Received: 16 June 2025 | Revised: 8 August 2025 | Accepted: 24 August 2025

DOI: 10.1002/alz.70719

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

PERSPECTIVE

An AI-first framework for multimodal data in Alzheimer's disease and related dementias

Varuna H. Jasodanand[1] | **Matteo Bellitti**[1] | **Vijaya B. Kolachalama**[1,2]

[1]Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA

[2]Department of Computer Science and Faculty of Computing & Data Sciences, Boston University, Boston, Massachusetts, USA

Correspondence
Vijaya B. Kolachalama, Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts 02118, USA.
Email: vkola@bu.edu

Funding information
National Institute on Aging's Artificial Intelligence and Technology Collaboratories, Grant/Award Numbers: P30-AG073104, P30-AG073105; American Heart Association, Grant/Award Number: 20SFRN35460031; National Institutes of Health, Grant/Award Numbers: R01-HL159620, R01-AG083735, R01-AG062109, R01-NS142076; National Institute on Aging, Grant/Award Number: R01-AG083735; National Heart, Lung, and Blood Institute, Grant/Award Number: R01-HL159620

Abstract

Advancing the understanding and management of Alzheimer's disease and related dementias requires integrating and analyzing diverse data modalities. Traditional diagnostic tools, like neuroimaging, provide valuable insights but are limited by accessibility and infrastructure demands. Meanwhile, emerging modalities, including wearable sensors and speech analysis, enable less invasive and more continuous data collection but introduce challenges related to standardization and privacy. The coexistence of these heterogeneous data streams complicates multimodal integration across cohorts, populations, and clinical settings. Current analytical approaches typically require modality-specific preprocessing pipelines and harmonization methods that were not designed to accommodate modern AI-based capabilities, such as multimodal fusion. In this perspective, we propose an "AI-first" strategy for multimodal data integration that aligns data structuring, harmonization, and modeling within a unified set of guiding principles to optimize modern AI development, while remaining flexible enough to support classical analytical approaches.

KEYWORDS
artificial intelligence, machine learning, multimodal data

Highlights

- Understanding and managing ADRD requires integrating biological, cognitive, and behavioral data across multiple modalities.
- Incorporating multiple modalities requires new standards for harmonization and interoperability.
- Current data platforms are not necessarily built to support multimodal fusion or generalizable AI models across diverse ADRD populations.
- Modern AI models are capable of learning from messy, multimodal, and incomplete data but require infrastructure designed for this purpose.
- We propose rethinking ADRD data systems to prioritize AI compatibility, enabling scalable tools for early diagnosis and longitudinal care.

1 | INTRODUCTION

Accurate diagnosis, monitoring, and management of Alzheimer's disease and related dementias (ADRD) require the integration of multimodal data spanning biological, cognitive, and functional domains (Figure 1A). No single test or assessment can independently capture the full complexity of the disease.[1] Instead, meaningful insight emerges when diverse signals, from neuroimaging and neuropsychological testing to blood-based biomarkers and real-world behavioral data, are interpreted together. This multimodal approach reflects the insidious, multifactorial nature of ADRD and offers a pathway toward more precise, timely, and individualized care, particularly when enabled by artificial intelligence (AI) systems capable of synthesizing complex inputs. This need for integration has become more urgent as the field enters a transformative era. The emergence of disease-modifying therapies brings the possibility of slowing neurodegenerative processes, but it also demands earlier detection, more dynamic disease monitoring, and personalized intervention strategies. These advances pose a fundamental challenge: How can we adapt our data structures and analytical methods to support clinically meaningful and generalizable insights?

Over the past two decades, the field has made considerable progress in collecting and analyzing rich, clinically validated data. Neuroimaging, cerebrospinal fluid (CSF) and plasma biomarkers, and structured cognitive assessments remain central to disease staging and progression modeling. More recently, a wave of emerging data sources, including wearable sensors, speech and language processing, and passive monitoring, has expanded opportunities to observe individuals in their everyday environments.[2] These technologies promise to complement traditional assessments by enabling continuous and ecologically valid measures of cognition and function. Yet, as the diversity of data modalities expands, so too does the complexity of integrating them. Each modality differs in collection methods, granularity, missingness patterns, and contextual constraints. Neuroimaging is high-resolution but infrequent and expensive; digital sensing is longitudinal but noisy and requires engagement from the patient; plasma biomarkers are scalable, but their performance varies depending on the analyte and assay technology used. Integrating these data streams for large-scale analysis remains a non-trivial, resource-intensive task that exposes fundamental limitations in how current data systems are designed.

Despite ongoing efforts to standardize and harmonize ADRD data, most data infrastructure remains siloed within individual cohorts or specific modalities and was not designed to support multimodal, multicohort analyses or modern AI-model development. Few support tasks like multimodal fusion, prediction under partial data missingness, or generalization across diverse populations and clinical settings. As richer and more varied data types are incorporated, it becomes insufficient to retrofit current integration solutions. In this evolving landscape, what is needed is not simply more data but infrastructure intentionally designed to make existing data useful. This is especially important given two realities. First, real-world data, which are often messy, incomplete, and multimodal, vastly outnumber curated research datasets. For instance, there were approximately one billion office-based physician visits in the United States in 2019 alone.[3] This stands in stark contrast to the tens of thousands of individuals enrolled in research cohorts, underscoring the broad reach of routine clinical care compared to research participation. Second, modern AI models have demonstrated a remarkable ability to learn from such complex and imperfect data.[4] Translating such approaches to ADRD populations, however, requires addressing unique ethical, technical, and infrastructural challenges: ADRD develops over long timescales and often without acute manifestations, in contrast to diseases in other branches of medicine, where AI adoption has been more widespread. For example, unlike oncology, where diagnosis and staging largely rely on tissue biopsies and well-characterized molecular markers that provide a clear reference standard, ADRD diagnosis relies on neuropsychological tests and expert clinical evaluations, which are inherently subjective. ADRD also encompasses a wide spectrum of disease subtypes and clinical presentations, often complicated by other age-related systemic disorders, thereby making standardized data collection and organization difficult. Among these considerations, the lack of interoperable, AI-ready data infrastructure stands out as the most actionable.

In this perspective, we introduce the tenets of an AI-first approach to multimodal data integration for ADRD. This approach aligns data structuring, harmonization, and modeling within a unified framework, defined as a set of guiding principles and requirements designed from the outset to meet the needs of modern AI systems while remaining flexible enough to accommodate traditional analytical approaches. An AI-first strategy emphasizes the intentional co-design of data organization, metadata annotation, and analytic workflows to ensure the resulting infrastructure natively supports AI applications across heterogeneous multimodal ADRD data. Existing solutions typically focus on either organizing and harmonizing specific data modalities or developing user-friendly platforms, often without resolving the cross-cohort semantic and structural mismatches that hinder true integration. In contrast, our perspective advocates for unifying these elements within a single, interoperable system.

