ELSEVIER

ISSN-L 2274-5807 – eISSN 2426-0266

JPAD

THE JOURNAL OF PREVENTION OF ALZHEIMER'S DISEASE

Special Article Collection

Leveraging Advances in Artificial Intelligence to Accelerate Alzheimer's Disease Research

CTAD 2025
Clinical Trials on Alzheimer's Disease

vol. 13 - 2026

https://www.sciencedirect.com/journal/the-journal-of-prevention-of-alzheimers-disease

JPAD
The Journal of Prevention of Alzheimer's Disease©

VOLUME 13, NUMBER 1, 2026

www.jpreventionalzheimer.com

Impact factor: 7.8 (2024)

The Journal of Prevention of Alzheimer's Disease is indexed and abstracted in the Directory of Open Access Journals (DOAJ), PubMed Medline and PubMed Central (PMC), the Science Citation Index Expanded (SCIE), the Essential Science Indicators, Scopus, SCImago Journal Rank (SJR), SNIP, and Embase.

This edition of JPAD The Journal of Prevention of Alzheimer's Disease© is published by Serdi Publisher
- Submit a manuscript:
https://www.editorialmanager.com/tjpad
- About the Journal:
https://www.sciencedirect.com/journal/the-journal-of-prevention-of-alzheimers-disease

SERDI Publisher cannot be held responsible for errors or for any consequences arising from the use of the information contained in this journal.

Instruction for Authors available on the journal website (www.editorialmanager.com/tjpad)
For the actual version of record please always check the online version of the publication.

ISSN 2274-5807

The Journal of Prevention of Alzheimer's Disease

JPAD

Special Article Collection

Leveraging Advances in Artificial Intelligence to Accelerate Alzheimer's Disease Research

- **Editorial**

- **Data Sharing, Research, and Models**

- **Diagnostics**

- **Upstream Drug Discovery**

- **Clinical Trials**

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Editorial

What can artificial intelligence bring to Alzheimer's disease clinical trials? A first perspective

Recent pharmaceutical advances and the development of sensitive and specific blood-based biomarkers are reinvigorating the field of Alzheimer's disease (AD) and related dementias (ADRD). The pharmaceutical research and development pipeline is expanding rapidly, encompassing both traditional targets, namely, the amyloid and tau pathways of AD, and a growing array of candidate pathways and novel mechanisms [1].

Concurrently, the repertoire of blood-based biomarkers continues to evolve. Emerging from diverse AT[N] assay platforms and both targeted and untargeted multi-omics technologies, these biomarkers hold the promise of enhancing the accuracy and precision with which the biological and clinical underpinnings of ADRD—and their potential subtypes—are characterized. Such progress represents an essential step towards the realization of evidence-based precision medicine. These developments pave the way for innovative preventative and therapeutic strategies, including combination therapies that have already demonstrated substantial benefit in the HIV/AIDS pandemic, as well as cancer and other complex, multifactorial diseases.

In parallel, several international data-sharing initiatives, such as the Global Neurodegeneration Proteomics Consortium, exemplify the commitment of key observational and interventional cohort studies to harmonize and make state-of-the-art datasets accessible to the broader scientific community. Such collaborations are critical to accelerating discovery and translation in the ongoing effort to address the significant challenges posed by ADRD and other neurodegenerative diseases.

At this juncture, there is increasing recognition across the AD research and clinical community—including academia, industry, and healthcare—of the transformative, multi-dimensional potential of artificial intelligence (AI) in discovery research and clinical development. AI offers powerful tools to enhance literature review processes, facilitate data harmonization, extract meaningful insights from high-dimensional digital and biomarker data, and aid in data interpretation. Moreover, AI holds promise in optimizing patient stratification and accelerating recruitment within clinical trials. These topics, alongside a critical appraisal of the potential risks and limitations of AI applications, are addressed throughout this special issue.

With these considerations in mind, we would like to express, on behalf of the Editorial Board, our sincere gratitude to the authors who have contributed to this special issue. Their diverse yet complementary perspectives provide valuable insights into how AI can advance disease understanding, refine diagnostic precision, and enable the development of effective preventative and therapeutic interventions. Collectively, these contributions mark a significant step forward as we enter a new era of research and clinical innovation in Alzheimer's disease and related dementias.

The Editors are in agreement with the view expressed by Moore [2] et al. that "AI is not an autonomous solution, but rather a powerful amplifier and accelerator of human expertise and its greatest value will come from fusing computational power with the insight, creativity, and compassion of the scientific and medical community".

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Cummings JC, Zhou Y, Lee G, Zhong K, Fonseca J, Leisgang-Osse AM, Cheng F. Alzheimer's disease drug development pipeline: 2025. Alzheimers Dement (N Y) 2025;11(2):e70098. https://doi.org/10.1002/trc2.70098. Jun 3eCollection 2025 Apr-Jun. PMID: 40463637. PMCID: PMC12131090.

[2] Moore GJ, Bose Niranjan, Manji HK, Reiman Eric M, Sperling Reisa. Editorial: artificial Intelligence and the acceleration of Alzheimer's research: from promise to practice. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100421.

Lefkos T Middleton[a,*], Sandrine Andrieu[b]

[a] *Imperial College London, London, United Kingdom*

[b] *IHU HealthAge, Toulouse III University-Paul Sabatier, INSERM CERPOP, Toulouse, France*

[*] Corresponding author.

E-mail address: l.middleton@imperial.ac.uk (L. T Middleton).

https://doi.org/10.1016/j.tjpad.2025.100423

Received 23 October 2025; Accepted 24 October 2025

The Journal of Prevention of Alzheimer's Disease 13 (2026) 100421

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Editorial

Artificial intelligence and the acceleration of Alzheimer's research - From promise to practice

Alzheimer's disease (AD) remains one of the most formidable challenges in modern medicine. Despite extraordinary advances in molecular neuroscience, imaging, and biomarker science, therapeutic progress has been painfully slow. Drug development timelines remain long, clinical trial costs remain high, and millions of patients and families continue to face the devastating impact of a disease for which cures remain elusive. Without new solutions, the global prevalence of AD and related dementias is projected to triple in the next quarter century [1], which makes the urgency of new solutions undeniable.

In April of this year, several of us contributed to a *Nature Medicine* perspective [2] outlining a call to action: artificial intelligence (AI) offers the potential to overcome entrenched bottlenecks in AD research and accelerate the path to effective prevention and treatment. This special issue of JPAD represents the next step in that journey. Across its pages, leading scientists and clinicians from around the world illustrate how AI is already reshaping the landscape of AD research — from biomarker discovery to trial innovation — and offer a glimpse of what lies ahead.

1. Biomarkers and early detection

One of the clearest opportunities for AI is in the development of scalable biomarkers for early detection. By leveraging multimodal data streams — from imaging and fluid biomarkers to speech and digital phenotyping — AI can detect subtle patterns that elude conventional analyses. In this issue, Wang et al.[3] illustrate how AI-enabled speech analysis can identify prodromal cognitive decline,while Au et al.[4] propose re-envisioning the widely adopted A–T–N framework [5] to integrate digital and AI-derived measures. These contributions exemplify how the combination of machine learning and novel data modalities can shift us toward earlier, more precise detection and stratification — a critical step for prevention trials and clinical care alike. Beyond detection, AI also offers opportunities to uncover causal and modifiable risk and resilience factors — from genetics to environmental exposures — which can serve as targets for risk-reduction and resilience-building strategies in prevention trials.

2. Drug discovery and knowledge integration

Equally transformative is AI's role in therapeutic discovery. The explosion of omics data, neuroimaging results, and real-world clinical observations has created a landscape that is too vast for any human researcher to navigate. AI systems can now synthesize these complex datasets into evolving models of AD biology. In this issue, Wittenberg et al. [6] describe how Big Data and AI are accelerating drug discovery pipelines, while Funk et al. [7] demonstrate how machine learning can build coherent, integrative models from noisy and even contradictory findings. Extending this vision, Roberts and Landsness et al. [8] advance the concept of an "AI biomedical scientist assistant" — a partner that augments human creativity in hypothesis generation, experimental design, and data interpretation. Collectively, these advances suggest a future in which AI not only accelerates discovery but fundamentally reshapes how we think about biomedical science.

3. Transforming clinical trials

Clinical trials remain among the most time- and cost-intensive aspects of AD research. Here, too, AI is beginning to make inroads. Yigamawano et al.[9] and Welchman & Kourtzi [10] describe how advanced machine learning methods can improve patient recruitment and stratification, reducing attrition and enhancing trial efficiency. Complementing these strategies, digital twin models — highlighted across multiple contributions in this issue — offer the potential to simulate disease trajectories and treatment responses before interventions are tested in vivo. In parallel, contemporaneous work outside of this special issue, such as Devanarayan et al. [11], has demonstrated that multimodal prognostic modeling of individual cognitive trajectories can substantially improve efficiency in prevention trials, with the potential to reduce required sample sizes by more than one-third. Together, these approaches point to a future in which both trial enrollment and trial durations are shorter, more predictive, and more patient centered. Together, these approaches point to a future in which both trial enrollment and trial durations are shorter, more predictive, and more patient centered.

4. Ethics, equity, and data sharing

The promise of AI will not be realized without attention to its risks. AI systems reflect the data on which they are trained, raising the specter of bias, inequity, and limited generalizability. Kolachalama et al. [12] remind us that reproducibility and fairness must remain central priorities, urging safeguards to prevent AI from amplifying existing disparities. Building on this, Adams et al. [13] demonstrate how large language models can be harnessed for semantic harmonization across diverse Alzheimer's cohorts, showing that harmonization is not just a technical advance but a cornerstone of equitable and scalable data use. More broadly, the contributions in this issue emphasize that success will depend on data-sharing frameworks that are both privacy-preserving and globally inclusive. Without such frameworks — and the diverse datasets and international collaboration they enable — AI-driven tools risk serving only a fraction of the world's patients and face the peril of overlooking critical insights from global populations.

https://doi.org/10.1016/j.tjpad.2025.100421
Received 9 September 2025; Accepted 22 October 2025

5. Looking ahead

The contributions in this special issue reflect both momentum and responsibility. The momentum is evident in the rapid emergence of AI applications across biomarker science, therapeutic discovery, and clinical research. The responsibility lies in ensuring that these advances are rigorously validated, ethically grounded, and equitably distributed.

The convergence of AI and neuroscience represents a pivotal opportunity to redefine how we study, diagnose, and ultimately treat Alzheimer's disease. Realizing this potential will also require careful attention to how these tools are integrated into clinical practice — from interoperability with health records to clinician and patient engagement, and the development of appropriate regulatory frameworks. AI is not an autonomous solution, but rather a powerful amplifier and accelerator of human expertise. Its greatest value will come from fusing computational power with the insight, creativity, and compassion of the scientific and medical community.

This issue of JPAD articulates the first wave of AI's impact in AD research. It also makes plain the imperative: to scale these approaches, sustain them across global contexts, and ensure they address the urgent needs of patients and families. With shared purpose and deliberate action, the field can seize this opportunity to move from incremental progress to transformative change — and ultimately to a future in which Alzheimer's is preventable and curable.

Declaration of generative AI and AI-assisted technologies in the writing process

The manuscript was originally drafted without the use of AI technologies; and subsequently AI (ChatGPT) was used to improve readability and formatting of the manuscript and its associated references. Subsequently the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Gregory J. Moore: Conceptualization, Writing – original draft. **Niranjan Bose:** Writing – review & editing. **Husseini K. Manji:** Writing – review & editing. **Eric M. Reiman:** Writing – review & editing. **Reisa Sperling:** Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gregory J. Moore MD, PhD reports financial support was provided by Gates Ventures. Gregory J. Moore MD, PhD reports a relationship with Gates Ventures that includes: consulting or advisory. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Nichols E, Steinmetz JD, Vollset SE, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022;7(2):e105–25. https://doi.org/10.1016/s2468-2667(21)00249-8.

[2] Andrieu S, Bateman RJ, Bereczki E, et al. Harnessing artificial intelligence to transform Alzheimer's disease research. Nat Med 2025;31:1384–5. https://doi.org/10.1038/s41591-025-03632-8.

[3] Wang L, et al. Multi-Modal Data Analysis for Early Detection of Alzheimer's Disease and Related Dementias. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100399.

[4] Au R, et al. Reinventing "N" in the A/T/N Framework: The Case for Digital. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100395.

[5] Jack Jr CR, Bennett DA, Blennow K, et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. JAMA 2018;319(13):1225–34. https://doi.org/10.1016/j.jalz.2018.02.018.

[6] Wittenberg G, Elwood F, Houghton A, et al. The Evolution of Alzheimer's Target Identification: Towards a Fusion of Artificial and Cellular Intelligence. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100417.

[7] Funk C, et al. Mining the Gaps: Deciphering Alzheimer's Biology through AI-Driven Reconciliation. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100402.

[8] Roberts KF, Landsness E, et al. Towards an AI Biomedical Scientist: Accelerating Discoveries in Neurodegenerative Disease. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100398.

[9] Yigamawano FK, Odom AR, Xue C, et al. AI-Augmented Frameworks for Enhancing Alzheimer's Disease Clinical Trials: A Memory Clinic Perspective. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100396.

[10] Welchman AE, Kourtzi Z. Solving the Goldilocks Problem in Dementia Clinical Trials with Multimodal AI. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100397.

[11] Devanarayan V, et al. Multimodal prognostic modeling of individual cognitive trajectories to enhance trial efficiency in preclinical Alzheimer's disease. Alzheimers Dement 2025;21(9):e70702. https://doi.org/10.1002/alz.70702. Sep.

[12] Kolachalama VB, Sureshkumar V, Au R. AI models, bias, and data sharing efforts to tackle Alzheimer's disease and related dementias. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100400.

Adams K, Salimi Y, Can AyM, et al. A Benchmark of Text Embedding Models for Semantic Harmonization of Alzheimer's Disease Cohorts. J Prev Alzheimers Dis 2026. https://doi.org/10.1016/j.tjpad.2025.100420.

Gregory J. Moore[*a], Niranjan Bose[b], Husseini K. Manjic[c], Eric M. Reiman[d], Reisa Sperling[e]

[a] Gates Ventures, Kirkland, USA

[b] Alzheimer's Disease Data Initiative, Kirkland, USA

[c] Oxford University, Oxford, United Kingdom

[d] Banner Health Alzheimer's Institute, Tucson, USA

[e] Massachusetts General Hospital and Harvard Medical School, Boston, USA

[*] Corresponding author.
E-mail address: prof.neuro@gmail.com (G.J. Moore).

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

A benchmark of text embedding models for semantic harmonization of Alzheimer's disease cohorts

Tim Adams [a,1], Yasamin Salimi [a,1], Mehmet Can Ay [a], Diego Valderrama [a], Marc Jacobs [a], Holger Fröhlich [a,b,c,*]

[a] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt, Augustin, 53757, Germany
[b] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
[c] Institute for Digital Medicine, University Hospital Bonn, Bonn, Germany

ARTICLE INFO

Keywords:
Harmonization
Alzheimer's disease
Text-embeddings
Large language models

ABSTRACT

Background: Harmonizing diverse healthcare datasets is a challenging task due to inconsistent naming conventions. Manual harmonization is time- and resource-intensive, limiting scalability for multi-cohort Alzheimer's Disease research. Large Language Models, or specifically text-embedding models, offer a promising solution, but their rapid development necessitates continuous, domain-specific benchmarking, especially since general established benchmarks lack clinical data harmonization use cases.
Objectives: To evaluate how different text-embedding models perform for the harmonization of clinical variables.
Design and setting: We created a novel benchmark to assess how well different Language Model embeddings can be used to harmonize cohort study metadata with an in-house Common Data Model that includes cohort-to-cohort mappings for a wide range of Alzheimer's Disease cohorts. We evaluated five different state-of-the-art text embedding models for seven different data sets in the context of Alzheimer's disease.
Participants: No patient data were utilized for any of the analyses, as the evaluation was based on semantic harmonization of cohort metadata only.
Measurements: Text descriptions of variables from different modalities were included for the analyses, namely clinical, lifestyle, demographics, and imaging.
Results: Our benchmark results favored different models compared to general-purpose benchmarks. This suggests that models fine-tuned for generic tasks may not translate well to real-world data harmonization, particularly in Alzheimer's disease. We propose guidelines to format metadata to facilitate manual or model-assisted data harmonization. We introduce an open-source library (https://github.com/SCAI-BIO/ADHTEB) and an interactive leaderboard (https://adhteb.scai.fraunhofer.de) to aid future model benchmarking.
Conclusions: Our findings highlight the importance of domain-specific benchmarks for clinical data harmonization in the field of Alzheimer's disease and motivate standards for naming conventions that may support semi-automated mapping applications in the future.

1. Introduction

As data availability in healthcare continues to expand, so does access to diverse, large-scale datasets collected across various institutions and populations. This growing wealth of information presents a unique opportunity to advance data-driven research and improve clinical decision-making. Published cohort studies, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [1], have been shown to accelerate research [2] and enhance our understanding of disease progression by providing open access to high-quality, longitudinal data. However, individual cohorts are often biased toward specific demographic, geographic, or clinical characteristics. This bias can limit generalizability and reduce predictive model robustness in broader populations [3].

Training predictive models across multiple, heterogeneous cohorts has the potential to mitigate these limitations by increasing sample

[*] Corresponding author at: Schloß Birlinghoven, Sankt, Augustin, 53757, Germany.
E-mail address: holger.froehlich@scai.fraunhofer.de (H. Fröhlich).
[1] Equal contribution.

https://doi.org/10.1016/j.tjpad.2025.100420
Received 3 September 2025; Received in revised form 20 October 2025; Accepted 22 October 2025

diversity and improving generalizability across populations [4]. Approaches such as Federated Learning (FL) offer a promising solution by enabling collaborative model training across distributed datasets without requiring data to be centrally aggregated [5]. However, FL requires a common data model (CDM) to which all local cohorts are harmonized.

Harmonizing multiple cohorts is an ongoing struggle: different cohorts, even when recording the same variables, rarely adhere to standard naming conventions, which often requires substantial effort to harmonize to a uniform standard manually. Such manual curation is time and resource-intensive and requires domain-specific experts to ensure semantic consistency and accuracy. This bottleneck not only slows down research workflows but also introduces variability depending on the expertise and interpretations of individual curators.

Numerous initiatives have been undertaken to address data harmonization, often through a CDM or standard variable schema. For instance, the Alzheimer's Disease Data Initiative offers a standard variable system comprising 124 common variables within Alzheimer's disease (AD) cohorts. Using these variable mappings, users can harmonize cohort variables to the standard term for performing cross-cohort investigations [6]. Similarly, tranSMART is an open-source data warehouse and analytics platform that supports integration, harmonization, and analysis of translational research data using a CDM and controlled vocabularies [7]. More recent work has introduced AD-specific harmonization frameworks. AD-Mapper builds a CDM from 20 AD cohorts complemented by external CDMs and ontologies, totaling over 1200 reference variables; it leverages a BioBERT-based model to map new cohort variables and demonstrates performance improvements over simple string matching [8]. Another study harmonized ADNI and National Alzheimer's Coordinating Center (NACC) datasets via the Alzheimer's Disease Element Ontology (ADEO) to enable unified data element definitions and cross-cohort querying [9]. In addition, a Dutch consortium working with nine dementia cohorts applied an ETL pipeline to map local datasets into the OMOP CDM under a federated learning setup, reporting substantial benefits but also highlighting challenges with cohort-specific fields and vocabulary mismatches [10]. These approaches typically involve some degree of manual curation by the users and require an established user profile prior to harmonization. While valuable, the manual curation could potentially be further expedited through artificial intelligence (AI).

Recent advancements in language processing - particularly the rapid development and continuous improvement of large language models (LLMs) - may offer a promising solution to the challenge of labor-intensive manual curation. The application of LLM or transformer-based text embeddings for harmonization tasks has gained growing attention and has shown promising results in recent studies [11–15].

The development of new and improved LLM-based text embedding models in this field is rapid - new LLMs are published and released almost monthly. Continuous benchmarking of such models is therefore essential to ensure the best possible performance of applications that utilize them.

One of the largest and most comprehensive benchmarks to assess the performance of such embedding models is the Massive Text Embedding Benchmark (MTEB) [16]. While this benchmark includes a wide range of classification, clustering, and ranking tasks for a vast number of languages and diverse text sources, it does not cover tasks involving the automatic harmonization of clinical data. Clinical data descriptions pose a unique set of challenges due to their diverse and highly specialized terminology and vocabulary, which makes clinical data harmonization particularly complex. Accordingly, it is not clear whether LLMs performing well for various generic tasks outside the medical domain are well-suited for the harmonization of clinical data.

To fill this gap, we developed a specialized benchmark for multi-cohort variable alignment in the domain of AD, which constitutes the core contribution of this work. We evaluate five state-of-the-art text embedding models that are currently high ranking in the general MTEB embedding benchmarks, using a custom benchmark consisting of seven different AD cohorts that we map to a previously established AD CDM [8]. We discuss the challenges and limitations of automated harmonization in this domain and propose a framework of rules for clinical study metadata to guide and standardize future harmonization efforts, aiming to improve consistency, interoperability, and the quality of integrated clinical data across AD research studies.

2. Methods

To assess the feasibility of harmonizing variables across cohorts, we collected metadata (i.e., data dictionaries), consisting of variable names and variable descriptions. This was done for seven different cohort datasets in the context of AD. We evaluated five of the currently best-performing language models to benchmark their ability to automatically match cohort variable descriptions to a ground-truth description provided in a manually curated CDM based on their semantic similarity. The general approach to how matches were evaluated based on their similarity is shown in Fig. 1.

2.1. *Cohort data*

We collected cohort data from a total of seven different studies, which we will briefly describe in this section.

The Pre-symptomatic Evaluation of Experimental or Novel Treatments for Alzheimer's Disease (PREVENT-AD) [17] cohort study is an open science dataset that collected measurements from cognitively impaired participants with a family history of AD, particularly parents or siblings. The study resulted in the collection of five years of measurements, including imaging, cerebral fluid, genetic, and clinical information [17].

The European Medical Information Framework (EMIF) [18] cohort was established to identify noninvasive biomarkers for the diagnosis of AD. Various clinical measurements, including neuropsychological tests, medication use, and comorbidities, as well as demographic variables and clinical information, were collected [18].

The GERAS cohort studies, including GERAS I, GERAS II, GERAS JAPAN, and GERAS EU, are large, prospective, multicenter observational studies designed to evaluate the clinical, social, and economic impact of AD on patients and caregivers across Europe and Japan [19–22]. These studies collected comprehensive data on patient demographics, cognitive and functional status (e.g., Mini-Mental State Examination (MMSE)), behavioral symptoms, medication use, caregiver characteristics, healthcare resource utilization, and quality of life, with assessments conducted at baseline and multiple follow-up time points.

The PREVENT Dementia programme [23] is a multi-centre, prospective cohort study conducted across five sites in the UK and Ireland, designed to examine midlife risk factors for dementia and to identify the earliest indices of neurodegenerative disease development. The study recruited cognitively healthy participants, collecting deeply phenotyped baseline data that includes demographic information, biological samples (e.g., blood, saliva, urine, and optional cerebrospinal fluid), detailed lifestyle and psychological questionnaires, a comprehensive cognitive test battery, and multi-modal 3T MRI scans with both structural and functional sequences.

2.2. *Common data model*

To assess variable harmonization performance across various language models, we collected data dictionaries from seven different AD cohort studies. We used a previously established and publicly available AD CDM (i.e., AD-Mapper CDM), which defines 1300 core variables commonly collected in AD studies, including demographics, clinical assessments, biomarkers, and imaging data. The AD-Mapper CDM comprises the variable naming conventions of 23 distinct cohort studies that were harmonized against one another. Additionally, the CDM

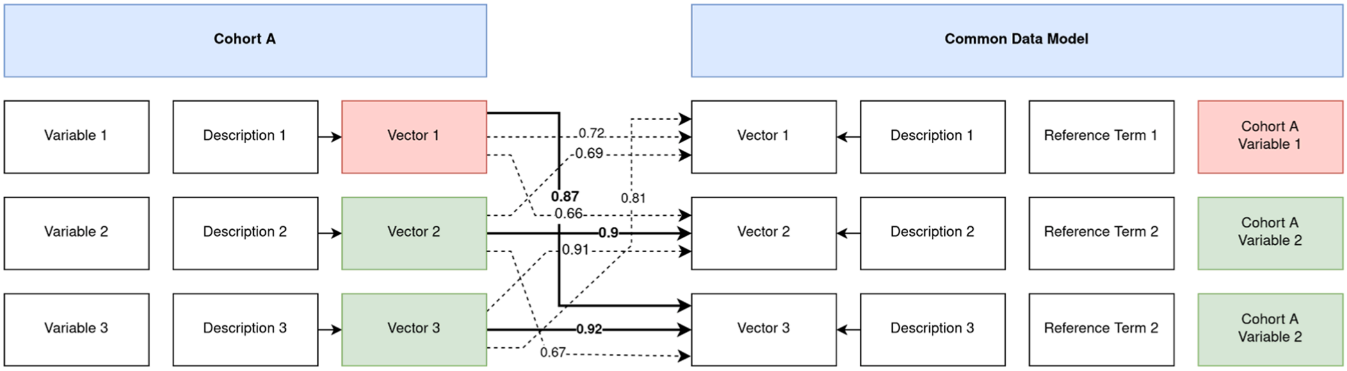


Fig. 1. Benchmarking workflow: For each variable in each cohort, as well as each reference term in the CDM, we compute vector embeddings of the respective models. After matching them based on their cosine similarities, we compare against the respective ground truth mapping in the CDM.

includes reference terms to which all variables were mapped, along with a definition for each reference term extracted from well-established ontologies, such as the Systemized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) or Logical Observation Identifiers Names and Codes (LOINC) [8].

Given that evaluating automatic harmonization requires a ground-truth, we initially manually harmonized the variables from the seven cohorts to the AD CDM. Two variables were considered a match when both the variable description and value ranges were comparable. To ensure correctness, each mapping was validated by inspecting the patient-level data and corresponding value distributions. The harmonization was carried out independently by two curators, and any discrepancies were resolved through discussion until consensus was reached.

These variable mappings were then used as ground truth for evaluating the correctness of the matches suggested by different models. Since the harmonization task relied on variable descriptions provided in the metadata, we included only variables with available descriptions, as some lacked this information. Due to a low variable count for each individual GERAS study, we combined them into a single cohort study for benchmarking, which we denote as GERAS.

To evaluate the models on independent studies that may follow entirely different naming conventions, we selected two cohorts we had harmonized in our earlier work [14]. First, we identified cohorts that had been ranked as "excellent" based on a manual assessment of their metadata quality. Previously, we implemented a three-category ranking system, namely, “poor,” “adequate,” and “excellent.” The ranking was performed by two independent curators based on the clarity and comprehensiveness of the variable descriptions. For instance, a cohort data dictionary was ranked “poor” when the descriptions did not clarify what measurement had been collected or to which modality it belonged (for example, cerebrospinal fluid (CSF) vs. blood biomarker). A dictionary was ranked “adequate” when descriptions were recorded, but lacked sufficient detail to avoid ambiguity, such as “memory score” without specifying the cognitive test used. Finally, a dictionary was ranked “excellent” when the descriptions were clear, informative, and not misleading, explicitly indicating the nature of the measurement and the modality (for example, “CSF Aβ42 concentration”).

Second, we narrowed down the selection to cohorts with a comparable number of variable mappings accompanied by available variable descriptions (i.e., between 30 and 50 variables). This selection aimed to ensure comparability between cohorts and to minimize the potential bias of one cohort's performance disproportionately influencing the results. The PREVENT-AD and EMIF cohorts met these criteria and were included in our analyses. In addition, we included the PREVENT-Dementia cohort, which had not been part of our previous metadata quality ranking, but whose variable descriptions were deemed sufficiently clear and whose variable count fell within the target range.

The selected cohorts were chosen to provide a representative benchmark for evaluating variable harmonization. They encompass a range of study designs, participant populations, and data modalities, including cognitive assessments, biomarkers, imaging, and lifestyle measures. PREVENT-AD includes participants with a family history of AD, EMIF focuses on biomarker discovery across multiple clinical variables, and PREVENT-Dementia targets cognitively healthy midlife individuals. This diversity ensures that the benchmark captures heterogeneity commonly observed in AD studies and allows the assessment of harmonization approaches across different variable naming conventions and data structures. Collectively, these cohorts provide a rigorous and generalizable benchmark for evaluating harmonization performance.

An overview of the number of included variables from each cohort is shown in the supplementary material in Figure S1.

2.3. Large language model-based variable embeddings

We evaluated five language models, three of which ranked among the top models on the MTEB benchmark as of August 2025. We additionally evaluated OpenAI's most recent model as one of the leading proprietary competitor models in the field, as well as MiniLM as the currently most widely used lightweight open-source baseline for embedding tasks. The five evaluated models, the number of parameters, and their respective ranks in the MTEB leaderboard are shown in Table 1.

For every model, we calculated vector embeddings for each cohort variable description as well as for each feature description in the CDM. Vectors were L2 normalized to unit length to ensure consistent scaling across different models. The normalized vectors were then matched based on their cosine similarity to each possible CDM vector (see Fig. 1).

We measured model performances based on:

Table 1
Evaluated text-embedding models with corresponding MTEB leaderboard rank and model parameter size. Neither OpenAI nor Google discloses the size of their models.

Model	MTEB Leaderboard Rank	Number of Parameters
Google gemini-embedding-001 [24]	1	Undisclosed
Qwen3-Embedding-8B [25]	2 (3 + 4 for smaller variants)	8B
Linq-AI-Research/Linq-Embed-Mistral [26]	5	7B
OpenAI text-embedding-3-large [27]	16	Undisclosed
all-MiniLM-L6-v2 [28]	117	22M

a) Accuracy of Zero-Shot classification, defined as the proportion of variables correctly matched to their corresponding feature in the CDM based on their highest cosine similarity (Table 2).
b) Area Under Precision Recall Curve (AUPRC) for all variable mappings depending on the similarity range of each mapping (Fig. 2, Table 2).

Precision and *Recall* were calculated for 100 thresholds of vector cosine similarities from 1.0 to 0.01. This approach yielded 100 data points for each precision and recall. Included variables that correctly matched any one-to-one or upper-level concept in the CDM were considered as *True Positives (TP)*. Although the automatic mapping procedure was performed on a one-to-one basis, the manual curation occasionally harmonized multiple cohort variables to the same reference concept. For example, for a cohort where the apolipoprotein E (APOE) genotype was recorded separately, two cohort variables were mapped from that cohort to the reference term "APOE" in the CDM. In this case, if either of those variables were mapped to the "APOE" variable, we counted that variable as a correct mapping. Variables included in the threshold that did not correctly match any CDM variable were considered *False Positives (FP)*; variables excluded due to the respective threshold that had a potential match in the CDM were considered *False Negatives (FN)*. We excluded any variables that did not have any potential match in the CDM prior to vector computation.

We computed precision and recall for each threshold using the standard formulas:

$$Precision = \frac{TP}{TP+FP} Recall = \frac{Tp}{TP+FN}$$

Table 2

AUPRC, zero-shot accuracy, and weighted composite scores for each model across cohorts. Zero-shot accuracy measures correct CDM matches without prior training. The composite score combines metrics weighted by cohort size. Cohorts in the table are ordered based on the rank in the MTEB.

Model	Cohort	AUPRC	Zero-Shot	Composite Score
Google: gemini-embedding-001	GERAS	**0.43**	0.62	0.40
	PREVENT Dementia	0.29	0.48	
	PREVENT-AD	0.28	0.28	
	EMIF	0.21	0.42	
Qwen: Qwen3-Embedding-8B	GERAS	0.29	0.46	0.30
	PREVENT Dementia	0.29	0.36	
	PREVENT-AD	0.21	0.22	
	EMIF	0.11	0.35	
Linq-AI-Research: Linq-Embed-Mistral	GERAS	0.39	0.35	0.37
	PREVENT Dementia	**0.42**	0.36	
	PREVENT-AD	**0.29**	0.25	
	EMIF	**0.35**	0.54	
OpenAI: text-embedding-3-large	GERAS	0.35	**0.66**	**0.43**
	PREVENT Dementia	0.31	**0.52**	
	PREVENT-AD	0.28	**0.39**	
	EMIF	0.32	0.52	
UKP Lab: all-MiniLM-L6-v2	GERAS	0.35	0.56	0.40
	PREVENT Dementia	0.31	0.44	
	PREVENT-AD	**0.29**	0.34	
	EMIF	0.3	**0.58**	

While a high minimum similarity threshold (i.e., only consider vectors of "perfect" similarity of 1.0) will likely result in high precision and low recall, the opposite will likely lead to high recall but lower precision. When plotting these two measures against each other, the AUPRC can be used to determine how well a model is able to match similar descriptions other than the best match, since the correct match may not always have the highest similarity score, but could still be among the most similar candidates.

To reach a statistically meaningful assessment of language model capabilities, results included a total of eight scores per model, with the two metrics introduced above for each of the four evaluated cohorts. To enable a comprehensive comparison across all metrics and cohorts, we additionally computed a weighted composite score as follows:

$$score = \sum\nolimits_{cohort}(0.5 \cdot AUPRC_{cohort} + 0.5 \cdot ZeroShotAccuracy_{cohort}) \cdot \frac{\#Vars_{cohort}}{\#Vars_{total}}$$

We assigned equal weights to AUPRC and zero-shot accuracy because they capture complementary but equally important aspects of model performance: AUPRC reflects ranking quality across different thresholds, while zero-shot accuracy measures direct classification success based on the highest similarities without threshold tuning. Giving them equal contribution ensures that the composite score balances both threshold-independent ranking and practical accuracy. Cohort-level weighting by variable count was applied to ensure that cohorts contributing more data have a proportionally larger influence on the composite score, reflecting the practical impact of model performance on the overall harmonization task. This approach prevents smaller cohorts with few variables from disproportionately affecting the composite score, which could misrepresent overall performance. The final composite score ranged from 0 to 1, and we report this score for all models in Table 2, alongside the individual AUPRC and zero-shot accuracy values.

OpenAI's *text-embedding-3-large* model produced slightly different embedding vectors across runs when retrieved via the Python API, due to floating-point precision differences. This occasionally led to flips in similarity rankings for individual comparisons. To mitigate this, we averaged the results over 10 runs. These variations affected only isolated mappings and did not influence the model's overall aggregated performance score.

3. Results

We first evaluated the total performance of each individual benchmarked model based on its combined performance across both metrics (zero-shot accuracy, AUPRC) over all cohorts using the composite score introduced in the previous section. In terms of their total score, OpenAI's *text-embedding-large* performed best with a total score of 0.43, followed closely by Google's *gemini-embedding-001* model and all-MiniLM tied with a score of 0.40 (Table 2).

Out of the open-source models, *all-MiniLM-L6-v2* performed best with a composite score of 0.40. *Linq-Embed-Mistral* followed closely with a composite score of 0.37, with only the *Qwen3-Embedding-8B* model falling short with an overall score of 0.30.

In terms of zero-shot classification performance, OpenAI's *text-embedding-large* model outperformed all other evaluated models for all evaluated cohorts except for EMIF, with a total zero-shot accuracy of 0.66 for the combined GERAS cohorts, an accuracy of 0.52 for PREVENT Dementia, and 0.39 for PREVENT-AD. For the EMIF cohort, the all-MiniLM-L6-v2 model performed best with a zero-shot accuracy of

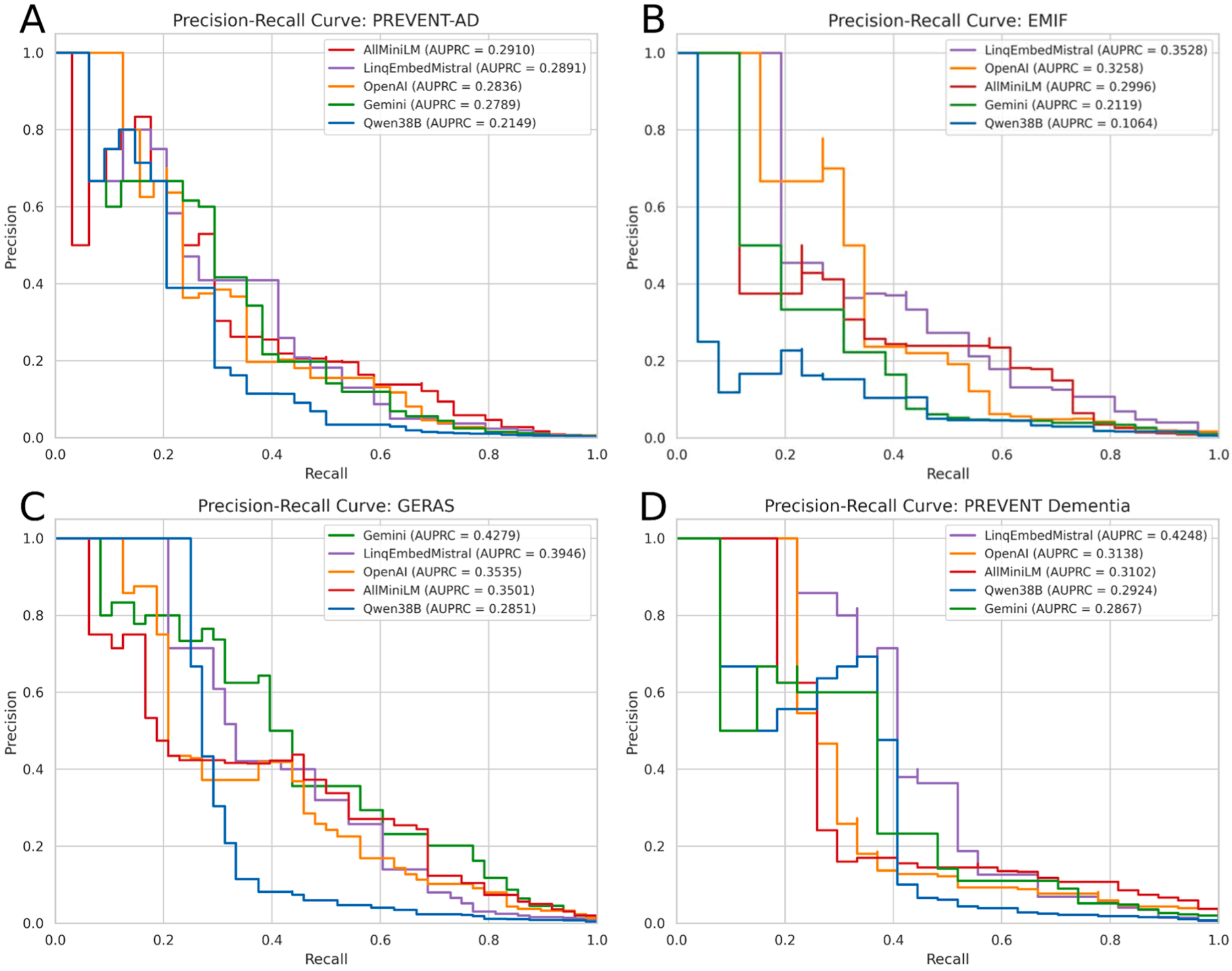


Fig. 2. Precision-Recall curves and computed AUPRC for all models per cohort. Evaluated cohorts are PREVENT-AD (A), EMIF (B), a combination of GERAS studies (C), and PREVENT Dementia (D).

0.58, closely followed by the Linq-Embed-Mistral model with an accuracy of 0.54.

Google's *gemini-embedding-001* model performed slightly worse than OpenAI's model for the first two cohorts (both −0.04). Though being a significantly smaller model (less than 1/300 of parameters) than the other two tested open-source models, as well as ranking the lowest (rank 117) on the MTEB, the all-MiniLM model outperformed both the LinqEmbedMistral and Qwen3 models for all cohorts in terms of zero-shot classification accuracy and even the leading OpenAI model specifically for the EMIF cohort.

While there were significant differences in terms of zero-shot accuracy for the individual models, there was also a high overlap in cohort variable descriptions that could not be correctly mapped by any of the evaluated models. We show a visualisation of incorrect variable overlaps for each evaluated model for each cohort in Fig. 3. For the GERAS studies, a total of 6 variables could not be correctly mapped by any model based on the most similar description vector, which corresponds to a fraction of 12.5 % of all recorded variables. Other cohorts show a bigger relative and absolute number of variables that were incorrectly matched based on the zero-shot approach. EMIF had a total of 8 or 30.8 % of variables incorrectly matched based on the most similar embedding for all models, and PREVENT Dementia had a total of 10 variables corresponding to 40 % of all variables. PREVENT-AD showed the highest misclassification rate, with 15 variables (46.9 %) mapped incorrectly by all models.

A closer inspection of these errors reveals common patterns. For example, variables related to age and time points, such as "Age when assessed" (EMIF), "Age in months at the time of test" (PREVENT-AD), or "Age of participant" (PREVENT Dementia), were frequently mismatched due to subtle differences in phrasing. Long or complex variable descriptions, such as "Long-form variable for disease code for any selected diagnoses…" (GERAS) or "Has the participant shown a low cognitive performance…determined by clinical consensus" (PREVENT-AD), also led to misclassifications by the models. In addition, modality-specific biomarkers like "Neurofilament light values of central CSF analyses" (EMIF) or "Beta-amyloid 1–42 concentration in CSF" (PREVENT-AD) were often mismatched, reflecting difficulties in distinguishing between similar biological measurements across different sample types. These examples illustrate that systematic mismatches often arise for variables with subtle wording differences, unusually detailed descriptions, or modality-specific context, highlighting the limitations of current text embedding models in capturing nuanced clinical semantics. We propose technical guidelines on data descriptions based on commonly mismatched variables in the Discussion section.

The AUPRC (Fig. 2) showed mixed results depending on both the specific models and cohorts; Google's *gemini-embedding-001* performed the best for the GERAS cohorts with an AUPRC of 0.43. For the other 3 cohorts, the Linq-Embed-Mistral model showed the highest AUPRC of

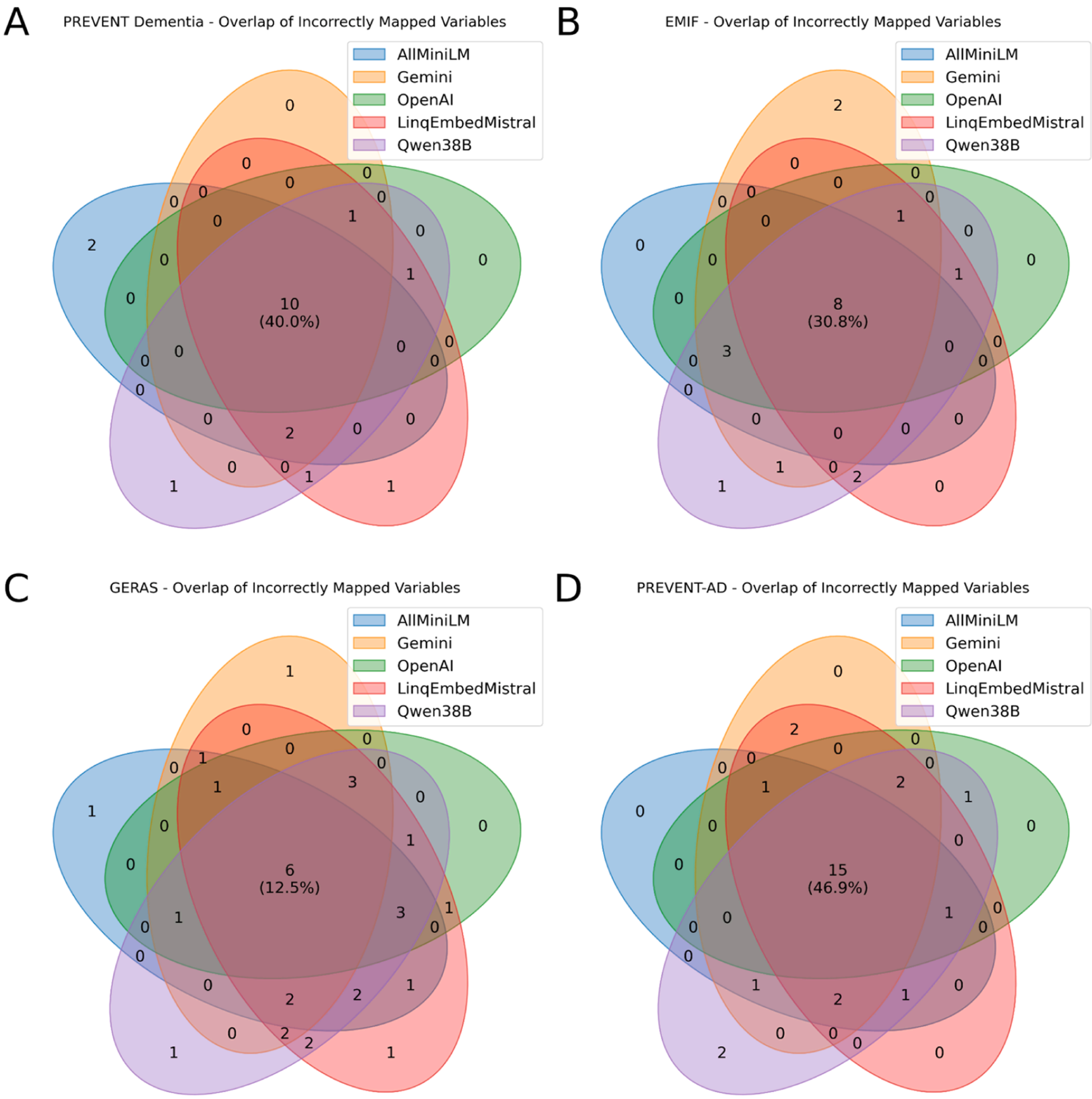


Fig. 3. Overlap of incorrectly mapped variables (zero-shot) across models for each cohort. Panels A–D correspond to the PREVENT Dementia, EMIF, GERAS, and PREVENT-AD cohorts, respectively. The Venn diagrams show the number of shared and unique mapping errors among the five models (AllMiniLM, Gemini, OpenAI, LinqEmbedMistral, Qwen38B).

0.42 for PREVENT Dementia, 0.35 for EMIF, and 0.29 for PREVENT-AD, tied with the all-MiniLM model.

As new text-embedding models with improved capabilities are emerging, with increased frequency, in the public domain, choosing the right model, especially for time and resource-intensive tasks such as data harmonization, can be challenging. To facilitate the benchmarking of new models, we provide a Python library named "Alzheimer's Disease Harmonization Text Embedding Benchmark" (ADHTEB) [29] to enable benchmarking of future models. We also publish an interactive leaderboard[2] showcasing top-performing models. Users can benchmark their own custom models against this novel benchmark with minimal effort, whilst retaining the option to publish their results to the leaderboard with a single line of code.

