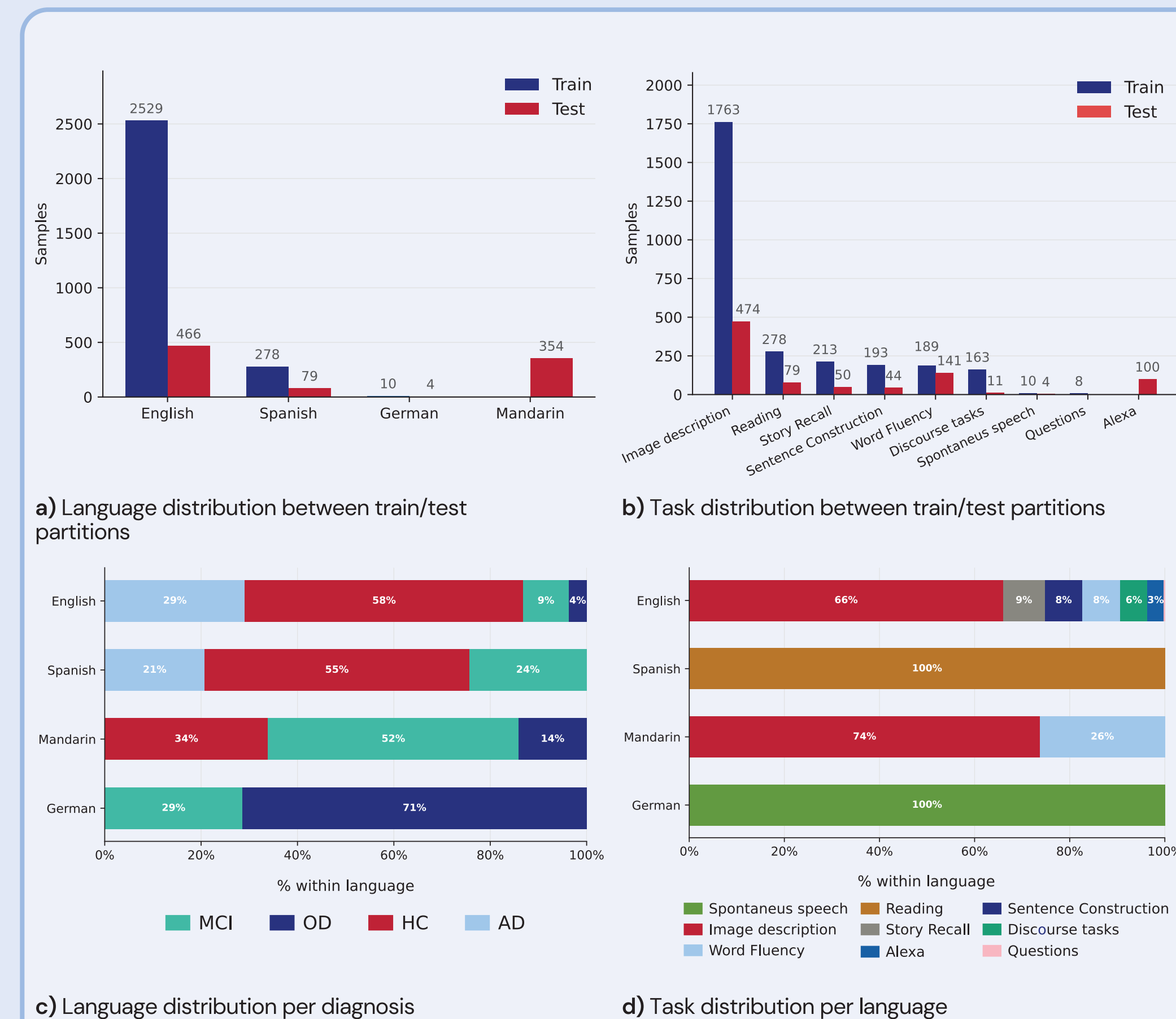
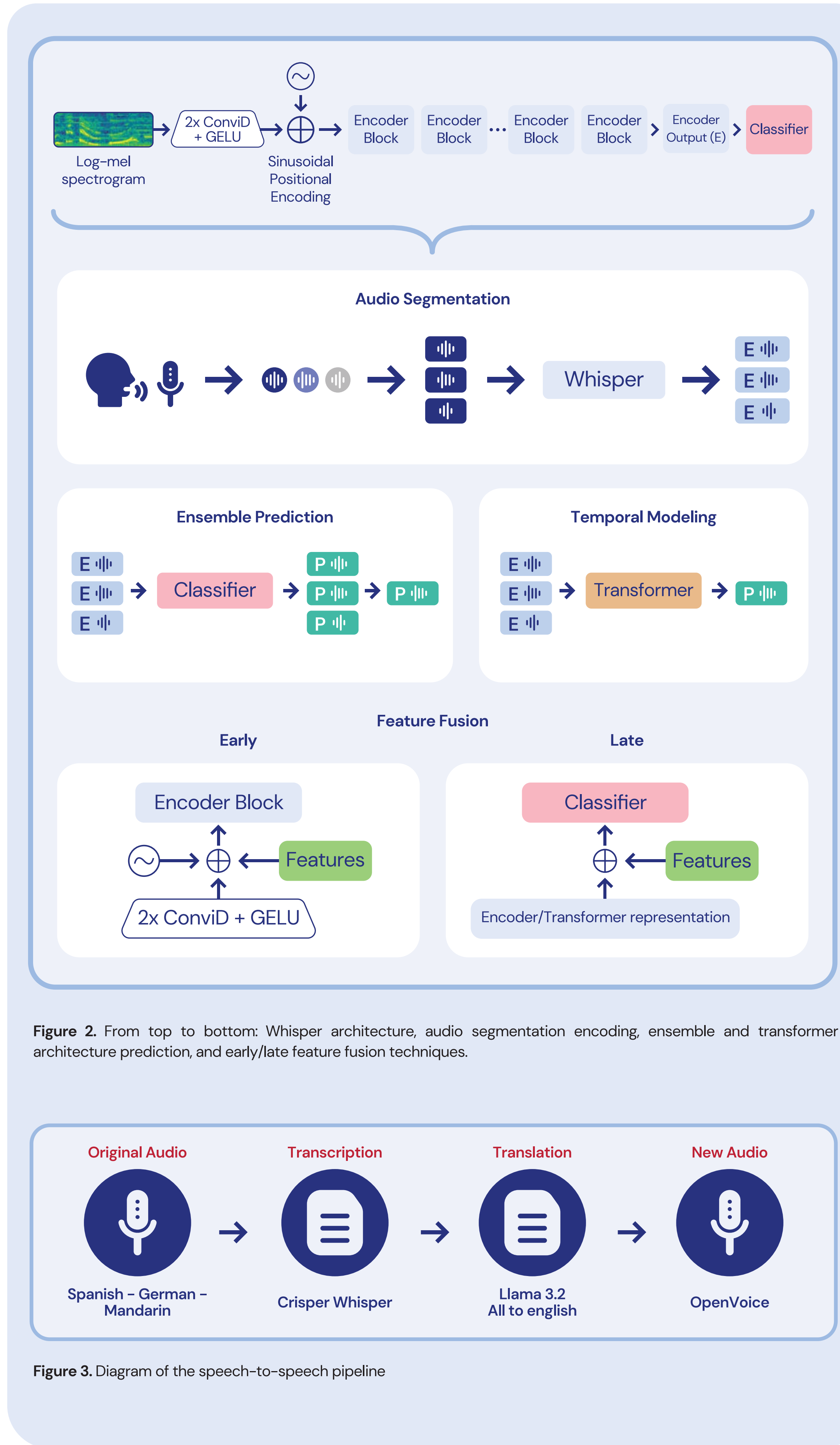