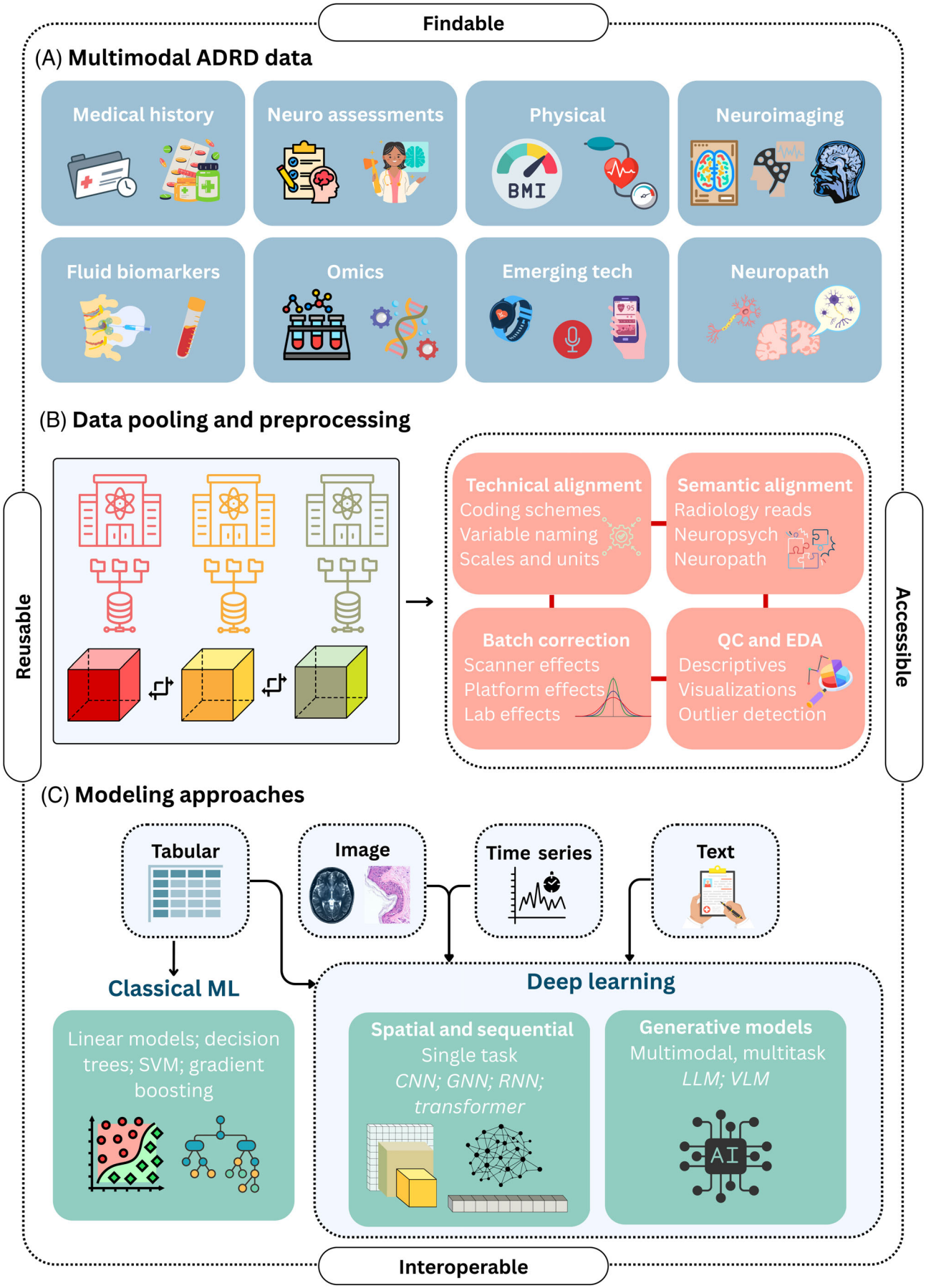
Here, we explore the technical and clinical challenges of this shift and outline how it can enable the development of scalable AI tools for early detection, diagnosis, and long-term care in real-world settings. We envision a future where rich ontology-driven metadata simplify harmonization, annotated missingness allows models to reason about data gaps, and data systems become adaptive, co-evolving with the capabilities of AI systems themselves. To familiarize the reader with AI-related jargon, we present a collection of terms with definitions in Table 1.

2 | DATA LANDSCAPE IN ADRD

2.1 | Clinical practice: sequential acquisition and pragmatic constraints

In clinical settings, ADRD diagnostic tests are typically administered sequentially and selectively, guided by clinical judgment, the patient's

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION



presentation, financial considerations such as insurance barriers, and logistical constraints such as limited availability of specialized testing.[5] A primary-care provider (PCP) may first suspect cognitive impairment, triggering a referral to a neurologist, who then determines the need for further evaluations, including neuropsychological testing, neuroimaging, and biomarker testing, as indicated. Each step is contingent on prior findings, insurance coverage, and patient or family preferences. For example, a patient with prior spinal surgery or significant agitation may be unable to undergo CSF collection, while another with contraindications to magnetic resonance imaging (MRI) may forgo neuroimaging. This pragmatic, stepwise workflow prioritizes clinical actionability over data comprehensiveness by design. Although optimized for clinical purposes, the fragmented and unstructured nature of these data typically makes it hard to use traditional statistical techniques for analysis.

2.2 | Research cohorts: deep phenotyping versus ecological validity tradeoffs

In contrast, large cohort studies or clinical trials, including the National Alzheimer's Coordinating Center (NACC),[6–8] the Framingham Heart Study (FHS),[9] the Alzheimer's Disease Neuroimaging Initiative (ADNI),[10–12] and the Anti-Amyloid Treatment in Asymptomatic Alzheimer's (A4) study, collect data through rigid, protocol-driven workflows, designed for cohort-level data standardization and scientific rigor. Participants undergo comprehensive neurological and physical exams, neuropsychological assessments, and functional evaluations to establish clinical diagnoses based on the consensus of dementia experts. High-resolution MRI scans are often collected to quantify various neuroimaging features. Additionally, amyloid, tau, and fluorodeoxyglucose positron emission tomography (PET) provide insights into in vivo disease pathology.[13] These modalities are often complemented by fluid biomarkers, such as CSF and plasma measurements, and multi-omic profiles, all of which offer extensive biological characterization of disease.[14–16] The result is deeply phenotyped datasets that enable precise modeling of disease mechanisms. However, the tight inclusion criteria and controlled protocols that produce these data often exclude participants with comorbidities or atypical presentations, creating "clean" datasets that poorly represent the heterogeneity of clinical populations.

2.3 | Emerging data streams: attempting to bridge the translational divide

Models developed on homogeneous cohorts often fail to generalize to clinical settings where missingness is unpredictable, comorbidities are common, and workflows are sequential. Emerging digital health technologies (DHTs) offer an opportunity to bridge this divide, introducing continuous, real-world data streams in the research landscape. Wearable devices can capture gait patterns, physical activity levels, and physiological measures such as sleep cycles and heart rate. Speech analysis can identify subtle changes in cognition.[17,18] Self-administered digital cognitive assessments can also allow remote, repeated monitoring of participants,[19,20] paving the way for acquiring data from a broader, more diverse population that is traditionally not well represented in research cohorts. Integrating these novel data streams creates a multidimensional view of ADRD and bridges the granularity of research-grade biomarkers with more ecologically valid data. However, their promise will only be realized if models are trained and validated on integrated datasets that reflect the diversity and complexity of real-world populations.

3 | DATA PREPARATION APPROACHES IN ADRD

3.1 | Data pooling is useful

To address the reproducibility and generalizability limitations of siloed datasets, pooling multimodal data across ADRD studies is essential. Data pooling addresses these challenges through several mechanisms. First, combining datasets increases statistical power: Studies relying on small sample sizes are more likely to suffer from sampling variability, making the strength and direction of reported associations difficult to replicate. Second, pooling enhances sample diversity across demographic subgroups and medical comorbidities. Pooling is especially valuable for studying rare diseases, such as Huntington's, and less common presentations, such as the logogenic variant primary progressive aphasia, which would be infeasible to investigate at scale in single-cohort analyses. Ultimately, pooled datasets better reflect the underlying variability in ADRD, thus improving external validity and generalizability. However, pooling alone is insufficient: Naive

FIGURE 1 Framework to support AI-driven research and translation in ADRD. To develop scalable, trustworthy, and clinically actionable AI models that reflect the multifactorial complexity of dementia, data systems must be findable, accessible, interoperable, and reusable. (A) Modern ADRD research relies on a range of heterogeneous data types, including traditional clinical inputs (e.g., medical history, neuroimaging). Additional insights can be captured through omics technologies and emerging data streams such as speech, sensory-derived activity metrics, and digital cognitive assessments. *Post mortem* data remain essential for validation. (B) Multimodal ADRD data integration requires systematic alignment of heterogeneous data from multiple cohorts, addressing four critical challenges: technical alignment, semantic alignment, batch correction, and QC paired with EDA. This systematic approach transforms siloed, cohort-specific datasets into integrated, analysis-ready data structures. (C) Effective AI systems for ADRD must accommodate varied data modalities, including tabular data, images, time-series data, and unstructured text. Classical machine learning methods are typically modality-specific and sensitive to data missingness (SVM). Deep learning approaches enable greater representational power, with some methods (CNN, GNN, RNN) being useful for single-modality tasks, while others can integrate multimodal and multitask inputs (LLM, VLM). ADRD, Alzheimer's disease and related dementias; AI, artificial intelligence; CNN, convolutional neural network; EDA, exploratory data analysis; GNN, graph neural network; LLM, large language model; RNN, recurrent neural network; SVM, support vector machine; VLM, vision-language model; QC, quality control.