4. Discussion

Based on the results of the previous section, the resulting ranking of our benchmarks differs substantially from the general benchmark results shown in the MTEB. While we also found Gemini's embedding model performing consistently well across all tasks, it was outperformed by OpenAI's model for all cohorts for the zero-shot classification task, as well as two of four cohorts based on the overall AUPRC.

Two of the recently published, high-ranked open-source models

[2] https://adhteb.scai.fraunhofer.de/

performed substantially worse than in the MTEB rankings. Notably, the Qwen3 model that held the second to fourth spot in the MTEB ranking performed worst in terms of the computed overall score. The widely used, but arguably older model, all-MiniLM, interestingly performed best out of the tested open-source models, despite its small parameter size. This could be potentially explained by either:

a) **Overfitting to MTEB:** Some models may have been fine-tuned to perform well specifically on MTEB tasks, which could limit their generalization to other applications, such as the harmonization task simulated in our benchmark.
b) **Task-specific differences:** Our harmonization benchmark likely differs fundamentally from the tasks included in MTEB, both in terms of domain and structure, which may favor different model capabilities.

In either case, these findings strongly support the relevance and potential value of our proposed benchmark. Whether the performance gap arises from MTEB-specific fine-tuning or from fundamental differences in task requirements, it highlights that general-purpose benchmarks may not adequately capture performance in real-world, domain-specific applications such as data harmonization. These observations also highlight the potential benefits of developing specialized or fine-tuned AI models for dementia-related data, as further discussed in Section 4.1.

The results obtained from different models and evaluated across distinct cohorts indicate that harmonization accuracy is strongly influenced by the quality of variable descriptions. In some instances, variables were mapped to incorrect modalities. For example, the variable "Neurofilament light values of central CSF analyses" was incorrectly harmonized to "Neurofilament Light Chain (NfL) in Plasma" in the EMIF cohort by all evaluated models. Moreover, in many cases where harmonization failed, the absence of shared terminology across variable descriptions was a contributing factor.

A key factor for mismatches in automated mappings is the general lack of variable naming conventions in clinical metadata. Establishing and adhering to consistent naming conventions when defining cohort metadata and variable descriptions can not only facilitate manual harmonization but also enhance the performance of embedding-based approaches for automated data harmonization. Based on our observations obtained in this benchmark study, we propose the following general recommendations to improve metadata quality and semantic clarity:

- Metadata should be provided in a machine-readable format (e.g., JSON, CSV with standardized headers) to facilitate automated parsing and harmonization. Providing structured metadata allows both human users and computational tools to access and process variable information consistently.
- A common character encoding (e.g., UTF-8) should be used for all metadata to ensure consistent interpretation across systems. This helps avoid issues with special characters, symbols, or accented letters that may otherwise cause parsing errors or mismatches during harmonization.
- Descriptions should always specify the modality from which the variable originates (e.g., CSF). Otherwise, for variables that can be measured across different modalities, such as Aβ, harmonization may inadvertently group distinct measurements together.
- Variables intended to be mapped to higher- or lower-level concepts (e.g., hippocampus volume vs. left hippocampus volume) should be described in sufficient detail to prevent higher-level measurements from being mistakenly mapped to more specific ones.
- Unnecessary wording can lead to misleading matches. For example, the variable "Age in months at the time of test" was incorrectly mapped to "The month in which a person was born." Since measurement units are often recorded in a separate column, excluding them from the variable description may improve mapping accuracy and help avoid such errors.

Based on the zero-shot accuracies of the evaluated cohort to CDM mapping, even though text-embedding models may facilitate the task of data harmonization, it is apparent that they cannot fully replace human curation. Although models may not always match the correct variables based on the description with the closest cosine similarity, the correct match is, in most cases, still among the most similar variable matches. Harmonization workflows that utilize text-embedding similarities can thus be better applied to enable *semi-automatic* harmonization, where a human in the loop may choose from a list of variables that are high-ranking in terms of their semantic similarity. By narrowing down potential correct matches using a ranked list of promising terms, the overall curation effort is significantly expedited by reducing the search space and guiding human experts toward the most semantically relevant candidates. For example, when harmonizing a cohort with 100 variables, instead of reviewing all possible matches individually, the user could focus on the top 10 most semantically similar candidates, greatly reducing effort while maintaining accuracy.

The computed AUPRC provides a way to assess how well variables are ranked, in our implementation, based on their absolute cosine similarities. Although we can see trends in the results that match those of the zero-shot classification performance, the results for the different models vary between cohorts. A possible explanation is the different wording for different cohorts; some cohorts may follow different naming conventions that favor different models trained on different sets of data. Another potential reason for the performance discrepancy in terms of AUPRC when compared to zero-shot accuracies, particularly for the PREVENT Dementia and EMIF cohorts, could be a low number of variables. To address this, future studies will include additional cohorts, which should also help reduce variance caused by differences in variable descriptions across cohorts.

4.1. Limitations and future work

Our benchmark exclusively contains cohorts in the domain of AD, which may in itself be biased toward certain naming conventions or variable formulations. Hence, we now plan to explore extending it with cohorts in other related neurodegenerative diseases. A natural next step would be to include Parkinson's disease, as our previous work has shown that there is a substantial overlap in variables while introducing new domain-specific elements [14].

In our current approach, we evaluate the capabilities of LLMs with regard to text embeddings. A possible extension could involve leveraging LLMs not only for embedding generation but also for actively selecting the most appropriate matches among candidate variables. Instead of relying solely on cosine similarity to determine the best match, a more advanced harmonization pipeline could employ the full reasoning capabilities of LLMs in a second stage. This two-stage approach would enable the model to go beyond surface-level similarity and incorporate domain-specific logic and latent cues present in variable descriptions, leading to more robust and interpretable harmonization decisions. LLMs could also be further utilized to interface between data curators and harmonization outputs by generating natural language explanations for mapping decisions or giving feedback on inconsistent manual mappings. Additionally, when individual AI models perform sub-optimally, researchers could leverage complementary strategies, such as ensemble approaches that combine outputs from multiple models or cross-validation with manually curated reference mappings. Incorporating human-in-the-loop curation at critical decision points can further improve reliability. Moreover, model performance could be enhanced by iterative fine-tuning using domain-specific examples or by integrating structured ontologies as additional guidance. These strategies collectively allow AI to support harmonization more effectively, even when single-model performance is limited.

5. Conclusion

While AI and especially LLM-assisted systems can potentially facilitate and accelerate the manual curation effort, it is important to promote transparency both in the model selection itself as well as in the choices made by the model during harmonization. Our benchmark contributes to this goal by systematically evaluating model behavior in a domain-specific setting, thereby offering insight into each model's strengths, limitations, and suitability for real-world harmonization tasks. Nonetheless, a human expert should always remain in the loop to review and validate model outputs, ensuring that final decisions are made by curators rather than delegated entirely to automated systems.

In alignment with these considerations, our work introduces a benchmark that operationalizes transparency in the context of variable harmonization. A domain-specific benchmark, as presented, is essential to address the unique challenges posed by harmonization workflows of general-purpose benchmarks, such as the MTEB, in AD. We outline key limitations of automated data harmonization and propose best practices for naming conventions to be able to still leverage embedding-based harmonization workflows. Finally, we present our benchmark approach as an accessible, open-source Python package that can be used to evaluate new and upcoming models in the future.

Glossary

AD (Alzheimer's Disease): A neurodegenerative disease characterized by rapid cognitive decline.

ADEO (Alzheimer's Disease Element Ontology): Standard vocabulary for Alzheimer's disease research data.

ADNI (Alzheimer's Disease Neuroimaging Initiative): A large, multicenter study collecting clinical, imaging, genetic, and biomarker data to investigate Alzheimer's disease progression and improve early diagnosis.

AI (Artificial Intelligence): An application or software able to perform tasks or produce output that would usually require some degree of human intelligence.

AUPRC (Area Under Precision Recall Curve): A metric that measures model performance based on precision and recall across different thresholds.

CDM (Common Data Model): A standardized framework for structuring and describing data elements to enable interoperability across datasets.

CSF (Cerebrospinal Fluid): Clear fluid of the brain and spinal cord.

CSV (Comma Separated Values): A common file format for tabular data storage.

EMIF (European Medical Information Framework): European cohort combining clinical and biomarker data to study Alzheimer's disease progression and risk factors.

ETL (Extract, Transform, Load): A common data processing practice, consisting of data extraction from multiple sources, transformation into a standard format, and storage into a database or data repository.

FL (Federated Learning): An application of several decentralized machine learning models that is trained to produce one centralized prediction or output.

GERAS: Observational studies of Alzheimer's disease and dementia.

JSON (JavaScript Object Notation): A text-based data transfer format, commonly used to transfer data between applications.

LLM (Large Language Model): A transformer-based Neural Network with a high amount of parameters, trained on a large corpus of texts to understand human language.

LOINC (Logical Observation Identifiers Names and Codes): Standard system for identifying and coding laboratory tests and clinical measurements.

MMSE (Mini-Mental State Examination): A test assessing cognitive function and screening for cognitive impairment.

MTEB (Massive Text Embedding Benchmark): A generalized, public, and well-established benchmark of different text embedding models.

NACC (National Alzheimer's Coordinating Center): A consortium collecting and sharing standardized clinical and neuropathological data from Alzheimer's disease research centers.

OMOP (Observational Medical Outcomes Partnership): A standardized Common Data Model used primarily in health care and patient-related data processing.

PREVENT-AD (Pre-symptomatic Evaluation of Experimental or Novel Treatments for Alzheimer's Disease): A longitudinal study focused on identifying early biomarkers and testing preventive interventions in individuals at risk for Alzheimer's disease.

SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms)**:** An international standardized terminology for indexing of medical terms.

Funding sources

SYNTHIA: Synthetic Data Generation framework for integrated validation of use cases and Al healthcare applications.

This project is supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No 101172872. The JU receives support from the European Union's Horizon Europe research and innovation programme, COCIR, EFPIA, Europa Bío, MedTech Europe, Vaccines Europe and DNV. The UK consortium partner, The National Institute for Health and Care Excellence (NICE) is supported by UKRI Grant 10132181.

Disclaimer

Funded by the European Union, the private members, and those contributing partners of the IHI JU. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used OpenAI's ChatGPT in order to proofread and stylistically refine the text. The tool was used solely for spelling, grammar correction, and rephrasing assistance. All scientific content, data interpretation, and conclusions were generated independently by the authors without AI involvement. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Tim Adams: Conceptualization, Formal analysis, Investigation, Methodology, Software, Software, Validation, Visualization, Writing – original draft. **Yasamin Salimi:** Writing – original draft, Visualization, Resources, Methodology, Investigation, Formal analysis, Data curation. **Mehmet Can Ay:** Writing – review & editing, Visualization, Software, Investigation, Formal analysis, Data curation. **Diego Valderrama:** Writing – review & editing, Resources, Data curation. **Marc Jacobs:** Writing – review & editing, Conceptualization. **Holger Fröhlich:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Holger Froehlich reports that financial support was provided by Gates Ventures. Tim Adams, Yasamin Salimi, and Diego Valderrama report that administrative support was provided by Gates Venture. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Greg Moore, Prarthna Desai, and the Alzheimer's Disease team at Gates Ventures for their administrative support and review.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tjpad.2025.100420.

References

[1] Mueller SG, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). Alzheimers Dement J Alzheimers Assoc 2005;1(1):55–66. https://doi.org/10.1016/j.jalz.2005.06.003.

[2] Weiner MW, et al. Impact of the Alzheimer's Disease neuroimaging initiative, 2004 to 2014. Alzheimers Dement 2015;11(7):865–84. https://doi.org/10.1016/j.jalz.2015.04.005.

[3] Birkenbihl C, Salimi Y, Fröhlich H. Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. Alzheimers Dement 2022;18(2):251–61. https://doi.org/10.1002/alz.12387.

[4] Schinkel M, Bennis FC, Boerman AW, Wiersinga WJ, Nanayakkara PWB. Embracing cohort heterogeneity in clinical machine learning development: a step toward generalizable models. Sci Rep 2023;13(1):8363. https://doi.org/10.1038/s41598-023-35557-y.

[5] Zhang F, et al. Recent methodological advances in federated learning for healthcare. Patterns 2024;5(6). https://doi.org/10.1016/j.patter.2024.101006.

[6] Bauermeister S, et al. Research-ready data: the C-Surv data model. Eur J Epidemiol 2023;38(2):179–87. https://doi.org/10.1007/s10654-022-00916-y.

[7] Szalma S, Koka V, Khasanova T, Perakslis ED. Effective knowledge management in translational medicine. J Transl Med 2010;8(1):68. https://doi.org/10.1186/1479-5876-8-68.

[8] Wegner P, et al. Semantic harmonization of Alzheimer's disease datasets using AD-mapper. J Alzheimers Dis 2024;99(4):1409–23. https://doi.org/10.3233/JAD-240116.

[9] Hao X, et al. An ontology-based approach for harmonization and cross-cohort query of Alzheimer's disease data resources. BMC Med Inform Decis Mak 2023;23 (S1):151. https://doi.org/10.1186/s12911-023-02250-z.

[10] Mateus P, et al. Data harmonization and federated learning for multi-cohort dementia research using the OMOP common data model: a Netherlands consortium of dementia cohorts case study. J Biomed Inform 2024;155:104661. https://doi.org/10.1016/j.jbi.2024.104661.

[11] Yang D, et al. Robust automated harmonization of heterogeneous data through ensemble machine learning: algorithm development and validation study. JMIR Med Inform 2025;13(1):e54133. https://doi.org/10.2196/54133.

[12] Verbitsky A, Boutet P, Eslami M. Metadata harmonization from biological datasets with language models. bioRxiv 2025. 2025–01.

[13] X. Zhou, L.S. Dhingra, A. Aminorroaya, P. Adejumo, and R. Khera, "A novel sentence transformer-based natural language processing approach for schema mapping of electronic health records to the OMOP common data model," 2024. doi: 10.1101/2024.03.21.24304616.

[14] Y. Salimi, T. Adams, M.C. Ay, H. Balabin, M. Jacobs, and M. Hofmann-Apitius, "On the utility of large language model embeddings for revolutionizing semantic data harmonization in Alzheimer's and Parkinson's disease," 2024. doi: 10.21203/rs.3.rs-4108029/v1.

[15] Adams T, Aborageh M, Salimi Y, Fröhlich H, Jacobs M. INDEX: the intelligent data steward toolbox. In: International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences; 2024. p. 2024. Accessed: Jun. 04, 2025. [Online]. Available, https://publica.fraunhofer.de/entities/publication/28a63af5-7ddf-4267-91ac-79c341794e52/details.

[16] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: massive text embedding benchmark," Mar. 19, 2023, arXiv: arXiv:2210.07316. doi: 10.48550/arXiv.2210.07316.

[17] Tremblay-Mercier J, et al. Open science datasets from PREVENT-AD, a longitudinal cohort of pre-symptomatic Alzheimer's disease. NeuroImage Clin 2021;31:102733. https://doi.org/10.1016/j.nicl.2021.102733.

[18] Bos I, et al. The EMIF-AD multimodal biomarker discovery study: design, methods and cohort characteristics. Alzheimers Res Ther 2018;10(1):64. https://doi.org/10.1186/s13195-018-0396-5.

[19] "The GERAS study: a prospective observational study of costs and resource use in community dwellers with Alzheimer's Disease in three European countries – study design and baseline findings - Anders Wimo, Catherine C. Reed, Richard Dodel, Mark Belger, Roy W. Jones, Michael Happich, Josep M. Argimon, Giuseppe Bruno, Diego Novick, Bruno Vellas, Josep Maria Haro. Accessed: Jul. 07, 2025. [Online]. Available, https://journals.sagepub.com/doi/abs/10.3233/JAD-122392; 2013.

[20] Costs and quality of life in community-dwelling patients with Alzheimer's disease in Spain: results from the GERAS II observational study | International Psychogeriatrics | Cambridge Core. " Accessed, https://www.cambridge.org/core/journals/international-psychogeriatrics/article/abs/costs-and-quality-of-life-in-communitydwelling-patients-with-alzheimers-disease-in-spain-results-from-the-geras-ii-observational-study/90921EE600D3DA33DD77790D83813DE2; 2025.

[21] Bruno G, Mancini M, Bruti G, Dell'Agnello G, Reed C. Costs and resource use associated with Alzheimer's disease in Italy: results from an observational study. J Prev Alzheimers Dis 2018;5(1):55–64. https://doi.org/10.14283/jpad.2017.31.

[22] Robinson RL, et al. Observation of patient and caregiver burden associated with early Alzheimer's disease in the United States: design and baseline findings of the GERAS-US cohort Study1. J Alzheimer's Dis 2019;72(1):279–92. https://doi.org/10.3233/JAD-190430.

[23] Ritchie CW, et al. The PREVENT dementia programme: baseline demographic, lifestyle, imaging and cognitive data from a midlife cohort study investigating risk factors for dementia. Brain Commun 2024;6(3). https://doi.org/10.1093/braincomms/fcae189. p. fcae189.

[24] Lee J, et al. Gemini embedding: generalizable embeddings from Gemini," arXiv. Org. Accessed, https://arxiv.org/abs/2503.07891v1; 2025.

[25] Y. Zhang et al., "Qwen3 embedding: advancing text embedding and reranking through foundation models," Jun. 10, 2025, *arXiv*: arXiv:2506.05176. doi: 10.48550/arXiv.2506.05176.

[26] C. Choi et al., "Linq-Embed-Mistral technical report," Dec. 04, 2024, arXiv: arXiv:2412.03223. doi: 10.48550/arXiv.2412.03223.

[27] Model - OpenAI A.P.I. Accessed, https://platform.openai.com; 2025.

[28] Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers," in advances in neural information processing systems, Curran Associates, Inc. Accessed: Jun. 11, 2025. [Online]. Available, https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html; 2020.

[29] "SCAI-BIO/ADHTEB. ADHTEB - Alzheimers disease harmonization text embedding benchmark." accessed [Online]. Available, https://github.com/SCAI-BIO/ADHTEB; 2025.

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

Towards an AI biomedical scientist: Accelerating discoveries in neurodegenerative disease

Kaleigh F. Roberts[a,aa,1], Eric C. Landsness[b,aa,1], Justin Reese[c,aa], Donald Elbert[d,aa], Gabrielle Strobel[e,aa], Elizabeth Wu[e,aa], Yixin Chen[f,aa], Albert Lai[f,g], Zachary B. Abrams[g,aa], Mingfang Zhu[f], Justin Melendez[b,z,aa], Srinivas Koutarapu[b,aa], Sihui Song[b,aa], Yun Chen[b,aa], Robert Lazar[h], Payam Barnaghi[i], John F. Crary[j], Sergio Pablo Sardi[k,aa], Marc D. Voss[k], Rajaraman Krishnan[k], Joel W. Schwartz[l], Ron Mallon[m,aa], Gustavo A. Jimenez-Maggiora[n,aa], Chenguang Wang[o,aa], Thomas Sandmann[p,aa], Niranjan Bose[q,x], Mukta Phatak[r,aa], Gayle Wittenberg[s], Yannis G. Kevrekidis[t,aa], Cassie S. Mitchell[u], Ludovico Mitchener[v], Towfique Raj[w], Luca Foschini[x,aa], Gregory J. Moore[y,aa], Randall J. Bateman[b,z,aa,*]

[a] *Department of Pathology & Immunology, Washington University, St. Louis, MO, USA*
[b] *Department of Neurology, Washington University, St. Louis, MO, USA*
[c] *Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
[d] *Department of Neurology, University of Washington, Seattle, WA, USA*
[e] *Alzforum Foundation Inc., MA, USA*
[f] *Department of Computer Science & Engineering, Washington University, St. Louis, MO, USA*
[g] *Institute for Informatics, Data Science and Biostatistics (I2DB), Washington University, St. Louis, MO, USA*
[h] *Booz Allen Hamilton*
[i] *Department of Brain Sciences, Imperial College London, London, England United Kingdom*
[j] *Department of Pathology, Nash Family Department of Neuroscience, Department of Artificial Intelligence & Human Health, Neuropathology Brain Bank & Research CoRE, Ronald M. Loeb Center for Alzheimer's Disease, Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, NY, NY, USA*
[k] *Sanofi, Cambridge, MA, USA*
[l] *Bristol Myers Squibb, Cambridge, MA, USA*
[m] *Department of Philosophy, Washington University, St. Louis, MO, USA*
[n] *Alzheimer's Therapeutic Research Institute, Keck School of Medicine of University of Southern California, San Diego, CA, USA*
[o] *Department of Computer Science and Engineering, University of California Santa Cruz, CA, USA*
[p] *Denali Therapeutics Inc., South San Francisco, CA, USA*
[q] *Alzheimer's Disease Data Initiative, USA*
[r] *The 10,000 Brains Project, USA*
[s] *Johnson & Johnson, USA*
[t] *Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, USA*
[u] *Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA*
[v] *FutureHouse, USA*
[w] *Department of Neuroscience, Icahn School of Medicine at Mount Sinai, NY, NY, USA*
[x] *Sage Bionetworks, USA*
[y] *Gates Ventures, USA*
[z] *Tracy Family SILQ Center, Washington University, St. Louis, MO, USA*
[aa] *Consortium for Biomedical Research and Artificial Intelligence in Neurodegeneration (C-brAIn)*

ARTICLE INFO

Keywords:
Artificial intelligence
Alzheimer's disease

ABSTRACT

Despite major advances in Alzheimer's disease and related diseases (ADRD) research, the translation of discoveries into impactful clinical interventions remains slow. Overwhelming data complexity, fragmented

* Corresponding author.
E-mail address: batemanr@wustl.edu (R.J. Bateman).

[1] Equal contribution.

https://doi.org/10.1016/j.tjpad.2025.100398
Received 4 September 2025; Received in revised form 14 October 2025; Accepted 20 October 2025

Large language model
Biomedical AI
Multimodal datasets

knowledge, and prolonged research cycles hinder progress in understanding and treating neurodegenerative diseases. Artificial intelligence (AI) offers a promising path forward, particularly when developed as a scientist-in-the-loop system that collaborates with researchers throughout the scientific discovery process. This paper introduces the concept of an AI Biomedical Scientist, an intelligent platform designed to support literature synthesis, hypothesis generation, experimental design, and data interpretation. This platform aims to function as a holistic scientific partner, integrating diverse biomedical data and expert reasoning to accelerate discovery. We review commercial and academic efforts and introduce targeted Minimum Viable Products (MVPs) needed for general biomedical research lab utilization of AI, such as robust and accurate tools for literature and data analysis, negative data models, and virtual peer review, with a longer-term vision of foundation models trained directly on biomedical datasets. In AD and neurodegeneration research, such tools are anticipated to deliver efficiency gains ranging from modest improvements in specific research tasks to potential multi-fold accelerations in discovery workflows as systems mature and scale. This review examines the technical foundations, challenges, and anticipated impacts of AI and aims to inform and engage researchers in utilizing these systems to transform biomedical discovery, starting with AD and extending to other complex conditions.

1. Introduction: the need for innovation in Alzheimer's research

Alzheimer's disease (AD) is a common neurodegenerative disorder, affecting an estimated 50 million people worldwide and contributing substantially to human, social, and economic burdens at an immense scale. More than a century after its initial description by Alois Alzheimer, research has advanced significantly, leading to the development of highly accurate diagnostic biomarkers and modestly effective disease-modifying therapies [1–3]. However, for most patients, treatments that meaningfully alter disease progression or outcomes remain limited [4].

A key challenge in AD research is the biological complexity of the disease. Studies have demonstrated that AD pathology begins up to 25 years before symptom onset, initiating a prolonged and dynamic pathogenic process involving compensatory mechanisms, evolving cell states, and interacting molecular pathways [5]. The extended course of Alzheimer's disease, combined with the rapid growth and complexity of biomedical data, has outpaced the capacity of traditional research methods to effectively synthesize all information to generate maximally informed actionable insights. Although over two million biomedical papers are published annually, fewer than 0.1 % of discoveries have a direct impact on human health outcomes [6]. Translational barriers, including fragmented knowledge across disciplines, slow cycles of hypothesis testing, and challenges in integrating diverse data types, are particularly pronounced in neurodegenerative diseases [7]. Moreover, more than 97 % of drugs entering AD clinical trials do not achieve approval (data from https://www.alzforum.org/therapeutics), underscoring inefficiencies in current discovery pipelines [8].

The challenges of AD research are compounded by the vast and accelerating amounts of biomedical data generated, including published data and "dark data" hidden in inaccessible sources or unpublished findings, such as negative results [9–13]. The huge number of publications (currently > 200,000 for AD based on PubMed query for "Alzheimer's disease") make it impossible for researchers to stay abreast of findings outside their specialization, further exacerbating siloed understanding across domains and hampering disruptive science that fundamentally shifts current understanding [14]. As a consequence, the timelines to train researchers have also substantially increased, leading to a relative shortage in the number of qualified scientists needed for the challenge [15].

Artificial intelligence (AI) has the potential to improve how researchers navigate and interpret complex biomedical data [16–20]. While expert scientists possess highly refined reasoning skills, they are fundamentally limited by cognitive bandwidth, i.e. the amount of literature, data modalities, and prior findings they can hold in mind and integrate at once. AI systems can help overcome this limitation by surfacing relevant knowledge from across vast datasets and literature corpora, organizing connections, and enabling hypothesis generation that is informed by a broader and more comprehensive information space than a human could synthesize alone [21,22]. In this way, AI acts as a contextual amplifier, allowing scientists to apply their expertise more effectively across the full scope of available evidence.

Despite this potential, many current AI applications in biomedicine remain narrowly focused on specific tasks such as literature mining, diagnostic image analysis, risk prediction, or natural language summarization. While valuable, these tools stop short of supporting the more integrative and iterative processes of scientific reasoning required for foundational discovery. There remains a need for AI systems that can assist with the cognitive and analytical tasks central to discovery, including synthesizing knowledge, generating hypotheses, designing experiments, and learning from new data. Toward this end, the Consortium for Biomedical Research and AI in Neurodegeneration (c-brAIn) has been launched to accelerate health impactful basic science discoveries through the use of AI tools.

In this paper, we describe the concept of an AI Biomedical Scientist, a collaborative, scientist-in-the-loop system intended to support researchers throughout the scientific process. We focus on AD as an initial use case, given its clinical relevance, extensive datasets, and established research infrastructure. We outline the technical foundations of this approach, the early-stage tools currently in development, and the potential for AI to enhance research efficiency and outcomes in neurodegenerative disease studies.

2. What is an AI biomedical scientist?

An AI Biomedical Scientist is designed to support researchers across the biomedical research process, including literature and data review, hypothesis generation, experimental design, and data interpretation. Unlike traditional AI tools focused on single tasks, the AI Biomedical Scientist is designed to iterate through the entire scientific pipeline, including identifying biological questions, synthesizing literature, generating hypotheses, designing experiments, analyzing data, and interpreting results, by integrating publications, biomedical data, critical reasoning, and domain expertise to accelerate discovery and generate insights that improve human health.

A defining feature of this approach is its scientist-in-the-loop design, in which scientific experts remain involved in developing, refining, and interpreting the system's outputs to ensure scientific rigor and contextual relevance [23]. Rather than aiming to replace researchers, the AI scientist is intended to augment human work by improving efficiency, reproducibility, and the capacity to quickly and meaningfully explore new scientific questions.

The AI scientist is designed to integrate data from multiple domains, including genomics, proteomics, lipidomics, metabolomics, imaging, electronic health records, and behavioral measures, which are often distributed across institutions and research silos. Scientists can conduct analyses using local or external data sources through natural language interfaces. The system automates analyses through machine learning and deep neural net approaches, accelerating the time from measurement to interpretation.

Together, these capabilities position the AI Biomedical Scientist as a valuable tool for advancing research in complex areas such as AD, underscoring the need to understand the specific AI technologies that make such a system possible.

3. Core AI technologies

The AI Biomedical Scientist combines several artificial intelligence technologies to support various aspects of biomedical research. To better understand how these technologies contribute, it is helpful to clarify key terms that are sometimes used interchangeably but refer to distinct concepts (Fig. 1).

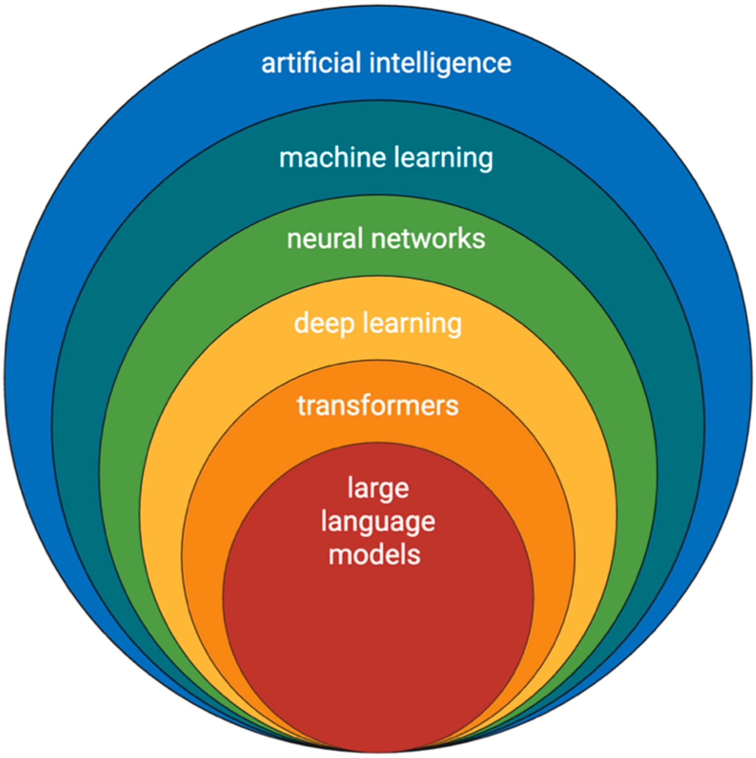
Artificial intelligence broadly describes computational systems that perform tasks typically requiring human intelligence, such as understanding language, recognizing patterns, or making decisions. Within AI, *machine learning* refers to algorithms that improve performance by learning from data rather than following explicitly programmed rules. *Deep learning* is a specialized subset of machine learning that relies on large *neural networks*, which are computational models inspired by the structure and function of neurons in the brain. These networks consist of layers of interconnected nodes ("neurons") that process input data through weighted connections, enabling the system to learn complex patterns. *Transformers* are a specific class of deep neural network architectures that excel at processing sequential data and have become central to many modern AI applications.

A key component of the AI Biomedical Scientist is the *large language model* (LLM) (Fig. 2). LLMs are transformer-based neural networks trained on extensive textual corpora, including scientific literature, to generate human language [24]. They can summarize information, answer questions, and suggest hypotheses. However, general-purpose LLMs, such as GPT-5, Gemini, Claude, and Grok4, while effective in conversational tasks, often lack the precision, domain-specific knowledge, and interpretability needed for rigorous biomedical research [25]. Additionally, these models can produce "hallucinations" (or more accurately, confabulations), plausible but incorrect information, which poses challenges for their use in scientific contexts, where accuracy and truth are paramount [26].

An alternative approach to improve domain relevance is the development of biological *specialist LLMs*, models pre-trained or fine-tuned specifically on biomedical text. Examples include BioBERT[27], BioGPT[28], BiomedLM[29], Med-Gemini[30], and Med-PaLM[31]. These models can offer enhanced vocabulary coverage, improved factual recall, and greater alignment with domain-specific language. However, specialist LLMs face several limitations. First, their knowledge is static and can quickly become outdated in fast-evolving fields. Second, they are often narrow in scope, performing well in specific biomedical subdomains but struggling when tasks require broader reasoning or interdisciplinary integration. Finally, they still retain the inherent limitations of LLMs, including susceptibility to hallucinations and opaque decision-making.

To address these limitations, the AI Biomedical Scientist incorporates *retrieval-augmented generation (RAG)* architectures. RAG systems enhance LLMs by connecting them to external databases or curated literature, helping ensure that generated responses are grounded in factual sources [32,33]. This design allows us to circumvent some of the limitations of static pretrained models by deferring knowledge retrieval to inference time, increasing flexibility, and reducing the need for frequent retraining. A further refinement of this approach is integrating *knowledge graphs* with RAG (e.g. GraphRAG), to link textual outputs directly to structured evidence nodes [34]. Knowledge graphs represent biomedical concepts and the relationships among them and offer a way to organize information and support reasoning that goes beyond simple keyword matching.

In addition, the proposed AI Biomedical Scientist leverages emerging concepts from *agentic AI* [35,36]. Unlike traditional models that only respond to queries, agentic AI systems are designed to autonomously plan, reason, and take actions toward defined goals. Agentic AI can iteratively break down complex research tasks, select appropriate tools (such as querying databases, running analyses, or simulating models), and adapt based on intermediate results. This agent-like behavior shifts AI from a reactive assistant toward a proactive collaborator capable of



artificial intelligence (AI): computational systems that perform tasks typically requiring human intelligence

machine learning: algorithms that learn patterns from data to make predictions without explicitly being programmed for each task

neural network: machine learning model inspired by the human brain composed of interconnected nodes that process inputs through weighted connections enabling learning of complex patterns

deep learning: subset of machine learning relying on neural networks with many layers

transformer: class of deep neural network architecture designed to process sequential data using self-attention mechanisms to capture relationships between elements

large language model (LLM): transformer-based neural networks trained on massive textual corpora to generate human language

specialist LLM: LLM trained or fine-tuned on domain-specific data

retrieval-augmented generation (RAG): framework that combines search and generation by retrieving relevant documents and using them to produce informed, context-aware responses

knowledge graph: structured representation of information that connects entities and their relationships, enabling machines to reason over data and draw meaningful inferences

agentic AI: systems designed to autonomously plan, make decisions, and take actions toward goals, often by coordinating multiple tools or models in a task-oriented workflow

Fig. 1. Glossary of AI terms.

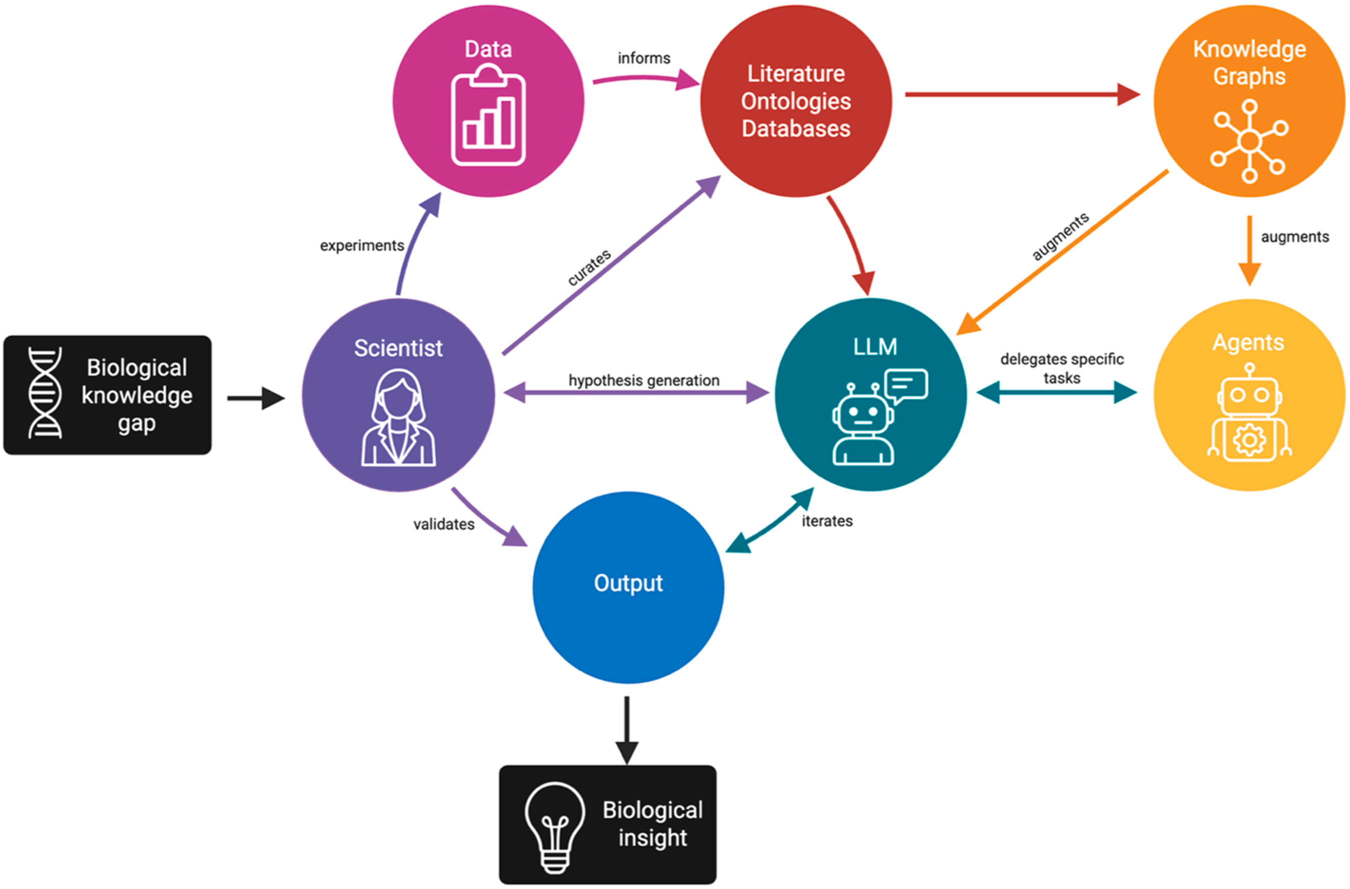


Fig. 2. Workflow diagram for AI Biomedical Scientist. A biological knowledge gap is first identified by a scientist, who conducts experiments to generate raw data. These data are then captured in published literature, ontologies, and databases curated by human experts. Organized knowledge frameworks derived from these sources can be transformed into knowledge graphs and incorporated into large language models (LLMs). Through an iterative process, human scientists interact with the LLM, which assigns specific tasks to specialized AI agents such as data analysis, hypothesis generation, literature retrieval, and critical review. This human-AI collaboration supports the identification of new biological insights, their contextualization within existing data and literature, and expert validation to ensure scientific relevance.

orchestrating multi-step workflows. However, while these systems introduce powerful automation capabilities, they are not intended to operate in isolation. A key design principle is maintaining an appropriate balance between automated task execution and human oversight. Researchers remain essential in setting goals, curating inputs, interpreting results, and determining when to trust or override AI-driven decisions.

Another envisioned capability of the AI Biomedical Scientist is *multimodal data integration*. Modern biomedical research generates diverse datasets across domains, scale, and time, including genomics, proteomics, and other biomolecular measures across atomic, molecular, cellular, tissue, and organ scales with human imaging, clinical, vocal, and behavioral measurements that are often stored in separate systems. The AI Biomedical Scientist is designed to utilize these heterogeneous data sources, helping researchers examine relationships across different domains and develop integrated models of disease processes.

Finally, the AI Biomedical Scientist’s architecture is designed to operate in *federated environments*, allowing analyses to be performed across multiple institutions without requiring the sharing of raw data. In a federated approach, data remain securely within each institution while algorithms or models are shared and run locally, and only the aggregated results are combined. This can be especially important for working with proprietary pharmaceutical data, sensitive patient records, or other data subject to privacy regulations. Examples of successful federated environments in multiomic workflows include PPML-Omics[37], DataSHIELD[38] and OmicSHIELD[39]. Within the AD research community, the Alzheimer’s Disease Data Initiative (ADDI) utilizes the Federated Data Sharing Appliance (FDSA), a secure data application that enables multiple organizations to securely share data without the need for centralization in a single repository [40]. Federated training of AI models is achieved by distributing global model parameters to local sites, training on private local data, and then returning updated model parameters to the centralized server [37,41–44]. Such methods help address privacy, sovereignty, and regulatory requirements while enabling collaboration at a scale needed for research in AD and other complex conditions.

Together, these technologies can create an assistive framework designed to complement scientific expertise. By combining natural language processing, structured knowledge representation, data integration, agentic AI, and customized foundational models, the AI Biomedical Scientist aims to accelerate how researchers generate hypotheses, design experiments, and analyze results. In our own development efforts, we have implemented and evaluated early prototypes of this system, specifically focusing on literature retrieval and synthesis using RAG architectures trained on curated Alzheimer’s disease corpora [45], and also testing prototype multi-agentic systems. These systems represent the foundation for future more advanced capabilities such as multimodal integration and added-value agentic workflows.

4. Current AI tools and systems

The use of artificial intelligence to support scientific research is gaining momentum in both commercial and academic settings. Several initiatives are exploring LLM-based systems designed to function as AI scientists, capable of parsing scientific literature, suggesting hypotheses, or assisting with experimental planning [46–49]. These efforts have emerged in response to the growing challenge that individual researchers face in keeping up with the rapidly expanding volume of scientific publications and data.

Examples of specialized AI tools for scientific applications include Google's Co-Scientist [50], FutureHouse's Robin [51], and AI2's Asta. These systems differ in their scope, underlying technologies, and the extent to which they incorporate expert scientific input (also discussed in the paper by Funk et al. in this special edition of JPAD)., [ref] While these systems represent significant progress and are at the current cutting edge of applying AI to research tasks, we believe significant additional validation is required to demonstrate their reliability and impact in real-world scientific research. A non-exhaustive list of AI biomedical scientific platforms is summarized in Table 1.

These initiatives represent important early progress in applying generative AI and agentic tools to biomedical research, and many offer capabilities that will likely inform and complement future systems. Our proposed AI Biomedical Scientist builds on this growing foundation, with a particular focus on addressing the specific challenges of AD and biomedical research. Rather than replacing or competing with existing platforms, it seeks to integrate and extend their capabilities within a disease-focused, scientist-guided framework. Key distinguishing features include its explicit design for AD and neurodegeneration, enabling deeper incorporation of domain-specific knowledge, and its emphasis on high-quality multimodal "dark" data fusion across genomic, proteomic, imaging, clinical, and behavioral domains that are otherwise not accessible. Importantly, the system is developed around a scientist-in-the-loop model, in which domain experts play an active role in curating inputs, interpreting outputs, and refining system behavior to ensure scientific rigor and practical relevance.

5. Why Alzheimer's disease is the ideal proving ground

AD is a particularly suitable area for developing and implementing an AI Biomedical Scientist [63]. Decades of both broad and deep research have produced enormous datasets, from molecular characterization, cellular and animal models, through human pathologic, imaging, genomic, proteomic, and clinical data from initiatives such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [64], Dominantly Inherited Alzheimer Network (DIAN) [65], DIAN-TU [66], and national and international observational cohorts and clinical trials summarized in Table 2. These resources provide a strong basis for developing and validating AI models.

Beyond data availability, the field also benefits from structured knowledge platforms like Alzforum and data aggregators such as the AD Data Initiative (ADDI). These curated resources are valuable for training AI systems designed to work with the complexities of neurodegenerative biology.

Furthermore, the AD research ecosystem is characterized by strong cross-sector research groups in pharmaceutical and biotechnology companies, academic institutions, and patient advocacy groups, with thousands of researchers working to decipher the causes and pathophysiology of AD. This large and diverse group of researchers will enable scientist-in-the-loop training and integration and evaluation of AI-driven research tools.

Taken together, diverse datasets, structured knowledge platforms, and strong research collaboration across academia and industry make AD an ideal domain for developing AI tools. These resources create opportunities for practical systems that support researchers in addressing complex scientific questions, a goal that begins with developing focused initial solutions.

Table 1
Examples of AI biomedical scientific platforms currently in development including notable features and links for access.

System	Use	Access
Readily accessible via web user interface		
FutureHouse Robin [51]	Multi-agent system for literature search, hypothesis generation, experimental design, data analysis, figure generation, and experimental planning	Free partial web access to agents at platform.futurehouse.org, code available at https://github.com/Future-House/robin
Ai2 Asta	Multi agent system for literature search and summarization	Free, https://asta.allen.ai/chat
BenchSci Ascend	Multi agent system using proprietary multimodal LLMs supported by a knowledge graph and ontology knowledge base	https://knowledge.benchsci.com/home/platform-fundamentals, free access to Selector Tool for academics, most tools require paid subscription
AlzAssistant	Literature-based Q&A using PaperQA2, curated AD paper corpus, and Alzheimer's knowledge graph	Free, https://chat.alzassistant.org/
AlzheimerRAG[52]	Q&A using multimodal RAG pipeline with AD PubMed corpus	Free, https://tinyurl.com/AlzheimerRAG
Biomni[22]	Generalist agentic architecture that integrates LLM reasoning with retrieval-augmented planning and code-based execution, enabling complex biomedical workflows	Free, https://biomni.stanford.edu/ https://github.com/snap-stanford/biomni
DORA[53]	Multi-agent scientific exploration and draft outline research assistant for automated or semi-automated research studies and report generation	https://dora.insilico.com
Code available, but no web interface		
Sakana The AI Scientist[48]	Idea generation, computational experiment conduction, paper writing and review	https://github.com/SakanaAI/AI-Scientist/tree/main/ai_scientist
SemNet[54]	Literature-based discovery system enabling PubMed relationship literature mining	https://github.com/pathology-dynamics/semnet-2
The Virtual Lab[55]	LLM principal investigator agent guiding a team of LLM agents with different scientific backgrounds (e.g., a chemist agent, a computer scientist agent, a critic agent), with a human researcher providing high-level feedback	https://github.com/zou-group/virtual-lab
Data-to-paper[56]	Automation platforms that guides LLM agents starting with annotated data through hypothesis generation, data analysis, results interpretation, and manuscript preparation	https://github.com/Technion-Kishony-lab/data-to-paper
RBio by CZI[57]	Reasoning model combining virtual cell models with chat interface of LLMs to predict how cells will behave in experiments	https://github.com/czi-ai/rbio
X-Master[58]	Tool-augmented reasoning agent designed to emulate human researchers by interacting flexibly with external tools during its reasoning process	https://github.com/sjtu-sai-agents/X-Master

(continued on next page)

Table 1 (*continued*)

System	Use	Access
BioResearcher[59]	Modular multi-agent architecture integrating search, literature synthesis, experimental design, and programming	https://github.com/XMUDM/BioResearcher
BioDiscoveryAgent [60]	Agent for designing genetic perturbation experiments	https://github.com/snap-stanford/BioDiscoveryAgent
Not publicly available		
Google Co-Scientist [50]	Multi-agent system focused on literature search and iterative hypothesis refinement	Not publicly available, paid institutional access
Lila.ai	Platform announced by Flagship Pioneering to develop "superintelligence in science," integrating LLMs, reasoning systems, and autonomous laboratory platforms to accelerate discovery	Not publicly available
PROTEUS[61]	Fully automated scientific discovery system for hypothesis generation from raw proteomic data	Not publicly available
STELLA[62]	Multi-agent architecture that self-evolves reasoning strategies and discovers and integrates bioinformatic tools	Not publicly available

Table 2
Examples of Multimodal Data Resources.