Figure 1. Dataset language, task and diagnosis distributions

We evaluated several models, including distilled and non-distilled Whisper⁴ variants and a hybrid transformer architecture, for temporal modeling, with optional early/late fusion of demographic metadata and audio features. Figure 2 shows a schema of the original Whisper architecture, and the audio segmentation and feature fusion approaches. Recordings were segmented into 30-second chunks to match Whisper encoder requirements. However, the audios were also split into 10-second chunks for the transformer architecture. Participant-level predictions were computed by averaging segment-level class probabilities for Whisper models, while the hybrid architecture produced participant-level predictions directly. To address language imbalance and improve generalizability, we introduced a speech-to-speech translation pipeline as shown in Figure 3.



RESULTS

The most relevant experiments are displayed in Table 1. Whisper Large achieved the best overall performance across diagnostic classes (HC, MCI, AD, and OD), particularly when intermediate layers were frozen. AUPRC and accuracy were 0.799/0.757 (HC), 0.440/0.677 (MCI), 0.798/0.728 (AD), and 0.399/0.899 (OD), respectively. Incorporating demographic metadata improved OD classification, suggesting utility for certain clinical subtypes. Our hybrid architecture effectively captured temporal speech dynamics and achieved the strongest OD performance, potentially due to longer recordings in this group. The speech-to-speech translation approach improved predictions for most non-healthy classes in Mandarin and Spanish, with notable gains for AD, as shown in Table 2.