TABLE 1 Glossary of technical terms.

API	A set of rules and communication standards that allow different software systems to communicate and share data.
CNN, RNN, GNN	Popular deep learning architectures specialized in processing images (CNN), sequential data (RNN), or graphs (GNN).
Data infrastructure	Systems and resources that enable the collection, storage, management, integration, processing, and accessibility of data. This includes hardware (servers, storage devices), software (databases), standards, and governance policies.
Data lake	Storage system designed to hold large amounts of data in their native format. It allows flexible data access and analysis without predefined organization.
Data warehouse	Storage system designed to store and organize structured data. Unlike a data lake, it uses a predefined schema to ensure consistency and faster querying.
Deep learning	A subset of ML that uses neural networks with many layers. It excels in tasks like image recognition, speech processing, and natural language understanding.
FAIR data principles	Guidelines to make data Findable, Accessible, Interoperable, and Reusable. They ensure that data can be effectively used by both humans and machines.
Feature engineering	The act of building variables out of simpler ones (e.g., a risk score). A key advantage of deep learning is that it requires only minimal feature engineering compared to traditional methods.
Federated learning	Computational approach where models are trained in a decentralized way without sharing source data.
Generalizability	A result is generalizable if it applies to both the sample under study and the population it is from, or similar populations.
Generative AI	Collection of AI techniques capable of creating new content, such as text or images, by learning patterns from existing data (e.g., large language models).
Graphical processing unit (GPU)	Specialized processor, designed for applying the same operation across multiple data elements at the same time. Enables ML analysis of large datasets (e.g., omics, imaging) with greater efficiency.
Interpretable AI	AI systems that are designed to explain not only what their prediction is, but how they reached it.
Masked modeling/feature masking	A technique where parts of input data are hidden, and the model learns to predict the missing pieces. It helps models learn from unlabeled data and become robust against missingness.
Multilabel classification	A type of classification where each data point can belong to multiple categories at once. For example, a medical image might be labeled with multiple diagnoses.
Multimodal fusion	The act of combining qualitatively different data (e.g., imaging and text). The fusion is "early" if the modalities are combined before any significant processing and "late" if they undergo notable transformation independently.
Ontology	A structured framework that organizes knowledge into categories and defines relationships between them. In AI, it helps machines interpret and use complex information consistently.
Regularization	ML technique that reduces overfitting by adding a penalty to a model's complexity. It helps improve the model's performance on new data.
Scalability	The ability of a system to handle increased workload efficiently, without prohibitive cost or waiting times.
Structured querying	The process of using a structured language (e.g., SQL) to retrieve and manage data from databases.

Note: This glossary outlines key technical concepts relevant to AI-powered, multimodal frameworks for ADRD.
Abbreviations: ADRD, Alzheimer's disease and related dementias; AI, artificial intelligence; API, Application Programming Interface; CNN, convolutional neural network; FAIR, Findable, Accessible, Interoperable, and Reusable; GNN, graph neural network; ML, machine learning; RNN, recurrent neural network.

aggregation of heterogeneous data creates new challenges that require systematic technical and semantic alignment.

3.2 | Barriers to large-scale multimodal integration

Multimodal data integration processes face foundational harmonization challenges, ranging from simple syntax differences in cohort-specific data dictionaries to protocol misalignments and operationalization differences in clinical definitions. These barriers are exacerbated by non-biological variability, such as site effects, and extensive data missingness,[21] particularly for biomarker and neuropsychological tests. This presents a challenge and an opportunity: While misalignment across cohorts complicates predictive modeling efforts, it also enables researchers to innovate on methods that robustly handle heterogeneous missingness patterns and protocol variations, yielding models with greater real-world applicability. Successfully leveraging this opportunity, however, requires tackling the cross-cohort data's "Tower of Babel" through carefully designed data management strategies.

Harmonization challenges are particularly pronounced in quantitative analyses of neuroimaging data. Scanner models,

acquisition protocols, reconstruction algorithms, and image processing pipelines all influence quantitative imaging measures in unique ways.[22] For instance, PET imaging is affected by variations in radiotracer properties, reference regions, and partial volume correction methods. While standardization efforts such as the Centiloid and CenTaur scales for amyloid and tau PET have improved cross-study comparability,[23,24] they still cannot fully capture non-linear biological variation or harmonize across all technical parameters. Similarly, methodological variability in fluid biomarker measurements affects multicohort analyses. CSF and blood measurements, for example, are highly sensitive to assay technology, analyte properties, and study protocols.[25]

The emergence of digital technologies introduces yet another layer of complexity. Actigraphy devices differ in step-count algorithms, speech analysis tools vary in sampling rates,[26] and cognitive assessments administered via tablet versus personal computer are influenced by hardware-specific latencies. Further variability arises from differences in software versions, environmental factors, and adherence rates, which impact data quality and performance measures. Differences in data formats and limited interoperability between digital health platforms hinder seamless integration, making it difficult for clinicians and researchers to access and meaningfully interpret data from multiple sources.

3.3 | DIY data integration

Integrating multimodal, multicohort data into an analysis-ready format requires extensive preprocessing (Figure 1B). This effort demands not only domain knowledge but also data science expertise and access to computational resources. Broadly, the process involves two overarching phases: technical alignment, which ensures data from different sources are interoperable, and semantic alignment, which ensures that the meaning and context of variables are consistent. In the technical alignment phase, the goal is to address compatibility across cohorts by unifying variable naming conventions, standardizing coding schemes, and ensuring consistent scales and units. For instance, a researcher might want a variable indicating cognitive status to be labeled consistently, with values such as "0" always referring to normal cognition and "1" to impairment. In contrast, semantic alignment tackles the conceptual definitions underlying different variables. For example, definitions of "cognitive impairment" may vary across cohorts: Some may include amnestic mild cognitive impairment, others non-amnestic forms or even early dementia. These semantic mismatches extend beyond cognitive assessments. Differences in radiological reads, biomarker thresholds, or neuropathological grading systems can further complicate integration. Without resolving such semantic inconsistencies, model development risks generating invalid or non-generalizable findings. Additionally, batch effects, defined as the artificial variability introduced by hardware or software differences across cohorts, are also common in neuroimaging and omics data. Statistical harmonization methods like ComBat are frequently used to mitigate these effects while preserving meaningful biological variation.[27]

While these alignment efforts are critical, they also represent a bottleneck in multicohort ADRD research. Manual mapping of taxonomies and ontologies is time-consuming and requires collaboration between clinicians, informaticians, and statisticians. However, once achieved, these harmonized datasets offer a strong foundation for developing analytical strategies that can more effectively uncover patterns related to ADRD mechanisms and heterogeneity.

4 | MODELING APPROACHES IN ADRD RESEARCH

4.1 | Inferential statistics

Traditional analytic approaches in ADRD research have largely relied on inferential statistics, with separate processing pipelines developed to extract information from each data modality independently. For example, voxel- and surface-based morphometry have been standard methods for quantifying brain structure and pathology from MRI scans.[28,29] Genetic data are typically analyzed through genome-wide association studies[30] and calculation of polygenic risk scores,[31] while cognitive performance is often summarized using composite cognitive domain *z*-scores that can abstract performance across multiple tests.[32] Despite their clinical relevance, variables such as medical history and medication use remain underutilized, with biomarker-focused analyses dominating recent literature.

Statistical models serve as foundational tools for hypothesis testing and population-level inference. Their interpretability is a strength: Model coefficients can be tied to specific predictors, and confidence intervals provide estimates of uncertainty. Linear and logistic regression, as well as mixed-effects models, perform well when assumptions about variable distributions hold and data dimensionality is modest. However, their limitations become evident when applied to modern ADRD datasets that are high-dimensional, multimodal, and incomplete. They struggle to accommodate irregular longitudinal follow-ups, non-linear associations, or dependencies among diverse data types. Currently, there is no consensus on optimal analytic strategies for datasets with irregular sampling and complex missingness patterns, which underscores the need for innovative approaches. Richly annotated, flexible data structures that preserve temporal granularity and explicitly capture reasons for missingness are essential for enabling novel analytic methods that can operate under less restrictive assumptions and harness the full complexity of contemporary ADRD data.