Resource Type	Examples	Types of Data	Utility/Relevance
AD Data repositories	AD Knowledge Portal[67], ADDI Repository[68], GNPC[69], IDA [70], NACC[71], NIAGADS[72]	Clinical, cognitive, imaging, biomarkers, genomic, transcriptomic, proteomic, metabolomic	International harmonized datasets over multiple studies
AD-specific observational cohorts	ADNI[64], ADRCs, ADSP[73], AIBL [74], BioFINDER [75], DELCODE [76], DIAN[65], EPAD[77], ROSMAP[78]	Clinical, cognitive, imaging, biomarkers, genomic, transcriptomic, proteomic, metabolomic	Imaging, CSF/ biomarkers, cognitive assessments, and multi-omic data across diverse AD cohorts
AD-specific interventional studies	A4/LEARN[79], AHEAD 3–45[80], APEX, DIAN-TU [66]	Clinical, cognitive, imaging, biomarkers, genomic, transcriptomic, proteomic, metabolomic	Highly phenotyped AD cohorts with therapeutic interventions
General population and longitudinal aging cohorts	100-plus Study [81], All of Us Research Program [82], BLSA[83], CAMCAN[84], FinnGen[85], HASD/ACS[86], Human Connectome Project[87], MCSA[88], RESILIENT[89], UK Biobank[90]	Genomics, imaging, clinical	Large-scale data integrating genetic profiles, imaging, and health records across the lifespan
Real-world clinical datasets	Optum Clinformatics, TriNetX[91]	Federated EHR networks, claims data, and clinical datasets	Real-world data capturing clinical heterogeneity and operational variability not present in curated or protocol driven datasets
Digital biomarker studies	mPower[92], RADAR-AD[93], TIHM[94]	Wearable sensor data	Digital monitoring of daily living to capture dynamic changes missed in episodic clinical assessments
Not publicly available datasets	Proprietary Pharma data, individual laboratory research data	Clinical trial results, experimental data, negative results	Often unpublished, but critical for hypothesis generation and reducing duplication of effort
Curated knowledge	Alzforum	Curated literature, structured knowledge	AD-focused commentary, news, and community consensus

A4: Anti-Amyloid Treatment in Asymptomatic Alzheimer's, ACTC: Alzheimer's Clinical Trials Consortium, ADDI: Alzheimer's Disease Data Initiative, ADNI: Alzheimer's Disease Neuroimaging Initiative, ADRC: Alzheimer's Disease Research Center, ADSP: Alzheimer's Disease Sequencing Project, AIBL: Australian Imaging, Biomarkers and Lifestyle Study, APEX: Alzheimer's Plasma Extension Study, BioFINDER: Biomarkers For Identifying Neurodegenerative Disorders Early and Reliably, BLSA: Baltimore Longitudinal Study of Aging, CAMCAN: Cambridge Centre for Ageing and Neuroscience, DELCODE: DZNE Longitudinal Cognitive Impairment and Dementia Study, DIAN: Dominantly Inherited Alzheimer Network, DIAN-TU: DIAN Trials Unit, EPAD: European Prevention of Alzheimer's Dementia, GNPC: Global Neurodegeneration Proteomics Consortium, HASD/ACS: Healthy Aging & Senile Dementia/The Adult Children Study, IDA: Image & Data Archive at LONI, LEARN: Longitudinal

6. Phase 1: minimum viable products (MVPs) in development

A practical step toward applying the AI Biomedical Scientist in AD research is the development of targeted Minimum Viable Products (MVPs) that address specific challenges in the research process. These initial tools aim to test technical approaches and support AI-driven research amid large datasets, expanding experimental findings, and the complex biology of neurodegeneration and AD. One of the first areas of development is creating tools for literature search and synthesis. Early work across various domains suggests that traditional search is still better than AI-based search tools, however as AI continues to improve the advantage it gives in speed will become more relevant [95–97]. We have been developing AI literature search systems that integrate RAG architectures with LLMs trained on AD-focused scientific texts and connected to knowledge graphs [45]. The goal is to help researchers quickly identify, compare, and summarize relevant information from both published literature and internal datasets, making it easier to navigate the existing scientific knowledge. Importantly, domain-specific scientists are actively integrated into the developmental process to ensure accuracy and relevance of AI-generated responses. Additional emphasis on the curation of a trustworthy, high-quality corpus of training literature ensures a solid central knowledge foundation.

A second MVP focuses on addressing the challenge of negative and unpublished results in biomedical research. Negative findings are often absent from published literature, which can contribute to redundant research efforts and leave important areas of biological understanding unexplored [13]. The proposed "Negative Data Analyzer" would aim to process data from both published studies and non-public sources, including data held in federated environments such as pharmaceutical company datasets and unpublished results in labs. By integrating information from studies with non-significant or null results, this tool is intended to help reduce unnecessary duplication of experiments and identify conditions that influence biological mechanisms.

A third MVP, referred to as "Reviewer Three," is envisioned as a virtual scientific reviewer and research advisor trained specifically on a corpus of documents and paired reviews. Its purpose would be to provide feedback on grant applications, experimental designs, and manuscripts. Initial evaluations of LLMs as scientific reviewers have demonstrated substantial overlap between human and AI generated

Evaluation of Amyloid Risk and Neurodegeneration, MCSA: Mayo Clinic Study of Aging, mPower: Mobile Parkinson Disease Study, NACC: National Alzheimer's Coordinating Center, NIAGADS: National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site, RADAR-AD: Remote Assessment of Disease And Relapse – Alzheimer's Disease, ROSMAP: Religious Orders Study and Rush Memory and Aging Project, TIHM: Technology Integrated Health Management.

reviews and overall positive user perceptions of usefulness [98]. By simulating different review styles, from constructive mentoring to critical peer evaluation, this system is intended to help researchers refine hypotheses, strengthen experimental plans, and anticipate potential reviewer concerns.

Taken together, these MVPs represent a practical, additive approach to applying AI in AD research. Each is designed to address specific challenges faced by researchers and to serve as an early demonstration of how AI tools can integrate into scientific workflows. While these initial efforts are focused and targeted, they lay the foundation for more comprehensive systems in the future, where larger gains in efficiency and scalability may be possible through broader applications of AI technologies.

7. Phase 2: toward an integrated platform and foundation models for scientific discovery

While our initial focus is on developing targeted tools to address specific challenges in AD research, the longer-term vision extends beyond individual solutions. In Phase 2, the aim is twofold: to create a unified platform that integrates these capabilities into a cohesive system and to develop foundation models trained directly on biomedical data.

A key objective of Phase 2 is to combine the functions demonstrated in the initial MVPs into a single, integrated platform. Such a system is intended to help researchers navigate the entire scientific process more efficiently, providing interconnected tools for the iterative cycle of questioning, analysis, and discovery. The second major goal in Phase 2 involves developing foundation models specifically trained on large-scale biomedical data. Unlike general-purpose language models trained primarily on internet text, these biomedical foundation models would be developed using domain-specific datasets such as genomic and proteomic profiles, neuroimaging data, longitudinal clinical records, and results from both published and non-public studies. Unlike textual corpora, these raw biomedical data types are less prone to becoming outdated, offering a more durable substrate for foundational learning. Our strategy focuses on leveraging these resilient data sources to build more robust and broadly applicable models. These models are intended to capture complex patterns and relationships within the data, potentially enabling the generation of new hypotheses, predictions about disease mechanisms, or identification of biomarkers associated with disease progression and therapeutic response. Initial efforts have demonstrated that LLMs can generate novel and valid hypotheses even when tested on literature unrelated to the training data [99].

Overall, Phase 2 represents a transition from testing individual tools to building a unified, scalable AI system intended to support the entire biomedical research process. While data-related, methodological, and practical challenges remain, the potential for improved efficiency and deeper scientific insights underscores the importance of this next stage of development. Realizing this vision will require addressing these challenges, which are discussed in the following section.

8. Challenges and mitigation strategies

The application of AI in biomedical research, while promising, presents several important challenges that must be addressed to ensure scientific integrity and responsible use [100–102]. From a technical perspective, developing and deploying advanced AI models requires significant computational resources and specialized infrastructure, which can pose practical barriers for many current research groups.

Data privacy and security also remain critical concerns, particularly given the sensitive nature of biomedical information and the ethical and legal frameworks (e.g., HIPAA, GDPR) that govern its use [42,43,103,104]. In addition, intellectual property and copyright restrictions can limit the use of scientific publications and proprietary datasets for training AI models or deploying research tools, due to licensing agreements and data ownership concerns. A related concern is the risk of data duplication across repositories, which can lead to biased analyses, redundant processing, and inflated sample sizes. Strategies for mitigating data duplication include use of persistent universally unique identifiers (UUIDs) for participant-level tracking where available, and matching algorithms to identify likely duplicate records in datasets without direct identifiers [105,106].

Beyond infrastructure and governance, adoption within the scientific community presents its own set of challenges. Many scientists, especially those less familiar with AI methods, may have valid concerns about the reliability, transparency, effectiveness, and interpretability of current generative AI outputs, including the risk of generating hallucinations [26]. Addressing these concerns will require clear communication about both the capabilities and limitations of AI systems, along with careful validation and demonstration of their practical value in scientific contexts.

The rapid pace of advancement in AI further complicates adoption. New tools, models, and best practices evolve quickly, making it difficult for biomedical researchers to stay current. At the same time, biomedical data itself is constantly growing, posing logistical and organizational challenges for version control, reproducibility, and integration with existing systems. Ensuring that models remain both accurate and aligned with the latest knowledge will require adaptive infrastructure, modular system designs, and ongoing collaboration between AI experts and domain scientists.

A particularly important set of challenges centers on ethical concerns, specifically algorithmic bias, accountability, and potential misuse [103,101,107]. Like all data-driven tools, AI models are susceptible to biases embedded in their training data, which can result in unequal performance across populations or misleading conclusions when trained on limited or biased data. The types, quality, amount, and range of biological data will completely change what an AI system can identify and discover at all levels. Many of the strategies used to detect and mitigate human bias in science, such as disaggregated analyses, dataset balancing, and transparency in decision-making, can and should be adapted to AI development and evaluation. Establishing accountability frameworks is also essential. These may include audit trails, logging systems, and traceable outputs that allow researchers to understand how a model arrived at a conclusion, assess its reliability, and flag potential errors. Such infrastructure is especially important as AI becomes more integrated into workflows, where overreliance or automation bias may lead to uncritical acceptance of flawed results. Misuse can also take more subtle forms, such as reinforcing low-quality analyses or contributing to inefficiencies. These risks highlight the importance of a scientist-in-the-loop design, in which both users and developers share responsibility for evaluating outputs, selecting appropriate tools, and maintaining high standards of scientific integrity throughout the AI development process.

Responsible development of AI tools for biomedical research depends on maintaining high scientific standards [108]. Important principles include careful validation to understand performance and limitations, transparent reporting of how models are developed and assessed, and clear definitions of the roles and boundaries of human oversight. In particular, expert oversight is critical for curating high-quality input data and literature, filtering out low-quality or misleading information, and validating AI-generated outputs before they inform scientific conclusions. This close involvement of domain experts will help develop, judge, and rate AI tools, to maintain scientific credibility and relevance. The c-brAIn brings together a broad network of biomedical researchers to enable collective expert review of both inputs and outputs.

By thoughtfully addressing these challenges, AI has the potential to become a valuable tool for biomedical research, supporting scientists in navigating complex data and generating new insights, particularly in fields like AD. Maintaining scientific rigor, transparency, and human oversight will be essential to ensuring its responsible and effective use.

9. Anticipated impact on Alzheimer's research

Integrating AI into biomedical research has the potential to offer measurable benefits for the study of AD. The envisioned AI Biomedical Scientist, with its capacity to synthesize complex data and literature, is expected to support researchers in accelerating key scientific processes. Early estimates suggest that AI-assisted literature review could reduce the time required by ~25 % compared to traditional manual approaches [97]. Such improvements are particularly relevant in AD research, where timely insights can contribute to advancing therapeutic development. Looking ahead, broader integration of AI systems in Phase 2 may enable even greater efficiencies, potentially achieving gains of two- to ten-fold in certain research workflows [51].

Beyond improving efficiency, the assistant is intended to support the rigor and reproducibility of scientific research. By systematically integrating information from diverse sources, the AI system may help researchers identify consistent patterns and avoid pursuing directions less likely to yield meaningful results. This capability could contribute to more effective target identification and validation, potentially supporting higher success rates in experimental studies and subsequent clinical trials, although precise estimates of such impacts remain uncertain.

The potential benefits of AI tools in AD research extend beyond individual laboratories. Advanced analytics and knowledge synthesis capabilities, which have typically been available to large research institutions with substantial resources, could become more accessible to smaller labs and early-career investigators. Increasing the availability of these tools may help reduce disparities in research capacity and promote contributions from a more diverse scientific community.

In addition, AI systems could play a supportive role in training the next generation of biomedical researchers. By providing examples of literature analysis, hypothesis development, and experimental critique, the AI Biomedical Scientist may help shorten learning curves for trainees and early-career scientists. Access to such resources could enhance scientific literacy and build confidence in working with complex, data-driven questions, potentially contributing to a more skilled and adaptable research community.

While precise metrics will continue to emerge as these systems are further developed and tested, integrating AI into AD research offers meaningful opportunities to improve scientific workflows, enhance reproducibility, and expand access to advanced analytical capabilities.

10. Future vision

The longer-term vision for AI in biomedical research extends beyond developing individual tools or even unified platforms. As foundational models and integrated systems mature, one key aspiration for future development lies in enabling semi-autonomous experimental design. In this scenario, AI systems could propose experimental protocols, suggest statistical methodologies, and forecast potential outcomes by synthesizing prior literature and integrated data models. While ultimate decision-making and oversight would remain in scientists' hands, the ability of AI to rapidly generate and evaluate experimental scenarios has the potential to increase research efficiency and facilitate exploration of more diverse scientific hypotheses. This capability would effectively yield a multiplier effect on capacity for research studies in the lab.

Expanding the application of AI tools beyond AD represents another important frontier. Technical architectures and methodological insights developed through neurodegenerative research are anticipated to be adaptable to other biomedical domains, including oncology, rare diseases, immunology, and complex chronic conditions. For example, AI systems capable of integrating diverse datasets, such as genomic profiles, imaging, and real-world clinical data, could help uncover shared pathways across diseases or identify patient subgroups more likely to benefit from specific therapies. Such cross-disciplinary insights might accelerate progress in fields where traditional research approaches have faced persistent challenges.

Beyond individual research programs, the broader vision for AI in biomedical science envisions a fundamental evolution in how discoveries are made. The traditional paradigm, characterized by linear hypothesis testing and siloed data sources, could increasingly give way to iterative, data-driven exploration powered by AI systems able to synthesize information across domains and scales. While realizing such capabilities will require significant advances in AI technologies, robust validation, and sustained collaboration between computational scientists and biomedical experts, this evolution holds the promise of accelerating the translation of basic research into clinical advances, ultimately contributing to improved diagnostics, therapies, and preventive strategies across a wide range of diseases.

11. Call to action

The development of an AI Biomedical Scientist marks a step toward transforming scientific discovery in complex fields like AD. The urgency of this challenge and the scale of resources and expertise it demands has led to the formation of a dedicated consortium committed to designing, building, and refining this new class of scientific tools.

A white paper outlining the scientific and technical vision for this initiative has been published and is available to the research community. But moving from vision to practical reality requires deep collaboration across disciplines, institutions, and sectors. The AI Biomedical Scientist is being developed as a tool built by scientists, for scientists. Its success will depend on diverse input and engagement from those who understand the complexities of biomedical research and the pressing need for new solutions.

For those interested in contributing or learning more, additional information is available at: https://c-brain.org. We invite researchers, clinicians, data scientists, and technology developers to join this effort. There are many ways to participate, from contributing domain expertise and engaging in pilot projects, to building and testing emerging tools, offering feedback on usability and performance, and sharing perspectives on critical scientific questions where AI could make a difference. Funders and philanthropic organizations interested in accelerating scientific progress are also encouraged to explore ways to support the development and broad dissemination of these tools. All interested parties can begin by filling out the survey on the c-brAIn website.

AD poses enormous challenges, but it also presents a profound opportunity: to harness innovative technologies and collaborative spirit to unlock new understanding and improve outcomes for patients and families affected by these devastating conditions. We invite the scientific community to help shape and build the next generation of tools that could redefine how discovery happens.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve language and readability. After using this tool/service, the authors reviewed and heavily edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Randall J. Bateman reports financial support was provided by Avid

Radiopharmaceuticals Inc. Randall J. Bateman reports financial support was provided by Janssen Pharmaceuticals Inc. Randall J. Bateman reports financial support was provided by Roche Diagnostics GmbH. Randall J. Bateman reports financial support was provided by Genentech Inc. Randall J. Bateman reports financial support was provided by Eli Lilly and Company. Randall J. Bateman reports financial support was provided by Eisai Inc. Randall J. Bateman reports financial support was provided by Biogen Inc. Randall J. Bateman reports financial support was provided by AbbVie Inc. Randall J. Bateman reports financial support was provided by Bristol-Myers Squibb Company. Randall J. Bateman reports financial support was provided by Novartis Pharmaceuticals Corporation. Randall J. Bateman reports a relationship with C2N Diagnostics, LLC that includes: board membership, consulting or advisory, and equity or stocks. Randall J. Bateman reports a relationship with Roche that includes: board membership. Randall J. Bateman reports a relationship with Biogen Inc that includes: board membership. Robert Lazar reports a relationship with Booz Allen Hamilton Inc that includes: employment. Sergio Pablo Sardi reports a relationship with Sanofi that includes: employment. Marc D. Voss reports a relationship with Sanofi that includes: employment. Joel W. Schwartz reports a relationship with Bristol Myers Squibb Co that includes: employment. Thomas Sandmann reports a relationship with Denali Therapeutics Inc that includes: employment. Gayle Wittenberg reports a relationship with Johnson & Johnson Inc that includes: employment. Niranjan Bose reports a relationship with Gates Ventures that includes: employment. Gregory J. Moore reports a relationship with Gates Ventures that includes: independent contract. Ludovico Mitchener reports a relationship with FutureHouse that includes: employment. Gabrielle Strobel reports a relationship with Alzforum that includes: employment. Elizabeth Wu reports a relationship with Alzforum Foundation Inc that includes: employment. Luca Foschini reports a relationship with Sage Bionetworks that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This initiative is supported by major consortium partners including Washington University, Gates Ventures, Alzforum, AD Data Initiative, Alzheimer's Association, Anonymous Foundation, Bristol Myers Squibb, Sanofi, Tau Consortium, The Dolby Family, The 10,000 Brains Project, The Robertson Foundation, Eisai, Johnson & Johnson, and over 100 engaged biomedical researchers. Figures were created in BioRender. https://BioRender.com/gq38ex5 and https://BioRender.com/tw98m0y

References

[1] Barthélemy NR, Salvadó G, Schindler SE, He Y, Janelidze S, Collij LE, et al. Highly accurate blood test for Alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. Nat Med 2024;30(4):1085–95. Apr.

[2] Bateman RJ, Li Y, McDade EM, Llibre-Guerra JJ, Clifford DB, Atri A, et al. Safety and efficacy of long-term gantenerumab treatment in dominantly inherited Alzheimer's disease: an open-label extension of the phase 2/3 multicentre, randomised, double-blind, placebo-controlled platform DIAN-TU trial. Lancet Neurol 2025;24(4):316–30. Apr 1.

[3] Sims JR, Zimmer JA, Evans CD, Lu M, Ardayfio P, Sparks J, et al. Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. JAMA 2023;330(6):512–27. Aug 8.

[4] Vigneswaran S, Vijverberg EGB, Barkhof F, van de Giessen E, Lemstra AW, Pijnenburg Y, et al. Real-world" eligibility for anti-amyloid treatment in a tertiary memory clinic setting. Alzheimers Dement 2025;21(6):e70375. June 12.

[5] Li Y, Yen D, Hendrix RD, Gordon BA, Dlamini S, Barthélemy NR, et al. Timing of biomarker changes in sporadic Alzheimer's disease in estimated years from symptom onset. Ann Neurol 2024;95(5):951–65. May.

[6] Ioannidis JPA. Why most clinical research is not useful. PLoS Med 2016;13(6): e1002049. June 21.

[7] Myszczynska MA, Ojamies PN, Lacoste AMB, Neil D, Saffari A, Mead R, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol 2020;16(8):440–56. Aug.

[8] Yiannopoulou KG, Anastasiou AI, Zachariou V, Pelidou SH. Reasons for failed trials of disease-modifying treatments for Alzheimer disease and their contribution in recent research. Biomedicines 2019;7(4):97. Dec 9.

[9] Pfeffer C, Olsen BR. Editorial: journal of negative Results in biomedicine. J Negat Results Biomed 2002;1(1):2. Nov 12.

[10] Bik EM. Publishing negative results is good for science. Access Microbiol 2024;6 (4):000792. Apr 2.

[11] Fanelli D. Negative results are disappearing from most disciplines and countries. Scientometrics 2012;90(3):891–904. Mar 1.

[12] Kearns WG, Stamoulis G, Glick J, Baisch L, Benner A, Brough D, et al. The application of knowledge engineering via the use of a biomimetic digital twin ecosystem, phenotype-driven variant analysis, and exome sequencing to understand the molecular mechanisms of disease. J Mol Diagn 2024;26(7): 543–51. July.

[13] Brazil R. Illuminating 'the ugly side of science': fresh incentives for reporting negative results. Nat [Internet] 2024. May 8 [cited 2025 Aug 22]Available from, https://www.nature.com/articles/d41586-024-01389-7.

[14] Park M, Leahey E, Funk RJ. Papers and patents are becoming less disruptive over time. Nature 2023;613(7942):138–44. Jan.

[15] Hanson MA, Barreiro PG, Crosetto P, Brockington D. The strain on scientific publishing. Quant Sci Stud 2024;5(4):823–43. Nov 1.

[16] Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al. Scientific discovery in the age of artificial intelligence. Nature 2023;620(7972):47–60. Aug.

[17] Tu T, Fang Z, Cheng Z, Spasic S, Palepu A, Stankovic KM, et al. Genetic discovery enabled by A large language model [Internet]. bioRxiv, https://www.biorxiv.org/content/10.1101/2023.11.09.566468v1; 2023.

[18] Hulsen T. Literature analysis of artificial intelligence in biomedicine. Ann Transl Med 2022;10(23):1284. Dec.

[19] Athanasopoulou K, Daneva GN, Adamopoulos PG, Scorilas A. Artificial intelligence: the milestone in modern biomedical research. BioMedInformatics 2022;2(4):727–44. Dec.

[20] Kwa T, West B, Becker J, Deng A, Garcia K, Hasin M, et al. Measuring AI ability to complete long tasks [Internet]. arXiv, http://arxiv.org/abs/2503.14499; 2025.

[21] Gao S, Fang A, Huang Y, Giunchiglia V, Noori A, Schwarz JR, et al. Empowering biomedical discovery with AI agents. Cell 2024;187(22):6125–51. Oct 31.

[22] Huang K., Zhang S., Wang H., Qu Y., Lu Y., Roohani Y., et al. Biomni: a general-purpose biomedical AI agent. bioRxiv. 2025 June 2;2025.05.30.656746.

[23] Shah C. From prompt engineering to prompt science with Human in the loop [Internet]. arXiv, http://arxiv.org/abs/2401.04122; 2024.

[24] Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models [Internet]. arXiv, http://arxiv.org/abs/2307.06435; 2024.

[25] Pantha N, Ramasubramanian M, Gurung I, Maskey M, Ramachandran R. Challenges in guardrailing large language models for science [Internet]. arXiv, http://arxiv.org/abs/2411.08181; 2024.

[26] Massenon R, Gambo I, Khan JA, Agbonkhese C, Alwadain A. My AI is lying to me": user-reported LLM hallucinations in AI mobile apps reviews. Sci Rep 2025; 15(1):30397. Aug 19.

[27] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40. Feb 15.

[28] Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Br Bioinform 2022;23(6): bbac409. Nov 1.

[29] Bolton E, Venigalla A, Yasunaga M, Hall D, Xiong B, Lee T, et al. BioMedLM: a 2.7B parameter language model trained on biomedical text [Internet]. arXiv, http://arxiv.org/abs/2403.18421; 2024.

[30] Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of Gemini models in medicine [Internet]. arXiv, http://arxiv.org/abs/2404.18416; 2024.

[31] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80. Aug.

[32] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey [Internet]. arXiv, http://arxiv.org/abs/2312.10997; 2024.

[33] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 9459–74. NIPS '20.

[34] Han H, Wang Y, Shomer H, Guo K, Ding J, Lei Y, et al. Retrieval-augmented generation with graphs (GraphRAG) [Internet]. arXiv, http://arxiv.org/abs/2501.00309; 2025.

[35] Nisa U, Shirazi M, Saip MA, Pozi MSM. Agentic AI: the age of reasoning—A review. J Autom Intell [Internet] 2025. Aug 28 [cited 2025 Oct 6]Available from, https://www.sciencedirect.com/science/article/pii/S2949855425000516.

[36] Acharya DB, Kuppan K, Divya B. Agentic AI: autonomous intelligence for complex goals—A comprehensive survey. IEEE Access 2025;13:18912–36.

[37] Zhou, J., Chen, S., Wu, Y., Li, H., Zhang, B., Zhou, L., et al. PPML-Omics: a privacy-preserving federated machine learning method protects patients' privacy in omic data. Sci Adv. 2024, 10 (5):eadh8601.

[38] Avraam D, Wilson RC, Aguirre Chan N, Banerjee S, Bishop TRP, Butters O, et al. DataSHIELD: mitigating disclosure risk in a multi-site federated analysis platform. Bioinform Adv 2025;5(1). Mar 10vbaf046.

[39] Escriba-Montagut X, Marcon Y, Anguita-Ruiz A, Avraam D, Urquiza J, Morgan AS, et al. Federated privacy-protected meta- and mega-omics data

analysis in multi-center studies with a fully open-source analytic platform. PLoS Comput Biol 2024;20(12):e1012626. Dec 9.

[40] Federated Data Sharing Appliance full guide - v1.3.32 - resources - Federated Data Sharing Appliance (FDSA) - AD Connect [Internet]. https://community.addi.ad-datainitiative.org/fdsa/m/resources/557; 2024.

[41] McMahan B, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics [Internet]. PMLR; 2017. p. 1273–82 [cited 2025 Oct 6]Available from, https://proceedings.mlr.press/v54/mcmahan17a.html.

[42] Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020;10(1):12598. July 28.

[43] Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research with federated learning. NPJ Digit Med 2021;4(1):132. Sept 7.

[44] Zhang F, Kreuter D, Chen Y, Dittmer S, Tull S, Shadbahr T, et al. Recent methodological advances in federated learning for healthcare. Patterns [Internet] 2024;5(6). June 14 [cited 2025 Oct 6]Available from, https://www.cell.com/patterns/abstract/S2666-3899(24)00131-4.

[45] Xu T, Feng J, Melendez J, Roberts K, Cai D, Zhu M, et al. Addressing accuracy and hallucination of LLMs in Alzheimer's disease research through knowledge graphs [Internet]. arXiv, http://arxiv.org/abs/2508.21238; 2025.

[46] Skarlinski MD, Cox S, Laurent JM, Braza JD, Hinks M, Hammerling MJ, et al. Language agents achieve superhuman synthesis of scientific knowledge [Internet]. arXiv, http://arxiv.org/abs/2409.13740; 2024.

[47] Pu K, Feng KJK, Grossman T, Hope T, Mishra BD, Latzke M, et al. IdeaSynth: iterative research idea development through evolving and composing idea facets with literature-grounded feedback. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems [Internet]; 2025. p. 1–31 [cited 2025 July 18]Available from, http://arxiv.org/abs/2410.04025.

[48] Lu C, Lu C, Lange RT, Foerster J, Clune J, Ha D. The AI scientist: towards fully automated open-ended scientific discovery [Internet]. arXiv, http://arxiv.org/abs/2408.06292; 2024.

[49] Wei J, Yang Y, Zhang X, Chen Y, Zhuang X, Gao Z, et al. From AI for science to agentic science: a survey on autonomous scientific discovery [Internet]. arXiv, http://arxiv.org/abs/2508.14111; 2025.

[50] Gottweis J, Weng WH, Daryin A, Tu T, Palepu A, Sirkovic P, et al. Towards an AI co-scientist [Internet]. arXiv, http://arxiv.org/abs/2502.18864; 2025.

[51] Ghareeb AE, Chang B, Mitchener L, Yiu A, Szostkiewicz CJ, Laurent JM, et al. Robin: a multi-agent system for automating scientific discovery [Internet]. arXiv, http://arxiv.org/abs/2505.13400; 2025.

[52] Lahiri AK, Hu QV. AlzheimerRAG: multimodal retrieval augmented generation for clinical use cases using PubMed articles [Internet]. arXiv, http://arxiv.org/abs/2412.16701; 2025.

[53] Naumov V, Zagirova D, Lin S, Xie Y, Gou W, Urban A, et al. DORA AI scientist: multi-agent virtual research team for scientific exploration discovery and automated report generation [Internet]. bioRxiv, https://www.biorxiv.org/content/10.1101/2025.03.06.641840v1; 2025.

[54] Sedler AR, Mitchell CS. SemNet: using local features to navigate the biomedical concept graph. Front Bioeng Biotechnol 2019;7:156.

[55] Swanson K, Wu W, Bulaong NL, Pak JE, Zou J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. Nature 2025:1–3. July 29.

[56] Ifargan T, Hafner L, Kern M, Alcalay O, Kishony R. Autonomous LLM-driven research — From data to Human-verifiable research papers. NEJM AI 2025;2(1): AIoa2400555. Jan.

[57] Istrate AM, Milletari F, Castrotorres F, Tomczak JM, Torkar M, Li D, et al. rbio1-training scientific reasoning LLMs with biological world models as soft verifiers [Internet]. bioRxiv, https://www.biorxiv.org/content/10.1101/2025.08.18.670981v2; 2025.

[58] Chai J, Tang S, Ye R, Du Y, Zhu X, Zhou M, et al. SciMaster: towards general-purpose scientific AI agents, part I. X-Master Found: Can We Lead Humanity 19s Last Exam? [Internet] 2025. arXiv[cited 2025 Aug 29]Available from, http://arxiv.org/abs/2507.05241.

[59] Luo Y, Shi L, Li Y, Zhuang A, Gong Y, Liu L, et al. From intention to implementation: automating biomedical research via LLMs. Sci China Inf Sci 2025;68(7):170105. June 23.

[60] Roohani Y, Lee A, Huang Q, Vora J, Steinhart Z, Huang K, et al. BioDiscoveryAgent: an AI agent for designing genetic perturbation experiments [Internet]. arXiv, http://arxiv.org/abs/2405.17631; 2025.

[61] Ding N, Qu S, Xie L, Li Y, Liu Z, Zhang K, et al. Automating exploratory proteomics research via language models [Internet]. arXiv, http://arxiv.org/abs/2411.03743; 2024.

[62] Jin R, Zhang Z, Wang M, Cong L. STELLA: self-evolving LLM agent for biomedical research [Internet]. arXiv, http://arxiv.org/abs/2507.02004; 2025.

[63] Andrieu S, Bateman RJ, Bereczki E, Bose N, Brookes AJ, Doraiswamy PM, et al. Harnessing artificial intelligence to transform Alzheimer's disease research. Nat Med 2025;31(5):1384–5. May.

[64] Veitch DP, Weiner MW, Miller M, Aisen PS, Ashford MA, Beckett LA, et al. The Alzheimer's Disease Neuroimaging Initiative in the era of Alzheimer's disease treatment: a review of ADNI studies from 2021 to 2022. Alzheimers Dement 2024;20(1):652–94. Jan.

[65] Daniels A.J., McDade E., Llibre-Guerra J.J., Xiong C., Perrin R.J., Ibanez L., et al. 15 Years of longitudinal genetic, clinical, cognitive, imaging, and biochemical measures in DIAN. medRxiv. 2024 Aug 9;2024.08.08.24311689.

[66] Wagemann O, Liu H, Wang G, Shi X, Bittner T, Scelsi MA, et al. Downstream biomarker effects of Gantenerumab or Solanezumab in dominantly inherited Alzheimer disease: the DIAN-TU-001 randomized clinical trial. JAMA Neurol 2024;81(6):582–93. June 1.

[67] Greenwood AK, Montgomery KS, Kauer N, Woo KH, Leanza ZJ, Poehlman WL, et al. The AD Knowledge Portal: a repository for multi-omic data on Alzheimer's disease and Aging. Curr Protoc Hum Genet 2020;108(1):e105. Dec.

[68] McHugh CP, Clement MHS, Phatak M. AD Workbench: transforming Alzheimer's research with secure, global, and collaborative data sharing and analysis. Alzheimers Dement 2025;21(5):e70278. May 19.

[69] Lovestone S, Imam F. The GNPC provides a proteomic resource for biomarker discovery and mechanistic insight in neurodegenerative disease. Nat Aging 2025; 5(7):1181–5. July.

[70] Crawford KL, Neu SC, Toga AW. The Image and Data Archive at the Laboratory of Neuro Imaging. Neuroimage 2016;124:1080–3. Jan 1Pt B.

[71] Beekly DL, Ramos EM, Lee WW, Deitrich WD, Jacka ME, Wu J, et al. The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. Alzheimer Dis Assoc Disord 2007;21(3):249–58.

[72] Leung YY, Lee WP, Kuzma AB, Nicaretta H, Valladares O, Gangadharan P, et al. Alzheimer's Disease Sequencing Project release 4 whole genome sequencing dataset. Alzheimers Dement 2025;21(5):e70237. May.

[73] Beecham GW, Bis JC, Martin ER, Choi SH, DeStefano AL, van Duijn CM, et al. The Alzheimer's Disease Sequencing Project: study design and sample selection. Neurol Genet 2017;3(5):e194. Oct.

[74] Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 2009;21(4):672–87. Aug.

[75] Pichet Binette A, Gaiteri C, Wennström M, Kumar A, Hristovska I, Spotorno N, et al. Proteomic changes in Alzheimer's disease associated with progressive Aβ plaque and tau tangle pathologies. Nat Neurosci 2024;27(10):1880–91. Oct.

[76] Jessen F, Spottke A, Boecker H, Brosseron F, Buerger K, Catak C, et al. Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE). Alzheimer 19s Res Ther 2018;10(1):15. Feb 7.

[77] Saunders S, Gregory S, Clement MHS, Birck C, der Geyten S van, Ritchie CW. The European Prevention of Alzheimer's Dementia Programme: an Innovative Medicines Initiative-funded partnership to facilitate secondary prevention of Alzheimer's disease dementia. Front Neurol [Internet] 2022:13. Nov 22 [cited 2025 Aug 29]Available from, https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2022.1051543/full.

[78] Pérez-González AP, García-Kroepfly AL, Pérez-Fuentes KA, García-Reyes RI, Solis-Roldan FF, Alba-González JA, et al. The ROSMAP project: aging and neurodegenerative diseases through omic sciences. Front Neuroinform 2024;18: 1443865. Sept 16.

[79] Sperling RA, Donohue MC, Raman R, Rafii MS, Johnson K, Masters CL, et al. Trial of Solanezumab in preclinical Alzheimer's disease. N Engl J Med 2023;389(12): 1096–107. Sept 20.

[80] Rafii MS, Sperling RA, Donohue MC, Zhou J, Roberts C, Irizarry MC, et al. The AHEAD 3–45 study: design of a prevention trial for Alzheimer's disease. Alzheimers Dement 2023;19(4):1227–33. Apr.

[81] Holstege H, Beker N, Dijkstra T, Pieterse K, Wemmenhove E, Schouten K, et al. The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. Eur J Epidemiol 2018;33(12):1229–49.

[82] Bianchi DW, Brennan PF, Chiang MF, Criswell LA, D'Souza RN, Gibbons GH, et al. The All of Us Research Program is an opportunity to enhance the diversity of US biomedical research. Nat Med 2024;30(2):330–3. Feb.

[83] Cai Y, Zhou J, Scott PW, Tian Q, Wanigatunga AA, Lipsitz L, et al. Physical activity complexity, cognition, and risk of cognitive impairment and dementia in the Baltimore Longitudinal Study of Aging. Alzheimers Dement (N Y) 2025;11(2): e70077.

[84] Shafto MA, Tyler LK, Dixon M, Taylor JR, Rowe JB, Cusack R, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC Neurol 2014;14:204. Oct 14.

[85] Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature 2023;613(7944):508–18. Jan.

[86] Fernandez MV, Liu M, Beric A, Johnson M, Cetin A, Patel M, et al. Genetic and multi-omic resources for Alzheimer disease and related dementia from the Knight Alzheimer Disease Research Center. Sci Data 2024;11(1):768. July 12.

[87] Bookheimer SY, Salat DH, Terpstra M, Ances BM, Barch DM, Buckner RL, et al. The Lifespan Human Connectome Project in Aging: an overview. Neuroimage 2019;185:335–48. Jan 15.

[88] Schwarz CG, Kremers WK, Prakaashana CM, Przybelski SA, Christenson LR, Wiliams JM, et al. A large public release of clinical and imaging data from the Mayo Clinic study of aging. Alzheimers Dement 2025;20(Suppl 9):e093966. Jan 9.

[89] Nilforooshan R., Barnaghi P. The RESILIENT dataset: multimodal monitoring of ageing-related comorbidities and cognitive decline [Internet]. Zenodo; 2025 [cited 2025 Aug 29]. Available from: https://zenodo.org/records/16755408.

[90] Huang X, Han X, Chang H, Yu T, Dong Y, Mao M, et al. Associations between trajectories of plasma biomarkers for Alzheimer's disease, brain structures, and cognitive function: a prospective cohort study in the UK Biobank. Mol Psychiatry 2025. Aug 28.

[91] Palchuk MB, London JW, Perez-Rey D, Drebert ZJ, Winer-Jones JP, Thompson CN, et al. A global federated real-world data and analytics platform for research. JAMIA Open 2023;6(2). Julyooad035.

[92] Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data 2016;3: 160011. Mar 3.

[93] Lentzen M, Vairavan S, Muurling M, Alepopoulos V, Atreya A, Boada M, et al. RADAR-AD: assessment of multiple remote monitoring technologies for early detection of Alzheimer's disease. Alzheimers Res Ther 2025;17(1):29. Jan 27.

[94] Palermo F, Chen Y, Capstick A, Fletcher-Loyd N, Walsh C, Kouchaki S, et al. TIHM: an open dataset for remote healthcare monitoring in dementia. Sci Data 2023;10(1):606. Sept 9.

[95] Tomczyk P, Brüggemann P, Mergner N, Petrescu M. Are AI tools better than traditional tools in literature searching? Evidence from E-commerce research. J Librariansh Inf Sci 2024. Nov 1509610006241295802.

[96] Lau O, Golder S. Comparison of elicit AI and traditional literature searching in evidence syntheses using four case studies. Cochrane Evid Synth Methods 2025;3 (6):e70050. Nov.

[97] Wang Z, Cao L, Jin Q, Chan J, Wan N, Afzali B, et al. A foundation model for human-AI collaboration in medical literature mining [Internet]. arXiv, http://arxiv.org/abs/2501.16255; 2025.

[98] Liang W, Zhang Y, Cao H, Wang B, Ding DY, Yang X, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI 2024;1(8):AIoa2400196. July 25.

[99] Qi B, Zhang K, Tian K, Li H, Chen ZR, Zeng S, et al. Large language models as biomedical hypothesis generators: a comprehensive evaluation [Internet]. arXiv, http://arxiv.org/abs/2407.08940; 2024.

[100] Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare—Perspectives on trustworthiness, explainability, usability, and transparency. npj Digit Med 2020;3(1):47. Mar 26.

[101] Gabriel I, Manzini A, Keeling G, Hendricks LA, Rieser V, Iqbal H, et al. The ethics of advanced AI assistants [Internet]. arXiv, http://arxiv.org/abs/2404.16244; 2024.

[102] Tang X, Jin Q, Zhu K, Yuan T, Zhang Y, Zhou W, et al. Risks of AI scientists: prioritizing safeguarding over autonomy [Internet]. arXiv, http://arxiv.org/abs/2402.04247; 2025.

[103] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1(9):389–99. Sept.

[104] Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25(1): 37–43. Jan.

[105] McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, et al. Identifiers for the 21st century: how to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol 2017;15(6):e2001414. June 29.

[106] Davis KR, Peabody B, Leach P. Universally unique IDentifiers (UUIDs) [Internet]. Internet Eng Task Force 2024. May [cited 2025 Oct 6]. Report No.: RFC 9562. Available from, https://datatracker.ietf.org/doc/rfc9562.

[107] Hofmann B. Biases in AI: acknowledging and addressing the inevitable ethical issues. Front Digit Health 2025;7:1614105.

[108] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3(1):160018. Mar 15.

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

AI models, bias and data sharing efforts to tackle Alzheimer's disease and related dementias

Vijaya B. Kolachalama [a,b,c,*], Vijay Sureshkumar [d], Rhoda Au [a,e,f,g]

[a] Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA, 02118
[b] Department of Computer Science, Boston University, MA, USA, 02215
[c] Faculty of Computing & Data Sciences, Boston University, MA, USA, 02215
[d] Gates Ventures, Seattle, WA, USA, 98033
[e] Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA, 02118
[f] The Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA, 02118
[g] Departments of Anatomy & Neurobiology and Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, 02118

ARTICLE INFO

Keywords:
Artificial intelligence
(AI)
Bias
Data sharing

ABSTRACT

Artificial intelligence (AI), often seen as a harbinger of future innovation, also presents a dilemma: it can perpetuate existing human biases. However, this issue is not novel or unique to AI. Humans have long been the progenitors of biases, and AI, as a product of human creation, often mirrors these inherent tendencies. Here, we present a perspective on the development and use of AI, recognizing it as a tool influenced by human input and societal norms, rather than an autonomous entity. Modern efforts to technologically enabled data collection approaches and model development, particularly in the context of Alzheimer's disease and related dementias, can potentially reduce bias in AI. We also highlight the importance of data sharing from existing legacy cohorts to help accelerate ongoing AI model development efforts for greater scientific good and clinical care.

1. Introduction

The study of biases, specifically perceptual biases, which are systematic patterns of deviation from normative or rational judgment, has been a significant area of research for centuries, spanning disciplines like psychology, sociology, and behavioral economics. These biases manifest in various forms, ranging from those characterized by noise to motivational biases influenced by experientially reinforced associations. B.F. Skinner famously put forward the notion that positive and negative reinforcements shaped human behavior [1–4]. Divergent from Skinner's supposition is the premise that this shaping process leads to intrinsic bias that influences human judgments and decisions simultaneously. In the context of artificial intelligence (AI), the challenge lies not within its own inherent nature but in how it reflects and amplifies our own prejudices. This realization makes necessary a balanced perspective that neither blindly dismisses AI as fundamentally flawed nor uncritically heralds it as an unblemished force for good.

Bias in AI predominantly originates from the data that these systems are trained on. This data is generated through humanly designed methods and sources, leading to permeation of the prevailing prejudices and disparities that have historically always shaped human societies. In the context of Alzheimer's disease (AD) and related dementias (ADRD), these biases can manifest in cohort selection, diagnostic labeling, and access to healthcare resources, potentially skewing the development and performance of AI models intended to detect or monitor cognitive impairment [5,6]. For instance, imbalanced diagnostic imaging datasets that lack population representation may lead to underdiagnosis in underrepresented groups, delaying treatment and worsening prognosis, as seen in models that perform poorly on non-White populations. Similarly, in biomarker analysis, AI might overestimate risk based on genetic factors like APOE4 that vary by ethnicity, leading to inequitable treatment planning. In cognitive testing, biased algorithms could misclassify symptoms in diverse linguistic or socioeconomic groups, impacting early intervention and clinical outcomes. If left unaddressed, such biases risk perpetuating disparities in diagnosis and treatment, especially among underrepresented populations. When AI algorithms process and interpret this data, they risk not only adopting these biases but also amplifying them, inadvertently reinforcing societal inequalities they might be employed to mitigate. This amplification is particularly concerning given AI's wide-reaching impact and its role in decision-making

[*] Corresponding author.
E-mail address: vkola@bu.edu (V.B. Kolachalama).

https://doi.org/10.1016/j.tjpad.2025.100400
Received 3 September 2025; Received in revised form 8 October 2025; Accepted 19 October 2025

processes across various sectors. It is important to note that the presence of bias in AI outcomes does not necessarily indicate that AI systems are themselves inherently biased. Instead, it reflects the biases that are already embedded in the data sources used for training AI. These biases may arise from a range of sources: from skewed sampling methods that overlook different segments of the population to historical data that inherently carries the biases of past societies. The recognition of AI's tendency to reflect and magnify existing biases presents a paradoxical yet unique opportunity. AI, with its advanced data analysis capabilities, can serve as a powerful tool in identifying and dissecting the biases embedded within our systems and decision-making processes [7,8]. AI can process vast quantities of data, identify patterns and correlations that might be imperceptible to human analysis, and provide an objective view of systemic biases.

AI's potential extends beyond merely reflecting human biases; it also possesses the capacity to challenge and rectify them. For example, in the context of ADRD research and care delivery, AI can identify diagnostic disparities, such as lower accuracy in detecting AD among Hispanic populations in the Health and Retirement Study (HRS) [9], and propose adjustments to clinical algorithms to enhance equity in cognitive screening and referrals. Additionally, by analyzing longitudinal data, AI can reveal racial biases, like overestimating dementia risk in African American cohorts due to unrepresentative training data. It can also examine loan approval rates to detect biases in financial services that are barriers to long-term economic growth [10]. By bringing these insights to the forefront, AI can enable organizations and societies to confront and address these biases more effectively. In the context of science, AI can serve as a catalyst for positive transformation but realizing this opportunity hinges on developing and steering AI with ethical considerations at its core. AI as a catalyst for positive transformation of science is contingent upon the development and guidance of AI with a strong emphasis on ethical considerations. Precision medicine heralds a new era of personalized healthcare but can only happen if the data that is used to generate solutions is sufficiently reflective of the population it seeks to serve. By embedding these principles of ethics and representativeness at the core of AI's design and application, we can harness its transformative capabilities for greater scientific good. To facilitate user comprehension, we present a list of technical terms with brief explanations in Table 1.