CONCLUSIONS

In summary, our results show that larger pretrained models, careful adaptation via partial fine-tuning, and modeling speech patterns over time are key to improving the accuracy and reliability of speech-based AD/ADRD prediction. Higher capacity models showed gains from large audios which points to two viable strategies for future data collection: maximizing variability through short recordings when using small models or leveraging longer recordings to enrich sequential speech modeling. Finally, translating non-English speech, while preserving speaker characteristics, can effectively reduce the language domain gap and improve model performance.

DISCLOSURE OF FUNDING

This research was funded by IGC Pharma, LLC.

REFERENCES

- Pastoriza-Domínguez, P. et al. Speech pause distribution as an early marker for Alzheimer's disease. *Speech Commun.* 136, 107–117 (2022).
- Pulido, M. L. B. et al. Alzheimer's disease and automatic speech analysis: A review. *Expert Syst. Appl.* 150, 113213 (2020).
- I, H. et al. Temporal parameters of spontaneous speech in Alzheimer's disease. *Int. J. Speech Lang. Pathol.* doi:https://doi.org/10.3109/17549500903137256.
- Radford, A. et al. Robust Speech Recognition via Large-Scale Weak Supervision. Preprint at https://doi.org/10.48550/arXiv.2212.04356 (2022).



SCAN ME

BEYOND COGNITIVE TESTS: DOMAIN-WISE ABLATION FOR INTERPRETABLE PREDICTION OF DEMENTIA SCREENING SCORES IN AGING COHORTS

Juanita Puentes, Daniel Crovo, Néstor F. González García, María Margarita Venegas, María Alejandra Tangarife, Paola Ruiz Puentes, Pablo Arbeláez, Ram Mukunda
IGC Pharma, Inc.



BACKGROUND

- Scalable dementia screening depends on finding survey-derived predictors of cognition that generalize across cohorts, especially when direct cognitive testing is missing or inconsistent.
- Beyond clinical and cognitive measures, aging surveys capture life-course, social, economic, and household factors that may serve as proxies for cognitive reserve, functional decline, care access, and social support.
- The key unanswered question is whether these non-cognitive structural domains provide independent predictive value for AD/ABR-related cognitive outcomes.

RESULTS

- Strong predictive signal emerged from non-cognitive domains**, indicating that AD/ABR-related information extends beyond variables traditionally used in dementia screening.
- Employment History provided the best standalone MMSE prediction** (RMSE ≈ 3.16), followed by **Family Structure** (≈ 3.20) and **Health Care Utilization and Insurance** (≈ 3.23); all achieved similar CSI-D error (≈ 0.61).
- These domains performed comparably to the **complete multi-domain model** (MMSE RMSE ≈ 3.32 ; CSI-D RMSE ≈ 0.61), suggesting that life-course, household, and care-access variables capture meaningful proxies of cognitive status (Table 1).

Section	Description	MMSE ↓	CSI-D ↓	CSI-D Informant ↓
H	Employment History	3.164	0.611	5.205
G	Family Structure	3.199	0.605	5.121
C	Health Care Utilization and Insurance	3.230	0.608	5.210
L	Assistance and Caregiving	3.325	0.610	5.108
O	End of Life Planning	3.317	0.611	5.193
O	Psychosocial	3.308	0.611	5.113
F	Income and Consumption	3.286	0.608	5.186
B	Health	3.338	0.612	5.198
J	Pension	3.343	0.606	5.203
K	Physical Measures	3.348	0.613	5.195
E	Financial and Housing Wealth	3.416	0.620	5.251
A	Demographics, Identifiers, and Weights	3.469	0.619	5.260
I	Retirement Plans, Expectations	3.469	0.611	5.237
All Sections	Combined	3.316	0.611	5.179

Table 1. Section-wise ablation results for predicting MMSE, CSI-D (self), and CSI-D Informant. Each row reports performance when the model is restricted to variables from a single questionnaire section, and the final row reports performance using all sections jointly. Lower RMSE indicates better performance.

- Attention and predictive performance were not aligned**, indicating that highly attended domains were not necessarily the most predictive and that attention alone is insufficient to assess domain importance.