4.2 | Classical machine learning in cohort studies

Machine learning (ML) methods have increasingly been used in ADRD research to support classification, subtyping, and disease progression modeling.[33] These approaches are attractive because they can accommodate "wide" data, which are datasets with many variables per sample, and process heterogeneous inputs that do not meet the assumptions of classical statistical models. This is especially

important in ADRD, where diagnosis is complicated by symptom overlap across dementias, within-disease heterogeneity, and comorbid conditions that obscure disease-specific signals.[34] With access to large, deeply phenotyped cohorts like NACC, which includes hundreds of variables ranging from neuropsychological testing to autopsy findings, ML offers a way to integrate multifaceted data into actionable outputs.

Derived measures from imaging, cognitive, and demographic data have been used in traditional supervised ML models to improve ADRD classification and progression prediction. Unsupervised approaches such as Subtype and Stage Inference (SuStaIn) have also been employed to characterize disease heterogeneity and dynamics across patients.[35] However, traditional ML models face limitations. Most are trained on AD-centric "clean" datasets, leading to suboptimal performance in real-world applications involving mixed pathologies and frequently incomplete or missing data.[36] Additionally, these models are vulnerable to overfitting in high-dimensional spaces, necessitating complex feature selection, regularization, and tuning strategies.

4.3 | Modern AI approaches

Recent AI advances, particularly deep learning (DL) and generative AI, are helping overcome many limitations of traditional analytic methods.[37–40] Unlike classical ML models, which rely heavily on engineered features and complete datasets, DL architectures can learn directly from raw or minimally processed inputs. Models such as convolutional neural networks, recurrent neural networks, transformers, and graph neural networks are now used to process diverse inputs, including MRI, PET, genetic data, voice recordings, wearable device data, and electronic health records (EHRs) (Figure 1C). A key innovation in these approaches is their ability to handle missing data natively. Techniques such as random feature masking, attention mechanisms, and uncertainty-aware inference allow these models to learn even when parts of the input are missing. This is crucial in ADRD, where missingness may not be random, as patients with cognitive decline are more likely to skip follow-up visits or certain assessments, making standard imputation approaches invalid.

Generative AI, particularly large language models (LLMs), is further expanding the landscape. LLMs are already being used in drug discovery to prioritize targets, in clinical trial design for cohort matching and safety monitoring, and in EHR systems to summarize clinical notes and suggest differentials. For example, OpenEvidence can automatically extract and synthesize guidelines and generate diagnostic hypotheses. A new frontier is emerging with agentic AI, as semi-autonomous, goal-oriented AI systems can plan, collaborate, and even drive scientific discovery with expert human guidance. A recent example includes Virtual Lab,[41] an AI–human collaborative platform to not only perform scientific research but also enable discovery. Such advanced AI systems are poised to become transformative tools in ADRD research as well.

However, this potential depends critically on well-designed data infrastructures. Despite recent AI advances in other fields, real-world applications in ADRD remain challenging. Current cohort infrastructures in ADRD are not built for AI: Data are often fragmented, poorly annotated, and siloed across institutions. Even the most advanced models will underperform if trained on biased or small datasets. Thus, a shift toward AI-first data architectures is essential, designing data systems from the ground up to support model training and the Findable, Accessible, Interoperable, and Reusable (FAIR) principles.[42,43]

5 | THE CASE FOR AN AI-FIRST DATA STRUCTURE

5.1 | Past and ongoing data integration efforts

To fully realize the potential of modern AI approaches, data integration is essential. Pooling, alignment, and harmonization are currently time-consuming steps that researchers must complete before unlocking the benefits of large-scale integrated datasets. Figure 2 presents the key distinctions between these processes, illustrating how each addresses different challenges in transforming heterogeneous data into analysis-ready formats. Encouragingly, significant progress has already been made in standardizing tabular and textual data through initiatives such as the Observational Medical Outcomes Partnership (OMOP),[44] Medical Subject Headings (MeSH), SNOMED CT,[45] Logical Observation Identifiers Names and Codes (LOINC),[46] and Health Level 7 (HL7). These efforts have established foundational taxonomies and ontologies that promise consistent data capture, sharing, and analysis across clinical and research domains.

In neuroimaging, the Brain Imaging Data Structure (BIDS) represents a major advancement.[47] It provides standards for organizing and sharing imaging data by enforcing metadata requirements and directory structures. Built upon widely adopted formats like Digital Imaging and Communications in Medicine (DICOM) and Neuroimaging Informatics Technology Initiative (NIfTI), BIDS has facilitated large-scale, multisite collaborations and streamlined data exchange. Although extensions of BIDS to other data types are emerging, its current use is still largely concentrated in imaging applications. As neuroscience research increasingly integrates diverse data types, the need for similarly robust standardization beyond neuroimaging is apparent.

In the context of ADRD, the Uniform Data Set by NACC provides a compelling example of clinical data harmonization[8] and support of longitudinal analyses at scale. Collaborations with the Alzheimer's Disease Sequencing Project (ADSP), the Phenotype Harmonization Consortium (PHC), and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) have enabled meaningful integration of clinical phenotypes with genomic and sequencing data. Still, as data types grow in diversity and complexity, additional considerations emerge, particularly when aggregating multimodal datasets from multiple sources. For example, integrating NACC and ADNI, two of the most widely used ADRD cohorts, presents both opportunities and challenges. Each cohort has developed its own infrastructure for data management and access, NACC offers consolidated CSV file downloads, while ADNI utilizes a distributed file system via the Laboratory of Neuro Imaging. These independent systems reflect thoughtful, long-standing design choices, but do not yet support a