Table 1
Glossary of technical terms.

Adversarial debiasing	A technique in AI training where an opposing model component removes correlations with biased attributes (like race or gender) to make predictions fairer.
Algorithmic transparency	The practice of making AI decision-making processes clear and understandable, so users can see how outcomes are reached.
Bias in AI	Systematic errors in AI models that lead to unfair outcomes, often reflecting prejudices in the training data.
CRediT taxonomy	A system for crediting contributors to research based on their specific roles, like data collection or analysis.
Data sovereignty	The right of individuals or communities to control how their data is used and shared.
Demographic distributions	Statistics showing how a dataset is divided by factors like age, gender, race, or location.
Differential privacy	A method to protect individual data by adding noise to outputs, allowing analysis without revealing personal information.
Explainability	The ability to understand and explain how an AI model arrives at its conclusions.
FAIR principles	Guidelines for data management: Findable, Accessible, Interoperable, and Reusable, to improve sharing and reuse.
Federated learning	A method where AI models are trained across multiple locations without sharing raw data, to protect privacy.
General Data Protection Regulation (GDPR)	A European law that sets rules for handling personal data to ensure privacy.
International Committee of Medical Journal Editors (ICMJE)	A group that sets guidelines for ethical publishing in medical journals, including authorship rules.
Missingness patterns	Trends in a dataset showing where data is incomplete or absent, which can indicate biases.
Privacy-preserving architectures	Systems designed to protect personal information while allowing data analysis, like federated learning.
Secure multi-party computation	A technique allowing multiple groups to compute results together without revealing their private data.
SHAP (Shapley Additive exPlanations)	A method to explain AI predictions by assigning importance values to each input feature.

2. The value and challenges of data sharing

Public data sharing as an expedient solution. Institutions, researchers, and stakeholders worldwide are increasingly recognizing the crucial role of data diversity in AI development [11]. They are raising awareness about how AI models trained on biased data can inadvertently perpetuate scientific insights that are limited in scope and utility [12]. This understanding has led to a concerted effort to collect data from diverse sources, encourage the recruitment of people from various backgrounds, and promote open-source data sharing of newly collected data. Programs like the All of Us Research Program exemplify these efforts by prioritizing representativeness of the U.S. population to create a more comprehensive and less biased dataset for further development [13–15]. However, while progress is being made in gathering and disseminating diverse datasets for future AI models, a significant challenge persists in the context of existing or legacy data. The underutilization of data collected over decades across many studies remains hindered by various barriers to current data sharing approaches. To realize the promise of these initiatives, practical limitations in current data sharing practices must be addressed.

Importance of legacy data. The inclusion of existing or legacy data in AI development is critical for advancing ADRD research. First, legacy datasets, such as those from the Alzheimer's Disease Neuroimaging Initiative (ADNI), [16] serve as a valuable source from which to identify historical biases and disparities and then correct them, enabling researchers to prevent perpetuation of inaccurate presumptions in AI models, which is essential to ensure AI fairness and neutrality. For instance, ADNI cohorts predominantly feature White participants, with biases including skewed educational status toward higher-status individuals, gender imbalances, genetic overrepresentation (e.g., of APOE4+ carriers), clinical preferences for MCI/dementia cases with fewer comorbidities, and selection/analytical flaws that overestimate performance. Second, integrating these datasets with diverse, newly collected data, e.g., from the AMP-AD multi-omics hub, [17] provides a comprehensive view, enabling AI to better reflect and serve underrepresented groups by illuminating long-term disease trajectories and evolving diagnostic thresholds. Third, decades-long prospective follow-up data, such as ADNI's verified incident outcomes, are invaluable for training prognostic models essential for ADRD prevention when combined with new collections. Fourth, combining legacy data with modern datasets fosters innovation, uncovering unique correlations, e.g., imaging-neuropsychology links in ADNI, that drive improved ADRD outcomes. Fifth, leveraging existing data optimizes resources by maximizing past investments reduces waste in data collection efforts. Finally, revisiting and managing legacy data ensures compliance with evolving privacy standards, safeguarding trust and mitigating legal risks, which is critical for datasets like ADNI. Together, these elements underscore the critical role of legacy data in building fairer, more informed, and efficient AI systems for ADRD.

3. Barriers to reuse of existing data

Resource constraints. In healthcare research and other fields, effective data sharing is frequently impeded by various factors including limited funding for data management infrastructure [18]. In research, there is a funding bias for the collection of new data relative to the funding of restructuring and reconstructing data collected using long-outdated measurement tools and stored in less secure and outdated useable formats. This financial scarcity impacts the ability to manage and share invaluable existing data effectively. Research funding also often prioritizes projects with immediate, tangible outcomes over more effortful and time-consuming infrastructural needs, such as updating old data management systems or developing more state-of-the art data sharing platforms. Additionally, the technical requirements for bringing secure and accessible data repositories to contemporary standards necessitate significant, sustained investments, often out of reach for smaller or under-funded projects. To overcome this, we recommend adopting FAIR (Findable, Accessible, Interoperable, Reusable) principles as a framework for data management, [19–24] which can guide efficient resource allocation and standardization. Funders could prioritize grants for FAIR-compliant infrastructure upgrades.

Direct access barriers. Another challenge in data sharing is the lack of incentives for researchers and institutions. Currently, career advancement and recognition are largely based on metrics like high-impact publications, grant funding, and individual contributions, often reflected in authorship position [25]. Sharing data, which involves significant effort to curate and maintain datasets, typically receives little comparable recognition, discouraging participation [26]. Additionally, even when data sharing initiatives are available, they often include restrictive conditions that pose obstacles. For example, contributors to data repositories may require researchers to navigate a multi-step permission process. Since approval often depends on human judgment beyond just data use agreements signed to ensure ethical use, it can reflect societal norms about acceptable science. Some researchers also hesitate to share datasets, treating them as valuable intellectual capital to maintain a competitive edge. The reliance on human-in-the-loop decisions for data access can further delay approvals, ultimately slowing scientific progress. Best practices include implementing automated, standardized access protocols aligned with FAIR principles, such as metadata templates for quick evaluation, and institutional policies that reward data sharing through metrics like data citation indices.

Data contributor recognition barriers. Many research groups that provide data also require authorship recognition. This stipulation typically takes two forms. First, the data access approval process may stipulate that contributing researchers be included as collaborators, and by extension, authors on any resulting publication. Second, some groups mandate pre-specified authorship inclusion, often requiring that all individuals involved in data collection be listed as co-authors, regardless of their involvement in conceptualization, study design, or analysis. In the first scenario, including data contributors may unintentionally perpetuate their scientific perspectives, particularly in hypothesis-driven research where investigators tend to pursue questions aligned with their prior beliefs. The second scenario reflects a broad authorship policy that conflicts with established standards [16]. Such practices run counter to the International Committee of Medical Journal Editors (ICMJE) authorship criteria, which require substantial contributions to the conception, design, analysis, or interpretation of data. Moreover, ethical authorship guidelines emphasize the importance of review and approval of the final manuscript, which can further reinforce the biases noted above. While recognizing contributions is important, these authorship requirements can complicate and deter broader use of the data. To address this, we propose adopting the CRediT taxonomy, ORCID-linked dataset DOIs, and data-citation indices to acknowledge data providers' efforts without mandating co-authorship, fostering equitable recognition and encouraging data sharing.

Platform utilization costs. Certain data repositories, such as the UK Biobank (UKBB), prohibit downloading data altogether, requiring researchers to use their platforms for analysis, [27] where high fees or limited functionality often monetize access and stifle open collaboration. However, the UKBB's new subsidized cloud-credit program now supports enhanced access, offering a step toward improvement. These practices, despite the subsidy, can still hinder the accessibility and equitable utilization of shared data, particularly for researchers in low resourced research environments, ultimately slowing the collective progress of AI in healthcare that will impact all. To address these challenges, it is essential to implement structural changes in the academic recognition system and establish clear, fair data-sharing policies that balance acknowledgment with usability. We suggest policies like open-access subsidies and standardized metadata reporting under FAIR guidelines to reduce costs and enhance interoperability across platforms.

4. Ethics, legal and social considerations

The data sharing process is further complicated by privacy and ethical considerations. Protecting participant/patient confidentiality while ensuring informed consent for data usage presents intricate challenges, particularly when maintaining the research utility of anonymized data. With the increasing capabilities of AI and other analytical tools, it is becoming more difficult to guarantee absolute privacy, and a residual risk of re-identification remains even with de-identified data [28]. Efforts are underway to enable data access while obfuscating identities; [29] however, this remains an ongoing and inherently imperfect process that may never fully eliminate the risk. Privacy-preserving architectures such as federated learning and secure multi-party computation offer promising solutions to mitigate these risks. Specifically, federated learning allows collaborative model training across decentralized datasets without exchanging raw participant/patient data, enabling institutions to contribute to AI development while keeping sensitive information local and secure. For example, the Alzheimer's Disease Data Initiative (ADDI) Workbench employs federated data sharing through its Federated Data Sharing Appliance, [30] facilitating ADRD research by aggregating model parameters rather than centralizing data, thus enhancing privacy and data sovereignty. Similarly, EU-funded projects like TRUMPET advance federated learning in healthcare ensuring compliance with GDPR and promoting equitable global collaboration [31]. To build on this, differential privacy can be integrated into federated setups by adding controlled noise to model updates, preserving individual data while maintaining utility [32, 33]. Federated multi-site training further enables cross-institutional collaboration for ADRD prediction, such as quantifying the extent of hippocampal atrophy in AD from MRI data across hospitals without data transfer, improving model generalizability and reducing bias amplification [34]. These methods have shown practical success, e.g., achieving comparable accuracy to centralized models while ensuring compliance in dementia cohorts [35]. Informed consent for reuse of data in the future, either in its raw or derived form, must include data analysis for scientific objectives that are unknown at the time of consent. Data sovereignty is a crucial factor in research involving marginalized communities, demanding ethical practices that respect these groups' participation and decision-making regarding their data. Navigating these ethical complexities alongside the goals of broad data sharing requires adherence to robust ethical guidelines and careful consideration. This is particularly relevant in ADRD research involving groups who may be underrepresented in neuroimaging and biomarker studies. Ensuring their inclusion in shared datasets must go together with ethical safeguards and community engagement to promote trust and transparency.

5. Designing equitable and transparent platforms

Equitable data collection, technological accessibility for global

reach. To promote inclusive and impactful research, platforms for participant data collection and data sharing must be guided by principles of accessibility, global reach, explainability, and transparency. Ensuring usability across diverse settings, especially in low- and middle-income countries, requires support for multi-lingual, multi-region, and low-literacy use cases that support equitable data collection and inclusion of underrepresented populations globally.

Technological accessibility can be strengthened by designing open-access platforms with no-code interfaces that lower technical barriers and broaden participation. Strategic partnerships and community engagement are critical to extending these platforms to underrepresented researchers and populations, thereby helping democratize AI and data tools. Equally important is embedding explainability and transparency into system design. AI outputs must be auditable and reproducible, particularly in sensitive domains like biomedical research, to identify and reduce bias. Platforms should also integrate metrics and evaluation tools that track inclusivity, equity, and diversity throughout their development and deployment. Together, these design principles support a more equitable and trustworthy research infrastructure.

Importance of transparency and responsible use of data during data sharing. As AI models increasingly rely on multimodal datasets for ADRD research, ensuring transparency and detailed documentation demonstrating responsible use during data sharing becomes critical to uncover and mitigate potential biases. Many datasets, particularly legacy cohorts, may contain imbalances related to gender, race and ethnicity, socioeconomic status, and geographic representation. Without explicit documentation of these attributes, there is a substantial risk that AI models will unknowingly perpetuate or even amplify these historical inequities. To promote transparency, datasets should be accompanied by comprehensive summary statistics and demographic distributions. Key variables to report include gender breakdown, race and ethnicity composition, age ranges, socioeconomic indicators when available, and geographic distribution across urban, rural, and regional settings. In addition, repositories should document missingness patterns across critical modalities such as neuroimaging, cognitive assessments, and biomarker data. By providing these metrics upfront, researchers can critically assess the diversity, strengths, and limitations of datasets before proceeding with model development, allowing for more informed decisions about data use, subgroup analyses, and fairness evaluations.

Importantly, transparency efforts should extend beyond the data level to the modeling process itself to address explainability. AI developers should carry forward demographic metadata into training, validation, and testing stages, systematically reporting model performance across key subgroups. This practice enables the identification of differential error rates, highlights potential disparities and strengthens the interpretability and fairness of model outcomes and makes the models explainable. Embedding demographic transparency throughout the data-to-model pipeline will not only improve scientific rigor but also foster the development of AI systems that are equitable, generalizable, and clinically responsible. Ultimately, addressing explainability at the outset of data sharing lays the foundation for developing AI models that better reflect and serve diverse populations. As the ADRD field embraces increasingly complex and heterogeneous datasets, maintaining visibility of underlying biases throughout the research lifecycle will be essential to ensure that technological advances translate into meaningful, equitable clinical improvements.

6. Towards ethical AI: addressing bias through better practices

The criticism aimed at AI for propagating bias underscores the necessity of comprehending its multidimensional role. AI's efficacy and fairness are contingent on the quality of the data it learns from and the intent behind its creation and application. Consequently, developing and implementing AI demands a deliberate focus on ensuring data representativeness, algorithmic transparency, and constant vigilance for biases. Confronting AI's complexities necessitates a collaborative effort encompassing a broad spectrum of stakeholders, including technologists, ethicists, policymakers, and the communities they impact. Jointly, these groups can establish ethical AI standards, advocate for diverse and inclusive data practices, and implement ongoing evaluation mechanisms to detect and mitigate biases.

While upstream data practices (e.g., representative data collection and low barrier data access) are essential, biases can persist or emerge during model training and deployment. To mitigate this, targeted techniques that intervene directly in the AI pipeline include: (a) Adjusting the influence of underrepresented groups in training data. For instance, oversampling minority cohorts (e.g., non-White participants in ADRD datasets) or assigning higher weights to their samples can balance model learning and reduce disparities in prediction accuracy. (b) Training models with an adversary component that penalizes predictions correlated with protected attributes (e.g., race, sex). This encourages the model to focus on disease-relevant features, such as neuroimaging patterns, rather than demographic proxies [36]. (c) Incorporating fairness metrics directly into the loss function during training, such as enforcing equalized odds (ensuring similar true positive rates across groups) or demographic parity (equal prediction rates across groups). These methods build on general data equity but are tailored to AI's iterative learning process, helping prevent amplification of historical biases in legacy ADRD cohorts.

To assess whether mitigation efforts are effective, rigorous evaluation is key. Post-training audits can include: (a) Quantitative measures like disparate impact (ratio of favorable outcomes between groups, ideally close to 1.0) or equalized odds/error rates to quantify bias. For example, in an ADRD diagnostic model, evaluate if false negative rates are higher for underrepresented ethnic groups. (b) Stratify performance by demographics (e.g., via confusion matrices per race/ethnicity) and test robustness to perturbations in input data, revealing hidden biases. (c) Employ techniques like SHAP (SHapley Additive exPlanations) to interpret feature importance, [37,38] identifying if biased variables (e.g., socioeconomic proxies in HRS data) unduly influence ADRD risk predictions.

To demonstrate how mitigation can lead to better clinical outcomes, such as earlier accurate diagnoses and equitable treatment planning, consider the following ADRD-specific examples:

- *Racial bias in diagnostic models:* In models trained on predominantly White cohorts (e.g., ADNI), algorithms may overestimate dementia risk in African American individuals due to unrepresentative biomarkers like APOE4 prevalence. One way to address this is to apply reweighting during training and evaluate with equalized odds [39].
- *Sex/gender bias in prognostic models:* Female participants might be underrepresented in neuroimaging subsets, leading to poorer prediction of outputs for women [40,41]. A strategy to address is via adversarial training to decorrelate gender from outputs, combined with subgroup analysis to ensure balanced sensitivity/specificity.
- *Socioeconomic bias in screening tools:* AI relying on electronic health records may underperform for low-socioeconomic status groups due to access disparities. To address this aspect, one could integrate fairness constraints and test with disparate impact metrics, then refine via federated learning across diverse institutions [42].

Understanding the dual nature of AI, both as a reflector of human biases and a potential tool for overcoming them, is crucial in shaping its future trajectory. Rejecting AI as inherently flawed or uncritically accepting it as a universal solution oversimplifies its impact. Instead, recognizing AI as a human-crafted tool with the dual capacity to both mirror and amend our biases is key. By embracing the complexity and potential of AI, we can steer its evolution towards a future where it serves not only as a technological advancement but to foster equity and deeper understanding, aligning with our aspirations for a fairer and more insightful society. We commend the recent initiatives undertaken by various agencies to encourage the sharing of diverse datasets.

Additionally, we advocate for institutions, researchers, and other stakeholders to actively support the sharing of historical data. This collaboration is crucial for meaningful contributions to the development of AI models, both ongoing and future. This comprehensive approach to data sharing, encompassing both contemporary and legacy datasets, is essential for creating AI models that are robust, representative, and effective [7].

7. Conclusion

As AI continues to transform scientific discovery and healthcare delivery, particularly in the context of ADRD, the way we collect, document and then share data will profoundly shape its fairness and utility. The study of bias that is rooted in centuries of behavioral research reminds us of these cognitive distortions, once solely considered human, are now embedded in the datasets that train machine intelligence. While AI has the potential to reflect and reinforce longstanding inequities, it also offers a powerful mechanism for identifying and correcting them, if developed with intention, transparency, and inclusive design. Moving forward, a core requirement for ethically sound AI will be transparency in data sharing. Publishing summary statistics on key demographic and geographic variables, alongside detailed documentation of data completeness and structure, is no longer optional. It is foundational. Equally, ensuring that these metadata persist through the full modeling lifecycle will allow for subgroup-specific performance analysis, surfacing hidden inequities and enabling their mitigation. This level of transparency not only enhances model generalizability and interpretability but also builds public trust in AI systems used in high-stakes contexts such as ADRD. The path ahead requires collaborative leadership. Institutions, funders, and researchers must work together to remove disincentives for sharing well-annotated, diverse datasets including legacy data, and develop clear standards that reward openness and fairness. Ethical AI will not emerge by default; it will result from deliberate choices made at every level of the data and model pipeline. By confronting the structural origins of bias and embedding equity as a design principle, the field can move toward an AI-enabled future that serves all populations accurately, responsibly, and justly.

8. Ethics declarations

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. R.A. is a scientific advisor to Signant Health and NovoNordisk. The remaining author declares no competing interests.

Declaration of competing interest

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. R.A. is a scientific advisor to Signant Health and NovoNordisk. The remaining author declares no competing interests.

Funding

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

References

[1] Skinner BF. Selection by consequences. Science 1981;213(4507):501–4. https://doi.org/10.1126/science.7244649. PubMed PMID: 7244649.

[2] Skinner BF. The evolution of behavior. J Exp Anal Behav 1984;41(2):217–21. https://doi.org/10.1901/jeab.1984.41-217. PubMed PMID: 6716037; PMCID: PMC1348035.

[3] Skinner BF. The phylogeny and ontogeny of behavior. Contingencies of reinforcement throw light on contingencies of survival in the evolution of behavior. Science 1966;153(3741):1205–13. https://doi.org/10.1126/science.153.3741.1205. PubMed PMID: 5918710.

[4] Skinner BF. The shaping of phylogenic behavior. J Exp Anal Behav 1975;24(1): 117–20. https://doi.org/10.1901/jeab.1975.24-117. PubMed PMID: 16811859; PMCID: PMC1333387.

[5] Magdamo C.G., He Y., Dickson J.R., Tyagi T., Westover M.B., Mukerji S., Ritchie C. S., Hyman B.T.T., Blacker D., Das S. Evaluating sociodemographic bias in an artificial intelligence algorithm to detect cognitive impairment in electronic health records. Alzheimer's & dementia. 2025;20(S4). doi: 10.1002/alz.093404.

[6] Yuan C, Linn KA, Hubbard RA. Algorithmic fairness of machine learning models for Alzheimer disease progression. JAMA Netw Open 2023;6(11). https://doi.org/10.1001/jamanetworkopen.2023.42203.

[7] Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. npj Digit Med 2025;8(1). https://doi.org/10.1038/s41746-025-01503-7.

[8] Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-centered design to address biases in artificial intelligence. J Med Internet Res 2023:25. https://doi.org/10.2196/43251.

[9] Ofstedal MB, Weir DR. Recruitment and retention of minority participants in the health and retirement study. Gerontologist 2011;51(1):S8–20. https://doi.org/10.1093/geront/gnq100.

[10] Purificato E, Lorenzo F, Fallucchi F, De Luca EW. The use of responsible artificial intelligence techniques in the context of loan approval processes. Int J Hum–Comput Interact 2022;39(7):1543–62. https://doi.org/10.1080/10447318.2022.2081284.

[11] Zowghi D, Bano M. AI for all: diversity and inclusion in AI. AI Ethics 2024;4(4): 873–6. https://doi.org/10.1007/s43681-024-00485-8.

[12] Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns 2021;2(10). https://doi.org/10.1016/j.patter.2021.100347.

[13] Amrollahi F, Shashikumar SP, Meier A, Ohno-Machado L, Nemati S, Wardi G. Inclusion of social determinants of health improves sepsis readmission prediction models. J Am Med Inf Assoc 2022;29(7):1263–70. https://doi.org/10.1093/jamia/ocac060. PubMed PMID: 35511233; PMCID: PMC9196687.

[14] Khoury MJ, Bowen MS, Clyne M, Dotson WD, Gwinn ML, Green RF, Kolor K, Rodriguez JL, Wulf A, Yu W. From public health genomics to precision public health: a 20-year journey. Genet Med 2018;20(6):574–82. https://doi.org/10.1038/gim.2017.211. Epub 20171214PubMed PMID: 29240076; PMCID: PMC6384815.

[15] Rasooly RS, Gossett DR, Henderson MK, Hubel A, Thibodeau SN. High-throughput processing to preserve viable cells: a Precision Medicine Initiative cohort program workshop. Biopreserv Biobank 2017;15(4):341–3. https://doi.org/10.1089/bio.2017.0016. Epub 20170425PubMed PMID: 28441039; PMCID: PMC5582583.

[16] Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, Weiner MW. Alzheimer's Disease Neuroimaging Initiative (ADNI). Neurology 2010;74(3):201–9. https://doi.org/10.1212/WNL.0b013e3181cb3e25.

[17] Ertekin-Taner N, Petanceska SS. Ten year anniversary of AMP AD: enabling a precision medicine approach to target and biomarker discovery for Alzheimer's disease. Alzheimer's Dement 2025;20(S1). https://doi.org/10.1002/alz.086464.

[18] Grinnell F, Anger M, Wendelborn C, Winkler EC, Schickhardt C. Neither carrots nor sticks? Challenges surrounding data sharing from the perspective of research funding agencies—A qualitative expert interview study. Plos One 2022;17(9). https://doi.org/10.1371/journal.pone.0273259.

[19] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J. Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3(1). https://doi.org/10.1038/sdata.2016.18.

[20] Barker M, Chue Hong NP, Katz DS, Lamprecht A-L, Martinez-Ortiz C, Psomopoulos F, Harrow J, Castro LJ, Gruenpeter M, Martinez PA. Honeyman T. Introducing the FAIR Principles for research software. Sci Data 2022;9(1). https://doi.org/10.1038/s41597-022-01710-x.

[21] Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, Pétavy F, Galvez J, Becnel LB, Zhou FL, Harmon N, Jauregui B, Jackson T, Hudson L. FAIR data sharing: the roles of common data elements and harmonization. J Biomed Inf 2020:107. https://doi.org/10.1016/j.jbi.2020.103421.

[22] Tai KH, Müller M, Mansmann U, Vieira Armond AC, Decullier E, Le Louarn A, Munung NS, Naudet F, Prasser F, Sax U. Key concepts in clinical epidemiology: fAIRification of biomedical research data. J Clin Epidemiol 2025;187. https://doi.org/10.1016/j.jclinepi.2025.111920.

[23] Fouad K, Vavrek R, Surles-Zeigler MC, Huie JR, Radabaugh HL, Gurkoff GG, Visser U, Grethe JS, Martone ME, Ferguson AR, Gensel JC. Torres-Espin A. A practical guide to data management and sharing for biomedical laboratory

researchers. Exp Neurol 2024;378. https://doi.org/10.1016/j.expneurol.2024.114815.

[24] David R, Rybina A, Burel JM, Heriche JK, Audergon P, Boiten JW, Coppens F, Crockett S, Exter K, Fahrner S, Fratelli M, Goble C, Gormanns P, Grantner T, Grüning B, Gurwitz KT, Hancock JM, Harmse H, Holub P, Juty N, Karnbach G, Karoune E, Keppler A, Klemeier J, Lancelotti C, Legras JL, Lister AL, Longo DL, Ludwig R, Madon B, Massimi M, Matser V, Matteoni R, Mayrhofer MT, Ohmann C, Panagiotopoulou M, Parkinson H, Perseil I, Pfander C, Pieruschka R, Raess M, Rauber A, Richard AS, Romano P, Rosato A, Sánchez-Pla A, Sansone SA, Sarkans U, Serrano-Solano B, Tang J, Tanoli Z, Tedds J, Wagener H, Weise M, Westerhoff HV, Wittner R, Ewbank J, Blomberg N, Gribbon P. Be sustainable": eOSC-Life recommendations for implementation of FAIR principles in life science data handling. EMBO J 2023;42(23). https://doi.org/10.15252/embj.2023115008.

[25] Borgman CL. The conundrum of sharing research data. J Am Soc Inf Sci Technol 2012;63(6):1059–78. https://doi.org/10.1002/asi.22634.

[26] Hrynaszkiewicz I, Altman DG. Towards agreement on best practice for publishing raw clinical trial data. Trials 2009;10(1). https://doi.org/10.1186/1745-6215-10-17.

[27] Watts G. UK Biobank opens its data vaults to researchers. Bmj 2012;344(2):e2459. https://doi.org/10.1136/bmj.e2459. -e.

[28] Scherer RW, El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS ONE 2011;6(12). https://doi.org/10.1371/journal.pone.0028071.

[29] Ahangaran M, Dawalatabad N, Karjadi C, Glass J, Au R, Kolachalama VB. Obfuscation via pitch-shifting for balancing privacy and diagnostic utility in voice-based cognitive assessment. Alzheimers Dement 2025;21(3):e70032. https://doi.org/10.1002/alz.70032. PubMed PMID: 40084735; PMCID: PMC12045024.

[30] McHugh CP, Clement MHS, Phatak M. AD Workbench: transforming Alzheimer's research with secure, global, and collaborative data sharing and analysis. Alzheimer's Dement 2025;21(5). https://doi.org/10.1002/alz.70278.

[31] Pedrouzo-Ulloa A., Ramon J., Péerez-González F., Lilova S., Duflot P., Chihani Z., Gentili N., Ulivi P., Hoque M.A., Mukammel T., Pritzker Z., Lemesle A., Loureiro-Acuña J., Martínez X., Jiménez-Balsa G. Introducing the TRUMPET project: tRUstworthy Multi-site privacy Enhancing Technologies. 2023 IEEE International Conference on Cyber Security and Resilience (CSR), 2023. p. 604–11.

[32] Shin H, Ryu K, Kim J-Y, Lee S. Application of privacy protection technology to healthcare big data. Digit Health 2024;10. https://doi.org/10.1177/20552076241282242.

[33] Adnan M, Kalra S, Cresswell JC, Taylor GW, Tizhoosh HR. Federated learning and differential privacy for medical image analysis. Sci Rep 2022;12(1). https://doi.org/10.1038/s41598-022-05539-7.

[34] Wu J, Dong Q, Zhang J, Su Y, Wu T, Caselli RJ, Reiman EM, Ye J, Lepore N, Chen K, Thompson PM, Wang Y. Federated morphometry feature selection for hippocampal morphometry associated beta-amyloid and tau pathology. Front Neurosci 2021;15. https://doi.org/10.3389/fnins.2021.762458.

[35] Mateus P, Moonen J, Beran M, Jaarsma E, van der Landen SM, Heuvelink J, Birhanu M, Harms AGJ, Bron E, Wolters FJ, Cats D, Mei H, Oomens J, Jansen W, Schram MT, Dekker A, Bermejo I. Data harmonization and federated learning for multi-cohort dementia research using the OMOP common data model: a Netherlands consortium of dementia cohorts case study. J Biomed Inf 2024;155. https://doi.org/10.1016/j.jbi.2024.104661.

[36] Correa R, Pahwa K, Patel B, Vachon CM, Gichoya JW, Banerjee I. Efficient adversarial debiasing with concept activation vector — Medical image case-studies. J Biomed Inf 2024:149. https://doi.org/10.1016/j.jbi.2023.104548.

[37] Xue C, Kowshik SS, Lteif D, Puducheri S, Jasodanand VH, Zhou OT, Walia AS, Guney OB, Zhang JD, Pham ST, Kaliaev A, Andreu-Arasa VC, Dwyer BC, Farris CW, Hao H, Kedar S, Mian AZ, Murman DL, O'Shea SA, Paul AB, Rohatgi S, Saint-Hilaire MH, Sartor EA, Setty BN, Small JE, Swaminathan A, Taraschenko O, Yuan J, Zhou Y, Zhu S, Karjadi C, Alvin Ang TF, Bargal SA, Plummer BA, Poston KL, Ahangaran M, Au R, Kolachalama VB. AI-based differential diagnosis of dementia etiologies on multimodal data. Nat Med 2024;30(10):2977–89. https://doi.org/10.1038/s41591-024-03118-z. Epub 20240704PubMed PMID: 38965435; PMCID: PMC11485262.

[38] Jasodanand VH, Kowshik SS, Puducheri S, Romano MF, Xu L, Au R, Kolachalama VB. AI-driven fusion of multimodal data for Alzheimer's disease biomarker assessment. Nat Commun 2025;16(1):7407. https://doi.org/10.1038/s41467-025-62590-4. Epub 20250811PubMed PMID: 40789853; PMCID: PMC12339743.

[39] Huang J, Galal G, Etemadi M, Vaidyanathan M. Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. JMIR Med Inf 2022;10 (5). https://doi.org/10.2196/36388.

[40] Akushevich I, Kravchenko J, Yashkin A, Doraiswamy PM, Hill CV. Expanding the scope of health disparities research in Alzheimer's disease and related dementias. Alzheimer's & Dementia: diagnosis. Assess Dis Monit 2023;15(1). https://doi.org/10.1002/dad2.12415.

[41] Dibaji M, Ospel J, Souza R, Bento M. Sex differences in brain MRI using deep learning toward fairer healthcare outcomes. Front Comput Neurosci 2024;18. https://doi.org/10.3389/fncom.2024.1452457.

[42] Hong J., Zhu Z., Yu S., Wang Z., Dodge H.H., Zhou J. Federated adversarial debiasing for fair and transferable representations. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining; 2021. p. 617–27.

The Journal of Prevention of Alzheimer's Disease 13 (2026) 100402

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

Mining the gaps: Deciphering Alzheimer's biology through AI-driven reconciliation

Cory C. Funk [a,b,*], Tom Paterson [b], Alex Bangs [b], David M. Cannon [f], George Savage [b], Eric Ringger [c], Lee Hood [a,b,d,e]

[a] Institute for Systems Biology, Seattle WA
[b] Fulcrum Neuroscience, Palo Alto, CA
[c] Brigham Young University, Provo, UT
[d] Phenome Health, Seattle, WA
[e] The Buck Institute, Novato, CA
[f] Provo, Utah, USA

ARTICLE INFO

Keywords:
AI
Machine learning
LLMs
Etiology
Personalization
Reconciliation

ABSTRACT

Alzheimer's disease remains one of the most complex and contested domains in biomedicine, characterized by fragmented findings, competing hypotheses, and limited translational success. We propose that AI can offer not just technical acceleration but a deeper epistemic contribution: reconciliation. Rather than optimizing predictive performance or replicating existing assumptions, the goal is to align disparate data, methods, and mechanistic insights into coherent models that explain how the disease emerges, progresses, and can be treated. This approach centers on digital twins, not as monolithic models, but as flexible, testable architectures grounded in homeostasis, destabilization, and multiscale coherence. Through an iterative, interoperable AI architecture, digital twins integrate evidence, resolve contradictions, and highlight where critical gaps remain. This framework moves beyond incremental progress within the prevailing model to catalyzing a paradigm shift in how Alzheimer's is understood. Reconciliation, in this sense, is not a method but a guiding principle for transforming both the science and its applications.

1. Introduction

Despite decades of intensive research, Alzheimer's disease (AD) remains without a cohesive, mechanistically grounded hypothesis of its etiology. The amyloid hypothesis has long shaped therapeutic development, and recent trials of lecanemab and donanemab have demonstrated modest cognitive benefits in early disease, in spite of both drugs significant reduction in amyloid plaques [1,2]. These results confirm that anti-amyloid therapies can produce incremental clinical effects, although side effects and cost limit their applicability in many patients. Notably, brain atrophy continues despite plaque clearance [3], raising the possibility that neuronal loss may precede or drive amyloid accumulation rather than follow it.

In parallel, billions of dollars in research funding have yielded rich datasets that document AD in extraordinary detail. Efforts such as Alzheimer's Disease Neuroimaging Initiative (ADNI) [4], the Dominantly Inherited Alzheimer Network (DIAN) [5,6], the Religious Orders Study and Memory and Aging Project (ROSMAP) [7], the Accelerating Medicines Partnership - Alzheimer's Disease (AMP-AD), and the ADDI Workbench [8] have collected genomic, transcriptomic, proteomic, metabolomic, imaging, and clinical data across thousands of individuals. ADNI alone has cost over $210 million and includes multi-modal, time-resolved data from thousands of participants [9]. Yet despite this scale, meaningful therapeutic breakthroughs have not followed. Like the parable of the blind men and the elephant, each dataset reveals one part of the story, but integration without reconciliation has left the whole picture incomplete.

This fragmentation extends beyond datasets to tools and standards. As molecular biologist Robert Tjian quipped, scientists would rather use each other's toothbrushes than each other's nomenclature. This simple aphorism reflects broader challenges, including conflicting analysis pipelines, incompatible data formats, and isolated computing environments. Similarly within transcriptomics, dozens of competing tools exist for RNA-seq alignment and analysis, each tailored for narrow use cases

* Corresponding author.
E-mail address: cfunk@isbscience.org (C.C. Funk).

https://doi.org/10.1016/j.tjpad.2025.100402
Received 4 September 2025; Received in revised form 14 October 2025; Accepted 20 October 2025

and often incompatible with others. Although major consortia are building interoperable platforms to support data harmonization, such efforts typically reinforce existing models rather than produce new insight.

What is needed is not just better integration, but a shift in the architecture of explanation. As Thomas Kuhn described in *The Structure of Scientific Revolutions*, science often progresses through long periods of stability punctuated by paradigm shifts that restructure the conceptual foundations of a field [10]. Clayton Christensen's theory of disruptive innovation makes a similar point in organizational settings: dominant actors tend to optimize within current frameworks, while transformative change requires the willingness to rebuild from first principles [11]. AD research shows symptoms of both stagnation and sunk-cost inertia, where existing investments make it harder to abandon familiar approaches even when they fall short. A similar dynamic is seen in evolutionary biology, where the theory of punctuated equilibrium describes long periods of stasis interrupted by bursts of genomic reorganization [12,13].

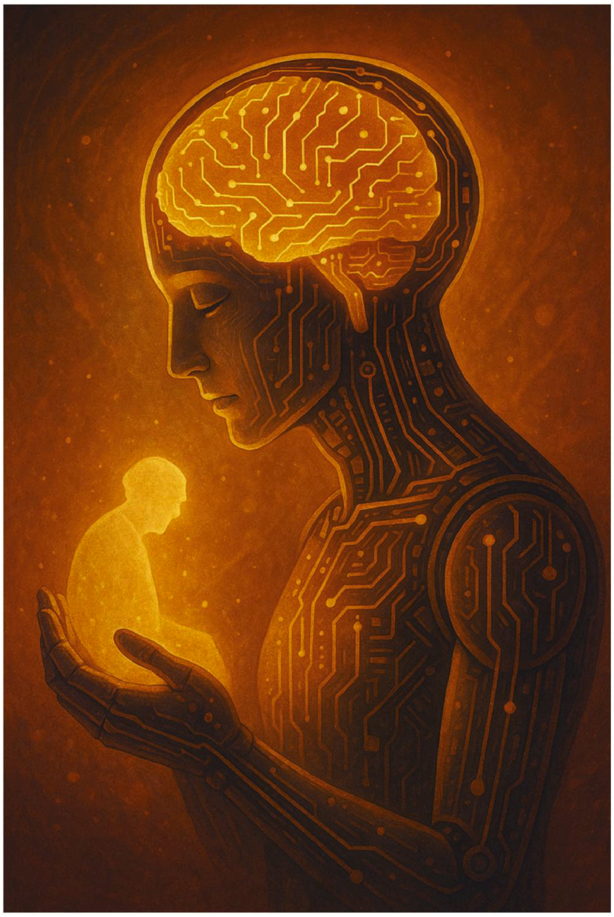
Artificial intelligence offers an opportunity to catalyze such a shift, but only if used strategically. Many applications of AI in AD focus on basic harmonizing of datasets or improving predictions, which are important but limited goals. The real opportunity lies in enabling reconciliation: aligning heterogeneous, multiscale data into causal, testable frameworks that explain rather than merely correlate. This is the promise of digital twins: mechanistic, data-driven models that can simulate biological systems, evaluate interventions, and generate new hypotheses with explanatory power.

In the sections that follow, we define reconciliation as a guiding principle for AI in AD research. We review how different AI frameworks, including language models, generative models, and digital twins, may help resolve contradictions, identify hidden variables, and support causal inference. Our aim is not to catalog all AI tools, but to show how select approaches can drive the kind of conceptual change that Alzheimer's research urgently needs.

2. Reconciliation as the central challenge

Progress in understanding Alzheimer's disease is now limited less by data availability and more by the challenge of reconciling diverse and sometimes conflicting findings. Contradictions, population heterogeneity, experimental artifacts, and static views of dynamic processes fragment our understanding. A metabolic shift in CSF, for instance, could signal pathology, compensation, or sampling error, each with different causal implications. Bridging these gaps requires more than pattern recognition; it calls for tools that integrate data within consistent causal frameworks.

We define reconciliation as the process of aligning and integrating disparate or seemingly conflicting data within a shared explanatory framework that preserves biological plausibility. This involves three essential elements: (1) integrating evidence across multiple scales and modalities, including quantitative, phenotypic, mechanistic, and systems-level; (2) making the causal logic linking these data transparent and open to scrutiny; and (3) ensuring the logic remains faithful to underlying biological reality. In Alzheimer's research, reconciliation means constructing interpretable models that can hold contradictory findings in view, explain variability, and evolve as new evidence emerges. It is not about declaring one pathway correct and discarding the rest, but about building frameworks that accommodate uncertainty while still supporting testable hypotheses, actionable interventions, and scientific trust.



ChatGPT-generated humanoid machine brain holding a fading human memory.

This challenge is made clearer by analogy to cellular automata, where simple local rules can generate highly complex global behavior. If the rules and initial conditions are known, predicting future states is straightforward. But working backward to infer the rules from observations is often computationally intractable, a problem known as computational irreducibility [14]. We see a parallel in Alzheimer's research. Even as data accumulates across scales, we still cannot explain stark phenotypic outliers. A particularly stark example involves the rare APOE Christchurch and RELN protective variants, both of which have been observed to mitigate the effects of early-onset PSEN1 mutations [15–17]. These individuals challenge prevailing causal models and offer a test for any mechanistic framework. Reconciling such cases is not optional; it is the benchmark for mechanistic understanding.

Although we have yet to see AI fully reconcile data into mechanistic explanations, biology offers precedents that demonstrate that reconciliation is possible and potentially powerful. The discovery of the Yamanaka factors, which reprogram somatic cells into induced pluripotent stem cells, reconciled 103 transcriptional profiles with functional assays to overturn assumptions about irreversible cell fate, is one notable example [18,19]. Eric Davidson's work on sea urchin development, integrating gene expression, cis-regulatory logic, and perturbation studies to infer gene regulatory networks to explain cell fate specification, is another [20]. In Drosophila embryogenesis, spatial gene expression patterns were linked to morphogen gradients like Bicoid through dynamical modeling to explain robust developmental patterning [21,22]. These efforts involved spatial, temporal, and functional data, resolved into causal models. They show that even for complex biological systems, simple rules may underlie seemingly intractable complexity. Alzheimer's, by definition, is a complex disease with a complex etiology, but this does not mean it lacks underlying structure. What's missing may be the tools to reconcile it.

One of the key requirements for reconciliation is interpretability. In many AI methods, interpretability and predictive accuracy turn out to be opposing goals. For Alzheimer's research, they must be deeply intertwined. Predictive models without explanation cannot build trust, and trust is essential for both clinical and scientific adoption. The disease's long history of failed trials suggests that predictive accuracy alone is not enough. Models must explain mechanisms to break the cycle of failed predictions and unsupported hypotheses.

Interpretability in this setting must go beyond local explanations of outputs to support epistemic transparency: the ability to reconstruct a model's internal logic, assumptions, and inference pathways so that scientists and clinicians can meaningfully engage with, evaluate, and build upon them [23,24]. This need for transparency is not just a philosophical preference, it is foundational to building trust. In clinical settings, where decisions directly affect patient well-being, the principle of do no harm demands caution [25]. Treatments like lithium or aspirin succeeded before mechanisms were fully understood, but this was only possible due to consistent empirical outcomes. AI models, in contrast, must justify their outputs to earn similar credibility AI models, in contrast, must justify their outputs to earn similar credibility. Without reconciliation of outputs to known biology, AI predictions risk leading to similar past outcomes, with limited or no benefit to patients.

This dynamic can be understood through the lens of the Hegelian dialectic. The *thesis* is interpretability: models whose structures and reasoning align with biological processes, enabling transparency, hypothesis generation, and scientific engagement. The *antithesis* is predictive accuracy: black-box models that achieve impressive results but resist explanation and may lack mechanistic grounding [26,27]. The *synthesis* we argue for is reconciliation. The most powerful are models that integrate predictive strength with epistemic transparency: they not only forecast outcomes but also explain mechanisms, integrate conflicting evidence, and generate new hypotheses. Interpretable AI systems thus become dialectical tools, helping researchers see how disparate observations cohere into a unified understanding, and enabling their reasoning processes to be interrogated, revised, and refined.

Mechanistic understanding can be advanced through both data and modeling. Human-relevant models such as organoids and organ-on-chip systems aim to capture biology that traditional animal models miss, and may help reduce reliance on animal testing [28]. However, whether animal models reflect human biology is itself contested. A prominent study once argued that mouse genomic responses fail to mimic human inflammation, though later analyses disputed this claim, highlighting the need to reconcile model systems themselves [29]. On the modeling side, digital twins and in silico trials can integrate biological constraints and simulate interventions. Such models have already reduced control-arm sizes by up to 33 % in Phase III trials, improving statistical power and reducing patient burden [30,31].

As George Box noted, all models are wrong, but some are useful. We would extend this: the most useful models are those that reconcile the most data across the most contexts, while remaining flexible enough to evolve. In Alzheimer's, where many observations remain unexplained or contradictory, reconciliation should not be an afterthought. It should be the central organizing activity of scientific inquiry. Through reconciliation, we can transform Alzheimer's research from a fragmented collection of signals into a coherent framework capable of explaining resilience, guiding intervention, and restoring scientific clarity.

3. Evaluating AI approaches for reconciliation

3.1. Language models: fluency without mechanism

Large Language Models (LLMs) have become widespread in research, offering new tools for summarization, hypothesis generation, and automation of routine tasks. These models are based on generative transformer architectures that excel at detecting and reproducing linguistic patterns across long sequences. While this makes them highly effective for contextual reasoning, they are optimized for fluency and statistical plausibility rather than factual accuracy or mechanistic understanding [27,32,33]. They do not possess an internal model of physical or biological systems and often fail when tasked with problems outside their training distribution. A concise glossary of these methods is provided in Box 1 for reference.

Recent LLMs are increasingly paired with tool-augmented systems that allow interaction with external resources such as code execution environments, retrieval modules, and search engines. These hybrid systems function as orchestrators, using natural language as the interface for integrating outputs from other tools that may offer greater factual precision or structured reasoning. One such extension is Retrieval-Augmented Generation (RAG), which improves factual grounding by linking responses to external documents. This is especially valuable in scientific domains where traceability and citation are essential. RAG systems enhance transparency by allowing users to verify claims against original sources.

Despite these enhancements, the core limitations of LLMs remain. As noted by Apple researchers in *The Illusion of Thinking* [*34*], LLMs still struggle to construct coherent causal chains, even when provided with the right data. They often falter not due to missing information, but because they lack the capacity to integrate knowledge into a structured, mechanistic understanding. This limitation is particularly problematic in biology, where reconciliation requires aligning data from heterogeneous, noisy, and sometimes contradictory sources. LLMs cannot simulate counterfactuals, weigh conflicting findings, or infer biological mechanisms. Even when they retrieve the correct papers, they frequently fail to interpret differences in experimental design, patient stratification, or underlying confounders [35]. The addition of additional context material to an LLM through RAG does not change these limitations.

LLMs and RAG systems represent one branch of a broader machine learning ecosystem. This ecosystem also includes neural networks for image recognition, causal inference frameworks, interpretable models, and structured causal representations. The strengths of LLMs lie in their ability to synthesize large volumes of literature, generate plausible hypotheses, and automate tasks involving pattern recognition and contextual reasoning. Their weaknesses are equally clear: they may sacrifice accuracy for fluency, lack grounding in biological mechanisms, and perform poorly when asked to generalize beyond their training data. In short, LLMs and RAG systems are effective tools for summarization and idea generation, but they remain inadequate for reconciliation tasks that require causal reasoning and mechanistic fidelity.

3.2. The case for interpretability

Machine learning models vary in how much insight they provide into their predictions. Many deep learning systems offer high performance but low transparency, leaving users with little understanding of how decisions are made. Interpretable machine learning (IML) methods aim to make the logic of a model accessible. These can be inherently interpretable models or post-hoc techniques such as SHAP values [36]. In Alzheimer's research, where understanding mechanism is essential, interpretability is not just a convenience but a requirement.