Figure 3. Attention matrix from the joint model across questionnaire sections. Color intensity reflects average model attention from the classification token to variables within each section across layers, summarizing which domains the model prioritizes when predicting cognitive scores.

METHODS

- Data:** Mexican Health and Aging Study (MHAS) and the 2021 Mex-Cog ancillary study.
- Inputs:** Tabular survey variables from 13 questionnaire domains; direct cognition variables from Section D were excluded (Figure 1).
- Variable representation:** Each variable encoded its identity, observed value, missingness, and questionnaire domain or segment.
- Domain encoding:** Attention-based pooling compressed variable-level features into fixed-length domain representations compatible with BERT's input limit.
- Model architecture:** Domain sequences were concatenated with special tokens and processed by BERT to capture cross-domain interactions (Figure 2).
- Prediction task:** A fine-tuned regression head jointly predicted MMSE, CSI-D self, and CSI-D informant scores.
- Experiments:** Models were compared with and without explicit domain information (Section D).
- Evaluation:** Predictive performance was measured using RMSE, while attention maps were used to assess the relative importance of each survey domain.

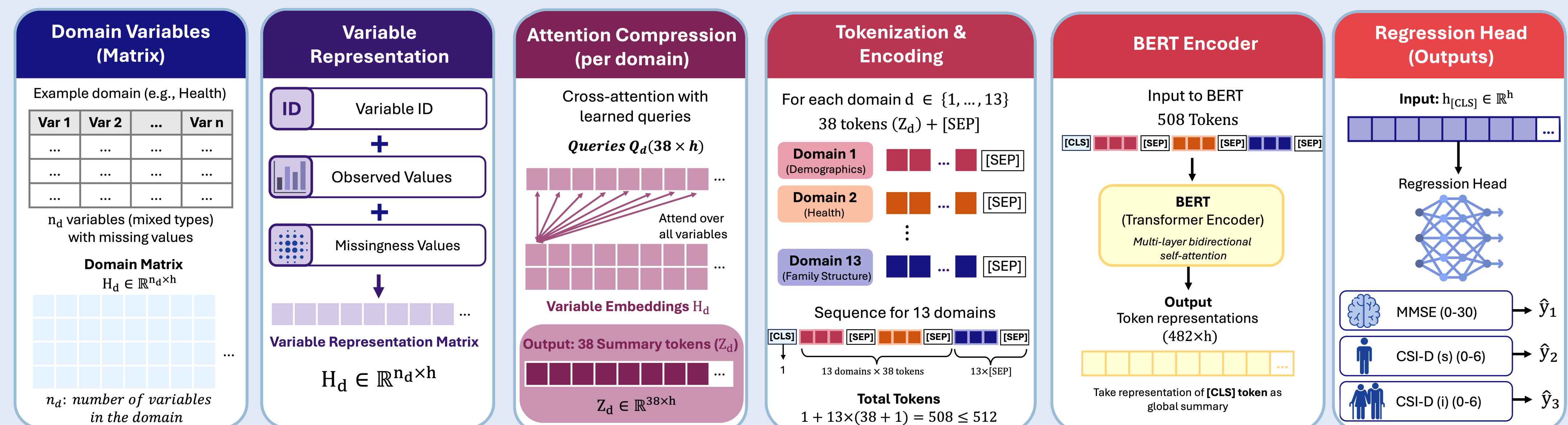


Figure 2. Architecture for domain-wise representation learning and cognitive score prediction. Within each questionnaire domain, variable-level representations integrate variable identity, observed value, and missingness information. These representations are compressed into 38 learned summary tokens through cross-attention pooling, enabling the 13 domain sequences, together with classification and separator tokens, to remain within BERT's 512-token input constraint. The resulting [CLS] representation is used as a global summary and passed to an MLP regression head to jointly estimate MMSE, CSI-D self, and CSI-D informant scores.

CONCLUSIONS

- Cognitive risk is distributed across multiple non-cognitive survey domains**, with employment history, family structure, and health care utilization providing predictive performance close to the full model.
- Non-traditional survey domains may serve as scalable proxies for cognitive risk when direct assessments are unavailable**, supporting dementia screening and cross-cohort integration in aging populations.