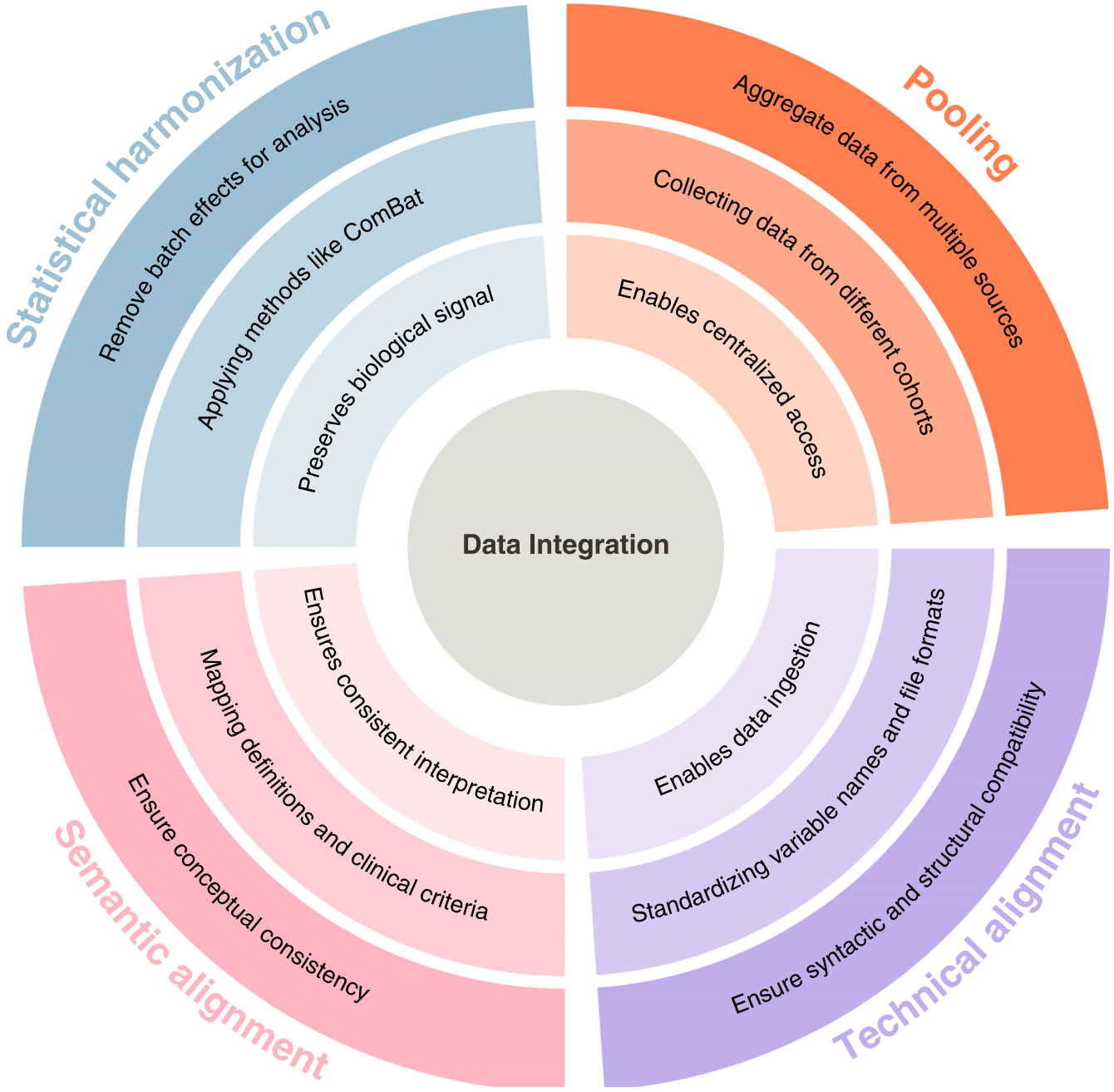


FIGURE 2 Key components in multimodal ADRD data integration. The framework illustrates four distinct but interconnected processes of data integration: organized in concentric layers from the processes' goals (outer ring) to specific examples (middle ring) and operational roles (inner ring) in the data pipeline. Pooling (orange) focuses on data aggregation from multiple sources to enable centralized access. Technical alignment (purple) addresses syntactic and structural compatibility by standardizing variable names and file formats to ensure data ingestion. Semantic alignment (pink) ensures conceptual consistency by mapping definitions and clinical criteria to ensure consistent interpretation across studies. Statistical harmonization (blue) removes systematic batch effects while preserving the biological signal for statistical analyses. Although these processes are complementary, each addresses distinct challenges in transforming heterogeneous multicohort ADRD datasets into analysis-ready formats. ADRD, Alzheimer's disease and related dementias.

unified, cohort-agnostic data access model. These challenges are not unique to ADRD research. Similar complexities arise in other scientific domains and the private sector, where integration strategies such as data lakes (for storing unstructured raw data), data warehouses (for structured querying), Application Programming Interfaces (APIs), and standardized exchange protocols have proven helpful. Such principles are increasingly reflected in emerging research platforms in the ADRD field. For instance, Synapse, the AD Workbench, and the Global Alzheimer's Association Interactive Network (GAAIN) offer credentialed access to curated datasets and support secure, cloud-based analysis. Another recent development is the Global Research and Imaging Platform (GRIP), which provides a modular cloud-based

platform for multimodal data analysis through tools such as Jupyter Notebooks and RStudio. They offer unified access and analysis within their respective ecosystems. Yet integrating data across platforms and studies still requires substantial effort, as no single, overarching data organization and access system has yet been widely adopted across all modalities. As a result, researchers often work with aggregated datasets that still necessitate cross-cohort mapping and harmonization.

Differences in data governance models also influence integration efforts. Each platform has varying policies on data access and usage, which may affect where and how analyses are conducted. Some platforms, like the AD Workbench, allow external dataset imports, while others, such as GAAIN, do not, reflecting differing priorities and institutional requirements. In addition, while these platforms support traditional statistical analyses, few currently provide access to graphical processing unit resources, which limits their utility for developing and deploying modern AI models. Taken together, these observations highlight the need to develop AI-first data structures that prioritize multimodal integration and cross-cohort interoperability from the outset. Finally, we acknowledge that while international collaboration is crucial for an unbiased understanding of ADRD and to prevent research fragmentation, it has challenges: Integrating data across national borders requires complying with multiple regulatory standards (i.e., HIPAA and GDPR), which adds another layer of complexity.

5.2 | Our perspective

To support effective applications of AI in ADRD, multimodal data integration must be grounded in an AI-first design philosophy. Rather than treating data integration and model development as separate processes, this perspective emphasizes the need for data structures that are co-designed with the requirements of modern AI systems in mind. For instance, using technologies such as REST APIs layered on traditional SQL databases allows researchers to flexibly and securely query distinct but linked data tables with only a few SQL commands (Figure 3). A researcher interested in examining cognitive scores and MRI scans for specific diseases over time would pull together neuroimaging metrics and cognitive test results, with specified longitudinal patterns through coordinated queries across SQL databases and object stores. SQL databases efficiently store structured tabular data, while object stores can handle large files, such as neuroimaging DICOMs, actigraphy, voice recordings, and omics data. Object stores can contain raw and processed data, as well as computed embeddings. By extending concepts from the *scverse*[48] ecosystem in single-cell biology, we propose organizing participant data into richly annotated, multimodal container objects, inspired by data structures such as AnnData and MuData.[49] These objects encapsulate not only raw and processed data across modalities, including imaging, omics, and clinical, but also detailed participant- and feature-level metadata on provenance, acquisition parameters, missingness reasons, and processing methods applied, if any. This layered, machine-readable organization enhances interoperability and supports flexible subgrouping of data based on scientific questions, while also facilitating scalable, reproducible analyses and AI model development.

Consider a 72-year-old patient enrolled in a longitudinal ADRD study (Figure 4): She provides a baseline MRI, plasma biomarkers, speech samples, and intermittent actigraphy via a smartwatch but declines lumbar puncture and misses a follow-up scan. In an AI-first structure, this patient's data are organized to preserve modality-specific detail, temporal context, and reasons for missingness. As described earlier, structured tabular data reside in SQL databases, while neuroimaging DICOMs, speech recordings, and actigraphy files are stored in object stores alongside computed embeddings and processed versions. The goal is to bridge the expanding diversity of ADRD-related data with models that can operate across modalities, handle partial or evolving inputs, and generalize to real-world environments. An AI-first data structure is not merely a cleaned or standardized dataset; it is a framework for organizing, encoding, and contextualizing information in ways that directly support learning from complex, multimodal data. What follows is an outline of its key characteristics.

1. *Privacy-aware by design.* An AI-first data structure for ADRD must incorporate privacy as a foundational design principle. As multimodal data in ADRD increasingly include sensitive personal information such as neuroimaging scans, passive behavioral metrics, speech recordings, actigraphy, and genomic profiles, ensuring privacy is not just a regulatory obligation but a prerequisite for trust, participation, and long-term scalability. Unlike traditional data models that separate privacy controls from the structure of the data themselves, an AI-first approach embeds privacy, consent, and governance metadata directly into the data architecture. For instance, in a MuData container, variable (var), and observation (obs) annotation fields can carry structured indicators of sensitivity, provenance, and permissible use. Such modular, policy-aware metadata enable dynamic control over what data can be accessed, by whom, and for what purpose. This is especially important in longitudinal ADRD studies, where consent may evolve as cognitive status changes. This design also supports scalable collaboration: By enabling fine-grained access control through the API layer, privacy-aware data architectures can facilitate secure multi-institutional research. Moreover, they lay the groundwork for future integration with privacy-preserving analytic techniques such as differential privacy or federated learning, without requiring datasets to be retrofitted or re-engineered. In practice, when a researcher queries the system for a patient's actigraphy and demographics data, the system should recognize their authorization level based on their access tokens and embedded metadata to provide streamlined approval to access the data. If their level of access later changes through an Institutional Review Board amendment or the patient revoking consent to sharing actigraphy data, the metadata should be updated to reflect such changes in the resulting multimodal object.
2. *Modality-agnostic, input-flexible, and clinically interoperable.* An AI-first data structure must be designed to accommodate the inherent variability of real-world ADRD data. Rather than requiring