Box 1
Glossary of AI Methods

Large Language Models (LLMs): Deep learning models trained on massive text corpora using transformer architectures to predict and generate language. LLMs can synthesize literature, generate hypotheses, and automate routine tasks, but they optimize for fluency rather than mechanistic truth, making them prone to errors and "hallucinations." Tool-augmented variants extend LLMs with external reasoning modules (e.g., retrieval, code execution, vision), enabling broader orchestration across AI approaches.

Retrieval-Augmented Generation (RAG): A hybrid approach that grounds LLM outputs in external documents by retrieving relevant references during generation. RAG enhances transparency and factual accuracy by linking outputs back to sources. Its strength lies in evidence retrieval, but it lacks deeper reasoning capacity, and struggles when data are inconsistent, noisy, or mechanistically incomplete.

Interpretable Machine Learning (IML): A set of methods designed to make model decision processes transparent. IML techniques, such as feature attribution, rule extraction, or inherently interpretable architectures, allow researchers to evaluate whether patterns reflect underlying mechanisms. These methods trade predictive accuracy for interpretability, which is essential in domains like Alzheimer's where mechanistic clarity and trust are critical.

Deep Learning Neural Networks (DNNs): Multi-layered computational architectures that learn hierarchical representations of data through successive transformations. DNNs underpin many modern AI systems, including LLMs, generative image models, and AlphaFold, by enabling powerful pattern recognition in high-dimensional spaces. Their strength lies in predictive accuracy and scalability across diverse modalities (text, images, omics), but they often operate as "black boxes," offering limited interpretability. This opacity makes them challenging to refine mechanistically, a key limitation in scientific domains where causal understanding is essential.

Reinforcement Learning (RL): An AI paradigm in which agents learn adaptive control policies through interaction, feedback, and iteration. RL excels at discovering strategies in dynamic systems without explicit supervision. In scientific domains, it offers a way to model adaptive responses, simulate interventions, and explore trajectories of system stability or breakdown. While not itself a neuro-symbolic method, RL integrates naturally within neuro-symbolic frameworks to probe feedback dynamics and intervention policies.

Neuro-Symbolic Reasoning (Umbrella): A hybrid paradigm that combines data-driven learning with structured knowledge (graphs, rules, and priors) to ensure predictions are both powerful and mechanistically grounded. Within this umbrella, specific modeling tools can be employed:

Structured Causal Models (SCMs): Directed acyclic graphs (DAGs) that encode causal assumptions, biological priors, and constraints. They clarify directionality (e.g., Mendelian Randomization), rule out confounders, and support counterfactuals. SCMs excel at testing conditional hypotheses but cannot represent feedback loops central to biological homeostasis.

Dynamic Models: Systems of equations that explicitly model time, feedback, and compensatory processes. They capture recursive regulation, nonlinear adaptation, and resilience/failure modes. Dynamic models are indispensable for simulating disease progression and for digital twins that integrate mechanistic priors with empirical data.

Digital Twins: Dynamic, continuously updated models that serve as reconciliation engines, integrating heterogeneous data, mechanistic constraints, and biological priors into evolving frameworks. Digital twins are capable of simulating homeostatic regulation and its breakdown, enabling a broader capability for causal inference, mechanistic explanation, and testing of interventions across scales.

Surveys by Leist et al. [37]., Freiesleben et al. [38]., and Roscher et al. [39]. emphasize the importance of interpretability for scientific discovery. However, IML also faces limitations. Interpretability can come at the cost of accuracy and may not scale well with high-dimensional biomedical data. Moreover, post-hoc explanations can create a false sense of understanding if they do not reflect the model's actual internal structure. For interpretability to support reconciliation, it must be validated against biological priors and experimental data.

Deep neural networks (DNNs) are the foundation of many high-performing AI systems, including LLMs, generative models, and AlphaFold. Built from layers of nonlinear transformations, they are capable of extracting complex, high-dimensional features from raw data, enabling remarkable predictive performance across fields ranging from natural language processing to protein structure prediction[40]. However, the same layered complexity that makes DNNs powerful also makes them opaque. Their internal representations are difficult to interpret, which limits transparency, reproducibility, and mechanistic insight, especially when data distributions shift[41]. This tension between predictive accuracy and scientific interpretability remains a central challenge in applying deep learning to biology, where causal understanding and experimental validation are essential.

3.3. *Explanation as iteration*

An alternative to static explanation is the view of explanation as an iterative process. In AI planning, Chakraborti and colleagues proposed that the explanation involves aligning an AI's internal model with the user's mental model through mutual adjustment[42,43]. Rather than delivering a final answer, the system engages in a process of interaction that corrects misconceptions and refines understanding, an approach reminiscent of the Hegelian dialectic, where *thesis* and *antithesis* converge into *synthesis*.

This framing is particularly relevant in biology, where reconciling models with human understanding is often the primary challenge. In biomedicine, the difficulty lies not just in explaining results but in identifying which assumptions are valid. Unlike AI planning, which starts from a well-defined model, biomedical science typically begins with fragmented or conflicting knowledge. Still, the iterative model offers a useful framework for how reconciliation might be operationalized: as a dynamic process of alignment, adaptation, and refinement.

3.4. *Neuro-symbolic reasoning as umbrella*

While LLMs are useful for generating hypotheses, they rely solely on language-based associations. In contrast, neuro-symbolic systems combine statistical learning with structured knowledge, allowing models to represent causal and mechanistic relationships explicitly. Researchers in this area argue that combining data-driven methods with symbolic reasoning supports better generalization, interpretability, and intervention [44,45].

In this framing, reconciliation involves integrating statistical inference with mechanistic priors to support causal understanding[46]. Structured causal models and dynamic models provide scaffolds that neuro-symbolic systems can use to reason across biological systems.

Reinforcement Learning (RL) adds the capacity to simulate how systems adapt over time, which complements but does not replace the role of structure in causal modeling. As we later argue, digital twins can be seen as a concrete instantiation of this neuro-symbolic vision: an architecture that couples mechanistic cores with adaptive learning to iteratively reconcile diverse data into coherent, testable narratives.

3.5. *Reinforcement learning: policies and control*

Reinforcement learning (RL) learns by interacting with its environment, adjusting its strategy based on feedback. It does not require labeled data and can discover control policies through experience. In biomedical contexts, RL can model how systems respond to perturbations or interventions over time. This makes it valuable for simulating dynamic responses to treatment or environmental change.

However, RL has limitations. It is resource-intensive, sensitive to reward specification, and prone to instability when feedback is delayed or noisy. RL is not inherently mechanistic, but when embedded in a neuro-symbolic framework, it can test adaptive behaviors while remaining grounded in known biology. Used in this way, RL adds flexibility without sacrificing structure.

3.6. *Structured causal models: clarifying directionality*

Structured causal models (SCMs) encode assumptions about cause and effect using directed acyclic graphs (DAGs). These models support hypothesis testing, intervention analysis, and the removal of confounding effects. Popularized by Judea Pearl in *The Book of Why*, they have been influential in fields like epidemiology, economics, and in biology where they underpin approaches like Mendelian randomization [47].

SCMs are not a subset of neuro-symbolic reasoning, but a distinct causal framework that can be integrated within neuro-symbolic systems to enhance grounding. SCMs are efficient tools when the causal structure is known or can be approximated from data. However, their acyclic nature means they cannot capture feedback loops or compensatory processes, which are central to biological systems. They can clarify directionality but cannot model the full dynamics of resilience and homeostasis.

3.7. *Dynamic models: Capturing feedback and adaptation*

Dynamic models extend causal approaches by explicitly representing time, feedback, and adaptation through mathematical formalisms. These models simulate how systems evolve, how they respond to internal or external perturbations, and how they maintain or lose stability. They are especially well-suited for studying diseases like Alzheimer's, where breakdowns in homeostasis occur gradually and are shaped by complex feedback mechanisms.

Dynamic models are interpretable by design and can be integrated into neuro-symbolic frameworks to enable iterative reconciliation. They allow researchers to test hypotheses about how diverse biological variables interact across time, and they support the simulation of how disease might progress or respond to intervention. This capacity makes them foundational to digital twin systems that aim to simulate both health and disease in mechanistic terms.

4. The AI scientist

Several groups have already advanced visions of an "AI scientist" that go beyond single task tools, creating systems that autonomously generate hypotheses, design experiments, and even debate or refine mechanistic models, including the paper by Landess and Bateman et al. in this special issue [REF]. Similar ambitions appear in projects across biomedical discovery and drug development [48–50]. As Demis Hassabis, Nobel laureate and CEO of DeepMind, has noted, the hardest frontier is not generating answers but identifying the right questions. The AI Futures Project's *AI 2027* roadmap envisions a "Superintelligent AI Researcher" capable of doing so at scale [51]. While such systems remain speculative, our framework offers a pragmatic roadmap for Alzheimer's that uses today's AI tools to orchestrate existing methods, reconcile fragmented evidence, and begin by asking the right questions needed to achieve true mechanistic understanding.

4.1. *Reckoning with failure, rethinking the questions that matter*

Framing better questions, which is central to the goal of an AI scientist, requires not just new tools but a shift in how we approach scientific complexity. This shift does not reject the field's prior successes. On the contrary, it reflects humility toward what decades of brilliant work have already uncovered. The only way forward is by standing on the shoulders of that work and being honest about the places where it has not yet translated to better clinical care.

Modern biology has advanced by dissecting complexity into tractable parts, producing detailed maps of genes, pathways, and molecular circuits. This reductionist strategy has powered transformative discoveries, including the identification of APOE as a key Alzheimer's risk gene and the development of CRISPR-based editing tools. But as Lazebnik warned with his "radio repair" analogy, understanding components in isolation can obscure the organizing principles of the system itself [52]. In Alzheimer's, as in many complex diseases, the result has been an explosion of specialized findings. Most of these findings are true and many are important, but they are often disconnected from a unifying explanation of system failure.

This fragmentation has encouraged a pattern some have called statistical storytelling, which involves weaving plausible narratives from correlational data in the absence of causal models. The reproducibility crisis reflects this broader problem [53]. Studies that initially appear compelling often fail to replicate, a pattern Ioannidis famously attributed to systemic biases and statistical misuse [54]. This is not due to bad science, nor is it exclusive to Alzheimer's [55]. It is often the predictable outcome of disconnected evidence, selective inference, and the lack of frameworks that can reconcile findings into robust, mechanistic insights. AI systems risk amplifying this pattern unless they are paired with models designed to integrate and interpret complexity.

In Alzheimer's research, this challenge is especially acute. Animal models have translated poorly, with over 99.6 % of drug candidates ultimately failing in clinical trials [56], and are often treated as black boxes: useful for producing pathology but poorly predictive of human outcomes. These models can reproduce plaque and tangle pathology but are unable to predict clinical course or therapeutic response. Meanwhile, nearly 100 independent GWAS loci have been linked to Alzheimer's [57, 58], yet aside from APOE, none currently inform diagnosis, prognosis, or treatment. This is not a failure of discovery. It is a failure to assemble discoveries into a cumulative, testable understanding.

A different framing is needed. Many Alzheimer's-associated loci already converge on interpretable biology, such as microglial function and cholesterol trafficking. The challenge is not the absence of meaningful discoveries but the inability to assemble them into a cumulative, testable understanding. Rather than beginning with isolated variables or model outputs, we must start with the goal of reconciling fragmented evidence into coherent, mechanistic understanding. This shift, grounded in past successes but honest about current limitations, is essential for asking better questions. It is the foundation for the AI scientist proposed here.

4.2. *No single tool is enough*

To achieve its goal, the AI Scientist must operate as an orchestrator across a wide range of methods. Predictive modeling, causal inference, dynamic simulation, symbolic reasoning, and mechanistic modeling all offer partial insights. Their true value emerges when used together to interrogate the same system from multiple perspectives. This orchestration is as much about judgment as computation: knowing which tool applies to which question, recognizing contradictions as informative, and adjusting models as knowledge evolves. We anticipate that many individual efforts will continue to use AI to harmonize datasets, identify features, and optimize predictions. These are important contributions, but without a unifying framework, they risk reinforcing the very fragmentation that is antithetical to reconciliation. The purpose of this paper is to outline how those incremental efforts can be directed by a higher-level orchestration, where the AI Scientist integrates them as components of a broader reconciliation strategy. As detailed in the sections that follow, we propose specific ways that AI can be used to define and constrain the solution space by treating reconciliation across biological scales as the central challenge. In Alzheimer's, this means integrating across scales (genes, cells, circuits, and populations) while staying focused on the deeper goal: not simply predicting decline, but uncovering how the brain's homeostatic balance destabilizes into disease, with the ultimate aim of enabling treatments, and ultimately cures, that restore stability.

4.3. Kind vs. wicked learning environments

A central challenge for the AI Scientist is recognizing the nature of the problem space. As David Epstein describes in *Range* [*59*], some domains are kind learning environments, where rules are explicit, feedback is consistent, and outcomes clearly reflect causes. Others are wicked, shaped by delayed feedback, hidden variables, and ambiguity. This distinction is crucial because it defines how AI systems can learn, iterate, and reconcile conflicting information.

Games like chess or Go are quintessentially kind: the rules are fixed, the objectives are unambiguous, and feedback is immediate. In fact, the real breakthrough for AlphaGo came not when it imitated human play, but when it moved beyond human examples and began generating novel strategies by exploring the game space under these transparent rules [60]. But biology, and Alzheimer's in particular, rarely offers such clarity. It is a wicked environment where data are sparse, signals are noisy, and feedback is often delayed. In this context, the goal is not to invent new capabilities, but to reverse-engineer mechanisms that nature has already solved and to align models with those underlying biological truths.

AlphaFold, an AI system developed by DeepMind, transformed structural biology by accurately predicting the 3D structures of proteins from their amino acid sequences, solving a problem that had eluded scientists for more than half a century [61]. Protein folding involves both kind and wicked elements. The kind aspects include well-defined physical constraints that make much of the problem tractable. The wicked aspects include intrinsically disordered regions, which make up approximately 40 percent of the human proteome and never resolve into a single stable structure [62]. AlphaFold succeeded in this mixed regime by embedding biophysical and evolutionary priors and applying iterative refinement to recycle its predictions until structure, constraint, and data aligned within the structured regions. Its success illustrates how AI can navigate partially understood systems by using known constraints while recognizing and respecting areas of unresolved complexity. As John Moult described, AlphaFold solved two problems simultaneously: finding the right solution and knowing when you're there [63].

This is the essence of reconciliation. In wicked or mixed domains, it is not enough to generate outputs that merely appear plausible. The AI Scientist must identify where the rules are well-defined, where uncertainty remains, and how advances in tractable areas can help constrain ambiguity elsewhere. Many current applications of agentic AI thrive in open-ended environments where the goal is to invent new capabilities, unconstrained by a single correct answer. But understanding biology is a fundamentally different challenge: it requires reverse-engineering mechanisms that nature has already solved, where success depends on aligning with those underlying truths. In this context, understanding the limits of current knowledge is itself a valuable scientific contribution, helping to guide discovery toward the questions that matter most.

4.4. Neuro-Symbolic reasoning as framework

To act as a scientist, AI must do more than fit patterns or generate predictions. It must also reason about mechanisms, test hypotheses, and update its beliefs in light of new evidence. This requires a framework that integrates both perception and inference. Neuro-symbolic reasoning provides such a framework by combining data-driven learners (such as deep nets) with structured knowledge (such as graphs, rules, and constraints), allowing inferences that are both powerful and checkable [44]. For Alzheimer's, this is especially important because the problem spans both kind and wicked learning environments. We need models that can learn from noisy, incomplete data while also asserting and testing mechanistic claims. In this framework, learned components handle perception and imputation, while the symbolic layer encodes biological priors, defines allowable transitions, and supports counterfactual reasoning. Structured Causal Models (SCMs) and dynamic systems naturally fit here. SCMs supply testable causal scaffolds, and dynamic models represent trajectories and feedback. Digital twins can instantiate this hybrid approach by embedding mechanistic cores (e.g., compartmental/ODE models, mass/energy/flux constraints) alongside learned modules and then updating as new data arrive. Used this way, neuro-symbolic reasoning transforms disparate data into transparent, testable narratives that not only predict outcomes but explain why, under what assumptions, and how an intervention might shift the course of disease.

SCMs help the AI scientist disentangle directionality by encoding assumptions as directed graphs, allowing for hypothesis formalization, confounder control, and causal inference. This supports systematic exploration of explanatory models, ruling some out while refining others. Mendelian Randomization exemplifies this, using genetic variants as natural experiments to probe causality [64]. But SCMs assume acyclicity and are limited in representing the feedback loops central to biological homeostasis. Used alone, they risk reducing complex dynamics to one-way arrows. Their strength lies in hypothesis narrowing and causal constraint, especially when paired with dynamic models that represent recursive regulation. For the AI scientist, SCMs are precision tools: valuable for pruning the explanatory space, but incomplete for modeling full systems.

Flux Balance Analysis (FBA) uses constraint-based optimization to infer metabolic fluxes under steady-state assumptions [65]. In Alzheimer's research, FBA helps test hypotheses about astrocyte–neuron metabolic coupling. Models show how astrocyte-produced lactate supports neuronal energy demands under aerobic glycolysis [66,67]. FBA enforces physical plausibility and checks biochemical consistency, offering more than statistical correlation. For the AI scientist, it's a principled method to assess whether observed metabolite patterns align with shuttle mechanisms like the ANLS. However, because FBA assumes steady state, it excels when conditions are stable but needs complementing with dynamic models in settings involving perturbations or time-dependent change.

Quantitative Systems Pharmacology (QSP) models use differential equations to simulate Alzheimer's-related pathways across compartments such as brain, CSF, and plasma. They encode production, aggregation, clearance, and drug responses (e.g., monoclonal antibodies) [68]. These models support hypothesis testing, dose optimization, and biomarker trajectory forecasting. For the AI scientist, QSP translates biological knowledge into simulation-ready form. However, QSP typically focuses on narrow pathways and lacks integration with broader homeostatic systems. As such, it is a powerful tool for scoped inquiries, but not a substitute for more comprehensive mechanistic models.

Reinforcement Learning (RL) enables machines to learn control policies through feedback—trial, correction, and policy refinement [69–71]. Layered RL architectures combine fast reflexes with slower strategic control [72], mirroring biological regulation across timescales from ionic shifts to transcriptional changes. For the AI scientist, RL offers a model for digital twins that adapt over time, not just predict. Alzheimer's pathology emerges from regulatory failure, making RL's adaptive framing essential.

While RL could be used to optimize treatment strategies (e.g., dosing), this risks shallow gains unless guided by deeper constraints. We envision RL agents operating within digital twins that embed homeostatic principles across scales. Here, the reward function prioritizes long-term system stability, penalizing destabilizing trajectories, such as impaired ANLS or microglial lipid overload. In this role, RL becomes not just an optimizer but a discovery engine, probing which control policies sustain resilience. It shifts from reactive adjustment to active inference of system-level scaffolds that underlie progression and recovery.

Box 2 outlines core use cases for AI in Alzheimer's research, illustrating how different methods—hypothesis generation, causal inference, dynamic modeling—offer complementary strengths. Rather than exhaustively listing tools, we suggest how combining these approaches, often within neuro-symbolic frameworks or digital twins, can shift the field from fragmented associations toward mechanistic, testable understanding.

4.5. *The AI scientist as conductor, digital twins as the orchestra*

Each AI approach illuminates only part of the Alzheimer's puzzle. Causal graphs clarify direction but miss feedback. Dynamic models capture adaptation but require constraint. Reinforcement learning discovers control policies but depends on environments that biology rarely provides. Language models synthesize evidence but often without mechanism. No single method is sufficient. The AI Scientist's role is conceptual: an orchestrator who selects, sequences, and integrates diverse methods to produce coherent, mechanistic hypotheses. This orchestration is similar to how current agentic AI systems coordinate multiple tools to accomplish goals. However, unlike today's agents, which operate without a world model, the AI Scientist must root its reasoning in causal and dynamic realities.

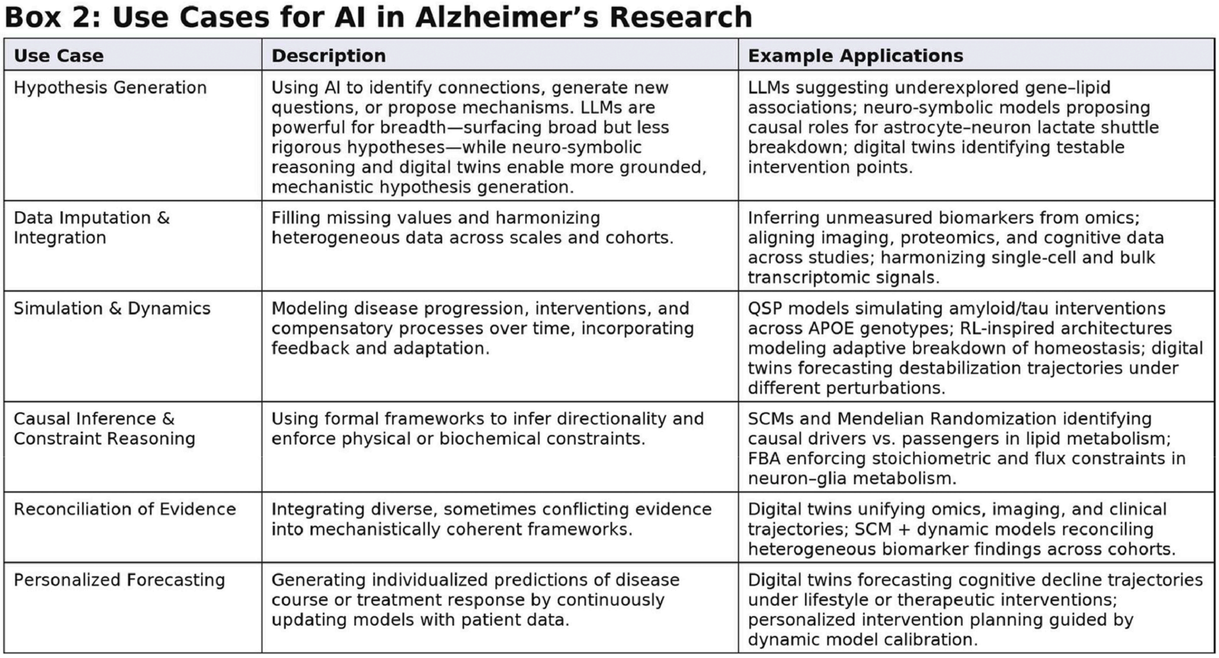
Digital twins provide the formal setting for this orchestration. They are not metaphorical scientists but structured environments where causal graphs, dynamic models, reinforcement learning policies, and statistical learners interact within a living representation of disease. Built on dynamic modeling, twins capture feedback, adaptation, and homeostasis, linking molecular to population scales. In this way, they bridge collective and individual variation, showing how mechanisms of risk or resilience emerge while keeping insight tethered to shared system dynamics. The following section details how such twins, grounded in systems engineering and multiscale integration, can supply the scaffolding needed to close persistent gaps in Alzheimer's research.

5. From orchestration to implementation: digital twins as the framework for reconciliation

5.1. *What we mean by a digital twin*

The idea of a "digital twin" traces its origins to aerospace and manufacturing: NASA's early use of simulators to mirror spacecraft behavior, especially during Apollo missions, laid the foundation for today's virtual replicas of complex systems. By the 2010s, digital twins had evolved into real-time, physics-based models used for predictive, "personalized" maintenance in jet engines and aircraft, continuously assimilating sensor data to forecast failures and optimize performance [73]. The American Institute of Aeronautics and Astronautics 2020 position paper reinforces this lineage, recognizing digital twins as integral decision-support tools in safety-critical environments [74]. In these settings, the analogy becomes clear: while every jet engine begins as a standardized design, each experiences different conditions (stress cycles,

Box 2
Use Cases for AI in Alzheimer's Research



Box 2: Use Cases for AI in Alzheimer's Research

Use Case	Description	Example Applications
Hypothesis Generation	Using AI to identify connections, generate new questions, or propose mechanisms. LLMs are powerful for breadth—surfacing broad but less rigorous hypotheses—while neuro-symbolic reasoning and digital twins enable more grounded, mechanistic hypothesis generation.	LLMs suggesting underexplored gene–lipid associations; neuro-symbolic models proposing causal roles for astrocyte–neuron lactate shuttle breakdown; digital twins identifying testable intervention points.
Data Imputation & Integration	Filling missing values and harmonizing heterogeneous data across scales and cohorts.	Inferring unmeasured biomarkers from omics; aligning imaging, proteomics, and cognitive data across studies; harmonizing single-cell and bulk transcriptomic signals.
Simulation & Dynamics	Modeling disease progression, interventions, and compensatory processes over time, incorporating feedback and adaptation.	QSP models simulating amyloid/tau interventions across APOE genotypes; RL-inspired architectures modeling adaptive breakdown of homeostasis; digital twins forecasting destabilization trajectories under different perturbations.
Causal Inference & Constraint Reasoning	Using formal frameworks to infer directionality and enforce physical or biochemical constraints.	SCMs and Mendelian Randomization identifying causal drivers vs. passengers in lipid metabolism; FBA enforcing stoichiometric and flux constraints in neuron–glia metabolism.
Reconciliation of Evidence	Integrating diverse, sometimes conflicting evidence into mechanistically coherent frameworks.	Digital twins unifying omics, imaging, and clinical trajectories; SCM + dynamic models reconciling heterogeneous biomarker findings across cohorts.
Personalized Forecasting	Generating individualized predictions of disease course or treatment response by continuously updating models with patient data.	Digital twins forecasting cognitive decline trajectories under lifestyle or therapeutic interventions; personalized intervention planning guided by dynamic model calibration.

maintenance regimes, environmental exposures) that gradually introduce variation. A fleet of engines thus becomes a distribution of outcomes, where twins both capture the shared physics of the system and track how individualized histories shape divergence over time.

Recent years have seen a surge of interest in digital twin technologies across biomedical domains, with a few broad classes beginning to emerge [75]. The first includes QSP-based digital twins, which model pharmacokinetics and pharmacodynamics at the individual level to simulate treatment effects and optimize dosing regimens. For example, Maharjan et al. describe how digital twins can transform pharmaceutical pipelines from discovery through post-market surveillance [76], while Susilo et al. demonstrate their utility in characterizing clinical dose-response relationships in rare diseases [77]. The second class focuses on neurology-specific digital twins, which aim to model dynamic disease trajectories for individuals. Fekonja et al. propose digital twins for personalized brain modeling [78], and Cen et al. show how such twins can track disease-specific atrophy in multiple sclerosis [79]. A third, increasingly visible application is the use of digital twins to optimize clinical trial design, where simulated populations are used to reduce the size of control arms, as noted in recent systems pharmacology literature [80,81].

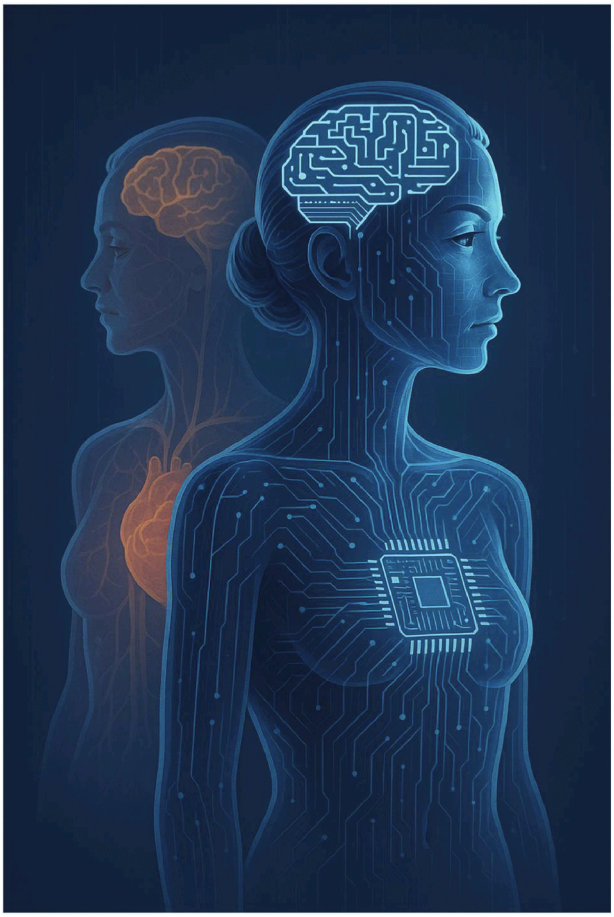
By contrast, the digital twin framework we propose differs in several critical respects. Rather than focusing on drug-specific responses or individualized prognostication, a reconciliation-centered model centers on homeostatic regulation and causal inference. Its primary goal is to reconstruct multiomic, imaging, and clinical data into mechanistic models of disease. We envision twins designed to not only simulate outcomes but also to test hypotheses about biological function. While QSP and neurology twins typically prioritize predictive accuracy or trial simulation, our approach aims to capture and resolve internal inconsistencies across diverse data modalities. This is key to uncovering the underlying rules that govern system-level dysfunction. Such a broad scope is especially critical in Alzheimer's disease, where diverse and potentially conflicting findings across data types, such as proteomics and imaging, may have limited individual contribution potential towards true causal understanding.

In the context of Alzheimer's, we adopt that best-practice foundation, but repurpose it for scientific reconciliation, not just operational forecasting. Our definition is more stringent: we envision digital twins as dynamic, mechanistic models that evolve with longitudinal data, enforce conservation and homeostatic constraints, and are built for causal inference rather than prediction alone. In doing so, we retain the proven architecture of adaptation and feedback from aerospace but deploy it as a reconciliation engine, integrating heterogeneous data and mechanistic priors into coherent, evolving models suited to unraveling the complexities of Alzheimer's disease.

5.2. Organizing principles

Homeostasis: Living systems evolved for stability, not disease. Homeostasis provides both the goal state and the constraints that digital twins must encode. This includes conservation laws (mass, energy, flux) and feedback loops governing lipid transport, neuronal excitability, and immune signaling. Without grounding in these regulatory architectures, models may generate statistically plausible but biologically invalid results.

Destabilization: Disease reflects progressive erosion of regulatory balance. In Alzheimer's, destabilization may stem from impaired lipid clearance, disrupted astrocyte–neuron energy coupling, or unchecked inflammation. Digital twins must represent not only intact feedback systems, but also how they degrade over time and stress. Capturing this dynamic misalignment enables models to explain how vulnerability accumulates and leads to pathology.



ChatGPT-generated representation of a digital twin.

Multiscale Reconciliation: Alzheimer's spans molecules, cells, circuits, and populations. Twins must integrate across these layers: molecular priors (e.g., APOE and lipid metabolism), cellular behaviors (e.g., microglial flux), systems physiology (e.g., glymphatic clearance), and population trajectories. This is more than model nesting—it requires coherence across scales, where cellular dynamics are constrained by cohort-level biomarkers and vice versa. Without this, the landscape fragments into disciplinary silos.

Temporal Alignment: Biological processes unfold on vastly different time scales: milliseconds (ion currents), hours (metabolism), days (immune shifts), and years (atrophy, plaques). Alzheimer's arises not from a single failure but from mismatches across these rhythms—-when fast neuronal needs outpace slower astrocyte support, or when debris accumulates over decades. Twins must simulate fast, medium, and slow processes together, showing how asynchrony drives system instability. Time becomes as central as scale.

5.3. Architectural roles

With foundational principles defined, we now explore how AI can operationalize them. In this framework, the AI Scientist coordinates a layered system of reasoning roles. The digital twin provides the structure within which this coordination unfolds. It is not a single model, but an environment composed of three core functions: the orchestrator, the enforcer, and the architect.

The orchestrator manages knowledge flow. It selects which tools to use, interprets their outputs, and ensures consistency across the system. This role can be performed by language models enhanced with retrieval tools and code execution, allowing them to synthesize literature, modeling results, and statistical outputs into coherent narratives. The orchestrator also tracks uncertainty, noting which results are supported by data, which rely on assumptions, and which remain unresolved. It routes this information between the enforcer and architect to support adaptive model refinement. While it does not generate mechanistic insight on its own, the orchestrator is essential for aligning evidence with evolving hypotheses.

The enforcer is responsible for simulation, optimization, and learning under constraint. This includes reinforcement learning, flux balance models, and dynamic systems. Its job is to evaluate how well candidate explanations hold up against biological constraints and available data. In wicked domains like biology, where signals are intermittent, indirect, or confounded by observational limits, the enforcer plays the critical role of pushing models until they either hold or break under the weight of evidence. In early stages, this is a human-guided process. Over time, the enforcer may gain autonomy, identifying new data sources or proposing targeted experiments to resolve conflict. It plays a central role in iterative discovery, using contradiction as a signal to refine or revise current understanding.

The architect designs and compares alternative model structures. Since no model can capture biology in full, the goal is to create multiple abstractions and evaluate how well each explains the data. This process draws on structured causal models, symbolic reasoning, and probabilistic tools. The architect also tracks assumptions and priors, helping clarify what each model implies. When gaps are exposed, the architect assists in generating targeted hypotheses that can be tested experimentally. While AI may suggest plausible experiments, matching those proposals to biologically meaningful systems remains a task that often requires domain expertise. A well-designed twin helps prioritize experiments by their likely impact and feasibility.

This system is designed to evolve. This progression aligns with systems engineering principles, where contradictions in the model prompt the acquisition of new, discriminative data. The goal is not to explain everything at once but to identify the most strategic gaps and design targeted experiments that reduce uncertainty. Reconciliation is a central function, allowing twins to integrate new data while also addressing the backlog of conflicting results. Taken together, the orchestrator ensures clarity, the enforcer grounds models in reality, and the architect explores structural alternatives. Their interaction forms a dynamic reasoning engine that transforms digital twins into living frameworks for discovery.

When reconciliation exposes a mechanistic gap, the next step is translating that gap into a tractable experiment. Today, this process is human-guided: researchers identify where a model lacks constraint and propose perturbations or observations to test it. LLMs can assist by suggesting plausible in vitro or in vivo experiments given sufficient context, and their utility in this domain is likely to grow. Still, designing testable hypotheses depends not just on mechanistic reasoning but on choosing a biologically relevant model system, which remains highly context specific. In fields like neurobiology, where cell type, spatial location, and timing influence results, domain expertise remains essential. A key function of digital twins is to help prioritize among candidate experiments by estimating their informativeness, feasibility, and relevance to the broader system.

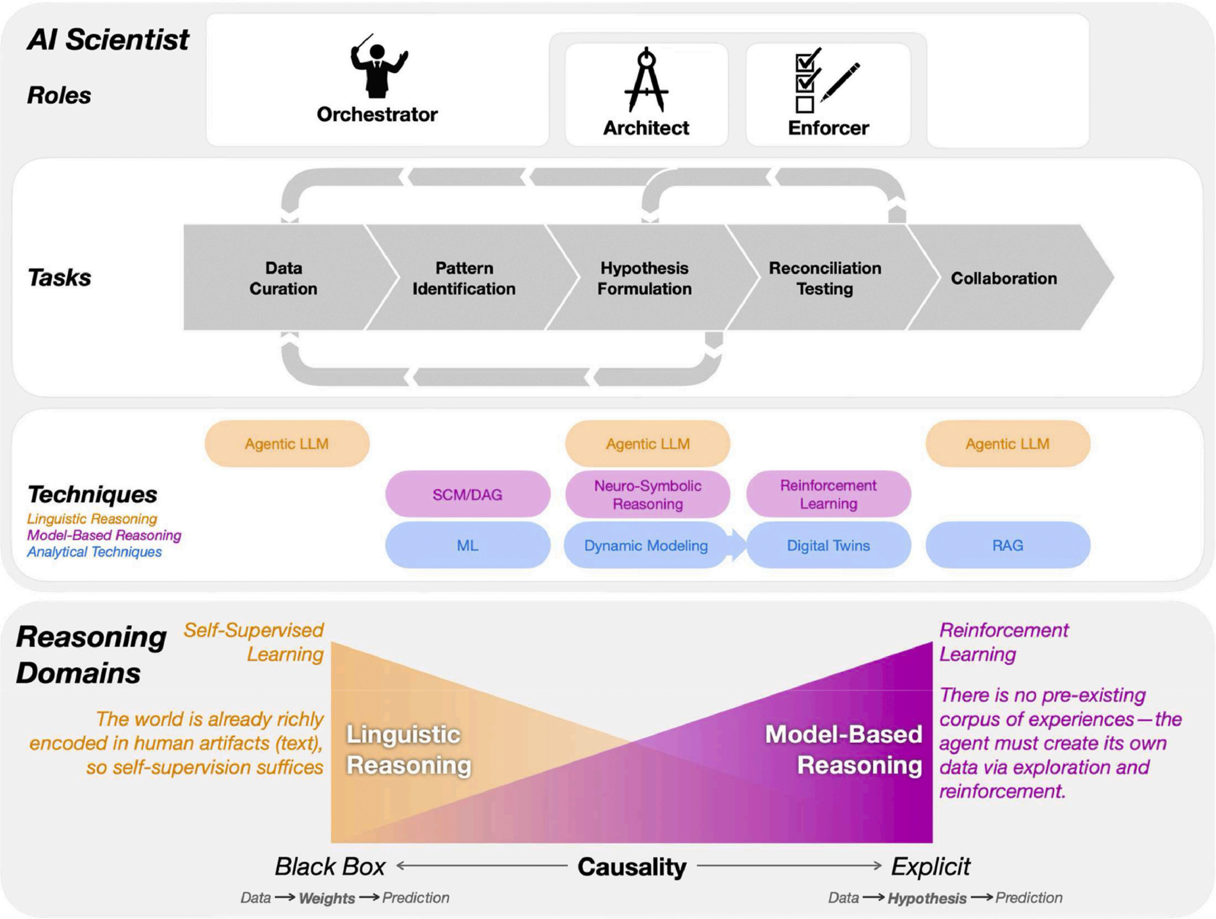


Fig. 1. Mapping AI Approaches for Reconciliation in Alzheimer's Research. Illustration of how diverse AI methods align with stages of the scientific process—from data curation and pattern identification to hypothesis formulation, reconciliation testing, and collaboration. Large language models, causal graphs, dynamic models, reinforcement learning, and neuro-symbolic reasoning each occupy different roles, but their greatest value emerges when coordinated iteratively. The framework highlights where linguistic reasoning suffices, where model-based reasoning and feedback are essential, and how iteration across methods enables reconciliation of partial evidence into mechanistic insight.

Just as experiments refine the model, models must be built to absorb those refinements. In our framework, this is not a technical hurdle but a foundational feature. Reconciliation-driven twins are structured to remain flexible at their edges, allowing new data to update causal links without full retraining. This refinement can be manual or automated depending on how constrained the relevant variables are. Yet the bigger challenge is not integrating future data, but absorbing the vast backlog of existing results. A reconciliation-first architecture helps triage which gaps warrant new experiments and avoids redundant or low-value inquiry.

These methods reinforce that the AI scientist's task is not simply tool selection but orchestration of iterative cycles of hypothesis, testing, and refinement. Fig. 1 outlines this process, showing how different AI approaches interact and where iteration drives progress toward mechanistic understanding.

The purpose of this two-level approach, which combines an AI scientist with a biological digital twin, is not simply to explain Alzheimer's in theory. It is designed to support better decisions about treatment, trials, and care. Mechanistic clarity has value only if it improves the ability to act under uncertainty. Decision theory provides the formal bridge between understanding and action by offering a framework to evaluate options through probabilities, outcomes, and utilities. In Alzheimer's, this means structuring an iterative relationship between scientific insight and clinical intervention. Mechanistic models inform therapeutic strategies, which then generate new data. That data refines the models, sharpening their predictions and explanations. This creates a self-reinforcing cycle connecting discovery, development, and practice.

In drug development, this approach enables rational trial design, more precise recruitment, and model systems that better reflect underlying biology. Instead of broad statistical averages, decision-aware twins can simulate how interventions affect specific genotypes or stages of disease, helping avoid costly failures and improving targeting. For clinicians, the same principles provide decision support grounded in mechanism, not just association, allowing for better anticipation of treatment outcomes. In this way, decision theory links the scientific goal of understanding disease with the practical goal of guiding action, making reconciliation both a scientific tool and a translational engine for progress.

6. Conclusion

6.1. Reconciliation as the central task

Artificial intelligence has the potential to transform Alzheimer's research, not through any single breakthrough, but by integrating diverse methods into a coherent system. Language models, retrieval tools, reinforcement learning, structured causal models, flux balance analysis, and quantitative pharmacology each illuminate a different aspect of disease. Yet individually, they remain partial and insufficient. What is needed is reconciliation: a way to align these tools into frameworks that are explanatory, testable, and faithful to biology.

Focusing on reconciliation introduces real challenges. Digital twins are computationally demanding and rely on rich, longitudinal data. Mechanistic models, while interpretable, can still embed flawed assumptions or biases. As AI-generated hypotheses enter clinical research, ethical and regulatory concerns increase, particularly around validation, transparency, and accountability. These challenges highlight the need for modular, interpretable systems that allow each component to be tested and trusted independently.

Digital twins provide the biological foundation for this reconciliation. Rather than static forecasts, they represent evolving, mechanistic models that span levels of scale and time: from cells to circuits, from patients to populations, from rapid feedback to long-term adaptation [82]. Guided by principles of homeostasis, destabilization, and multiscale coherence, digital twins operate as living systems—continually adjusting predictions, reconciling contradictions, and exposing knowledge gaps. Their effectiveness depends on the architecture built by the AI Scientist: orchestrators that manage reasoning, enforcers that stress-test hypotheses, and architects that explore and refine model structures. The AI Scientist supplies the reasoning infrastructure; the digital twins embody its evolving output.

The lesson of AlphaFold is that iteration, paired with constraint, can resolve seemingly intractable problems. The lesson of Alzheimer's is that no single method will succeed when biology is irregular in its signals, complex in its interactions, and shifting over time. By treating reconciliation as the central goal, and digital twins as the scaffolding that supports it, we can move from fragmented knowledge toward integrated understanding. If successful, this approach will not only close longstanding gaps in Alzheimer's but also offer a new model for scientific discovery—one in which AI acts not as a black box, but as a transparent, evolving system for mechanistic insight.

During the preparation of this work the authors used ChatGPT and Claude in order to research topics, gather information, and improve linguistic clarity and brevity. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

(Note on references: In the rapidly-developing field of AI, some papers published as preprints are highly influential and well-cited, and reveal important new insight into methods and capabilities. But publication in a reviewed journal may be done much later or never sought at all. We include some of these preprints in the following references.)

Declaration of competing interest

I, **Cory Funk**, am a cofounder of **Fulcrum Neuroscience**, a biotechnology company developing computational and mechanistic approaches for understanding and treating Alzheimer's disease. In this capacity, I hold an equity interest in the company, receive financial compensation, and am associated with intellectual property (patents, licenses, or royalties) related to its work.

These relationships represent potential financial conflicts of interest relevant to the subject matter of my research. I disclose them here in full recognition of the importance of transparency and to allow editors and reviewers to evaluate the manuscript with this context in mind.

References

[1] Dyck CH van, Swanson CJ, Aisen P, et al. Lecanemab in Early Alzheimer's Disease. New Engl J Med 2022;388:9–21.

[2] Mintun MA, Lo AC, Evans CD, et al. Donanemab in Early Alzheimer's Disease. New Engl J Med 2021;384:1691–704.

[3] Alves F, Kalinowski P, Ayton S. Accelerated brain volume loss caused by anti–β-amyloid drugs. Neurology 2023;100:e2114–24.

[4] Mueller SG, Weiner MW, Thal LJ, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). Alzheimer's Dement 2005;1:55–66.

[5] Morris JC, Aisen PS, Bateman RJ, et al. Developing an international network for Alzheimer's research: the Dominantly Inherited Alzheimer Network. Clin Investig 2012;2:975–84.

[6] Bateman RJ, Xiong C, Benzinger TLS, et al. Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease. N Engl J Med 2012;367:795–804.

[7] Bennett DA, Buchman AS, Boyle PA, et al. Religious orders study and rush memory and aging project. J Alzheimer's Dis 2018;64:S161–89.

[8] Imam F, Saloner R, Vogel JW, et al. The Global Neurodegeneration Proteomics Consortium: biomarker and drug target discovery for common neurodegenerative diseases and aging. Nat Med 2025;31:2556–66.

[9] Health F for the NI of. Alzheimer's Disease Neuroimaging Initiative (ADNI). fnih.org/our-programs/alzheimers-disease-neuroimaging-initiative-adni.

[10] Kuhn TS, Hawkins D. The Structure of Scientific Revolutions. Am J Phys 1963;31:554–5.

[11] Christensen CM, Raynor ME, McDonald R. What Is Disruptive Innovation? Harv Bus Rev 2015. https://hbr.org/2015/12/what-is-disruptive-innovation?.

[12] Koonin EV. Evolution of genome architecture. Int J Biochem cell Biol 2008;41:298–306.

[13] Heasley LR, Sampaio NMV, Argueso JL. Systemic and rapid restructuring of the genome: a new perspective on punctuated equilibrium. Curr Genet 2021;67:57–63.

[14] Wolfram S. *A New Kind of Science*. Champaign, IL, USA: Wolfram Media; 2002. https://www.wolframscience.com/nks/.

[15] Lopera F, Marino C, Chandrahas AS, et al. Resilience to autosomal dominant Alzheimer's disease in a Reelin-COLBOS heterozygous man. Nat Med 2023:1–10.

[16] Llibre-Guerra JJ, Fernandez MV, Joseph-Mathurin N, et al. Longitudinal analysis of a dominantly inherited Alzheimer disease mutation carrier protected from dementia. Nat Med 2025;31:1267–75.

[17] Arboleda-Velasquez JF, Lopera F, O'Hare M, et al. Resistance to autosomal dominant Alzheimer's disease in an APOE3 Christchurch homozygote: a case report. Nat Med 2019;25:1680–3.

[18] Buganim Y, Faddah DA, Jaenisch R. Mechanisms and models of somatic cell reprogramming. Nat Rev Genet 2013;14:427–39.

[19] Zhu F, Nie G. Cell reprogramming: methods, mechanisms and applications. Cell Regen 2025;14:12.

[20] Martik ML, Lyons DC, McClay DR. Developmental gene regulatory networks in sea urchins and what we can learn from them. F1000Res 2016;5:F1000. Faculty Rev-203.

[21] Schüpbach T. Genetic Screens to Analyze Pattern Formation of Egg and Embryo in Drosophila: a Personal History. Annu Rev Genet 2019;53:1–18.

[22] Lynch JA, Roth S. The evolution of dorsal–ventral patterning mechanisms in insects. Genes Dev 2011;25:107–18.