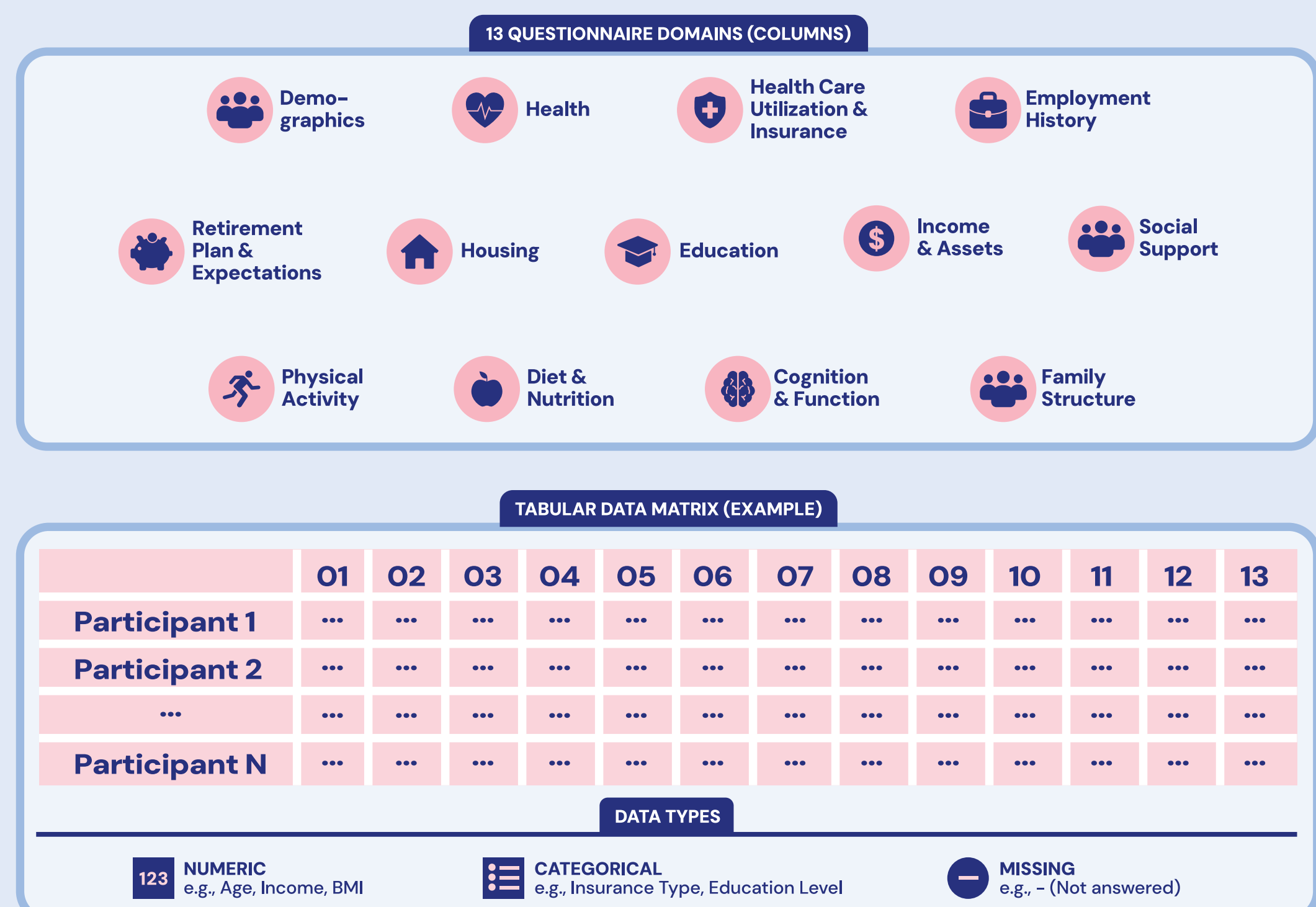
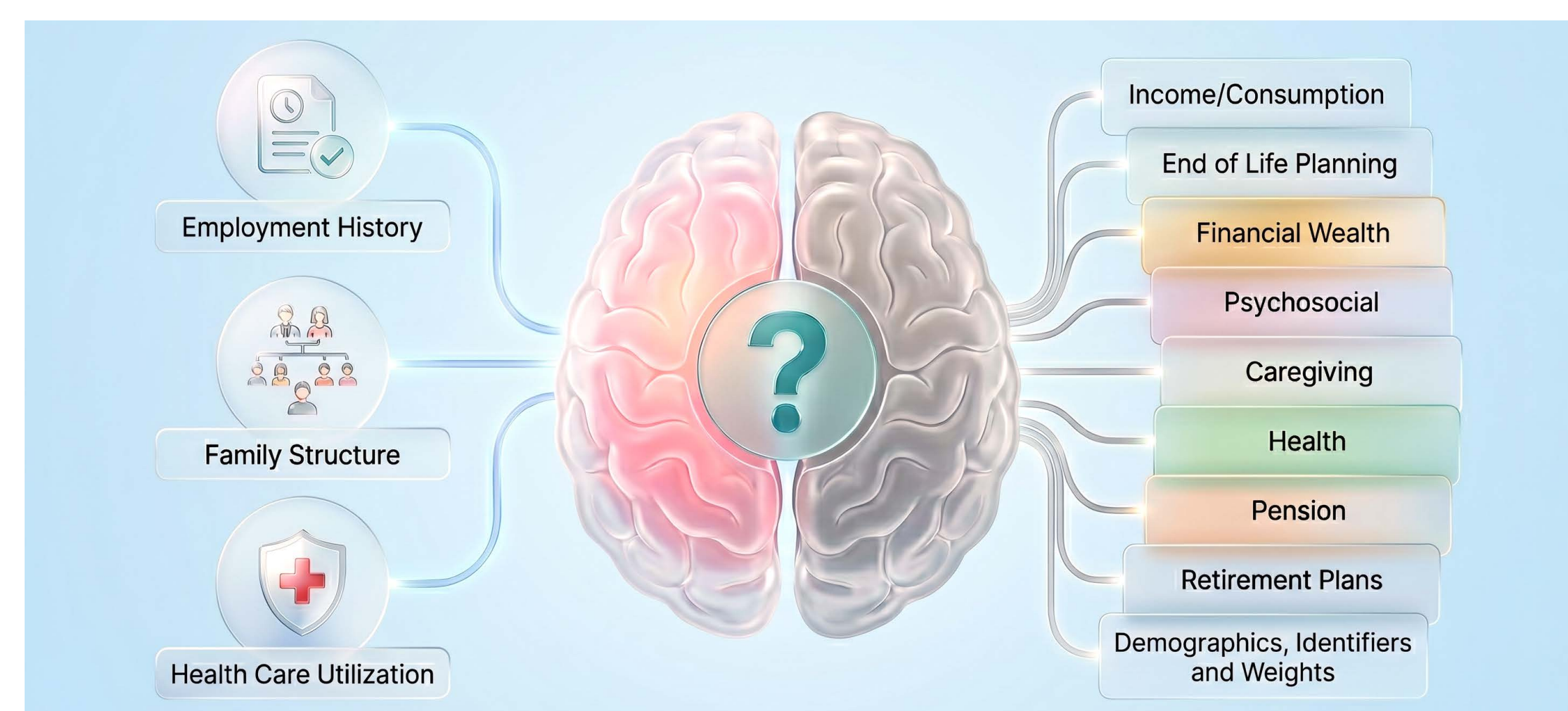