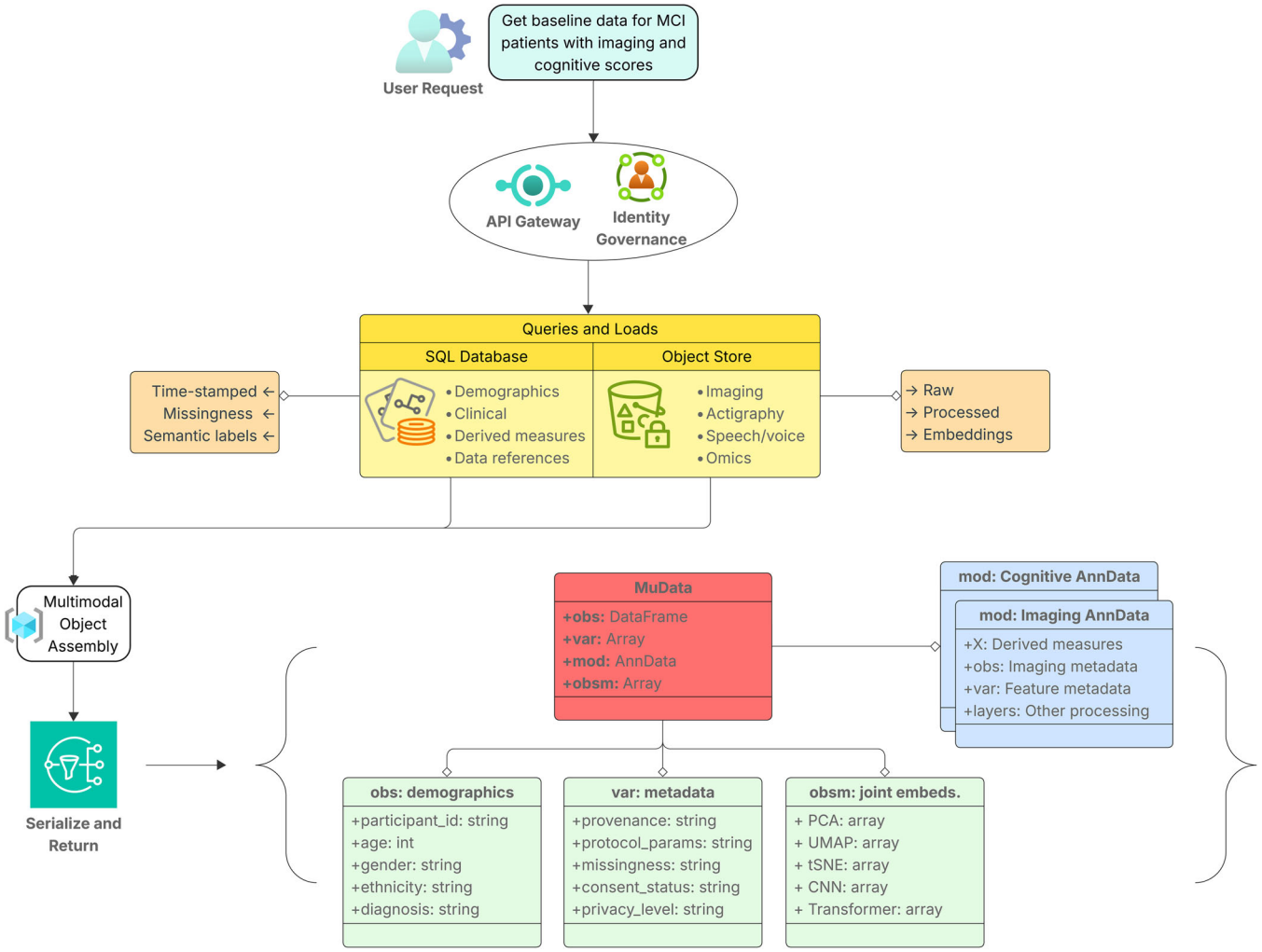


FIGURE 3 AI-first multimodal data infrastructure supporting flexible access to ADRD research data. This diagram illustrates the data flow in an AI-centric multimodal infrastructure tailored for ADRD research. User requests are routed through an API gateway with integrated identity governance for fine-grained access control and consent management. The system retrieves structured tabular data, such as demographics, clinical data, and derived imaging measures, from SQL databases and unstructured files, including imaging, actigraphy, speech, and omics sequencing, for example, from object stores, preserving temporal context, missingness metadata, and semantic annotations. Data progress through raw preparation, processed stages with algorithm application for insight extraction, and numerical embeddings (e.g., vectors for machine learning). They are then consolidated into multimodal objects like MuData, following standard conventions: participant-level identifiers in `obs`, cross-modal variable-level metadata (including provenance, protocol parameters, and privacy levels) in `var`, and joint embeddings (e.g., principal component analysis- or convolutional neural network-based representations) in `obsm`. Modality-specific AnnData objects, such as those for cognitive and imaging data, include measurements in X matrices (e.g., FreeSurfer brain volumes or cognitive test scores as derived measures), session metadata in `obs` (e.g., scan parameters, test versions), feature annotations in `var` (e.g., brain regions, test domains), and processing variants in `layers` (e.g., raw, normalized, harmonized). Final serialization outputs AI-ready formats (e.g., JSON, HDF5, Parquet) that maintain temporal alignment, explicit missingness encoding, and privacy metadata, enabling robust multimodal analysis, clinical interoperability, and scalable deployment in real-world scenarios. ADRD, Alzheimer's disease and related dementias; AI, artificial intelligence; API, Application Programming Interface.

complete, uniform data across all participants and modalities, the framework should treat each case as a flexible, modular data object with structured metadata indicating data availability and quality. In our example case, speech samples are available at irregularly spaced intervals, and the patient missed her follow-up MRI scan. A modality-agnostic and input-flexible design preserves all available data across both structured (SQL) and unstructured (object stores) components, regardless of completeness, enabling AI models leveraging these data to learn from partial, asynchronous, and population-specific combinations of inputs. Equally important is clinical interoperability. To support real-world deployment, the data structure should be compatible with standards such as HL7 FHIR or OMOP-CDM, allowing seamless integration with EHRs. This interoperability is necessary to enable AI systems to not only operate in research environments, but also at the point of care.

3. *Missingness is explicitly modeled, not hidden.* In an AI-first data structure, missing data are not treated as a nuisance to be imputed away but as a meaningful signal to be modeled. Each