[23] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. ArXiv 2017. https://doi.org/10.48550/arxiv.1702.08608. Epub ahead of print.

[24] Arrieta AB, Díaz-Rodríguez N, Ser JD, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115.

[25] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15.

[26] Johnston WJ, Fusi S. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. Nat Commun 2023;14:1040.

[27] Vafa K., Chen J.Y., Rambachan A., et al. Evaluating the World Model Implicit in a Generative Model. In: *Advances in neural information processing systems*. Curran Associates, Inc., pp. 26941–26975, 2024.

[28] Pound P, Bracken MB. Is animal research sufficiently evidence based to be a cornerstone of biomedical research? BMJ: Br Méd J 2014;348:g3387.

[29] Lancaster MA, Knoblich JA. Generation of cerebral organoids from human pluripotent stem cells. Nat Protoc 2014;9:2329–40.

[30] Wang Z, Gao C, Glass LM, et al. Artificial intelligence for in silico clinical trials: a review. ArXiv 2022. https://doi.org/10.48550/arxiv.2209.09023. Epub ahead of print.

[31] Sinisi S, Alimguzhin V, Mancini T, et al. Optimal personalised treatment computation through in silico clinical trials on patient digital twins. Fundam Informaticae 2020;174:283–310.

[32] McCoy RT, Yao S, Friedman D, et al. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. Proc Natl Acad Sci 2024;121:e2322420121.

[33] Hao S, Gu Y, Ma H, et al. Reasoning with language model is planning with world model. ArXiv 2023. https://doi.org/10.48550/arxiv.2305.14992. Epub ahead of print.

[34] Shojaee P, Mirzadeh I, Alizadeh K, et al. The illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity. Apple Machine Learn Res 2025. https://machinelearning.apple.com/research/illusion-of-thinking.

[35] Zhu Y, Yuan H, Wang S, et al. Large language models for information retrieval: a survey. ArXiv 2023. https://doi.org/10.48550/arxiv.2308.07107. Epub ahead of print.

[36] Ponce-Bobadilla AV, Schmitt V, Maier CS, et al. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin Transl Sci 2024;17:e70056.

[37] Leist AK, Klee M, Kim JH, et al. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. Sci Adv 2022;8:eabk1942.

[38] Freiesleben T, König G, Molnar C, et al. Scientific inference with interpretable machine learning: analyzing models to learn about real-world phenomena. Minds Mach 2024;34:32.

[39] Roscher R, Bohn B, Duarte MF, et al. Explainable machine learning for scientific insights and discoveries. IEEE Access 2020;8:42200–16.

[40] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

[41] Lipton ZC. The mythos of model interpretability. Commun ACM 2018;61:36–43.

[42] Chakraborti T, Sreedharan S, Zhang Y, et al. Plan explanations as model reconciliation: moving beyond explanation as soliloquy. Proc Twenty-Sixth Int Jt Conf Artif Intell 2017:156–63.

[43] Sreedharan S, Chakraborti T, Kambhampati S. Foundations of explanations as model reconciliation. Artif Intell 2021;301:103558.

[44] Garnelo M, Shanahan M. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. Curr Opin Behav Sci 2019;29:17–23.

[45] Lake BM, Ullman TD, Tenenbaum JB, et al. Building machines that learn and think like people. Behav Brain Sci 2017;40:e253.

[46] Tenenbaum JB, Kemp C, Griffiths TL, et al. How to grow a mind: statistics, structure, and abstraction. Science (1979) 2011;331:1279–85.

[47] Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books; 2018.

[48] Lu C, Lu C, Lange RT, et al. The AI scientist: towards fully automated open-ended scientific discovery. ArXiv 2024. https://doi.org/10.48550/arxiv.2408.06292. Epub ahead of print.

[49] Yamada Y, Lange RT, Lu C, et al. The AI scientist-v2: workshop-level automated scientific discovery via agentic tree search. ArXiv 2025. https://doi.org/10.48550/arxiv.2504.08066. Epub ahead of print.

[50] Gottweis J, Weng W-H, Daryin A, et al. Towards an AI co-scientist. ArXiv 2025. https://doi.org/10.48550/arxiv.2502.18864. Epub ahead of print.

[51] Kokotajlo D, Alexander S, Larsen T, et al. AI 2027. AI futures project. 2025. https://ai-2027.com/.

[52] Lazebnik Y. Can a biologist fix a radio?—Or, what I learned while studying apoptosis. Cancer Cell 2002;2:179–82.

[53] Resnik DB, Shamoo AE. Reproducibility and research integrity. Account Res 2017;24:116–23.

[54] Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2:e124.

[55] Nosek BA, Errington TM. What is replication? PLoS Biol 2020;18:e3000691.

[56] Drummond E, Wisniewski T. Alzheimer's disease: experimental models and reality. Acta Neuropathol 2017;133:155–75.

[57] Bellenguez C, Küçükali F, Jansen IE, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet 2022;54:412–36.

[58] Wightman DP, Jansen IE, Savage JE, et al. Largest GWAS (N=1126,563) of Alzheimer's disease implicates microglia and immune cells. medRxiv 2020. 2020.11.20.20235275.

[59] Epstein D. Range*: why generalists triumph in a specialized world*. New York: Riverhead Books, 2019.

[60] Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–9.

[61] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.

[62] Ruff KM, Pappu RV. AlphaFold and Implications for Intrinsically Disordered Proteins. J Mol Biol 2021;433:167208.

[63] Staff V. Scaling agentic ai safely — and stopping the next big security breach. 2025. https://venturebeat.com/ai/scaling-agentic-ai-safely-and-stopping-the-next-big-security-breach/.

[64] Boehm FJ, Zhou X. Statistical methods for Mendelian randomization in genome-wide association studies: a review. Comput Struct Biotechnol J 2022;20:2338–51.

[65] Thiele I, Swainston N, Fleming RMT, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol 2013;31:419–25.

[66] Çakır T, Alsan S, Saybaşılı H, et al. Reconstruction and flux analysis of coupling between metabolic pathways of astrocytes and neurons: application to cerebral hypoxia. Theor Biol Méd Model 2007;4:48.

[67] Kobayashi T, Yoshizawa K. Optimization algorithm for feedback and feedforward policies towards robot control robust to sensing failures. Robomech J 2022;9:18.

[68] Ramakrishnan V, Friedrich C, Witt C, et al. Quantitative systems pharmacology model of the amyloid pathway in Alzheimer's disease: insights into the therapeutic mechanisms of clinical candidates. CPT: Pharmacomet Syst Pharmacol 2023;12:62–73.

[69] Ali M, Giri S, Liu S, et al. Digital twin-enabled real-time control in robotic additive manufacturing via soft actor-critic reinforcement learning. ArXiv 2025. https://doi.org/10.48550/arxiv.2501.18016. Epub ahead of print.

[70] Kober J, Bagnell JA, Peters J. Reinforcement learning in robotics: a survey. Int J Robotics Res 2013;32:1238–74.

[71] Schena L, Marques PA, Poletti R, et al. Reinforcement Twinning: from digital twins to model-based reinforcement learning. J Comput Sci 2024;82:102421.

[72] Goel G, Chen N, Wierman A. Thinking Fast and Slow. ACM SIGMETRICS Perform Eval Rev 2017;45:27–9.

[73] Staff S. Digital twin evolution: a 30-Year journey that changed industry. 2025. https://www.simio.com/digital-twin-evolution-a-30-year-journey-that-changed-industry/?.

[74] Committee ADEI. *Digital Twin: Definition & Value*. American Institute of Aeronautics and Astronautics (AIAA) and AIA; 2020. https://aiaa.org/wp-content/uploads/2024/12/digital-twin-institute-position-paper-december-2020.pdf?.

[75] Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. npj Digit Med 2024;7:77.

[76] Maharjan R, NAh Kim, Kim KH, et al. Transformative roles of digital twins from drug discovery to continuous manufacturing: pharmaceutical and biopharmaceutical perspectives. Int J Pharm: X 2025;10:100409.

[77] Susilo ME, Li C, Gadkar K, et al. Systems-based digital twins to help characterize clinical dose–response and propose predictive biomarkers in a Phase I study of bispecific antibody, mosunetuzumab, in NHL. Clin Transl Sci 2023;16:1134–48.

[78] Fekonja LS, Schenk R, Schröder E, et al. The digital twin in neuroscience: from theory to tailored therapy. Front Neurosci 2024;18:1454856.

[79] Cen S, Gebregziabher M, Moazami S, et al. Toward precision medicine using a "digital twin" approach: modeling the onset of disease-specific brain atrophy in individuals with multiple sclerosis. Sci Rep 2023;13:16279.

[80] Mann DL. The Use of Digital Healthcare Twins in Early-Phase Clinical Trials Opportunities, Challenges, and Applications. JACC: Basic Transl Sci 2024;9:1159–61.

[81] Vidovszky AA, Fisher CK, Loukianov AD, et al. Increasing acceptance of AI-generated digital twins through clinical trial applications. Clin Transl Sci 2024;17:e13897.

[82] Barbiero P, Torné RV, Lió P. Graph representation forecasting of patient's medical conditions: towards a digital twin. 2020. https://arxiv.org/abs/2009.08299?. accessed October 9, 2025.

The Journal of Prevention of Alzheimer's Disease 13 (2026) 100399

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

Multi-modal data analysis for early detection of alzheimer's disease and related dementias

Liming Wang [a], Jim Glass [a], Lampros Kourtis [b], Rhoda Au [c,*]

[a] Massachusetts Institute of Technology, Cambridge, MA, USA
[b] Gates Ventures, Seattle, WA, USA
[c] Boston University Chobanian & Avedisian School of Medicine and School of Public Health, Boston, MA, USA

ARTICLE INFO

Keywords:
Dementia classification
Speech and language processing
State-space model
Multi-modal

ABSTRACT

Until recently, accurate early detection of clinical symptoms associated with Alzheimer's disease (AD) and related dementias (ADRD) has been difficult. Digital technologies have created new opportunities to capture cognitive and other AD/ADRD related behaviors with greater sensitivity and specificity. Speech captured through digital recordings has shown recent promise at feasible levels of scalability because of the widespread penetration of smartphones. One such study is described in detail to illustrate the depth in which artificial intelligence (AI) analytic approaches can be used to amplify the value of audio recordings. Another modality that has also attracted research interest are ocular scans that have near term potential for validation as a digital biomarker and a point of entry for clinical care workflows. Single modality measures, however, are rapidly giving way to multi-modality sensors that are embedded in all smartphones and other internet-of-things connected devices. Artificial intelligence (AI) driven analytic approaches are able to divine clinical signals from these high dimensional digital data streams. These data driven findings are setting the stage for a future state in which AD/ADRD detection will be possible at the earliest possible stage of the neurodegenerative process and enable interventions that would significantly attenuate or alter the trajectory, preventing disease from reaching the clinical diagnosis threshold.

1. Introduction

With United States' (US) Federal Drug and Administration (FDA) approval of lecanemab and donenamab to slow progression of Alzheimer's disease (AD) clinical symptoms, early diagnosis of AD has become increasingly important for initiating timely treatment, slowing disease progression, and improving patients' quality of life and life expectancy [1,2]. But determining who to treat and when remains a significant challenge. Widespread AD in vivo pathological detection is now possible through the FDA's more recent approvals of AD blood-based biomarkers [3]. But presence of AD pathology does not always lead to clinical expression of disease [4]. Further, clinical symptoms, particularly in the earliest stages are highly variable, resulting in a significant challenge of identifying *clinically meaningful* symptoms. Compounding the clinical detection objective is that AD is often co-morbid with other pathologies, such as vascular and/or Lewy body pathologies, which also result in cognitive and related behavioral indicators, some of which overlap with those of AD. The ethical dilemma is potential treatment of those who are AD biomarker positive but will never progress to clinically expressed disease given pharmacologically concomitant side effects. Thus, despite the exciting promise of AD blood biomarkers, the clinical utility of these blood tests will be constrained without confirmation that treatment is warranted.

Cognitive impairment is the most common clinical symptom of AD and related dementias (ADRD) and is a key primary outcome in AD clinical trials to determine intervention efficacy. Neuropsychological tests are widely used to assess cognitive state. They include multi-domain screening tests such as the widely used Mini-Mental State Examination (MMSE) [5] and domain specific assessments that typically rely on trained examiners administering standardized protocols. A comprehensive cognitive protocol is often labor-intensive to administer, comprised of tests that are culturally and educationally biased, and produce results that are variable making distinction from subtle or early-stage cognitive decline difficult to discern [6]. Manual scoring approaches also may miss nuanced indicators that signal cognitive impairment. Technological advances have emerged that can alleviate these issues. Digital capture of behaviors through smartphones, wearables and other internet of things (IoT) present an opportunity to

* Corresponding author.
E-mail address: rhodaau@bu.edu (R. Au).

https://doi.org/10.1016/j.tjpad.2025.100399
Received 4 September 2025; Received in revised form 15 October 2025; Accepted 20 October 2025

overcome the limitations of standardized neuropsychological testing and do so at a scale that has not been hereto for possible.

In this perspective, we highlight two areas in which digital technologies are being used in AD/ADRD because of their non-invasive and feasibly scalable nature. The first section provides a detailed description of research being done with speech as an illustration of how both technological advances are being applied to both data collection and analysis. The second section summarizes ocular scans research to introduce the potential realm of digital biomarkers. The final section looks beyond these highlighted efforts to present a vision of the future, well beyond the single modality approach and where much scientific discovery remains to be done.

2. Speech as a measure of early AD related symptoms

Many IoT devices have speech recording capacity and because speaking is a cognitively complex task, in the context of cognitive impairment detection have already led to automatic dementia classification (ADC) systems that infer cognitive state directly from digitally recorded speech of neuropsychological tests [6–8]. The potential of analyzing speech as an alternative approach to assessing cognitive state is particularly intriguing because it taps into multiple cognitive domains to produce intelligible content and most people speak, regardless of their education, culture, sex, or language.

Individuals with AD and other forms of dementia exhibit measurable changes in their speech production, seen in both the acoustic domain and the language domain [9]. These changes often precede noticeable cognitive related symptoms [10]. Digital recordings of structured speech, such as from neuropsychological testing, also allow for concurrent validation of speech markers as a surrogate measure of cognition compared to neuropsychological tests [11].

ADC systems that analyze speech recordings aim to detect subtle linguistic and paralinguistic cues (e.g., hesitations, disfluencies, semantic anomalies) indicative of a neurodegenerative disorder. In addition, these systems can also mitigate well-known test question biases [8,12,13] because the speech analysis can analyze all content and is not limited to scoring parameters that are impacted by test item relevance. Early work on ADC tasks for AD detection (ADD) used classical machine learning algorithms with hand-crafted speech and linguistic features [7,14]. More recent systems leverage deep learning architectures such as convolutional [11] and recurrent [15] neural networks and neural embeddings from pretrained speech representation models such as wav2-vec 2.0 [16,17] and Whisper [18,19] as well as text language models [18,20]. These speech processing methods have led to published studies that evidence the power of speech analysis across the dementia progression spectrum. Fraser et al. (2016), using a computational approach with 370 linguistic and acoustic features, achieved up to 82 % accuracy in classifying AD versus controls from picture descriptions [21]. Eyigoz et al demonstrated that linguistic variables derived from a pre-recorded picture description task could predict future AD onset (almost 15 years in advance) from a cognitively normal baseline with a significant area under the curve (AUC) of 0.74 and an accuracy of 0.70 [22]. Pan et al , exploring different automatic speech recognition (ASR) paradigms and bidirectional encoder representations from transformers (BERT)-based classification from the DementiaBank 2021 publicly available audio only speech dataset called Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS). They reported test results for their best acoustic-only model at 74.65 % accuracy and their best linguistic-only model at 84.51 % accuracy [23]. García-Gutiérrez et al. (2024), leveraging paralinguistic (acoustic) features combined with socio-demographic data, showed the ability to identify individuals with AD (F1-score = 0.92) and MCI (F1-score = 0.84). They found differentiating MCI from SCD (Subjective Cognitive Decline) yielded an AUC of 0.80, and MCI from AD had an AUC of 0.73 [24]. Agbavor and Liang (2022) found that GPT-3 based text embeddings notably outperformed conventional acoustic feature-based approaches for AD classification [25]. Their model achieved 80.3 % accuracy, 72.3 % precision, 97.1 % recall, and an 82.9 % F1-score on the ADReSSo unseen test set. For Mini Mental State Exam score prediction, a short dementia screening assessment, they found that a Ridge regression model (RMSE) using acoustic features had an RMSE of 6.250, while their GPT-3 Babbage model achieved a lower RMSE of 5.163. Heitz et al. (2024) leveraged GPT-4 to extract five semantic features from spontaneous speech transcripts with up to 93.1 % accuracy on manual transcripts and 90.5 % on ASR transcripts [26]. Amini et al. (2024) used natural language processing (NLP) and machine learning (ML) techniques on recorded neuropsychological test interviews to predict progression from MCI to AD within six years and achieved an accuracy of 78.5 % and a sensitivity of 81.1 %, with a moderate specificity of 75 % [27].

Despite progress in algorithmic design, existing work still focuses on sentence-level speech segments and small datasets such as ADReSS [6,14] and the Framingham Heart Study [7]. Their study surprisingly found that AD is possible with examiner speech only, indicating examiner bias in administration of standardized neuropsychological tests [8,12,13] that are presumed to follow prescriptive administration protocols . Existing speech-specific ADC systems are fundamentally limited in their ability to process neuropsychological test recordings that are typically long in duration, with many published protocols taking 1+ hours to administer. This constraint often forces segment-level inference using forced alignment or manual heuristics [7,17,18], leading to context fragmentation and a drop in fine-grained classification performance [28].

To date, many voice-related analyses have focused on acoustic features. The advantage is acoustic features can be easily extracted regardless of native language spoken. Available open-source automatic speech recognizer (ASR) such as Whisper have led to transcriptive based pipelines that can utilize natural language processing for analysis, but ASR accuracy is more variable outside of the most commonly spoken languages (e.g., English, Spanish, etc.). ASR suffer from loss of acoustic information as well as error propagation, especially in noisy, spontaneous and multi-speaker conversational settings, whereas acoustics only suffer from loss of language information, such as word choices, sentence structure and content richness. Further, speech generated in a natural context is generally longer in length and involves exchange between two or more speakers. To address the challenges of analyzing long-length digital recordings that include interactive speech, we have proposed leveraging state-space models (SSMs) [29,30], a family of architectures designed for efficient long-sequence modeling.

3. Joint acoustic and linguistic analysis of interactive long-length speech recordings

SSMs scale linearly in both space and time, making them ideal for modeling longer-length speech recordings without segmentation. For example, the dementia information in recording of neuropsychological test administration can be subtle and sporadic, with many conversational turns offering little diagnostic value [12]. SSMs' natural capacity for temporal compression allows them to distill salient patterns with minimal information loss, making them particularly well-suited for the ADC task. Below, we present *Demenba* [31], a memory- and compute-efficient architecture trained on over 1000 h of neuropsychological tests with balanced representation across dementia stages.

Fig. 1 shows the overall design of our multimodal dementia classification system, which combines both speech and language analysis. The system consists of four main components: (1) a speech segmenter, (2) an automatic speech recognizer (ASR), (3) an audio-based dementia classifier, and (4) a text-based dementia classifier.

4. Audio analysis

The **speech segmenter** takes an hour-long neuropsychological test recording and breaks it into shorter, more manageable segments. It

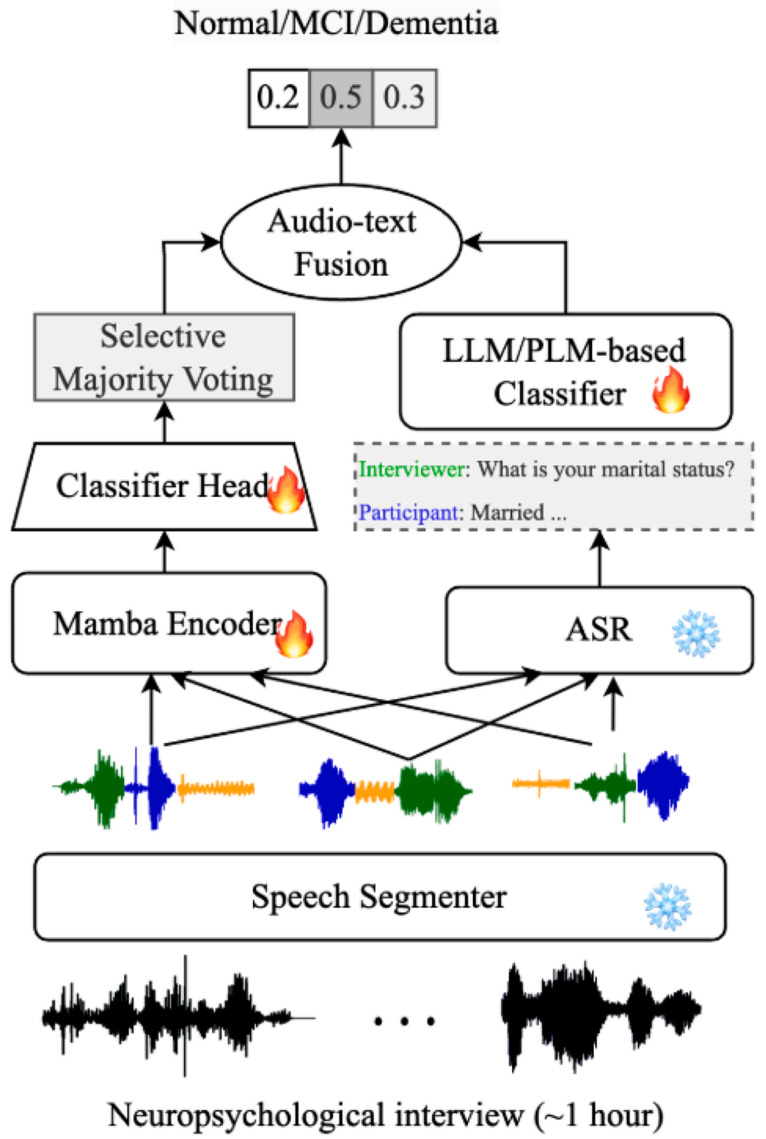


Fig. 1. Overall architecture of the proposed ADC model. The frozen speech segmenter divides the hour-long recording into shorter segments, a trainable SSM-based audio classifier and a trainable text- based text classifier. The predictions from the two classifiers are then combined via late fusion.

separates examiner speech, participant speech and pauses. This not only makes the system more efficient but also allows us to study how factors such as silence duration or examiner/participant speaking turns affect dementia classification.

Next, the **audio dementia classifier** evaluates how the participant sounds. It uses an advanced deep learning model [30,32–35] to capture long-range dependency in the speech. Each segment is assigned a probability of belonging to a dementia category. Since dementia-related cues may appear only in certain moments, the system highlights the most informative segments and weighs them more heavily in the final decision.

5. Text analysis

To complement the audio, the text dementia classifier looks at what the participant says. The ASR system, Whisper [19], automatically transcribes the entire recording into text. A large language model (LLM) [36,37] or pretrained language model (PLM) [38] then analyzes the transcripts, assessing whether the participant's responses are coherent, accurate, and consistent with normal cognition. Beyond simply assigning a diagnosis, the LLM can also generate a clinician-friendly narrative summary, improving interpretability for medical use.

6. Decision fusion

Lastly, the system integrates the predictions from both audio and text branches to determine overall dementia severity. This combined approach improves accuracy compared to using speech or text alone. In particular, for 2-class classification, the AUC of the best text-only and audio-only approaches are 91 % and 87 %, while the combined approach achieves an AUC of 95 %. The reason for this synergy may be that audio models tend to capture prosodic cues such as hesitation and intonation, whereas text models tend to capture lexical/linguistic patterns like filler words and semantic incoherence.

7. Clinical evaluation

We tested the system on the Framingham Heart Study (FHS) dataset [12], which includes about 11,000 h-long neuropsychological test recordings, 2058 of which have been adjudicated and labeled as cognitively intact ($n = 936$) and cognitively impaired ($n = 1122$). For 3-class fine-grained classification task, we further divide the cognitively impaired class into mild cognitive impairment (MCI) and dementia classes.

In the two-class setting (normal vs. dementia), our system (Demenba-medium) achieved an area under curve (AUC) of 92 %, surpassing a 6 % improvement over prior methods. For the three-class task (intact, MCI, dementia), performance remained strong with an AUC of 83 %, a 14 % AUC gain over the previous state of the art. Importantly, the advantage of our method grew when distinguishing more subtle differences between MCI and dementia, suggesting that the system is particularly sensitive to early signs of cognitive decline. Our method consistently outperformed prior methods, with the largest gains observed in the more challenging 3-class setting. Importantly, the system scales effectively, handling over 1000 h of training data and reliably analyzing speech segments up to 6 min long — capabilities essential for real-world clinical recordings. Details about the methods described above are provided in Wang et al. [31].

8. Concluding remarks regarding speech

Our Demenba analyses highlight the complementary roles of acoustic patterns (how someone speaks), speaker dynamics (who is speaking), and linguistic content (what is said) in dementia classification. These findings provide new insights into how different aspects of speech and language reflect cognitive status.Looking ahead to enhancing the potential of speech, we aim to enhance both the performance and interpretability of our models by integrating them with multimodal LLMs, which can combine speech, language and other clinical as well as digital data. Another challenge is to improve the accuracy of our approach by handling data variability introduced by factors such as accent, gender, age and additional health conditions.

In addition to the publicly available DementiaBank ADReSS, the emergence of other speech datasets such as that from the Alzheimer's Drug Discovery Foundation SpeechDx, a multi-center, longitudinal study that is collecting longitudinal voice recordings from approximately 2000 participants with strong clinical characterization profiles provide resources from which to accelerate the translation of audio recording research to clinical utility. Future directions enabled by these datasets include studying finer-grained classification of dementia subtypes and generalization performance on other dementia datasets, to ensure broader clinical applicability.

At the same time, the increasing diversity of datasets underscores important challenges the field must address. Harmonization across sites and recording protocols is critical to ensure that models trained on one cohort remain valid across others. Equally important are anonymization and de-identification strategies, which safeguard participant privacy while retaining the clinical richness of speech data. Developing methods that balance privacy with utility, while also addressing variability in recording conditions, accents, and disease progression, will be essential for making speech-based biomarkers both reliable and ethically deployable in real-world healthcare.

Further, while our method is a step toward more fine-grained categorization of dementia, it remains to be tested whether our method can be extended to detect more subtle cues of dementia in the early stages of dementia, which are much more challenging than differentiating MCI and AD. More clinically relevant metrics such as the RMSE of predicting the MMSE score can also be assessed, provided that the ground truth scores are available.

The broad penetration of smartphones provides a ubiquitous tool for capturing and processing natural speech during everyday phone conversations, offering a rich and passive means of collecting longitudinal speech and language data. This continual, real-world sampling enables for the detection of subtle changes in vocal, lexical, and syntactic patterns and identify early signs of cognitive decline if used over time.

9. Beyond speech: eye as a window to the brain and as a potential digital biomarker

Digital ocular image instruments can track eye movements, the analysis of which is playing an increasingly significant role in AD research due to its potential as a non-invasive tool for early detection and monitoring of the disease. However, like all other cognitive domain testing, the study of eye movements relies on the respective stimuli context such as smooth pursuit, scene viewing, visual search and other. Amongst the various stimuli, reading is a well-defined task that occurs numerous times per day on mobile devices without any prompt. The average person who is literate reads 3000–10,000 words per day on their mobile device. Reading metrics are a highly applicable biomarker in AD research due to the reading process' standardized nature as a well-defined, complex cognitive process whose underlying mechanisms are profoundly impacted by early AD-related cognitive and neurological changes, resulting in quantifiable alterations in eye movement patterns such as increased gaze duration, more fixations, and a loss of the contextual predictability effect.

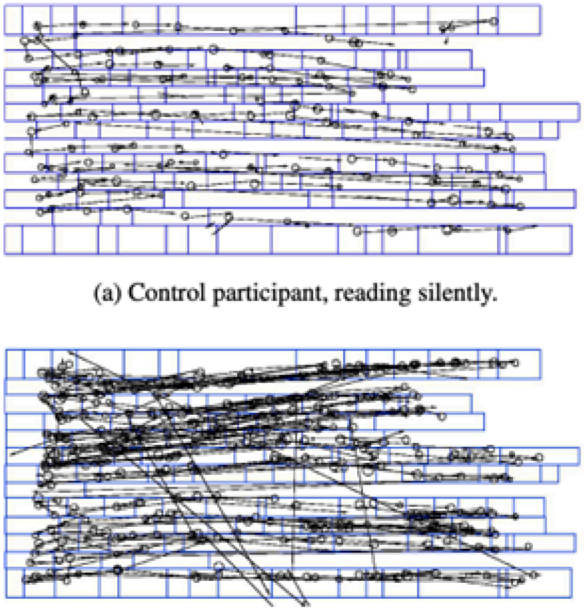
For those who are literate and are not visually impaired, reading is a complex cognitive activity that requires the fine integration of attention, ocular movements, word recognition, language comprehension, working memory, and semantic memory. Many of these cognitive processes, such as attention, inhibitory control, working memory, and decision-making, have been well-documented to be impaired in the early stages of AD/ADRD. Subtle alterations in movement coordination and planning, which are often unnoticed in early AD/ADRD when performing other fine motor tasks like writing, can be precisely detected through eye movement analysis during reading. This is because neurological connectivity changes occur early in the course of AD, disrupting controlled information processing that is critical for reading.

10. Previous studies demonstrating validity of ocular movements as a potential AD biomarker

The eye offers an intriguing opportunity to explore the concept of digital indices as digital biomarkers. The advantage of ocular research is the finite and well-defined measurements and the transparency in analysis. To move from a digital measure to a digital biomarker requires validation that can be easily reproduced or replicated. The few studies summarized below evidence the type of results that have much more direct clinical translation, which is necessary for FDA approval as a biomarker. There is also an existing clinical pathway for conducting eye scans, which would allow more easy integration of an ocular biomarker into the clinical workflow.

A series of studies have been able to show that ocular measures are able to capture in a quantifiable manner, cognitively related natural behaviors, such as reading. In an early study, Fernandez et al. 2013 found a sizeable difference of 23 % in outgoing saccade size between AD and Controls [39]. In another study, Fernandez et al. found that participants with mild AD did not show reduced gaze duration when reading highly predictable text as compared to cognitively intact controls, suggesting early impairments in memory-related mechanisms that support contextual word processing [40]. In another follow up study, Fernandez et al. 2015) further validated that participants with mild AD exhibited significantly more total, first-pass, and especially second pass fixations compared to cognitively healthy controls, indicating increased rereading behavior during both regular and high-predictability sentence reading [41]. They also showed fewer single fixations and shorter outgoing saccades, suggesting impaired contextual word processing (see Fig. 2). In general, cognitively healthy readers adjust their gaze based on the predictability of preceding and upcoming words, while those with AD do not, reflecting early deficits in semantic anticipation and memory-guided eye movement control.

Biondi et al. 2018 reported performance of a softmax classifier using a series of features like first pass fixations, unique fixations and multiple



(a) Control participant, reading silently.

(c) MCI participant, reading silently.

Fig. 2. Reproduced from Fraser et al. 2017, showing less scattered single fixations and outgoing saccades during silent reading from (a) control compared to (c) MCI participants.

fixations Refixations was 88.7 % to classify Controls and 91.0 % for AD patients [42]. Taken together, these studies show that ocular measures can differentiate with specificity between those with and without AD. These results are illustrative of the types of ocular studies underway that lend credence to the eventual validation of digital ocular AD biomarkers.

11. Digital interactions

Given that over 6.8 billion people are estimated to use a smartphones [43] it is worth emphasizing the unique opportunity they provide for continuous remote monitoring of AD/ADRD related behavioral changes. Embedded within each smartphone are the multiple sensors described above that can collect the raw 3-dimensional digital data streams that AI driven algorithms are interpreting into behavioral measures. Patterns of sensor rhythms reflect executive function, sleep-wake stability, and behavioral regularity, all of which deteriorate with mild cognitive impairment [44]. Kaye et al., (2011) used activity sensors and digital markers that included phone use to track daily regularity as a proxy for cognitive decline [45]. Previous studies have demonstrated that frequent app switching and short dwell times can reflect impulsivity or distractibility while reduced diversity may reflect narrowing interests or cognitive fatigue [46]. Changes in session structure can reflect attention span, fatigue, or executive function breakdown. Slower typing and higher error rates may correlate with motor or processing issues [47]. These studies in the aggregate provide further evidence that cognitive decline is associated with disrupted routines or decreased behavioral regularity, which can all be measured through sensors embedded in every smartphone [48].

12. Multi-Modal: the next frontier

Despite the promise of speech and ocular biomarkers, limiting early detection of AD/ADRD clinical symptoms to a single modality would be short-sighted. Cognition is reflected in virtually all bodily movement. Sensors embedded in smartphones, wearables and in home devices also collect behavioral movements. The different sensor modalities in combination, provide a comprehensive multi-modal assessment platform from which to detect early changes in cognitive and other related behaviors. Further, multi-modal assessments can help circumvent limitations of any single modality measurement. For example, despite the promise of reading-based ocular biomarkers, there are multiple factors beyond literacy levels that can impact accuracy of measurement including educational attainment levels, baseline or concomitant decline in visual acuity from cataracts, glaucoma, macular degeneration or other age-related eye disorders, whether reading materials are in the person's native language, language, etc. Other comorbid health conditions such as hearing loss, musculoskeletal related problems, breathing difficulties, are other examples of factors that can impact data collected from any single digital format, Successful interpretation of high volume and highly variable fluctuating patterns in different person-specific combinations of digital data will likely lead to much more accurate differentiation between normal behavioral fluctuations from those that are reflective of early neuropathological progression. While previous work centered on single sensor modality accounts for much of the current digital adoption into clinical research, trials and care, an inflection point is nearing where there will be a shift to multi-dimensional data streams uniquely customized to the individual, but fueled largely by AI analytic methods that are still able to effectively extract common meaningful information from them. If these AI solutions result in high sensitivity and specificity for AD/ADRD at much lower cost and far greater reach, they will accelerate the digital revolution that is already underway.

Other issues regarding digital biomarkers that needs significant consideration are the validation process and the transition to clinical care. Acquisition of data through digital devices does not automatically mean the resulting measurement is a "digital biomarker". Neither digital voice or eye movements can yet be considered digital biomarkers within the US until they go through the same rigorous validation process that both AD imaging and plasma biomarkers have gone through following FDA guidelines, which includes specific context of use [49]. Further FDA approval alone will not lead to clinical use in the US. Many factors impact the path post-validation such as whether the test is widely accessible, whether test results are easily interpretable or whether clear treatment guidelines are available, particularly when specialized expertise is unavailable [50]. Advances in research that capitalize on approaches that are ubiquitously obtainable can ensure this trend does not have to continue, despite the rise in the number of dementia cases around the world,. Combination therapies that are increasingly being touted [51], comprised of both pharmacologic and non-pharmacologic interventions, could have potential to disrupt the trajectory of AD/ADRD progression to the point that disease symptoms at the clinical diagnosis threshold is never reached. But this vision is contingent on detecting those changes many years if not decades earlier in the insidious onset process. Through smartphones and other IoT devices, single modality digital assessment tools are rapidly giving way to multi-modal ones. AI innovations today may soon be replaced by an even more powerful quantum computing environment. While much is unknown, what is certain is that technological advances are solving the challenge of early detection of AD/ADRD to the entire at risk global population, bringing with it great optimism in eradicating them.

Consent statement

Consent was not necessary for the purposes of this perspective piece.

CRediT authorship contribution statement

Liming Wang: Writing – review & editing, Writing – original draft. **Jim Glass:** Writing – review & editing, Writing – original draft. **Lampros Kourtis:** Writing – review & editing, Writing – original draft, Conceptualization. **Rhoda Au:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rhoda Au reports a relationship with Novo Nordisk Inc that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. Rhoda Au reports a relationship with Signant Health that includes: consulting or advisory. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was informed by research supported by the American Heart Association (20SFRN35360180), the National Institute on Aging (2P30AG013846, AG062109, AG068753, AG072654), the Alzheimer's Drug Discovery Foundation (201902–2017835) and Gates Ventures.

References

[1] Szekely CA, Thorne JE, Zandi PP, Ek M, Messias E, Breitner JC, Goodman SN. Nonsteroidal antiinflammatory drugs for the prevention of Alzheimer's disease: a systematic review. Neuroepidemiology 2004;23(4):159–69.

[2] Chuang YF, An Y, Bilgel M, Wong DF, Troncoso JC, O'Brien RJ, Breitner JC, Ferrucci L, Resnick SM, Thambisetty M. Midlife adiposity predicts earlier onset of Alzheimer's dementia, neuropathology and presymptomatic cerebral amyloid accumulation. Mol Psychiatry 2016;21(7):910–5. Jul.

[3] https://www.fda.gov/news-events/press-announcements/fda-clears-first-blood-test-used-diagnosing-alzheimers-disease#:∼:text=The%20Lumipulse%20G%20pTau217%2F%C3%9F,and%20symptoms%20of%20the%20disease.

[4] de Vries LE, Huitinga I, Kessels HW, Swaab DF, Verhaagen J. The concept of resilience to Alzheimer's Disease: current definitions and cellular and molecular mechanisms. Mol Neurodegener 2024 Apr 8;19(1):33. https://doi.org/10.1186/s13024-024-00719-7. PMID: 38589893; PMCID: PMC11003087.

[5] Kurlowicz L, Wallace M. The mini-mental state examination (mmse). J Gerontol Nurs 1999;25(5):8–9.

[6] Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. DLetecting cognitive decline using speech only: the ADReSSo Challenge. Interspeech 2021;2021:3780–4. https://doi.org/10.21437/Interspeech.2021-1220. Oct.

[7] Alhanai T, Au R, Glass J. Spoken language biomarkers for detecting cognitive impairment. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE; 2017. p. 409–16.

[8] Dawalatabad N, Gong Y, Khurana S, Au R, Glass J. Detecting de- mentia from long neuropsychological interviews. In: Findings of the Asso- ciation for Computational Linguistics: EMNLP 2022; 2022. p. 5270–83. Association for Computational Linguistics.

[9] van den Berg RL, de Boer C, Zwan MD, et al. Digital remote assessment of speech acoustics in cognitively unimpaired adults: feasibility, reliability and associations with amyloid pathology. Alz Res Ther 2024;16:176. https://doi.org/10.1186/s13195-024-01543-3.

[10] Young CB, Smith V, Karjadi C, Grogan SM, Ang TFA, Insel PS, Henderson VW, Sumner M, Poston KL, Au R, Mormino EC. Speech patterns during memory recall relates to early tau burden across adulthood. Alzheimers Dement 2024;20(4):2552–63. https://doi.org/10.1002/alz.13731. Apr; Epub 2024 Feb 13. PMID: 38348772; PMCID: PMC11032578.

[11] Ding H, Mandapati A, Karjadi C, Ang TFA, Lu S, Miao X, Glass J, Au R, Lin H. Association between acoustic features and neuropsychological test performance in the Framingham Heart Study: observational study. J Med Internet Res 2022 Dec 22;24(12):e42886. https://doi.org/10.2196/42886. PMID: 36548029; PMCID: PMC9816957.

[12] Al Hanai T, Au R, Glass J. Role-specific language models for processing recorded neuropsychological exams. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018. p. 746–52. Jun; Volume 2 (Short Papers). Association for Computational Linguistics.

[13] Pérez-Toro P, Bayerl S, Arias-Vergara T, Vásquez-Correa J, Klumpp P, Schuster M, Nöth E, Orozco-Arroyave J, Riedhammer K. Influence of the interviewer on the automatic assessment of Alzheimer's disease in the context of the ADReSSo Challenge. Proc Interspeech 2021;2021:3785–9.

[14] Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Alzheimer's Dementia Recognition through spontaneous speech: the ADReSS Challenge. In: Proceedings of INTERSPEECH 2020; 2020. p. 2172–6. Oct 25–29.

[15] Rohanian M, Hough J, Purver M. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. Proc Interspeech 2021;2021:3820–4.

[16] Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst 2020;33:12449–60.

[17] Balagopalan A, Novikova J. Comparing acoustic-based approaches for Alzheimer's disease detection. Proc Interspeech 2021;2021:3800–4.

[18] Li J., Song K., Li J., Zheng B., Li D., Wu X., Liu X., Meng H. Leveraging pretrained representations with task-related keywords for Alzheimer's disease detection. arXiv [Preprint]. 2023 Apr 12:arXiv:2304.06035.

[19] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning (ICML). PMLR; 2023 Jul 23-29. Honolulu, HI. Proc Mach Learn Res. 2023;202:28492-28518.

[20] Haulcy R, Glass J. Classifying Alzheimer's disease using audio and text-based representations of speech. Front Psychol 2021 Mar 26;11:62413721.

[21] Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimers Dis 2016;49(2):407–22. https://doi.org/10.3233/JAD-150520. PMID: 26484921.Eyigoz et al. (2020).

[22] Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine 2020 Oct 22;28:100583. https://doi.org/10.1016/j.eclinm.2020.100583. PMID: 33294808; PMCID: PMC7700896.

[23] Pan Y, Mirheidari B, Harris JM, Thompson JC, Jones M, Snowden JS, Blackburn D, Christensen H. Using the outputs of different automatic speech recognition paradigms for acoustic- and BERT-based Alzheimer's dementia detection through spontaneous speech. In: Proc. Interspeech 2021; 2021. p. 3810–4. https://doi.org/10.21437/Interspeech.2021-1519.

[24] García-Gutiérrez F, Alegret M, Marquié M, Muñoz N, Ortega G, Cano A, De Rojas I, García-González P, Olivé C, Puerta R, García-Sanchez A, Capdevila-Bayo M, Montrreal L, Pytel V, Rosende-Roca M, Zaldua C, Gabirondo P, Tárraga L, Ruiz A, Boada M, Valero S. Unveiling the sound of the cognitive status: machine Learning-based speech analysis in the Alzheimer's disease spectrum. Alzheimers Res Ther 2024 Feb 2;16(1):26. https://doi.org/10.1186/s13195-024-01394-y. PMID: 38308366; PMCID: PMC10835900.

[25] Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digit Health 2022 Dec 22;1(12):e0000168. https://doi.org/10.1371/journal.pdig.0000168. PMID: 36812634; PMCID: PMC9931366.

[26] Heita J, Scheider G, Lander M. The influence of automatic speech recognition on linguistic features and automatic Alzheimer's Disease detection from spontaneous speech. In: The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); 2024. p. 15955–69. Association for Computational Linguistics.

[27] Amini S, Hao B, Yang J, Karjadi C, Kolachalama VB, Au R, Paschalidis IC. Prediction of Alzheimer's disease progression within 6 years using speech: a novel approach leveraging language models. Alzheimers Dement 2024;20(8):5262–70. https://doi.org/10.1002/alz.13886. Aug; Epub 2024 Jun 25. PMID: 38924662; PMCID: PMC11350035.

[28] Bhati S, Gong Y, Karlinsky L, Kuehne H, Feris R, Glass J. DASS: distilled audio State space models are stronger and more duration-scalable learners. In: Proceedings of the 2024 IEEE Spoken Language Technology Workshop (SLT); 2024. p. 1015–22. Dec.

[29] Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. In: Int Conf Learn Represent (ICLR); 2022.

[30] Gu A., Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv [Preprint]. 2023 Dec 1:arXiv:2312.00752.

[31] Wang L, Bhati S, Karjadi C, Au R, Glass J. Recognizing dementia from neuropsychological tests with State space models. In: Automatic Speech Recognition and Understanding Workshop (ASRU); 2025.

[32] Liu Y., Tian Y., Zhao Y., Yu H., Xie L., Wang Y., Ye Q., Liu Y. Vmamba: Visual state space model. arXiv [Preprint]. 2024 Jan 23:arXiv:2401.10166.

[33] Shams S., Dindar S.S., Jiang X., Mesgarani N. Ssamba: Self-supervised audio representation learning with Mamba state space model. arXiv [Preprint]. 2024 May 19:arXiv:2405.11831.

[34] Zhang X, Zhang Q, Liu H, Xiao T, Qian X, Ahmed B, Ambikairajah E, Li H, Epps J. Mamba in speech: towards an alternative to self-attention. IEEE Trans Audio Speech Lang Process 2025;33:1933–48.

[35] Jiang X, Li YA, Florea AN, Han C, Mesgarani N. Speech Slytherin: examining the performance and efficiency of Mamba for speech separation, recognition, and synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025. p. 1–5.

[36] Touvron H., Lavril T., Izacard G., Martinet X., Lanchaux M.A., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E., Lample G. LLaMA: Open and efficient foundation language models. arXiv [Preprint]. 2023 Feb 27:arXiv:2302.13971.

[37] Yang A., et al. Qwen2 technical report. Technical Report. Qwen Team, Alibaba Group; 2024. arXiv [Preprint]. 2024 Jun 23.

[38] Devlin J, Chang MW, Lee K, Toutanova KBERT. 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 (NAACL-HLT); 2019. p. 4171–86.

[39] Fernández G, Mandolesi P, Rotstein NP, Colombo O, Agamennoni O, Politi LE. Eye movement alterations during reading in patients with early Alzheimer disease. Invest Ophthalmol Vis Sci 2013 Dec 30;54(13):8345–52. https://doi.org/10.1167/iovs.13-12877. PMID: 24282223.

[40] Fernández G, Manes F, Rotstein NP, Colombo O, Mandolesi P, Politi LE, Agamennoni O. Lack of contextual-word predictability during reading in patients with mild Alzheimer disease. Neuropsychologia 2014;62:143–51. https://doi.org/10.1016/j.neuropsychologia.2014.07.023. Sep; Epub 2014 Jul 28. PMID: 25080188.

[41] Fernández G, Schumacher M, Castro L, Orozco D, Agamennoni O. Patients with mild Alzheimer's disease produced shorter outgoing saccades when reading

sentences. Psychiatry Res 2015 Sep 30;229(1–2):470–8. https://doi.org/10.1016/j.psychres.2015.06.028. Epub 2015 Jun 27. PMID: 26228165.Biondi 2018- arXiv: 1702.00837v3.

[42] Biondi J., Fernandez J., Castro S., Agamennoni O., Eye-movement behavior identification for AD diagnosis. arXiv:1702.00837.

[43] The Raticati Group. Mobile Statistics Report, 2020-2024. https://www.radicati.com/wp/wp-content/uploads/2021/Mobile_Statistics_Report,_2021-2025_Executive_Summary.pdf.