Figure 1. Tabular survey data across 13 domains. Rows represent participants and columns represent numeric, categorical, or missing variables grouped by questionnaire domain.

DISCLOSURE OF FUNDING

This research was funded by IGC Pharma, LLC.

REFERENCES

- Wang, C., Tachimori, H., Yamaguchi, H., Sekiguchi, A., Li, Y., Yamashita, Y., & Alzheimer's Disease Neuroimaging Initiative. (2024). A multimodal deep learning approach for the prediction of cognitive decline and its effectiveness in clinical trials for Alzheimer's disease. *Translational psychiatry*, 14(1), 105.
- Joshi, P., Hendrie, K., Jester, D. J., Dasarathy, D., Lavretsky, H., Ku, B. S., ... & Tampi, R. R. (2024). Social connections as determinants of cognitive health and as targets for social interventions in persons with or at risk of Alzheimer's disease and related disorders: a scoping review. *International psychogeriatrics*, 36(2), 92-118.
- Wong, R., Michaels-Obregon, A., & Palloni, A. (2017). Cohort profile: the Mexican health and aging study (MHAS). *International journal of epidemiology*, 46(2), e2-e2.
- Mejia-Arango, S., Nevarez, R., Michaels-Obregon, A., Trejo-Valdivia, B., Mendoza-Alvarado, L. R., Sosa-Ortiz, A. L., ... & Wong, R. (2020). The Mexican cognitive aging ancillary study (Mex-Cog): study design and methods. *Archives of Gerontology and Geriatrics*, 91, 104210.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, volume 1 (long and short papers) (pp. 4171-4186).