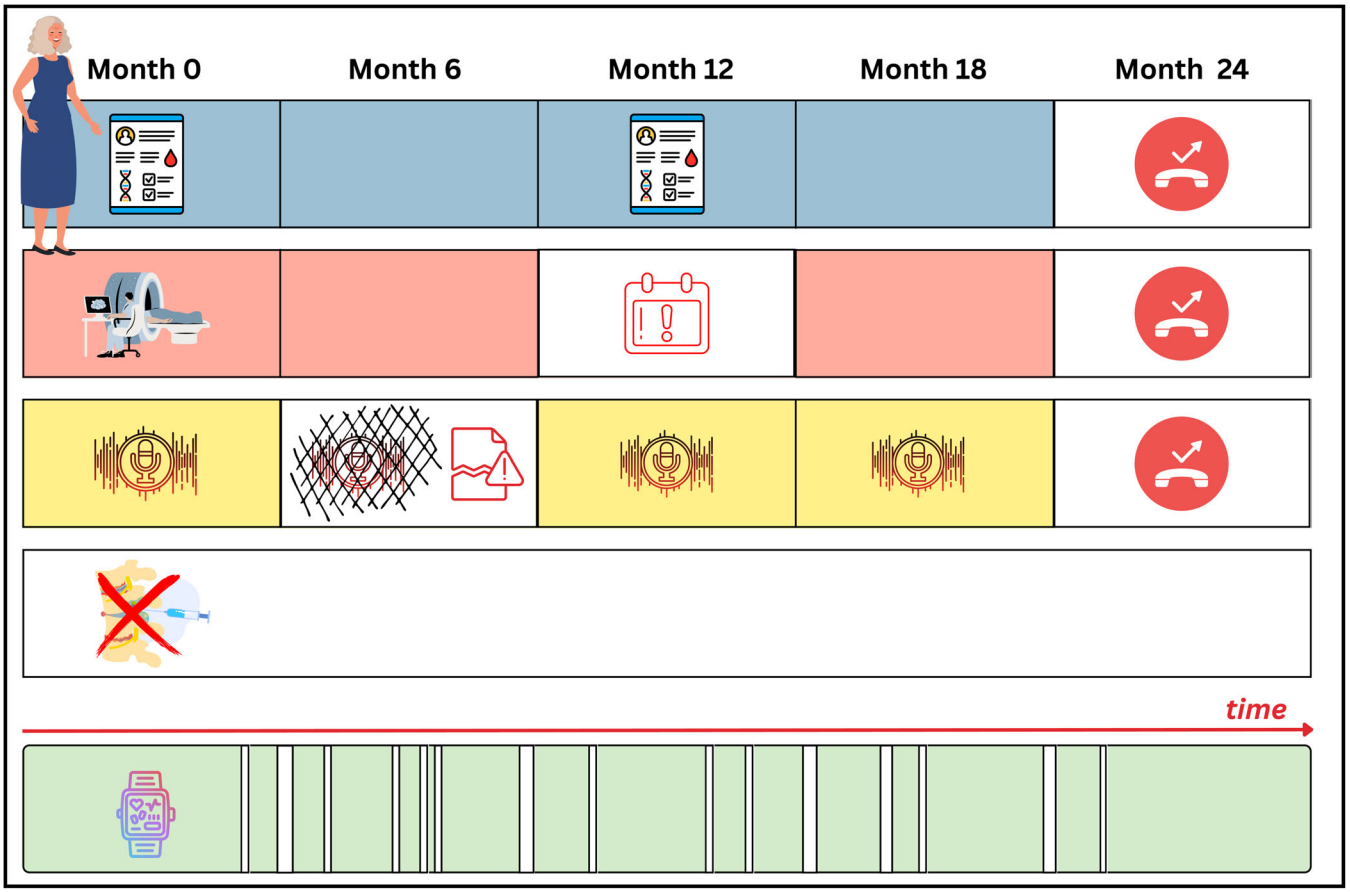


FIGURE 4 A concrete example: One participant, many modalities. This figure illustrates the longitudinal, multimodal data collected from a hypothetical 72-year-old participant in an ADRD study spanning 24 months. The participant contributes structured clinical, imaging, speech, and wearable data at various time points, along with instances of missingness common in real-world settings. At month 0 and month 12, key measures collected include demographics, APOE genotype, neuropsychological scores, and plasma biomarker levels (Aβ42/40 and p-tau217). A T1-weighted structural MRI is collected at baseline but skipped at month 12 due to a scheduling conflict. Speech samples are collected at months 0, 6, and 12, though the month 6 sample is corrupted. Daily actigraphy is captured consistently via a smartwatch with intermittent missingness, and the patient declines lumbar puncture, resulting in missing CSF data. At month 24, the participant cannot be contacted, though she continues to wear her smartwatch. In an AI-first approach, this patient is represented as a flexible multimodal object that includes structured inputs (e.g., actigraphy time series, MRI-derived features), metadata on data quality, and a missingness mask with reason codes. This representation can be passed directly into models such as multimodal transformers that learn to attend to available modalities, compensate for missing ones, and support downstream tasks like diagnosis, stratification, and progression prediction. This example underscores how an AI-first data structure enables learning from partially complete, heterogeneous data, rather than discarding them, thus improving robustness, inclusivity, and clinical relevance in ADRD modeling. ADRD, Alzheimer's disease and related dementias; AI, artificial intelligence; APOE, apolipoprotein E; CSF, cerebrospinal fluid; MRI, magnetic resonance imaging.

data field carries structured metadata indicating the reason for absence, whether due to technical failure, patient non-compliance, unavailability, or study design. This enables downstream models to distinguish between types of missingness and to apply appropriate strategies such as masked modeling, attention-based weighting, or uncertainty estimation. In the case of the 72-year-old patient introduced earlier, her data are represented as a multimodal object with machine-readable annotation layers that capture variable-level metadata about reasons for missingness, such as (`"MRI": "missing", "reason": "schedule_conflict"`) and `"CSF_biomarkers": "missing", "reason": "patient_preference"`). Temporal indices and data quality indicators, such as actigraphy sampling fidelity, are also provided in the variable-level metadata layer. Rather than being excluded, missing instances are represented appropriately, and models can be trained to attend to available inputs and learn to compensate for missing ones.

4. *The data structure preserves temporal resolution and context*. ADRD progression unfolds gradually and non-linearly, with individuals following diverse and often unpredictable trajectories. To capture this complexity, an AI-first data structure must preserve the temporal resolution and context of each data stream. It should retain timestamps for each data element and modality independently. AI models trained on this temporally aligned data can detect subtle patterns such as declining sleep quality preceding measurable cognitive decline or changes in speech coherence occurring before a

drop in test performance. Either of these may be identified in our example patient if the data are consistently timestamped using established standards, such as ISO 8601, to maximize compatibility and ease of collaboration. Such modeling is only possible if the temporal structure is preserved from the outset, allowing AI systems to reason not only about what was measured but also when and how it evolved over time.

5. *Co-designed with model architectures and task requirements.* An AI-first data structure is more than a container for multimodal inputs; it is intentionally aligned with how AI models consume, interpret, and learn from data. Each object is structured to match the input interfaces of modern model architectures and common software libraries (e.g., JSON, Apache Parquet, HDF5).[50] In addition to raw and embedded inputs, the data structure supports semantic labeling, including, for example, symptom categories or disease stages, for easier human readability and ontological alignment. Metadata such as MRI scanner strength, blood sample processing delay, or wearable device sampling rate are also encoded, enabling models to adjust for variation in data quality through uncertainty calibration or domain adaptation layers. By incorporating detailed and machine-readable annotations directly into the data structure, we minimize friction between data curation and AI deployment, supporting more robust, interpretable, and clinically useful systems.

The foregoing points are proposed as guiding principles for designing AI-first data structures, not as prescriptive solutions. The ultimate design should be co-developed through collaboration among relevant stakeholders. Although these principles can be implemented at a small scale by individual researchers, their true value emerges through upstream standardization. Ideally, large collaborative initiatives would collectively define and adopt data structures and interfaces aligned with these principles. This top-down alignment will provide clear expectations for individual researchers and facilitate broad adoption. Given the absence of a widely accepted approach to modeling such complex data, we intentionally refrain from specifying a single modeling approach, instead aiming to support a range of methods.

6 | FUTURE OPPORTUNITIES

Despite advances in data collection and AI-driven analytics for ADRD, a gap remains between research innovation and clinical implementation. Bridging this divide will require the development of a resilient, scalable data infrastructure that can address missing data, harmonize heterogeneous sources, support population diversity, and improve model interpretability. As the field shifts from controlled research environments to real-world deployment, the ability of AI systems to operate under conditions of uncertainty, partial input, and infrastructure variability becomes critical. One area of opportunity lies in leveraging existing large-scale infrastructures such as NACC. These resources already serve as cornerstones for ADRD research and could be expanded to support AI-first data integration. By layering AI-ready structures onto these initiatives, researchers can prototype and validate models that incorporate diverse data modalities while addressing real-world issues such as data sparsity and inconsistent sampling. Future systems must also be designed to integrate structured and unstructured data in ways that reflect clinical reality. AI-first data structures make this integration possible by embedding modality-specific detail, preserving temporal context, and explicitly modeling missingness. These capabilities enable AI models to operate flexibly, learning from what is available, down-weighting what is unreliable, and calibrating for what is missing, while still delivering clinically relevant insights. Importantly, such systems must be adaptable enough to scale from high-resource research environments to more variable settings such as primary care clinics. Equally important is the need for reproducibility, transparency, and alignment with FAIR data principles.[42,43] Incorporating these practices into ADRD research infrastructure is essential for fostering collaboration, regulatory compliance, and long-term sustainability. By grounding future efforts in robust data infrastructures, AI-aligned data design, and open science principles, the field is well positioned to accelerate discovery, personalize treatment, and extend the reach of ADRD care to more diverse and underserved populations.

7 | CONCLUSION

As the volume and complexity of ADRD data continue to grow, so does the urgency of building systems capable of integrating this information in meaningful ways. Existing data infrastructure efforts have laid a strong foundation for harmonizing and analyzing multimodal data, especially within well-curated cohorts. Building on this progress, we offer a complementary perspective that aligns data structure design with the unique demands of modern AI systems. An AI-first approach treats missingness as informative, embeds privacy by design, preserves temporal context, and supports direct ingestion of heterogeneous inputs. It moves beyond curated datasets to embrace the full spectrum of patient experiences, including those with partial or asynchronous data. By structuring data for how AI learns, we can build more robust, inclusive, and clinically useful tools capable of supporting diagnosis, monitoring, and care across diverse real-world settings.