[44] Satomi E, Apolinário D, Magaldi RM, Busse AL, Vieira Gomes GC, Ribeiro E, Genta PR, Piovezan RD, Poyares D, Jacob-Filho W, Suemoto CK. Beyond sleep: rest and activity rhythm as a marker of preclinical and mild dementia in older adults with less education. Neurobiol Sleep Circadian Rhythms 2024 Dec 25;18:100110. https://doi.org/10.1016/j.nbscr.2024.100110. PMID: 39834590; PMCID: PMC11745811.

[45] Kaye JA, Maxwell SA, Mattek N, Hayes TL, Dodge H, Pavel M, Jimison HB, Wild K, Boise L, Zitzelberger TA. Intelligent Systems for assessing aging changes: home-based, unobtrusive, and continuous assessment of aging. J Gerontol B Psychol Sci Soc Sci 2011;66 Suppl 1(Suppl 1):i180–90. https://doi.org/10.1093/geronb/gbq095. Jul; PMID: 21743050; PMCID: PMC3132763.

[46] Gordon ML, Gatys LA, Guestrin C, Bigham JP, Trister A, Patel K. App usage predicts cognitive ability in older adults. In: Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems (CHI '19), Glasgow, Scotland, UK; 2019. p. 12. https://doi.org/10.1145/3290605.3300398. 4–9 May.

[47] Alfalahi H, Khandoker AH, Chowdhury N, et al. Diagnostic accuracy of keystroke dynamics as digital biomarkers for fine motor decline in neuropsychiatric disorders: a systematic review and meta-analysis. Sci Rep 2022;12:7690.

[48] Kourtis LC, Regele OB, Wright JM, et al. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. npj Digit Med 2019;2:9.

[49] Vasudevan S, Saha A, Tarver ME, et al. Digital biomarkers: convergence of digital health technologies and biomarkers. npj Digit Med 2022;5:36. https://doi.org/10.1038/s41746-022-00583-z.

[50] Committee on Policy Issues in the Clinical Development and Use of Biomarkers for Molecularly Targeted Therapies; Board on Health Care Services; Institute of Medicine; National Academies of Sciences, Engineering, and Medicine Graig LA, Phillips JK, Moses HL. Biomarker Tests For Molecularly Targeted Therapies: Key To Unlocking Precision Medicine. Washington (DC): National Academies Press (US); 2016 Jun 30. p. 5. Processes to Improve Patient Care. Available from: https://www.ncbi.nlm.nih.gov/books/NBK379329/.

[51] Salloway SP, Sevingy J, Budur K, Pederson JT, DeMattos RB, Von Rosenstiel P, Paez A, Evans R, Weber CJ, Hendrix JA, Worley S, Bain LJ, Carrillo MC. Advancing combination therapy for Alzheimer's disease. Alzheimers Dement 2020 Oct 7;6(1): e12073. https://doi.org/10.1002/trc2.12073. PMID: 33043108; PMCID: PMC7539671.

Glossary of Key Terms

Automatic Dementia Classification (ADC): Computational systems that detect dementia from digital signals such as speech recordings.

Automatic Speech Recognition (ASR): Technology that converts spoken language into text.

Paralinguistic Features: Non-verbal aspects of speech such as pitch, pauses, hesitations, or intonation.

Linguistic Features: Verbal aspects of speech, including vocabulary, syntax, coherence, and semantics.

Deep Learning: Neural network-based machine learning models used for automated feature extraction and classification.

Convolutional Neural Network (CNN): A deep learning architecture effective for analyzing acoustic speech signals.

Recurrent Neural Network (RNN): A sequence-processing deep learning model used for temporal data like speech

State-Space Models (SSMs): Architectures designed for efficient modeling of long sequences, applied here for long speech recordings.

Language Models (LMs): Neural models trained on text data to capture linguistic patterns and meaning.

Area Under the Curve (AUC): A measure of a model's ability to distinguish between classes.

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

Reinventing “N” in the A/T/N framework: The case for digital

Rhoda Au [a,b,c,d,*], Zachary Popp [a,b,d,1], Spencer Low [a,b,d,1], Nicholas J. Ashton [e,f,g,h,i], Henrik Zetterberg [e,j,k,l,n,m]

[a] *Department of Anatomy & Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA 02118*
[b] *Boston University Alzheimer's Disease Research Center, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA 02118*
[c] *Departments of Neurology, Medicine, and Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine School of Medicine, Boston, MA, USA 02118*
[d] *Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA 02118*
[e] *Department of Psychiatry and Neurochemistry, Institute of Neuroscience & Physiology, the Sahlgrenska Academy at the University of Gothenburg, 405 30 Mölndal, Sweden*
[f] *Banner Alzheimer's Institute and University of Arizona, Phoenix, AZ, USA*
[g] *Banner Sun Health Research Institute, Sun City, AZ 85351, USA*
[h] *NIHR Biomedical Research Centre for Mental Health and Biomedical Research Unit for Dementia at South London and Maudsley NHS Foundation London SE5 8AZ, UK*
[i] *Centre for Age-Related Medicine, Stavanger University Hospital, 4068 Stavanger, Norway*
[j] *Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, 413 45 Mölndal, Sweden*
[k] *Department of Neurodegenerative Disease, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK*
[l] *UK Dementia Research Institute at UCL, London WC1N 3BG, UK*
[n] *Hong Kong Center for Neurodegenerative Diseases, Clear Water Bay, Hong Kong Science Park, Shatin, N.T., Hong Kong, China*
[m] *Wisconsin Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53792, USA*

ARTICLE INFO

Keywords:
A/T/N framework
Technology
Digital “N”
Alzheimer's disease
Biomarkers
Clinical meaningfulness

ABSTRACT

Breakthroughs in biomarkers for amyloid (A), tau (T), and neurodegeneration (N) have advanced the prospects of accurate Alzheimer's disease (AD) diagnosis. However, presence of pathology does not always translate into clinical expression and there are still clear knowledge gaps as to whether someone with AD biological indicators will lead to clinically apparent disease necessary to warrant drug treatments that carry toxicity risk. Reliance on decades-old assessment tools inhibits detection and monitoring at preclinical and early disease stages when new treatments could prove most effective. Evidence has accumulated that digital measures provide accurate detection of disease at early stages. We call for a re-evaluation of the A/T/N diagnostic framework, with digital evaluation measures complementing non-AD specific neurodegeneration markers, and even potentially replacing those non-specific to AD, to provide a clinically relevant feature critical to clinical trial advances and treatment decisions. Achieving this will only be possible if further research into novel digital evaluation tools is pursued with the same support and consideration as amyloid and tau.

Current state of the A/T/N framework

Accurate diagnosis at the primary care level has been one of the holy grails of Alzheimer's disease (AD) research. The United States' (US) Food and Drug Administration (FDA) 501(k) clearance of cerebrospinal fluid (CSF) biomarkers of amyloid (A) and tau (T) [1], the hallmark pathologies of AD, brought that goal within reach, but still within the remit of specialist care. Recent FDA clearance of the Lumipulse G pTau217/amyloid 1-42 ratio as a blood-based biomarker for AD, with more plasma AD platforms in the FDA approval pipeline [2], marks the remaining step to widespread clinical utility. AD blood-biomarkers will enable primary care health providers to identify patients at risk for AD [3] Accumulating data show CSF equivalence of many of the simplified blood tests in clinically relevant settings [4], and secondary analyses of key trials further demonstrating the promise of plasma biomarkers as pharmacodynamic markers of AD treatments, such as p-tau217 in the TRAILBLAZER study [5]. Simplified sampling and pre-analytical sample handling through dried plasma spot analysis create opportunities to

[*] Corresponding author at: 72 E. Concord Street, Boston, MA 02118.
E-mail address: rhodaau@bu.edu (R. Au).
[1] Contributed equally.

https://doi.org/10.1016/j.tjpad.2025.100395
Received 4 September 2025; Received in revised form 15 October 2025; Accepted 20 October 2025

deploy at scale [6]. Neurodegeneration (N) rounds out the A/T/N framework for diagnosis of AD distinct from other dementia subtypes [7], and can be indicated through magnetic resonance imaging (MRI) atrophy, fluorodeoxyglucose positron emission tomography (FDG-PET) metabolism, or several other markers, but obtaining these measures require access to facilities and are costly. Plasma neurofilament light (NfL) has been linked to AD neurodegeneration [8] and its FDA Breakthrough Device Designation status has led to NfL being a marker of N, that can much more easily be measured at scale. It should be noted, however, that NfL is also ubiquitously increased in most neurodegenerative disorders [9] and a primary outcome in multiple sclerosis (MS) and amyotrophic lateral sclerosis (ALS) drug efforts. Also of note is that the FDA- approved plasma biomarkers for specificity of AD do not include any general marker of neurodegeneration. Therefore, while NfL as well as glial fibrillary acidic protein (GFAP) have been promoted as part of the AD plasma 94 biomarker diagnostic panel, they are not critical for diagnosis of AD. Worldwide, given that people over the age of 65 now outnumber those under the age of 5, projected cases of AD are expected to greatly expand by 2050. Thus, the availability of AD plasma biomarkers makes possible an easily accessible and scalable AD diagnosis tool at greatly reduced costs, there democratizing AD research, and treatment care that currently suffers from bias towards the high income, highly educated and well-insured [10].

Spurred by the National Alzheimer's Project Act (NAPA) [11] established in 2012 under the Obama Administration, the secondary goal of finding effective treatments by 2025 is on track. Despite the many debated limitations, aducanumab was the first approved drug for treatment of AD since NAPA was constituted, and in the context of a precision medicine approach, appears to modify disease symptoms for a small subset of patients with early AD and PET-confirmed high amyloid [12]. Moreover, lecanemab received FDA accelerated approval in 2023 based on promising findings [13], and donanemab soon followed with its own FDA approval [14]. Readouts from other Phase 3 clinical trials are expected over the next few years setting the stage for other FDA approved AD treatment options in the near future. Further, the spate of Phase 1 and Phase 2 clinical trials has grown substantially, fueled in part by the additional investment in AD research by the National Institute on Aging. In combination, the progress that has been made in AD over the past decade is impressive. These successes demonstrated critical gaps that need to be addressed to maximize the benefit of recent scientific headways.

Limitations to the A/T/N framework

Autopsy-confirmed A/T/N is not always accompanied by concomitant clinical expression [15]. While it is widely presumed that antemortem A/T/N biomarker positivity is an indicator for preclinical or MCI due to AD detection, it is less clear whether presence of AD pathology will lead to clinically expressed disease. Much of the data available is on non-representative populations, whether it be from post-mortem study samples or from antemortem CSF, /PET/MRI studies. Given this known longstanding recruitment bias in AD research, stemming from a subset having the means or willingness to participate in studies that are based at a high resourced research environment, what remains unknown is how prevalent is AD pathology in the absence of clinical symptom in the general population. This gap in knowledge is important because recently FDA approved AD treatments carry significant toxicity risk that includes premature death. In general, risk of significant side effects is typical for any drug treatment. Thus, regardless of the number of AD drug treatments that make it to market, a perpetual question for any healthcare provider is under what circumstances would AD treatment be warranted, particularly if there is an absence of clinically apparent symptoms. Plasma AD biomarkers, which are highly sensitive to AD, still suffer from variability in specificity. If research documents even greater prevalence of AD biomarkers that is clinically silent in the general population than is currently known, the bar may likely be raised for evidence for clinically symptomatic indicators of disease [16]. Further, use cases for treatment and clinical trial eligibility are complicated by the heterogeneity identified in biomarker positivity and/or clinical presentation among patients with different demographic characteristics, including sex, education or racial and ethnic backgrounds [17]. Thus, despite the premise that early detection enabled by AD plasma biomarkers could potentially allow intervention at a time when treatment would be most effective, an ethical question is raised as to the appropriateness in trial enrollment, and eventually treatment, for those who are A/T/N biomarker-positive in the absence of any clinical indicators. We contend that detection of biomarkers must be accompanied by clinical symptoms to avoid treating individuals who may be misdiagnosed. For drug treatments that carry some level of toxic risk, reliance on biomarker positivity alone could be tantamount to giving drug treatment to any person who has any other risk factor in which AD risk is increased but does not result in AD that meets threshold for clinically diagnosed disease. For example, those who are Apolipoprotein e4 allele (APOE4) heterozygote or homozygote positive, their increased risk for AD is 2-3 or up to 10-15 fold higher that those who are *APOE4*-negative but clinical AD diagnosis is not 100 % inevitable. Similarly, those with high cardiovascular risk can be as high as triple the risk for AD compared to those with low cardiovascular risk but are not certain to developed clinically diagnosed AD.

This ethical issue is unlikely to be realized in the near future given the lack of available treatments at preclinical disease stage also known a mild cognitive impairment (MCI) due to AD. Recent advances, however, demonstrate how quickly new treatments may become available to patients. Solutions for detection at the preclinical disease stages would provide a new population for clinical trial enrollment with early intervention. An additional benefit is the potential to identify individuals to prioritize for non-pharmacological interventions. The current A/T/N framework does not have the specificity to detect the nuance of an inherently insidious clinical onset process that spans from presumably asymptomatic (using current clinical diagnostic tools) to the clinically symptomatic.

While A/T are largely accepted as specific to AD diagnosis, regardless of the why and how [18–21], the same cannot be said about N. NfL is just one of several different plasma biomarkers of neurodegeneration [22, 23], but given its widespread use, it has emerged as a well-accepted measure of neurodegeneration. NfL has been documented as an indicator of neuronal injury that is not only evident in multiple sclerosis and AD but also other neurodegenerative disorders, as well as in acute conditions, including cardiac arrest [24], stroke, brain trauma and encephalitis [25]. Further, blood NfL is also affected for those with chronic kidney disease and by BMI, [26,27] complicating the use of this more accessible marker in clinical practice. Given the non-specificity of any "N" biomarker, including NfL and the previously described, concerns around the utility of A/T/N for determining AD specificity, clinical trial eligibility and appropriate treatment options encompassing all individuals, an opportunity emerges to step out of conventional A/T/N biomarker lanes.

Importance of clinically meaningful endpoints

Clinical meaningfulness is generally applied as an outcome measure of import in clinical trials studies that seek FDA approval. For AD-related clinical trials, the administration of neuropsychological tests, such as the Mini-Mental State Examination (MMSE) or Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), are routinely used to generate metrics of cognition, one of the key performance indicators of clinical trial intervention effectiveness. Until the aducanumab approval, a component of the FDA AD drug treatment approval hinged on demonstrating either improvement or attenuated decline/no further decline in cognitive function. This determination was often made through a composite or domain specific neuropsychological test score (e.g., memory) as a proxy for clinical meaningfulness. The Centers for

Medicare and Medicaid Services (CMS) decision to restrict coverage of aducanumab, and its subsequent withdrawal from the market, was due to the lack of sufficient evidence related to neuropsychological measures of cognition despite its measured statistical significance (22 % reduction in decline of clinical dementia rating in high-dose arm vs placebo) [28, 29]. At the heart of the controversy related to FDA approvals for aducanumab is the recognition that clinical manifestation of AD is crucial and that cognition is a critical feature of clinical manifestation. The increased receptivity to lecanemab and donanemab both within and outside of the U.S., stems, in part, from the slowed progression of cognitive decline by 27 % and 33 %, respectively, within those in the treatment group compared to the placebo group [30] although to what extent this attenuation will persist beyond the original trial study period is still being determined.

Perpetuating assessment bias in clinically meaningful endpoints

The significance of cognition in any AD related clinical trial makes it all the more surprising how little has been done to advance the field of cognitive assessment. Clinical trials have relied on one or multiple tests that were developed decades ago (e.g., ADAS-cog, MMSE, Montreal Cognitive Assessment [MoCA], and CERAD [Consortium to Establish a Registry for Alzheimer's Disease]) [31], despite recognition of the constraints of these instruments [32]. Limitations include ceiling effects, invariances, reliability, and appropriateness for racially, ethnically, educationally or language diverse populations [33]. Advances in these instruments since their development are limited. Several large-scale trials have aimed to address the issues with traditional tools by adding supplementary tests [34] or exploring alternative scoring methods [35]. Alternatively, clinical trials have explored constructing composite batteries from subsections of the ADAS-cog and other instruments [36].

These efforts do not address the root cause of the shortcomings in cognitive assessment. ADAS-cog was developed in 1984 [37] the MMSE was developed in 1975 [38] and the CERAD in 1989 [39]. The inability to rethink effectively measuring cognition hinders the ability to detect efficacy in clinical trials and impacts who enrolls based on testing-based inclusion and exclusion criteria. High screen-fail rates have been addressed previously by swapping one traditional tool for another [40]. The minor iterations of decades-old instruments have resulted in heterogeneity of tests, with 31 known versions of the ADAS-cog instrument as of 2018 [32]. Additions of subscales further exacerbate the existing concerns of traditional assessment methods. The widespread practice of simply translating and making minor modifications to the assessments for use outside of North America and Western Europe has perpetuated the educational, cultural, and linguistic bias inherent in these tests. As efforts to develop effective translations have grown, there remains inconsistency in the standards for item-level modifications, and the lack of normative data continues to lead to cultural biases [41].

First call to action: invest in digital to develop more sensitive and less biased clinical measures

Measures currently described as "digital biomarkers" would be more accurately described as "digital phenotyping". Digital phenotyping describes the moment-by-moment quantification of behavior using embedded digital sensors, such as smartphones [42]. Many digital phenotyping approaches for cognitive assessments rely on active engagement assessments that require a person to respond to questions similar to standardized neuropsychological tests. For example the Framingham Heart Study deployed a smartphone application that collects voice and response to test stimuli or finger responses to screen-based tests. The Intuition Study, that was more nationally representative, including more geographically balanced, used the Cambridge Cognition smartphone application (CamCog) as a digital cognitive assessment. These and other studies are able to collect additional digital phenotyping measures through embedded sensors, supplementing more traditional neuropsychological test measures. This approach, however, does not provide more highly sensitive detection of subtle clinical changes that will more accurately differentiate those that might be AD biomarker positive and largely asymptomatic versus those who are AD biomarker position but showing early signs of MCI due to AD stage. This distinction between AD biomarker positive but asymptomatic versus MCI (or preclinical/prodromal AD) is important. The concept of "asymptomatic" itself is highly dependent on the sensitivity of the tools being used to measure symptom. Cognitive assessments acquired using digital tools show promise for detecting differences in cognitive performance to a greater extent than standardized test scores including through multimodal measurement, which brings together different digital data modalities to measure a series of interrelated functional and behavioral measures [43]. As digital phenotyping becomes more integrated into AD research, there is a likelihood that a recognized separation between "asymptomatic" versus "preclinical" will begin to emerge, just as the distinction between MCI and mild AD is now accepted. A growing number of studies are deploying digital tools to do more than derived measures that mimic well-established measures that detect cognitive AD-related manifestations. The undisputed pioneer in this realm has been the work of Kaye and colleagues, who were the first to test embedding sensors as an alternative to studying longitudinal trajectories of behavior change in the home [44]. They found that frequent sensor-based monitoring allowed for more accurate detection of the transition from normal cognition to MCI [44]. In addition, Kaye et al. monitored elderly participants answering online surveys and demonstrated slowed survey completion time preceded the onset of MCI by over a year [45]. In another study of 27 participants who were not demented, less daily computer use was associated with smaller hippocampal volume, a well-established neuroimaging biomarker of neurodegeneration associated with increased AD risk [46]. In the first study to harness the natural behavior of voice, Kaye et al. [47], utilized remote video telecommunication software to record average talk time per day to determine whether speech detection algorithms could determine normal cognitive aging to MCI transitions. They reported that MCI subjects spoke more words during conversations and exhibited longer daily talking times than normal subjects. They concluded that MCI subjects exhibit subtle language processing deficits that are sensitive to transitions to MCI. Home monitoring with infrared sensors has also been used to classify sleep, activity, gait, and behavioral changes relevant to neurodegeneration and traumatic brain injury. Through capturing novel metrics of behavior, use of mobile technologies affords the opportunity for longitudinal remote data collection with evaluation occurring more frequently and with greater granularity. Importantly, these measures can offset some of the active engagement assessment bias because they are collected from any person's use of these technologies' independent of their age, education, language or culture.

The unique importance of the early studies done by Kaye et al. is they repurposed technology to measure AD-related cognitive behaviors. while avoiding the inherent bias of traditional assessment tools through a focus on unstructured data streams like voice, and longitudinal behavioral change in a real-world setting. The evolution from computers to smartphones has lowered the barriers to assessment for those who live far from healthcare facilities/providers with the training or professional expertise to administer them. A notable effort by the Real-World Implementation, Deployment, and Validation of Early Detection Tools and Lifestyle Enhancement (AD-Riddle) project has reviewed and tested digital platform for multi-national deployment. While there are numerous similar efforts underway that are addressing the cross-cultural considerations, many assessments do not adequately address the importance of language on performance. There are an estimated 7159 languages in the world today, and no single cognitive assessment tool that is available in all these languages. With the anticipated increase in AD incidence and prevalence across the world, the lack of an assessment tool that could be applicable to anyone, anywhere will further exacerbate the significant global health inequities in clinical practice and research [48]. Building on the monitoring of health during daily behavior first demonstrated by Kaye et al. is expected to reduce biases in that evaluation of cognition and function need not rely on specific stimuli requiring specific languages for administration or recognition of culturally specific images and stories.

Second call to action: the case for digital as a clinical indicator of neurodegeneration

Long-used assessment tools continue to dominate cognitive outcomes in AD clinical trials even as digital tools offer an opportunity for lower burden, higher sensitivity, and reduced bias. Digital health tools encompass a broad swath of technologies including smartphone applications, wearable devices, computing platforms, software, and other sensors that can be applied to monitor health outcomes [49]. While the FDA has signaled strong support for digital indices that can serve as susceptibility/risk biomarkers, the pragmatic reality of current FDA approval tied to well-established clinical measures had led to a tepid response in investing in digital alternatives by both the pharmaceutical and scientific research community. In cardiovascular health, digital technologies have demonstrated utility for continuous monitoring of disease, improving patient outcomes and individual access to health data [50]. The Apple Heart Study provided evidence for the large-scale monitoring potential of digital technologies, with smartwatch-based irregular heartbeat notifications and electrocardiogram patches providing reliable home diagnoses [51]. Digital technologies can offer similar advantages in AD for large-scale monitoring [52], and may prove even more useful in evidencing the neurodegenerative process is clinically meaningful given the potential for granular capture of functional measures across domains of speech, gait, sleep, and activity that often precede clinical onset that meets diagnostic criteria [47]. While significant advances in imaging and fluid biomarkers of neurodegeneration have been developed over the past decade, despite their lack of disease specificity, identification and validation of digital clinical indicators of neurodegeneration are at an early, though promising, stage. A recent review by Polk et al. highlighted the growing body of evidence in support of the feasibility and validation of remote and unsupervised assessments for detection of subtle cognitive decline in preclinical AD. In their review, Polk et al., compared these digital assessments to more widely used cognitive assessments including the Preclinical Alzheimer's Cognitive Composite (PACC) and other standardized neuropsychological tests. Digital cognitive assessments have shown strong correlations with plasma biomarkers, including p-tau181, GFAP, and NfL. Accuracy in classification between MCI and healthy controls, and between MCI and AD has been demonstrated cross-sectionally and longitudinally over short time periods. Thus far, discriminative accuracy for long-term longitudinal characterization of AD risk for those with and without biomarkers has not been established, but combined digital cognitive assessments and blood-based biomarkers have demonstrated promise in predicting future cognitive decline. Given that having AD pathology does not always translate to clinically expressed disease coupled with the known bias of current assessment methods and the well-documented heterogeneity in clinically meaningful outcomes (e.g., cognition and cognitively-related behaviors), our second call to action is to leverage ongoing technological advances to develop better clinical measures that evidence progression in neurodegeneration.

Third call to action: avoid repeating history by expanding representation through scalable sensor-based devices

Our third call to action is to capitalize on continuous smartphone and wearable sensor-based assessments that can be done in a natural setting and increase inclusivity in clinical trials and other research studies [53]. Digital deployment alone will be insufficient to maximize inclusivity. Digital indices can be influenced by a multitude of sociocultural factors. Thus, it is a scientific imperative to prioritize global representation in the collection, identification, and validation of digital AD/AD and related dementias (ADRD) clinical measures. Failure to do so will perpetuate the limitations of research that has largely been conducted in high income countries, such as non-representative normative data or reliance on technologies widely inaccessible in low-resource areas. Continued proliferation of smartphones globally offers a unique opportunity to start from a globally inclusive baseline, so long as considerations of resource limitations (e.g., access to connectivity) and education/culturally/language agnostic stimuli are considered. Strategic use of technology will allow real-time monitoring of drug-treatment impact on a clinically meaningful digital "N", which can document symptom decline with replicability possible on a globally inclusive scale.

The cost and existing uptake of digital tools compared to blood, CSF, and imaging disease markers in low and middle-income countries offers an opportunity for expanding global research and clinical integration; however, many challenges remain in avoiding the pitfalls of developing globally representative digital markers. Current barriers to inclusion of digital trial endpoints include lack of standardization in digital monitoring and best practices for the storage, management, and analysis of high volumes of digital data [54]. The lack of standard digital tools is difficult to address given the constant evolution in health technologies. In clinical practice, this lack of standardization may also present an opportunity for increased patient autonomy. Heterogeneity in the market of available digital tools presents a multitude of options ranging from active gamified cognitive testing to passive engagement monitoring which can be conducted with no burden placed on the user. Transitioning from active engagement technologies that require varied levels of staff interaction, technological fluency, and participant burden to complete set tasks, towards passive engagement technologies that capture continuous streams of data from zero-touch home-based or device-embedded sensors could yield more robust data streams with low effort longitudinal participation [52]. The promise of passive monitoring for behavioral detection is limited, however, by unresolved questions about data privacy, security, and the challenges in data storage, processing, analysis, and interpretation given the unstructured nature of continuous digital data streams. Shifting the heterogeneity in digital monitoring from a barrier to an opportunity will depend on developing robust solutions around data management and analysis, as well as patient- and physician-centered research on the accessibility, security, and privacy needs for effective clinical implementation. To fully realize this future will require efforts to advance the harmonization and analysis of large-scale digital data streams that should aim to produce device-agnostic fluidic markers of health [55]. These efforts will depend on large digital data sources with well-validated endpoints, as well as interdisciplinary cooperation to foster inclusion of open data science and machine learning analysis. Harmonization is especially challenging in an area such as digital health, where new technologies are constantly arising and aiming to innovate towards new solutions; however, achieving harmonized protocols for data collection, storage, and analysis will be crucial to avoiding the pitfalls of culturally specific cognitive assessment characterized by fractured insights on non-generalizable samples. Importantly, in this call for action, the development of pre-competitive, open source digital data management and processing tools, as well as open access data resources and results will be critical if global inclusion is to be achieved. The current dominance of proprietary hardware and software risks perpetuating the same barriers to equal opportunity science that current methods have generated. Global digital phenotyping and increased uptake of non-proprietary digital secondary outcomes, with dedicated attention to both the advanced analysis of results and qualitative feedback of users, will be needed to move digital forward towards its potential as a low-cost, accessible, early detection tool.

Fourth call to action: reconceptualizing the A/T/N framework

Our fourth and final call to action is for a **reconceptualization of the A/T/N framework**. One that pushes to the forefront the objective of a maximally inclusive framework that is feasible worldwide, including in the lowest of research or clinical care settings. While A/T plasma biomarkers are specific to AD, complementing the current non-AD-specific "N" biomarker with a digital "N" as a clinical indicator of cognitively related behaviors and function (e.g., memory, speech, gait, balance) would attenuate the current uncertainty of whether A/T positivity will remain clinically silent. Longer-term, there is potential a digital "N" could obtain even better specificity in determining those who would most benefit from AD treatments compared to a biomarker "N". Most studies aim to distinguish MCI or dementia from control populations without distinguishing between disease stages or subtypes. Digital measurement in its current form provides a benefit over "N" biomarkers for its potential specificity to AD. Cognitively dependent functional and behavioral measures can be digitally monitored using smartphones or other mobile internet connected devices. The opportunity to continuously evaluate these outcomes would be potentially clinically meaningful to those at high AD risk and would likely be acceptable in lieu of costly and abstract blood biomarker measurement of neurodegeneration. Increasing specificity of diagnostic tools to identify those who have or will develop clinical symptoms could significantly improve screening for clinical trials for AD. Current screen-in confirmation of AD using PET and MRI scans have added not only significant burden in study execution, but also far greater costs and greater exclusion of the at risk population, particularly in low resourced regions. Current estimated costs to bring AD preclinical treatments to FDA approval is over $5 billion [56]. Low cost, ubiquitous sensors in the form of wearable devices and smartphone applications could serve as an initial screen for clinically detectable symptoms, substantially accelerating the pace and decreasing the cost of discovering new and better AD treatments.

AD is an insidious onset disease in which the demarcation of when pathological burden translates to clinical expression is highly dependent on the assessment modality. Galvanizing the research community to consider the role of digital will push the definition of clinical expression upstream in the disease course, (e.g., to stages that are currently considered "asymptomatic") potentially moving the focus of clinical trials that are currently centered on MCI to earlier stages. There is great excitement in the field for recently FDA approved drugs that slow the rate of cognitive decline. But how much more clinically meaningful would it be to someone at AD risk to intervene and delay or even prevent cognitive decline? Research has repeatedly indicated that delaying onset of symptoms by 5 years can reduce overall risk for disease by 50 % [57]. We argue that digital technologies offer an opportunity to re-evaluate current definitions of clinical meaningfulness and consider what new directions of intervention will emerge when at-risk patients and their physicians can consider prevention of disease as an alternative to treatment.

Further investment in and evaluation of digital markers could lead to highly accurate prognostic indications of AD and its etiological subtypes. Increased sensitivity and specificity of change far earlier in the AD disease pathway are ambitious but possible objectives. To meet this bold vision will require a transition from the practice of collecting single modality digital phenotypes that are often analyzed as relatively static derived measures towards collection of rich, multimodal digital data streams [58], that are analyzed in their native dynamic and fluid format. Rigorous and transparent testing for reproducibility and replicability will be needed, with prioritization of using foundational open-source tools. An early focus on pre-competitive approaches will be important because the considerable challenge of research keeping pace with the rapid and continuous evolution of digital technology and artificial intelligence (AI). These constant advances in digital measurement potential, and corresponding analytic capacity will lead to compressed timelines for monitoring disease indicators, providing clinical practice with a real time prognostic marker that has stronger sensitivity and specificity of clinically meaningful change. "N" markers today still provide value in the confirmation of conversion risk through structural change, although the cost of additional testing may not be justifiable if the potential for digital monitoring is fully realized.

Everything people do is through their brains, and cognitively related skills are continuously expressed through behaviors such as speech, gesture, and movement. It is time to harness the power of digital to capture these clinically meaningful measures that are critical to determining efficacy of any AD treatment. But to do so will require the same scientific creativity that has led to breakthroughs in AD imaging, CSF and plasma biomarkers of A/T and apply them to the development and validation of the digital "N".

Funding

This work was informed by research supported by the American Heart Association (20SFRN35360180), the National Institute on Aging (2P30AG013846, AG062109, AG068753, AG072654, the Alzheimer's Drug Discovery Foundation (201902-2017835), the TMCity Foundation, and the Davos Alzheimer's Collaborative.

Consent statement

Consent was not necessary for the purposes of this perspective piece.

CRediT authorship contribution statement

Rhoda Au: Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Zachary Popp:** Writing – review & editing, Writing – original draft. **Spencer Low:** Writing – review & editing, Writing – original draft. **Nicholas J. Ashton:** Writing – review & editing, Writing – original draft. **Henrik Zetterberg:** Writing – review & editing, Writing – original draft.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rhoda Au reports a relationship with Novo Nordisk Inc that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. Rhoda Au reports a relationship with Signant Health that includes: consulting or advisory. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Vaibhav Narayan, Ph.D., who first conceived the idea of an A/T/digital N and graciously let us go forward with his idea.

References

[1] Roche Alzheimer's disease cerebrospinal fluid (CSF) assays receive FDA clearance, supporting more accurate and timely diagnosis. | 2022| https://diagnostics.roche.com/us/en/news-listing/2022/roche-alzheimers-disease-cerebrospinal-fluid-assays-receive-fda-clearance.html.

[2] Quanterix's Alzheimer's blood test designated a breakthrough device | 2021-10-15 | FDAnews. https://www.fdanews.com/articles/204866-quanterixs-alzheimers-blood-test-designated-a-breakthrough-device.

[3] FDA clears first blood test used in diagnosing Alzheimer's disease | 2025 |. https://www.fda.gov/news-events/press-announcements/fda-clears-first-blood-test-used-diagnosing-alzheimers-disease.

[4] Ashton NJ, et al. Diagnostic accuracy of the plasma ALZpath pTau217 immunoassay to identify Alzheimer's disease pathology. medRxiv 2023. https://doi.org/10.1101/2023.07.11.23292493.

[5] Pontecorvo MJ, et al. Association of Donanemab treatment with exploratory plasma biomarkers in early symptomatic Alzheimer disease: a secondary analysis of the TRAILBLAZER-ALZ randomized clinical trial. JAMA Neurol 2022;79:1250–9.

[6] Simrén J, Ashton NJ, Blennow K, Zetterberg H. Blood neurofilament light in remote settings: alternative protocols to support sample collection in challenging pre-analytical conditions. Alzheimers Dement (Amst) 2021;13.

[7] Alcolea D, et al. Use of plasma biomarkers for AT(N) classification of neurodegenerative dementias. J Neurol Neurosurg Psychiatry 2021;92:1206–14.

[8] Lewczuk P, et al. Plasma neurofilament light as a potential biomarker of neurodegeneration in Alzheimer's disease. Alzheimers Res Ther 2018;10:1–10.

[9] Ashton NJ, et al. A multicentre validation study of the diagnostic value of plasma neurofilament light. Nat Commun 2021;12.

[10] Kawas CH, Corrada MM, Whitmer RA. Diversity and disparities in dementia diagnosis and care: a challenge for all of us. JAMA Neurol 2021;78:650–2.

[11] Khachaturian ZS, Khachaturian AS, Thies W. The draft "National Plan" to address Alzheimer's disease - National Alzheimer's Project Act (NAPA). Alzheimer's Dement. 2012;8:234–6.

[12] Cummings J, et al. Aducanumab: appropriate use recommendations. J Prev Alzheimers Dis 2021;8:398.

[13] van Dyck CH, et al. Lecanemab in early Alzheimer's Disease. N Engl J Med 2022. https://doi.org/10.1056/NEJMOA2212948/SUPPL_FILE/NEJMOA2212948_DATA-SHARING.PDF.

[14] FDA approves treatment for adults with Alzheimer's disease | 2025 | https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-treatment-adults-alzheimers-disease.

[15] Carandini T, et al. Testing the 2018 NIA-AA research framework in a retrospective large cohort of patients with cognitive impairment: from biological biomarkers to clinical syndromes. Alzheimers Res Ther 2019;11.

[16] Beker N, et al. Association of cognitive function trajectories in centenarians with postmortem neuropathology, physical health, and other risk factors for cognitive decline. JAMA Netw Open 2021;4. e2031654–e2031654.

[17] Wilkins CH, et al. Racial and ethnic differences in amyloid PET positivity in individuals with mild cognitive impairment or dementia: a secondary analysis of the imaging dementia–Evidence for amyloid scanning (IDEAS) cohort study. JAMA Neurol 2022;79:1139–47.

[18] Grøntvedt GR, et al. The amyloid, tau, and neurodegeneration (A/T/N) classification applied to a clinical research cohort with long-term follow-up. J. Alzheimer's Dis. 2020;74:829–37.

[19] Busche MA, Hyman BT. Synergy between amyloid-β and tau in Alzheimer's disease. Nat. Neurosci. 2020 23:10 2020;23:1183–93.

[20] Illán-Gala I, et al. Challenges associated with biomarker-based classification systems for Alzheimer's disease. Alzheimer's Dement.: Diagn. Assess. Dis. Monit. 2018;10:346–57.

[21] Seino Y, et al. Cerebrospinal fluid and plasma biomarkers in neurodegenerative diseases. J. Alzheimer's Dis. 2019;68:395–404.

[22] Moscoso A, et al. Longitudinal associations of blood phosphorylated Tau181 and neurofilament light chain with neurodegeneration in Alzheimer disease. JAMA Neurol 2021;78:396–406.

[23] Barro C, Zetterberg H. Neurological symptoms and blood neurofilament light levels. Acta Neurol Scand 2021;144:13–20.

[24] Ashton NJ, et al. Alzheimer Disease blood biomarkers in patients with out-of-hospital cardiac arrest. JAMA Neurol 2023;80:388–96.

[25] Yuan A, Nixon RA. Neurofilament proteins as biomarkers to monitor neurological diseases and the efficacy of therapies. Front Neurosci 2021;15:1242.

[26] Manouchehrinia A, et al. Confounding effect of blood volume and body mass index on blood neurofilament light chain levels. Ann Clin Transl Neurol 2020;7:139.

[27] Akamine S, et al. Renal function is associated with blood neurofilament light chain level in older adults. Sci Rep 2020;10.

[28] Aduhelm. Biogen abandons Alzheimer's drug after controversial approval left it unfunded by Medicare. BMJ 2024;384:q281.

[29] Budd Haeberlein S, et al. Two randomized phase 3 studies of Aducanumab in early Alzheimer's disease. J. Prev. Alzheimer's Dis. 2022;9:197–210.

[30] Espay A.J., Kepp K.P., Herrup K. Lecanemab and Donanemab as therapies for Alzheimer's Disease: an illustrated perspective on the data. ENeuro. 2024 Jul 1;11 (7):ENEURO.0319-23.2024. doi: 10.1523/ENEURO.0319-23.2024. PMID: 38951040; PMCID: PMC11218032.

[31] Bullain S, Doody R. What works and what does not work in Alzheimer's disease? From interventions on risk factors to anti-amyloid trials. J Neurochem 2020;155: 120–36.

[32] Kueper JK, Speechley M, Montero-Odasso M. The Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog): modifications and responsiveness in&nbsp; pre-dementia populations. A narrative review. J Alzheimer's Dis 2018;63:423–44.

[33] Du B, et al. Psychometric properties of outcome measures in non-pharmacological interventions of persons with dementia in low-and middle-income countries: a systematic review. Psychogeriatrics 2021;vol. 21:220–38. https://doi.org/10.1111/psyg.12647. Preprint at.

[34] Mohs RC, et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease assessment scale that broaden its scope. Undefined 1997;11:13–21.

[35] Weiner MW, et al. Recent publications from the Alzheimer's Disease Neuroimaging Initiative: reviewing progress toward improved AD clinical trials. Alzheimers Dement 2017;13:e1–85.

[36] Wang J, et al. ADCOMS: a composite clinical outcome for prodromal Alzheimer's disease trials. J Neurol Neurosurg Psychiatry 2016;87:993–9.

[37] Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry 1984;141.

[38] Folstein MF, Folstein SE, McHugh PR. Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12: 189–98.

[39] Morris JC, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology 1989;39:1159–65.

[40] Doherty T, et al. Increasing the Cognitive screening efficiency of global phase III trials in early Alzheimer disease: the cognitive task force. Alzheimer Dis Assoc Disord 2022;36:185–91.

[41] Khan G, Mirza N, Waheed W. Developing guidelines for the translation and cultural adaptation of the Montreal Cognitive Assessment: scoping review and qualitative synthesis. BJPsych Open 2022;8:e21.

[42] Onnela J-P. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 2021. https://doi.org/10.1038/s41386-020-0771-3.

[43] Cohen AS, Cox CR, Masucci MD, et al. Digital phenotyping using multimodal data. Curr Behav Neurosci Rep 2020;7:212–20. https://doi.org/10.1007/s40473-020-00215-4.

[44] Kaye JA, et al. Intelligent systems for assessing aging changes: home-based, unobtrusive, and continuous assessment of aging. *J Gerontol B Psychol Sci Soc Sci* 66B 2011:i180.

[45] Seelye A, et al. Weekly observations of online survey metadata obtained through home computer use allow for detection of changes in everyday cognition before transition to mild cognitive impairment. Alzheimer's Dement. 2018;14:187–94.

[46] Silbert LC, et al. Less daily computer use is related to smaller hippocampal volumes in cognitively intact elderly. J Alzheimers Dis 2016;52:713–7.

[47] Kaye J, et al. P4-348: social biomarkers for early signs of dementia: increased spoken word counts among older adults with Mild cognitive Impairment (MCI). Alzheimer's Dement. 2014;10.

[48] Chin AL, Negash S, Hamilton R. Diversity and disparity in dementia: the impact of ethnoracial differences in Alzheimer disease. Alzheimer Dis Assoc Disord 2011;vol. 25:187–95. https://doi.org/10.1097/WAD.0b013e318211c6c9. Preprint at.

[49] What is digital health | 2020 |. https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health.

[50] Zwack CC, Haghani M, Hollings M, et al. The evolution of digital health technologies in cardiovascular disease research. NPJ Digit, Med 2023;6:1. https://doi.org/10.1038/s41746-022-00734-2.

[51] Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM, Rajmane A, Cheung L, Hung G, Lee J, Kowey P, Talati N, Nag D, Gummidipundi SE, Beatty A, Hills MT, Desai S, Granger CB, Desai M, Turakhia MP. Apple heart study investigators. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 2019 Nov 14;381(20): 1909–17. https://doi.org/10.1056/NEJMoa1901183. PMID: 31722151; PMCID: PMC8112605.

[52] Popp Z, Low S, Igwe A, Rahman MS, Kim M, Khan R, Oh E, Kumar A, De Anda-Duran I, Ding H, Hwang PH, Sunderaraman P, Shih LC, Lin H, Kolachalama VB, Au R. Shifting from active to passive monitoring of Alzheimer disease: the State of the research. J Am Heart Assoc 2024 Jan 16;13(2):e031247. https://doi.org/10.1161/JAHA.123.031247. Epub 2024 Jan 16. PMID: 38226518; PMCID: PMC10926806.

[53] McCool J, Dobson R, Whittaker R, Paton C. Mobile health (mHealth) in low- and middle-income countries. Annu Rev Public Health 2022;vol. 43:525–39. https://doi.org/10.1146/annurev-publhealth-052620-093850. Preprint at.

[54] Landers M, Dorsey R, Saria S. Digital endpoints: definition, benefits, and current barriers in accelerating development and adoption. Digit Biomark 2021;vol. 5: 216–23. https://doi.org/10.1159/000517885. Preprint at.

[55] Au R, Kolachalama VB, Paschalidis IC. Redefining and validating digital biomarkers as fluid, dynamic multi-dimensional digital signal patterns. Front Digit Health 2022;3.

[56] Cummings JL, Goldman DP, Simmons-Stern NR, Ponton E. The costs of developing treatments for Alzheimer's disease: a retrospective exploration. Alzheimers Dement 2022 Mar;18(3):469–77. https://doi.org/10.1002/alz.12450. Epub 2021 Sep 28. PMID: 34581499; PMCID: PMC8940715.

[57] Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. Am J Public Health 1998;88:1337–42.

[58] Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. NPJ Digit Med 2019;2.

The Journal of Prevention of Alzheimer's Disease 13 (2026) 100417

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

The evolution of Alzheimer's target identification: Towards a fusion of artificial and cellular intelligence

Gayle Wittenberg[a], Fiona Elwood[b], Andrea Houghton[c], Tommaso Mansi[d], Bart Smets[e], Simon Lovestone[f,*]

[a] Neuroscience R&D, Johnson & Johnson Innovative Medicine, Titusville, NJ, USA
[b] Neuroscience R&D, Johnson & Johnson Innovative Medicine, Cambridge, MA, USA
[c] Neuroscience R&D, Johnson & Johnson Innovative Medicine, Spring House, PA, USA
[d] Data Science & Digital Health R&D, Johnson & Johnson Innovative Medicine, Titusville, NJ, USA
[e] Data Science & Digital Health R&D, Johnson & Johnson Innovative Medicine, Beerse, Belgium
[f] Neuroscience R&D, Johnson & Johnson Innovative Medicine, London, United Kingdom

ARTICLE INFO

Keywords:
Target identification
Artificial intelligence
Big Data
iPSCs

ABSTRACT

Decades of advances unfolding in parallel across diverse domains have delivered to science rapid rises in the scale of multiplexing, population-level cohort sizes, global computational capacity, massive-scale artificial intelligence (AI) models, and advanced human cellular modeling capabilities. These have generated unprecedented volumes of data, allowing researchers to explore Alzheimer's disease (AD) biology at a depth and scale never before possible. The explosion of multi-omics datasets and computational power heralds an era in which the complexity of AD can be meaningfully dissected and reconstructed leveraging AI. These can be applied to advance our understanding of the root causes of disease, fundamentally a forward problem, tracing how dysfunction emergence from interactions across genes, cells and environments over time. On the other hand, therapeutic discovery requires addressing the inverse problem, working back from the diseased state to pinpoint upstream interventions that restore health. Human induced pluripotent stem cells (iPSCs) and other human cell models play a pivotal role in this process, naturally computing the mapping from perturbation to phenotype at scale. By recreating human-relevant biology, this cellular intelligence enables validation of targets predicted by AI and testing of interventions that drive therapeutic progress. We look to the next horizon in Alzheimer's research as a collaboration, a convergence of three forms of intelligence: human, artificial and cellular. In unison, these complementary forces will shape a new frontier for AD research where scientific innovation and human ingenuity work together bringing hope for meaningful advances and new therapies.

1. Introduction

Biological systems are complex systems. Disease represents maladaptive perturbation to this system arising from genetic or environmental hits over time. Therapeutics seek to restore function by strategically modulating, typically, one or a small number of biological targets. Target identification is the nomination of such targets for therapeutic intervention; target validation is confirmation, through orthogonal approaches, that perturbation of the target restores the desired function. Together, a case could be made for target identification and target validation (TiTv) being the most critical hurdle in drug discovery.

Understanding the cause of disease, and how it emerges, is a *forward problem* - with an aim to understand the emergence of dysfunction from genetic variants or environmental exposures across the complexity of the 30 trillion cells in the human body, 20,000 genes, mRNAs, proteins, miRNAs, epigenetics, post-translational modifications, autoantibodies, and pathogen exposures. Target identification, by contrast, is best framed as an *inverse problem*: starting from an undesired system state - disease - we identify an upstream intervention, often involving one protein or pathway, that will restore healthy function (Fig. 1). Target identification demands not only biological insight but also a shift in perspective. This process aims to pinpoint molecular entities whose modulation can alter disease progression, providing a rational basis for

* Corresponding author.
E-mail address: SLovesto@its.jnj.com (S. Lovestone).

https://doi.org/10.1016/j.tjpad.2025.100417
Received 5 September 2025; Received in revised form 14 October 2025; Accepted 20 October 2025

therapeutic intervention.