ACKNOWLEDGMENTS

This project was supported by grants from the National Institute on Aging's Artificial Intelligence and Technology Collaboratories (P30-AG073104, P30-AG073105), the American Heart Association (20SFRN35460031), Gates Ventures, and the National Institutes of Health (R01-HL159620, R01-AG083735, R01-AG062109, and R01-NS142076).

CONFLICT OF INTEREST STATEMENT

V.B.K. is a co-founder and equity holder of deepPath Inc. and Cognimark, Inc. He also serves on the scientific advisory board of Altoida Inc. The remaining authors declare no competing interests. Author disclosures are available in the supporting information.

REFERENCES

1. Grande G, Valletta M, Rizzuto D, et al. Blood-based biomarkers of Alzheimer's disease and incident dementia in the community. *Nat Med*. 2025;31:2027-2035.
2. Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *NPJ Digit Med*. 2019;2:9.
3. Ashman JJ, Santo L, Okeyode T. *Characteristics of Office-based Physician Visits by Age, 2019*. National Health Statistics Reports; 2023:184.
4. Xue C, Kowshik SS, Lteif D, et al. AI-based differential diagnosis of dementia etiologies on multimodal data. *Nat Med*. 2024;30:2977-2989.
5. Frisoni GB, Festari C, Massa F, et al. European intersocietal recommendations for the biomarker-based diagnosis of neurocognitive disorders. *Lancet Neurol*. 2024;23:302-312.
6. Beekly DL, Ramos EM, Lee WW, et al. The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. *Alzheimer Dis Assoc Disord*. 2007;21:249-258.
7. Beekly DL, Ramos EM, van Belle G, et al. The National Alzheimer's Coordinating Center (NACC) Database: an Alzheimer disease database. *Alzheimer Dis Assoc Disord*. 2004;18:270-277.
8. Besser L, Kukull W, Knopman DS, et al. Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. *Alzheimer Dis Assoc Disord*. 2018;32:351-358.
9. Yang J, Ang TFA, Lu S, et al. Establishing cognitive baseline in three generations: Framingham heart study. *Alzheimers Dement*. 2023;15:e12416.
10. Jack CR Jr, Barnes J, Bernstein MA, et al. Magnetic resonance imaging in Alzheimer's Disease Neuroimaging Initiative 2. *Alzheimers Dement*. 2015;11:740-756.
11. Mueller SG, Weiner MW, Thal LJ, et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am*. 2005;15:869-877, xi-xii.
12. Mueller SG, Weiner MW, Thal LJ, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement*. 2005;1:55-66.
13. Márquez F, Yassa MA. Neuroimaging biomarkers for Alzheimer's disease. *Mole Neurodegen*. 2019;14:21.
14. Blennow K, Hampel H, Weiner M, Zetterberg H. Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nat Rev Neurol*. 2010;6:131-144.
15. Teunissen CE, Verberk IMW, Thijssen EH, et al. Blood-based biomarkers for Alzheimer's disease: towards clinical implementation. *Lancet Neurol*. 2022;21:66-77.
16. Bellenguez C, Küçükali F, Jansen IE, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet*. 2022;54:412-436.
17. Karjadi C, Xue C, Cordella C, et al. Fusion of low-level descriptors of digital voice recordings for dementia assessment. *J Alzheimers Dis*. 2023;96:507-514.
18. Xue C, Karjadi C, Paschalidis IC, Au R, Kolachalama VB. Detection of dementia on voice recordings using deep learning: a Framingham Heart Study. *Alzheimers Res Ther*. 2021;13:146.
19. Tsoy E, Zygouris S, Possin KL. Current state of self-administered brief computerized cognitive assessments for detection of cognitive disorders in older adults: a systematic review. *J. Prevent. Alzheimers Dis*. 2021;8:267-276.
20. Polk SE, Öhman F, Hassenstab J, et al. A scoping review of remote and unsupervised digital cognitive assessments in preclinical Alzheimer's disease. *NPJ Digital Med*. 2025;8:266.
21. Hardy SE, Allore H, Studenski SA. Missing Data: a special challenge in aging research. *J Am Geriatr Soc*. 2009;57:722-729.
22. Orlhac F, Eertink JJ, Cottereau A-S, et al. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J Nucl Med*. 2022;63:172-179.
23. Klunk WE, Koeppe RA, Price JC, et al. The Centiloid Project: standardizing quantitative amyloid plaque estimation by PET. *Alzheimer's & Dementia*. 2014;11.
24. Villemagne VL, Leuzy A, Bohorquez SS, et al. CenTauR: toward a universal scale and masks for standardizing tau imaging studies. *Alzheimers Dement*. 2023;15:e12454.
25. Giangrande C, Delatour V, Andreasson U, Blennow K, Gobom J, Zetterberg H. Harmonization and standardization of biofluid-based biomarker measurements for AT(N) classification in Alzheimer's disease. *Alzheimers Dement*. 2023;15:e12465.
26. Germine L, Reinecke K, Chaytor NS. Digital neuropsychology: challenges and opportunities at the intersection of science and software. *Clin Neuropsychol*. 2019;33:271-286.
27. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118-127.
28. Ashburner J, Friston KJ. Voxel-Based Morphometry—the methods. *Neuroimage*. 2000;11:805-821.
29. Fischl B. FreeSurfer. *Neuroimage*. 2012;62:774-781.
30. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nat Rev Meth Prime*. 2021;1:59.
31. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15:2759-2772.
32. Mukherjee S, Choi S-E, Lee ML, et al. Cognitive domain harmonization and cocalibration in studies of older adults. *Neuropsychology*. 2023;37:409-423.
33. Myszczynska MA, Ojamies PN, Lacoste AMB, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol*. 2020;16:440-456.
34. Beach TG, Monsell SE, Phillips LE, Kukull W. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. *J Neuropathol Exp Neurol*. 2012;71:266-273.
35. Young AL, Marinescu RV, Oxtoby NP, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun*. 2018;9:4273.
36. Rahimi J, Kovacs GG. Prevalence of mixed pathologies in the aging brain. *Alzheimers Res Ther*. 2014;6:82.
37. Wang D, Honnorat N, Toledo JB, et al. Deep learning reveals pathology-confirmed neuroimaging signatures in Alzheimer's, vascular and Lewy body dementias. *Brain*. 2025;148:1963-1977.
38. Romano MF, Zhou X, Balachandra AR, et al. Deep learning for risk-based stratification of cognitively impaired individuals. *iScience*. 2023;26:107522.
39. Qiu S, Miller MI, Joshi PS, et al. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nat Commun*. 2022;13:3404.
40. Qiu S, Joshi PS, Miller MI, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain*. 2020;143:1920-1933.
41. Swanson K, Wu W, Bulaong NL, Pak JE, Zou J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*. 2025. https://www.nature.com/articles/s41586-025-09442-9#article-info
42. Barker M, Chue Hong NP, Katz DS, et al. Introducing the FAIR Principles for research software. *Sci Data*. 2022;9:622.
43. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
44. Biedermann P, Ong R, Davydov A, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. *BMC Med Res Method*. 2021;21:238.
45. El-Sappagh S, Franda F, Ali F, Kwak K-S. SNOMED CT standard ontology based on the ontology for general medical science. *BMC Med Inf Decis Making*. 2018;18:76.

46. Baenziger J, Hutchins K, Tullis A, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem*. 1996;42:81-90.
47. Gorgolewski KJ, Auer T, Calhoun VD, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data*. 2016;3:160044.
48. Virshup I, Bredikhin D, Heumos L, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat Biotechnol*. 2023;41:604-606.
49. Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. anndata: access and store annotated data matrices. *J Open Source Software*. 2024;9:4371.
50. Wickham H. Tidy Data. *J Stat Softw*. 2014;59:1-23.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jasodanand VH, Bellitti M, Kolachalama VB. An AI-first framework for multimodal data in Alzheimer's disease and related dementias. *Alzheimer's Dement*. 2025;21:e70719. https://doi.org/10.1002/alz.70719