Historically, target identification relied on reductionism: isolating pathways, studying disease models, and gradually triangulating typically on a protein or gene believed to be central to pathology. Validation of a target could involve years of experimental work leveraging genetic knockouts, tool compounds, and animal models built around a carefully constructed mechanistic hypothesis. This approach has led to many notable successes in drug discovery and historically has been the driver of most therapeutic programs for Alzheimer's disease (AD) in the clinic and in clinical development today.

The emergence of big data and AI is enabling a new paradigm. With genome-wide association studies, whole exome and genome sequencing, and multi-omics data sets including emerging vast proteomics cohorts, the data landscape available for learning has exploded. This data explosion coupled with recent advances in artificial intelligence, such as generative AI, makes it possible to explore disease biology at unprecedented scale, depth and speed. These massive datasets and AI tools have broadened our view, enabling us to identify new targets, cluster patients by molecular subtype or project along different mechanistic dimensions of disease and uncover latent patterns driving progression.

In the face of optimism about these advances in technology, data, and AI, we must remember that these are tools to be applied judiciously. As our methods grow more powerful and our data more complex, it becomes even more critical to sharpen the questions we ask. Big data and AI are often used to search for biological causes of disease. It is easily assumed, without careful framing, that a causal gene will be a viable target. Dennis Noble reminds us that the concept of 'relativity' applies to biology, not just to physics [1]: in the reference frame of the drug developer, we are not searching for cause, but for cure – even if at times these may align. Below we outline the journey of TiTv in Alzheimer's disease, from empirical biochemistry, molecular and cellular biology through Big Data analytics to the potential of AI. We conclude that the potential of AI is enormous, but the role of the scientist remains paramount

2. TiTv before the age of 'big data'

Three key questions lie at the heart of selecting a novel target to treat disease: (1) which molecular target to modulate, (2) which therapeutic modality can be developed that best delivers the desired modulation of that target, and (3) which patients will benefit? While all three questions are important, the success of any discovery campaign relies on defining a strong link between a proposed target and disease. Before the advances in high-throughput omics and computational analytics, target identification relied primarily on a combination of human pathology, biochemical pathway elucidation, animal models, and genetic insights from rare familial disorders. These formed the bedrock of many successful drug development efforts.

Discovery and development of therapies typically have an increased chance of success when the biology translates well between preclinical species and man as it enables the drug discovery campaign significantly. In Alzheimer's disease (AD), translation between preclinical species and man is particularly challenging, given preclinical species do not succumb to the same pathophysiology. Therefore, it is unsurprising the earliest approved treatments modulated pharmacology known to be perturbed in disease: namely, acetylcholinesterase (AChE) inhibitors such as donepezil [2]. These treatments emerged from the observations of cholinergic deficits in the brains of patients with AD. Postmortem studies in the 1970s and 80 s revealed degeneration of cholinergic neurons in the basal forebrain, leading researchers to test therapeutics that could enhance acetylcholine signaling (see [3]). Furthermore, modulating acetyl choline transmission in rodents can improve cognition, building confidence in the approach, and enabling drug discovery programs. More specifically, the preclinical development of donepezil relied on *in vitro* assays - a rat brain homogenate assay of inhibition of AChE, a rat plasma assay of butyl cholinesterase – and *in vivo* assays of target modulation i.e. demonstration that donepezil inhibition of AChE in aged rat brains resulted in improved learning in rats. Much of this work also provided the basis for clinical development as the assays were translatable. For example, dose selection during clinical development was based on AChE inhibition in red blood cell membranes and plasma [4]. These treatments were successfully developed without molecular biomarkers or genetic stratification. This was possible because clinical trial duration for a symptomatic individual is relatively short (12 weeks), a sufficient percentage of the population responds to treatment, adverse events are monitorable, and patients with dementia of various

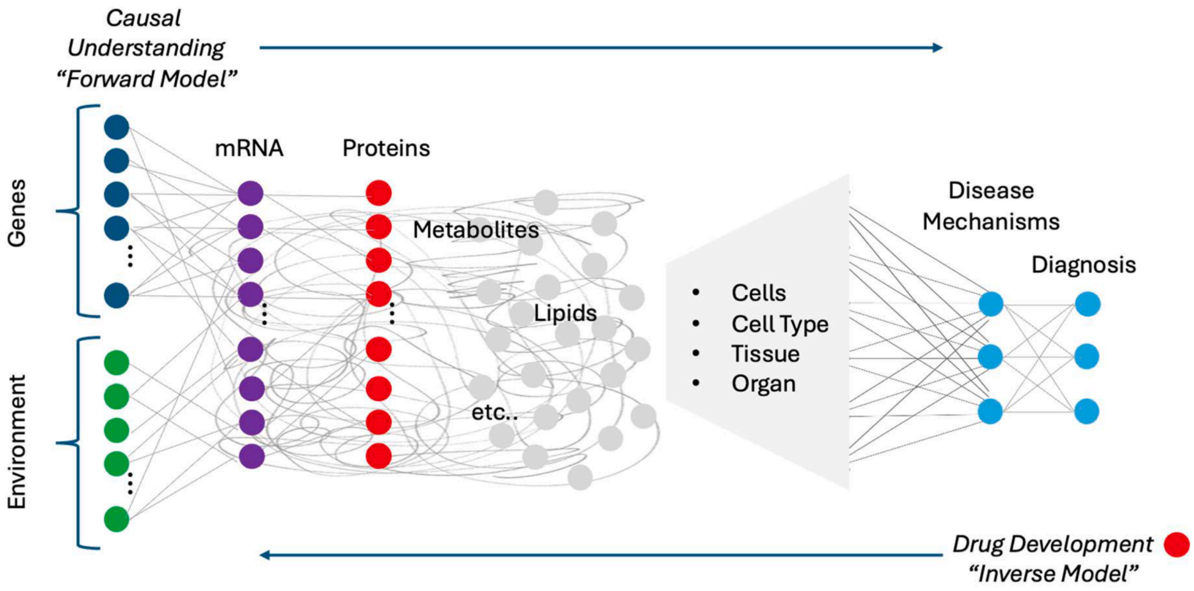


Fig. 1. Conceptual model of disease and treatment. Causal understanding of disease consists of tracing the biological cascade from genes and environmental inputs through the successive molecular, cellular and tissue-level networks towards disease mechanisms, and eventually clinical manifestations (e.g., ICD codes). This "forward model" reflects causal biological understanding – how perturbations in genes or environment drive disease. Drug development typically begins with the clinical labeling of disease and works backward, seeking to identify a target, often a protein, to manipulate to reverse or modulate disease associated signatures and disease.

etiologies respond to treatment with anti-AChE agents.

However, to move beyond symptomatic treatments, an understanding of disease pathology and underlying pathophysiology is required. The discovery that amyloid beta (Aß) was an integral part of plaques in postmortem brains from patients with AD in the mid 1980s [5] combined with the discovery that mutations in amyloid precursor protein (APP), presenilin 1 (PSEN1) and PSEN2 genes were associated with early onset familial Alzheimer's Disease (FAD), enabled scientists to focus on disease pathology in AD drug discovery programs. Importantly for drug discovery, knowledge of these genetic mutations enabled the development of preclinical models such as the Tg2576 mouse. This model has a Swedish FAD mutation (K670N/M671Lz), and the expression of the human APP is five-fold above the levels of endogenous APP; additionally, the expression of Aß1–40 and Aß1–42, and amyloid deposition increases with age, along with gliosis and dystrophic neuritis. Amyloid plaques appear in mice between 11 and 13 months of age. Furthermore, these mice also display spatial memory impairment by 9–10 months of age [6]. Having such models fueled investment in treatments for AD as there was viable platform to profile candidate molecules.

These models successfully recapitulated some pathological features but often failed to predict clinical efficacy in humans, contributing to the high attrition rate in Alzheimer's trials. For example, ß secretase (BACE) inhibitors were predicted to be efficacious based on transgenic mouse model data but failed in the clinic even though target modulation was demonstrated in preclinical species and in man (see Neuman et al., 2019 [7]). This is because these models do not recapitulate disease; rather, they model but one aspect of the biology. For AD research, most *in vivo* models are a model of amyloid pathology or a model of tau pathology. It is still unclear how to develop a mouse model in which these pathologies co-exist in a manner that more closely recapitulates the human disease state.

Thus, in recent years target identification and validation for drug discovery in AD has pivoted to a greater reliance on human data collected from living patients at different stages of disease. This has informed our understanding of disease evolution, led to new hypotheses to test (e.g. tau seeding) and provided data which we can exploit as we endeavor to build more human relevant models. Induced pluripotent stem cell (iPSC) technology has enabled the creation of human-based models for Alzheimer's disease (AD) drug discovery. These models, derived from patient cells, allow researchers to study disease mechanisms and test potential drug candidates in a human context, complementing traditional animal models. Human-induced pluripotent stem cell (hiPSC) models range substantially in complexity, from two-dimensional monolayers to three-dimensional organoids and from single purified cell types (e.g. neurons) to complex cultures with a mixture of cell types (e.g. neurons, microglia and astrocytes). Neurons can be derived from iPSCs of healthy control or AD-patient populations, and gene-editing tools such as CRISPR/Cas9 can be used to generate isogenic cell lines for the study of genetic variants associated with AD [8–10].

Researchers can leverage panels of hiPSCs derived from genetically diverse donors to study how genetic variants influence molecular and cellular phenotypes in multiple genetic backgrounds. These cellular-level association studies (e.g., eQTL or chromatin QTL mapping) complement large scale GWAS by helping to identify which variants have functional consequences in human cells. Epigenetic profiles can be generated for different cell types with single-cell resolution (e.g. single-cell ATAC-seq [11]). Sequencing can identify genes of relatively small effect size, provided the pools of *in vitro* samples are large enough. Emerging optical tools may further support pooled genetic screens with imaging and morphology-based phenotypes. When combined with CRISPR-based perturbation screens, these approaches can identify modifiers of disease-relevant cellular phenotypes in scalable target discovery efforts, while also providing compelling evidence to validate new targets in human cells.

Often, human biology used to identify and validate targets as well as enable drug discovery efforts can also be used as a translatable endpoint for clinical development. For example, CSF and plasma measures of phosphorylated tau have been used to monitor target engagement in anti-tau trials. Amyloid PET imaging has become a *de facto* standard for validating anti-amyloid therapies. Still, the disconnect between biomarker improvement and clinical benefit remains a central challenge in Alzheimer's drug development, underscoring the need for early validation of mechanistic links between target engagement and disease modification.

3. TiTv in the era of 'big data'

The advent of large-scale molecular datasets from human samples has transformed the landscape of target identification and validation (Fig. 2). Rather than relying largely on pre-defined hypotheses generated through painstaking empirical studies as described above, researchers can now interrogate high-dimensional data from genetics, transcriptomics, epigenomics, proteomics and other modalities to uncover novel putative targets. These data support a fundamentally different mode of discovery that is network-informed, context-sensitive, and population-aware, rather than linear or pathway-bound. It is also more successful. A target supported by genetic data is over twice as likely to yield a successful drug discovery programme, and a target supported by single cell sequencing expression data is similarly a better bet for success [12,13].

Nowhere has this shift been more profound than in neuroscience, where decades of research often failed to translate due to poor models and limited access to human brain tissue. Genome-wide association studies (GWAS) involving hundreds of thousands of individuals have now identified robust, replicable genetic risk loci across a range of neuropsychiatric and neurodegenerative disorders. Large-scale GWAS studies of late onset Alzheimer's disease have identified variants associated with altered risk near APOE, BIN1, CLU, and PICALM genes [14–16]. Beyond confirming the central role of amyloid processing, these studies highlighted the role of specific biological processes, such as neuroinflammation, lipid biosynthesis, and endocytosis in AD.

A primary challenge with GWAS data is translating statistical associations into a mechanistic understanding of disease and, ultimately, into viable drug targets. Many disease-associated single nucleotide polymorphisms (SNPs) reside in non-coding regions of the genome, making their functional consequences difficult to decipher [17,18].

To bridge this gap, GWAS data are being integrated with other omics layers to prioritize causal genes and pathways. A key strategy is the use of expression and protein quantitative trait loci (eQTLs and pQTLs), which link genetic variants to changes in gene expression and protein abundance, respectively. By co-localizing a GWAS risk signal with an eQTL or pQTL, researchers can form a hypothesis that the genetic variant influences disease risk by altering the expression of a specific gene or protein [19–21].

Statistical methods such as Mendelian randomization (MR) allow for more formal tests of causality. MR uses genetic variants as unconfounded proxies for an exposure (e.g., the level of a specific protein) to infer its causal effect on a disease outcome. Other powerful approaches include transcriptome-wide and proteome-wide association studies (TWAS/PWAS), which integrate GWAS summary statistics with gene expression or protein level data to identify genes whose expression levels are associated with disease risk.

The integration of eQTLs and pQTLs, derived from large-scale multi-tissue transcriptomics and proteomics datasets, with GWAS data using these approaches led to the prioritization of many genes and proteins causally implicated in AD and other neurological disorders [22–27]. These analyses are being further revolutionized by single-cell technologies, which allow for the investigation of the cell-type-specific effects of disease-associated variants [28].

These multimodal analyses are important because although the genome provides a static blueprint of inherited disease risk, it does not capture the dynamic changes that occur as a disease develops and

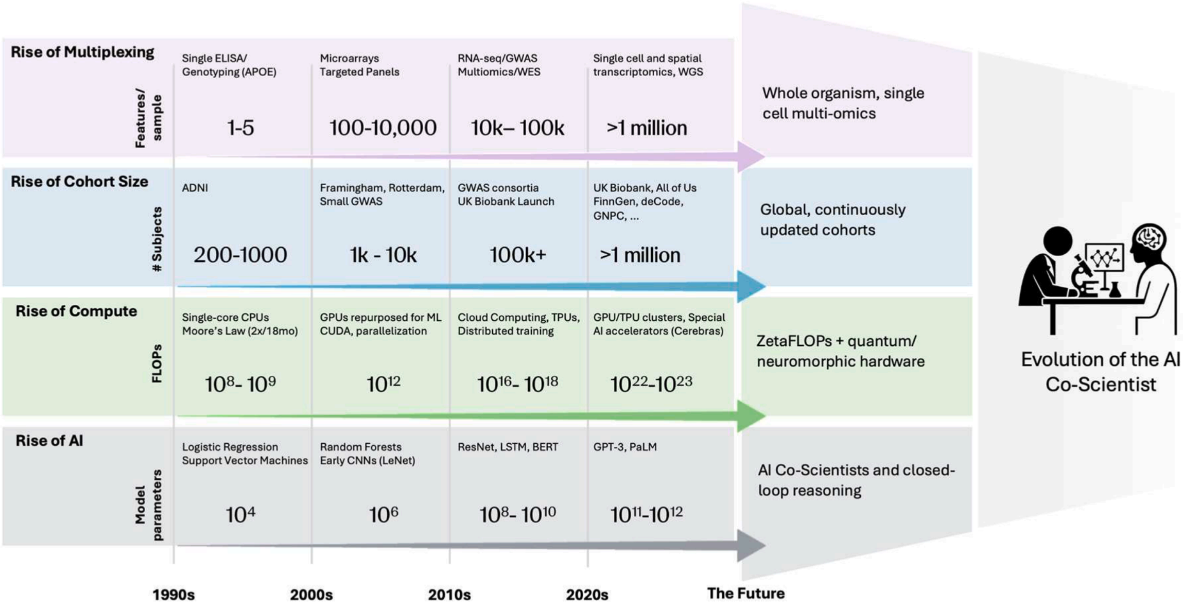

Rise of Multiplexing
Features/ sample
Single ELISA/ Genotyping (APOE)
1-5
Microarrays Targeted Panels
100-10,000
RNA-seq/GWAS Multiomics/WES
10k– 100k
Single cell and spatial transcriptomics, WGS
>1 million
Whole organism, single cell multi-omics
Rise of Cohort Size
# Subjects
ADNI
200-1000
Framingham, Rotterdam, Small GWAS
1k - 10k
GWAS consortia UK Biobank Launch
100k+
UK Biobank, All of Us FinnGen, deCode, GNPC, ...
>1 million
Global, continuously updated cohorts
Rise of Compute
FLOPs
Single-core CPUs Moore's Law (2x/18mo)
10^8- 10^9
GPUs repurposed for ML CUDA, parallelization
10^{12}
Cloud Computing, TPUs, Distributed training
10^{16}- 10^{18}
GPU/TPU clusters, Special AI accelerators (Cerebras)
10^{22}-10^{23}
ZetaFLOPs + quantum/ neuromorphic hardware
Rise of AI
Model parameters
Logistic Regression Support Vector Machines
10^4
Random Forests Early CNNs (LeNet)
10^6
ResNet, LSTM, BERT
10^8- 10^{10}
GPT-3, PaLM
10^{11}-10^{12}
AI Co-Scientists and closed-loop reasoning
Evolution of the AI Co-Scientist
1990s
2000s
2010s
2020s
The Future

Fig. 2. Trends Enabling AI-Driven TiTv and the Evolution of the AI Co-Scientist. This figure illustrates four parallel trends that are transforming precision neuroscience and biomedical discovery. (1) The rise of multiplexing (purple): from single-analyte assays to massively multiplexed, spatially resolved single-cell multi-omics. (2) The rise of cohort size (blue): from hundreds of individuals in early disease studies to global biobanks with > 1 million participants, moving towards continuously updated, real-world integrated cohorts bringing together medical records, clinical, biomarker and digital data. (3) The rise of compute (green): from single-core CPUs and Moore's Law to cloud-scale GPU/TPU clusters and specialized accelerators, with future prospects in zetaFLOPs and neuromorphic or quantum computing. (4) The rise of AI (grey): from classical machine learning to deep learning and foundational models, advancing towards agentic AI systems capable of reasoning, planning, and closed-loop experimentation. Together, these converging trajectories point towards the emergence and evolution of the AI Co-Scientist, partnering with human researchers to accelerate targeting discovery.

progresses. Transcriptomics, proteomics, metabolomics, and epigenomics offer a real-time snapshot of the molecular state of a cell or tissue, reflecting the integrated output of genetic risk, environmental influences and co-morbidities. Large-scale, spatiotemporal analyses of bulk brain transcriptomics data, utilizing resources from consortia such as the Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD), have facilitated the identification of co-expressed gene modules linked to AD pathology and cognitive decline [29,30]. These studies have uncovered both brain-region specific expression changes and coordinated alterations across different areas of the brain. While it is challenging to distinguish cause from consequence in transcriptomics data, the enrichment of top modules for AD genetic risk factors and the observation that many transcriptional changes occur early in the disease highlights potential pathways driving disease progression and points to possible targets for therapeutic intervention. Notably, these transcriptional modules are enriched for genes involved in processes such as nervous system development, axon growth, inflammation, and proteostasis.

A further leap in understanding has come from single-cell sequencing. Traditional bulk tissue analysis averages the molecular signals from millions of different cells, obscuring the contributions of rare cell types or specific cell states. Single-cell RNA sequencing (scRNA-seq) of post-mortem human brain tissue has enabled the deconvolution of this complexity, leading to the identification of specific disease-associated cell states, such as activated microglia (DAM), astrocytes (DAA), and inhibitory neuronal subtypes associated with resilience to AD pathology [31–34]. This level of resolution is critical for mapping genetic risk variants to the specific cell types in which they exert their effects and provides highly specific hypotheses for therapeutic intervention.

Recent advances in large-scale proteomics now bring the systems biology perspective of big data to the protein level, where most drug targets reside. High-throughput mass spectrometry and affinity-based platforms enable large-scale profiling of protein expression, post-translational modification and protein-protein interactions across tissues, cell types and disease states. In AD, proteomic analyses across multiple brain regions have uncovered disease-associated changes in protein co-expression networks, some of which were not observed at the RNA level, including a MAPK signaling module associated with cognition and a extracellular matrix proteins that showed a positive correlation with plaques and tangles [35–37].

The latter discovery underscores the complementarity of the different data layers and the critical importance of integrating multiple omics modalities to get the full picture of biology. By combining genomics, epigenomics, transcriptomics, proteomics and metabolomics from the same individuals, researchers are building more complete models of disease [38–40]. This will allow for the construction of a chain of evidence from a genetic risk variant to its functional consequence on the epigenome, transcriptome, proteome, and metabolome, providing a much richer understanding of disease mechanisms and a more solid foundation for target identification.

The immense scale of data required for robust omics-based target discovery necessitates a collaborative, open-science approach. Large consortia and public-private partnerships have been essential in generating and harmonizing the necessary multimodal datasets. Initiatives like the Accelerating Medicines Partnership for Alzheimer's Disease (AMP-AD), the Global Neurodegeneration Proteomics consortium [41], UK Biobank and FinnGen are creating invaluable resources that integrate deep clinical data with multi-omics profiles from thousands of participants. By making these data broadly available, these consortia are accelerating the pace of discovery and empowering researchers globally to identify and validate the next generation of therapeutic targets for Alzheimer's disease (Fig. 2).

Beyond identification of biologically effective targets, big data combined with advanced methods including AI is advancing the early identification of target-specific safety concerns. Human genetics and multi-omics integration now enable systematic prediction of on-target safety liabilities based on gene constraint metrics and natural loss-of-function variation [42,43]. Complementary transcriptomic and single-cell atlases allow tissue-specific expression mapping to anticipate unintended effects in critical organs [44,45]. Moreover, AI-driven systems toxicology approaches are beginning to model cross-pathway perturbations and highlight mechanisms of target-associated risk earlier in discovery [46]. Collectively, modern data science pipelines are reshaping how safety evaluation is incorporated into target prioritization.

Following the identification and validation of a target, opportunities need to be prioritized based on druggability and the latest understanding of the therapeutic modalities available to the scientist. Quantitative assessments of tractability now integrate structural features, ligandability scores, and prior success across protein families [47,48]. Different modalities will inherently bring different benefits and liabilities, and these need to be matched to the profile of the target and needs of the patient [49,50]. An integrated view of the end-to-end drug discovery and development process, in addition to an intimate understanding of the patient's experience, can lead to improved decision-making even at the earliest stages of target selection.

Altogether, this era of big data has not only enhanced the precision of target discovery but also expanded its scope, bringing molecular insights with disease phenotype, stages and therapeutic response to drive forward more personalized and effective drug development strategies.

4. Analysis before the age of AI

Before big data entered the scene, AI remained underdeveloped. Data analysis in target discovery was primarily rooted in classical statistical models, expert intuition and manual exploration. Linear regression, logistic models, survival analysis, ANOVA and PCA formed the core toolkit. In Alzheimer's disease, early biomarker studies applied these methods to investigate cerebrospinal fluid (CSF) levels of Aβ42, tau and phosphorylated tau. These methods prioritized transparency, interpretability, and hypothesis-driven reasoning; this produced many durable insights.

Researchers manually explored gene lists, drew networks, annotated pathways, and traced findings back to curated knowledge using tools like Gene Ontologies [51,52], KEGG [53], and Ingenuity Pathway Analysis [54]. These approaches lent interpretability and structure to transcriptomic studies in AD, for example, by highlighting disruptions in mitochondrial function, immune signaling or synaptic plasticity in postmortem brain tissue.

The rise of omics datasets exposed the limitations of these approaches. One of the most pervasive challenges was the 'curse of dimensionality': the number of features (e.g. genes, transcripts, proteins) far exceeded the number of samples, creating sparsity and instability in traditional statistical frameworks. This was especially stark in AD, where precious human biosamples were limited in availability and heterogeneity across individuals was high. In high-dimensional, low sample-size (HDLSS) regimes, traditional statistics became fragile – models overfit, false positives abounded, and biologically relevant signals could be easily missed [55].

To cope with complexity, analysts extended classical approaches using network models, weighted gene co-expression networks (WGCNA) [56], Bayesian frameworks [57], and dimensionality reduction techniques [58]. These methods edged closer to machine learning but remained grounded in static, human-interpretable paradigms. In AD, such methods helped uncover co-regulated gene modules associated with neuroinflammation and amyloid pathology. While useful, these approaches struggled to Nscale across conditions, datasets and modalities, often failing to capture nonlinear relationships, context-specific interactions or multivariate patterns that characterize neurodegenerative disease processes. This was particularly true in transcriptomics and proteomics, where multiple testing correction and arbitrary thresholds could obscure emergent biological signals.

Compounding these challenges was the heterogeneity of the data itself. Omics datasets are generated from a wide array of platforms and protocols, each with its own formats, distributions and artifacts. Batch effects, systemic, non-biological variations introduced during sample processing or sequencing, frequently masqueraded as biological signal. A lack of standardized pipelines and shared frameworks made it difficult to combine insights across studies or derive generalizable biological understanding.

Even when technical hurdles were overcome, interpretation often stopped short of biological meaning. Observational omics data could highlight correlations but rarely offered a clear path to causality or therapeutic intervention. The leap from statistical signal to functional relevance required time-intensive experimental validation; many promising findings failed to replicate. As datasets grew larger and more complex, the ability to reason through them manually, or to visualize and explore them directly reached a point beyond the limit of human intelligence, The sheer volume, diversity and complexity of modern datasets pushed conventional methods to their limits. Analyses remained constrained by what we knew how to model, which introduced biases and limited what we could discover. Human reasoning, while essential for grounding interpretation, could not keep pace with the full dimensionality of the data. These challenges set the stage for a transformation – a transition into the age of AI.

5. Analysis in the era of AI

Artificial intelligence (AI) has been undergoing a dramatic acceleration since the "revival" of neural networks and deep learning in the early 2010′s, powered by the exponential growth of data, through internet and compute through the use of GPU for AI (Fig. 2). Launched in the fields of vision, language and audio, AI approaches were quickly translated to the fields of biology and physics, making AI a foundational force in biomedical research. This rise has been propelled by several key enablers: the proliferation of high-throughput, high-dimensional biological data sets described above, advances in machine learning architectures [59], modern training approaches (like self-supervision), and unprecedented access to advanced hardware like GPUs or TPUs through cloud providers. Together, these advances have unlocked the potential to analyze not only large volumes of unstructured data, but also highly heterogeneous and noisy data at scale.

One of the most important paradigm-shifts in AI is the advent of self-supervised learning and resulting foundation models in the late 2010′s [60]. Traditionally, machine learning systems are trained from scratch from well-curated, small databases. Methods were tailored to extract insights from these isolated databases and studies or combined through meta-analyses. Self-supervised learning changed the approach [61]. Enabled by the proliferation of data, deep neural networks are pre-trained on a very large amount of unlabeled, heterogeneous datasets, leading to a so-called foundation model (FM). Intuitively speaking, the models are trained to recover perturbations made to the original data, enabling training at scale. Foundation models are then used as back-bones to more specific, task focused predictive models [62] or generative AI models [63]. By design, a FM thrives on heterogeneity, pulling strength from patterns distributed across vast, diverse corpora. Moreover, FMs enable principled ways of integrating multiple modalities [64], like text and images, learning across modalities and across studies.

For complex diseases like Alzheimer's, which defy simple causal models and often suffer from fragmented evidence across data types and cohorts, this shift in the power of AI, coupled with the release of large cohort data, holds transformational potential. For example, foundational models trained on single-cell transcriptomic data across development and disease stages have revealed cell-state transition in microglia, astrocytes, and neurons that mark early divergence from healthy aging [65,66]. These insights, elusive in smaller datasets, suggest new axes of stratification and windows for therapeutic intervention.

FM are able to integrate data across omics modalities and generate latent representations of biological structure that have the potential to capture the relationships between genes, proteins, cell types, pathways and disease states [67,68]. These methods enable not only higher accuracy, but a different kind of *reasoning* that is less dependent on predefined hypotheses and more able to surface unexpected emergent patterns. For instance, genetic variation, transcript abundance, protein levels, post-translational modifications, and epigenetic markers can be analyzed jointly, creating a layered understanding of disease, in a hypothesis-generating setting and purely driven by data. This multi-modal integration has the potential to enable mechanistic tracing from inherited risk alleles through cellular dysfunction to clinical outcomes, and can illuminate synergistic or antagonistic interactions that would remain invisible if each data type were analyzed in isolation.

When foundation models are used in their generative form, they add an additional capability: the ability to simulate new data instances that reflect learned distribution from the training datasets. In practice, this means AI can generate synthetic omics profiles within the manifold of the disease states observed in the training data, potentially offering new hypotheses about disease progression or subtypes, hypotheses that would still need to be verified experimentally of course. In Alzheimer's research this capacity is particularly valuable in the preclinical space, where early disease signatures are subtle and underpowered in any single dataset. Generative models can also propose novel peptides or molecules optimized for binding to pathologically relevant targets, including tau aggregates or synaptic receptors, accelerating iterative loop of design, synthesis and testing.

It seems likely that these advances in AI will increasingly dominate in our analysis of very large, and very complex datasets such as the multilayered bio-ome, for insights leading to novel TiTv. However, the form of AI that has so stunned the world in the last few years has been large language models (LLMs), and there are many ways in which this form of generative AI will impact on TiTv. Recent developments in LLM architectures, particularly those incorporating test-time compute strategies and reinforcement learning fine-tuning methodologies [69] have demonstrated enhanced reasoning capabilities in what are now characterized as large reasoning models. Emerging systems integrate tools or models together with retrieval-augmented generation frameworks; these systems exhibit capacity for executing complex tasks. Such approaches, often referred to as "AI co-scientists", are being explored to augment biological analysis and TiTv [70–72], with similar platforms designed to accelerate hypothesis generation (Fig. 2). These systems typically employ sophisticated orchestration mechanisms integrating literature retrieval, automated peer review protocols, and ranking systems, collectively contributing to expedited discovery workflows.

While the transformative potential of deep learning and LLMs is evident, their application to Alzheimer's disease and related neurodegenerative conditions remains constrained by data and domain limitations. Model performance scales with data volume, yet biological data sets are historically small, heterogeneous and context-dependent, limiting generalizability [73,74]. Moreover, phenomena such as "hallucinations" or spurious correlations can compromise interpretability and reliability when models are applied to clinical hypotheses [75,76]. Emerging Large Reasoning Models (LRMs) also exhibit inconsistency in producing accurate and coherent reasoning traces, with performance declining as task complexity increases [77,78]. Overcoming these challenges may require hybrid AI frameworks that integrate data-driven learning with causal and symbolic reasoning ground in biological and mechanistic knowledge [79,80].

The primary value proposition of these computational frameworks lies in their potential to not only accelerate discovery, but also to mitigate the cognitive biases and subjective preferences inherent in traditional scientific methodologies, while simultaneously identifying previously unconsidered research directions and providing more comprehensive analytical assessments. Concurrent developments in multi-modal agentic systems have enabled the integration of specialized models, such as large-scale single-cell analysis platforms, which can be dynamically activated by LLM orchestrators to execute targeted computational tasks with automated result interpretation and presentation. Finally, "vibe coding", where users describe what they want in natural language and the system automatically generates appropriate source code, is gaining traction in the world of computational biology, with the promise to bring complex data analysis to the fingertips of biologists, accelerating the hypothesis – experiment – analyze cycle for TiTv.

These tools, taken together, expand not just the scale but the scope of what can be discovered and the ease and accessibility of use. Ultimately, what distinguishes analysis in the age of AI is not only its power or precision, but its ability to tackle more and more complex data and scientific questions. Where traditional methods were optimized for curated datasets and well-defined questions, recent developments in AI are enabling more robust data quantification, accurate prediction, and more sophisticated hypothesis-generating experiments to discover novel biology (Fig. 3). Through the power of foundation models, AI allows weak signals across noisy datasets to reinforce one another. In doing so, it creates new opportunities for inference. This is particularly advantageous in diseases like AD where causality is diffuse, clinical manifestations are delayed, and success has long been hindered by the mismatch between mechanism and measurement.

6. Current limitations of AI and big data

Despite the rapid evolution and promise of AI in biomedical research, significant limitations remain. Most limitations are not rooted in the algorithms themselves, but in the data, systems, and human structures that surround them. These limitations are particularly salient in neuroscience and other complex therapeutic areas where disease heterogeneity, sparse signals and incomplete understanding of biology compound the inherent challenges.

One of the most commonly cited limitations is the quality and nature of the underlying data. Even with vast quantities of omic data, much of it is noisy, incomplete, inconsistently annotated, and riddled with batch effects. In such environments, AI systems are susceptible to the classic problem of 'garbage in, garbage out'. Quantity does not equal quality, and poor-quality data can mislead even the most sophisticated models. These issues are amplified when data are drawn from different platforms, time points or patient populations.

Adding to this challenge is the scarcity of data *for specific problems*. While the overall volume of biological data is increasing, well-annotated, disease-specific, or subgroup-specific datasets remain limited. This is especially true in fields like neurodegeneration, where relevant patient cohorts may be small, diverse and difficult to access – and where longitudinal data, paired with imaging and other biomarkers may be needed to facilitate interpretation, yet take years to generate. AI models trained on narrowly scoped or homogeneous data risk learning features that fail to generalize to broader populations, or worse, perpetuate biases embedded in the training data. This risk is acute when datasets underrepresent certain demographics, genotypes or disease stages raising both scientific and ethical concerns [81–83].

Even when data are sufficient in volume and quality, the models themselves may introduce their own limitations. Many of the most powerful AI systems, deep neural networks and large foundational models, operate as black boxes. They generate highly accurate predictions, but the logic behind those predictions often remains opaque. This lack of interpretability poses a challenge for clinicians and regulators in high-stakes domains like drug development where decisions must be both evidence-based and explainable [84]. Without understanding *why* a model has arrived at a given conclusion, it becomes difficult to validate or trust its output, particularly when it is used to justify experimental therapeutics or diagnostic decisions.

This limits biological insight. In drug discovery, it is often not enough to know that something works. We must understand why it works in order to improve it, avoid side effects and design next-generation

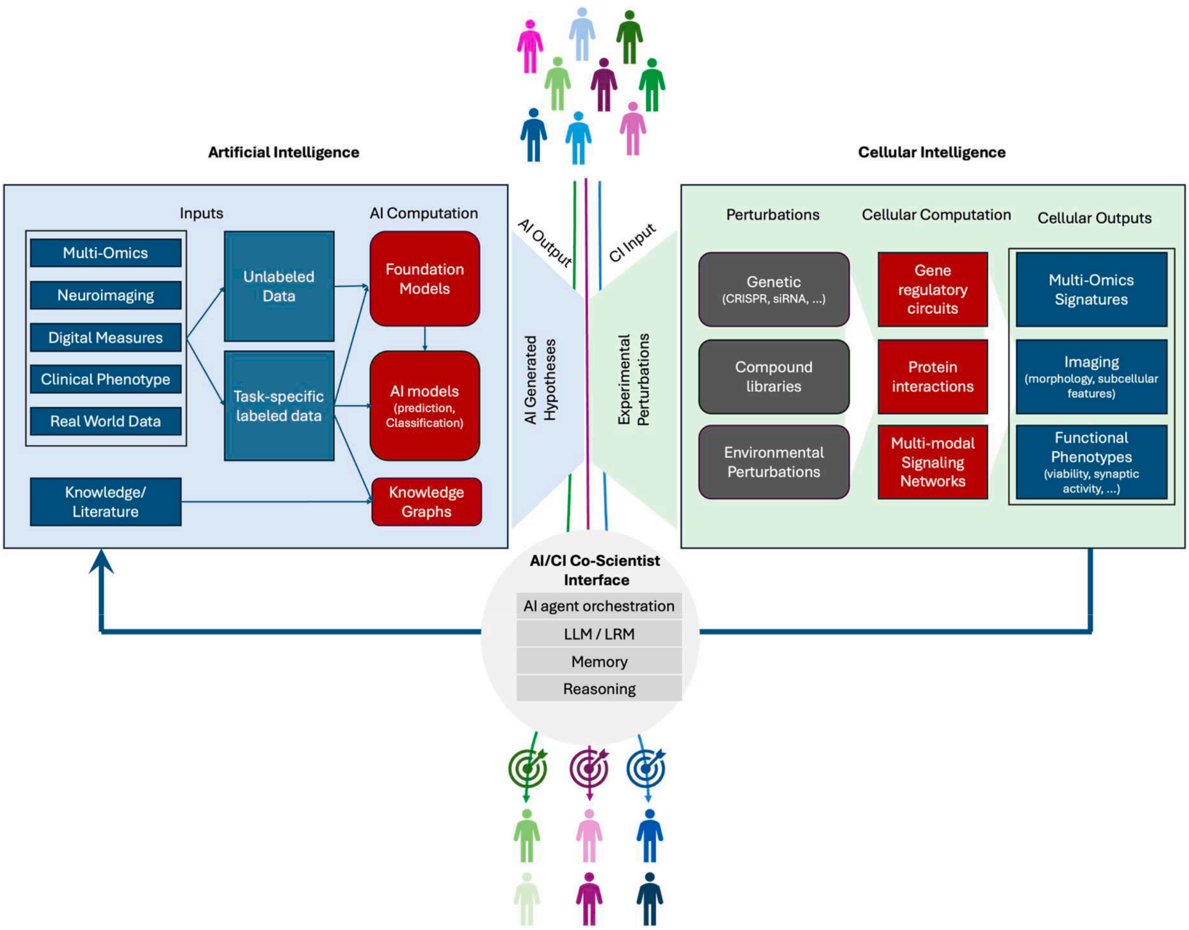


Fig. 3. Interplay of Artificial and Cellular Intelligence in target identification and validation. Artificial intelligence (left) integrates diverse inputs – including multi-omics, neuroimaging, digital and clinical measures, real-world data, and literature into broader foundation models, task-specific predictive models, and knowledge graphs. These models generate putative target hypotheses. These are encoded as perturbations (e.g. CRISPR edits, drug or environmental challenges). Cellular Intelligence (right) executes these perturbations through intrinsic biological computation (gene regulatory networks, protein interactions, and signaling networks) producing measurable outputs such as multi-omics signatures, morphological and imaging features, and functional phenotypes. These experimental results feed back into AI systems, refining the next cycle of hypothesis generation and refinement. Together this loop defines and AI-cell co-scientist paradigm, where artificial and cellular computation operate in tandem to accelerate discovery, and identify the patient population most likely to benefit from a novel intervention.

interventions. Models that obscure biological reasoning may succeed in pattern recognition but fail to advance mechanism-based understanding or rational design.

Reproducibility and generalizability further complicate the landscape. AI models can overfit to training data, particularly when data are sparse or biased, leading to performance drops on new or external datasets. Combined with the inherent variability of biological experiments, this contributes to the reproducibility crisis that has long plagued the biomedical sciences. Without shared benchmarks, rigorous standards and transparency around methods and results it is difficult to compare models or build on prior work.

Finally, technical and infrastructure-related challenges should not be underestimated. Integrating multimodal data across genomics, imaging, clinical records and other modalities remains a daunting task. Standards for data formatting, normalization and metadata are often inconsistent. Training advanced AI models requires significant computational resources: cloud computing, GPUs, and expert teams spanning biology, data science and engineering.

While AI offers unprecedented power, its impact today is bounded by the limitations of the data it consumes, the transparency of its operations and the structure within which it is deployed. Addressing these limitations is essential if we are to realize the full potential of AI in neuroscience drug discovery.

Many complementary efforts are addressing challenges that remain outside the reach of AI. Advances in human stem-cell-derived organoids and micro-physiological systems are providing experimentally tractable models that better capture both cellular and circuit-level context not yet represented *in silico* [85]. Large-scale longitudinal data initiatives like UK Biobank [86] and AllofUs [87] are improving diversity and generalizability while mitigating bias inherent in training data sets. In parallel, hybrid approaches that integrate mechanistic modeling with data-driven inference are emerging to bridge causal understanding with predictive power [88,89]. We focus here on the power of coupling AI with computations performed by the biological system, which we call "cellular intelligence".

7. Cellular intelligence

Data mining of very large and increasingly multi-modal molecular datasets ('multi-omics'), although demonstrably successful, has limitations, as noted above, which might be mitigated in part through a combined use of AI and innovation in experimental design.

Two examples to illustrate the point include the problem of the dependent variable or outcome, and the challenge of interpretation or how to use the results. In this section we discuss both, with some examples of how they might be addressed including with advanced analytics such as AI and ML. Importantly, the point is that simply using ever larger datasets combined with AI, whilst enormously valuable, is not enough. The scientist is still an important actor in this play.

In observational studies it is the disease itself that is the dependent variable or outcome, and in the case of molecular data mining, it is the omics that is the independent variable being measured. This is challenging when it comes to AD and other neurodegenerative disease. AD is a common disorder of the elderly with a long preclinical phase and because of this, unaffected individuals are a less-than-optimal control or comparison group. Many elderly people will have disease pathology even if apparently unaffected and others will already be on course to do so. In case-control studies, the controls may not be so different from the cases. Various approaches can be used to circumvent this limitation; the dependent variable can be switched to age of onset [90], or some other clinical phenotype of interest, for example, comparing people with AD with slow versus rapid decline [91,92]. As biomarkers for pathology become available, the pathology itself can be used as the dependent variable [93]. The latter is highly attractive when it comes to target identification for drug discovery as it provides the potential to launch a precision intervention accompanied by biomarkers. Taking a precision approach further, sub-groups of AD might be identified such as those relatively resilient or vulnerable to pathology. As examples of precision sub-group creation, using GNPC and other data-sources, Oh et al. [94] identify markers of cognitive resilience; using a very deeply phenotyped cohort, Ng et al. [95] identify, and then validate *in vitro*, sub-groups of people with AD relatively resilient or vulnerable to amyloid pathology. Dolan et al. show that *in vitro* iPSC derived microglia have a validated disease transcriptional phenotype when challenged [96]. Combining these experimental approaches that go beyond the AD case / age matched control together with AI/ML in very large datasets seems a promising approach for future target identification. Especially when, as in some of the examples given, it seems possible to replicate *in vitro* disease relevant phenotypic response to challenges.

A second limitation of large-scale data mining comes after the successful delivery of results—how to interpret these results? Typically, the outcome of the data mining will come in the form of a list of potential targets ranked in some way to reflect their contribution to the differentiation of the dependent variable. It is in the nature of biology that this ranking is subtle – the difference between the top of the list and the middle of the list might be relatively small. Furthermore, it is in the nature of the analysis that repeating the exercise with the same data and the same analytical approach often yields a different list with a different ranking. This is to be expected; if a number of variables contribute equally to the differentiation, unless prevented, the model will represent all these variables with a single one. And perhaps a different one on repetition. In effect, the ranking of the list of variables contributing to the outcome is by itself a poor identifier of targets. Taking this into account, bioinformaticians will frequently represent the outcome list with a pathway nomination using GO terms or some similar approach to interpretation. Whilst this can be helpfully suggestive, all such bioinformatics tools have their limitations [97]. Alternatively, lists of targets can be parsed using a druggability assessment or validation from existing literature, both approaches being made easier using large language models to analyze the scientific literature.

An alternative mitigation for the challenge of interpretation is to not interpret but to instead to use the outcome of the analysis *in its entirety*. To use all of the omics signature detected, rather than trying to pick out targets. This was the in effect the driver behind the impactful NIH funded connectivity map (cMap) generated by the Broad Institute [98]. In the first iteration of this, the transcript map of cells perturbed with each of some 1500 compounds was generated and made available on a platform together with analytical software designed to allow researchers to compare expression signatures from disease to those generated by drugs. A number of studies have used the cMAP Gene Set Enrichment Analysis and other tools, to compare lists of genes differentially expressed in AD to the effect of compounds, seeking a signature counter-match (i.e. similar genes ranking but in opposite direction) as part of drug repurposing efforts [99–101].

Given that the targets of these compounds, representing those in clinical use, are known [102], then such approaches can be used to support novel compound discovery programs, as much as for repurposing. The cMAP and its successor, the LINCS program [103] have been followed by another Broad/MIT led program, The Joint Undertaking in Morphology and Cell Painting (JUMP-CP) in collaboration with a number of pharma, that instead of using expression analysis used morphological profiling. Essentially generating an image of cells perturbed by compounds, this public-private initiative was very high throughput and generated, and has made available, data on over 100k compounds [104]. Given the power of generative AI to analyze imaging data, this is an obviously rich source for deep learning on the effects of compounds on cells and the identification of targets. However, in contrast to the cMAP, there is no ready source of cell morphology data to compare to the compound perturbation. But clearly opportunities exist to develop such data – cell lines could be engineered to carry AD related genetic variants to identify morphological signatures which in turn could be matched to compound signatures. Such an *in silico* agnostic phenotypic screen might generate useful packages for drug discovery including compounds for target deconvolution and which might be used as tools for validation, as starting chemical matter for discovery programs or for repurposing efforts (Fig. 3).

There are many approaches to target identification using large data sets that could be enhanced using AI methods. One such is the use of real-world clinical data (RWD) either to identify or to validate targets. Whether from administrative data or from electronic medical records, the amount of RWD is steadily increasing as is the community of scientists using it; most obviously as represented in the Observational Health Data and Informatics initiative (https://www.ohdsi.org/). In neurodegenerative diseases such data was used to validate targets for Parkinson's disease identified through a screening program [105] although varying results from replication RWD studies demonstrate that interpreting such findings can be complicated [106,107]. Combining real world clinical data with genomic and other molecular data might add confidence in the findings, with an example from a public-private consortium study being the nomination of JAK-STAT signaling participants as targets for AD using a combination of real world clinical evidence, GWAS, expression data together with experimental data from preclinical models [108]. Given that so much real-world data is contained in text and given the explosive advances in the ability of AI to derive information from language inputs it seems very likely that using AI together with ML will significantly enhance the combined use of very large clinical and molecular data including electronic medical records (EMRs) augmented by biology.

8. The future of drug discovery

If drug discovery for Alzheimer's disease started with the cholinergic hypothesis and the identification of the protein forming the core of plaque pathology approximately a half-century ago, then it has to be acknowledged that the paradigm for TiTv in use for most of that time has been pretty successful. Using post-mortem studies of human brain together with hypothesis driven cell and molecular biology and biochemistry, a generation of scientists have identified targets that have fed today's rich and diverse drug discovery portfolio [109]. Now supplemented by ever larger datasets of the layers of biology from genomes through transcriptomes to proteomes and not forgetting lipidomes, metabolomes, microbiomes and so on and so forth, the cell and molecular biologist, the physiologist and biochemist have become increasingly adept in utilizing advanced analytics to derive information from such data and use this for precision neuroscience target identification and validation.

However, in the last few years with the sudden arrival – an instant success built on several lifetimes of work – of foundational models, large language models and generative AI, it seems as if the world of TiTv has just shifted, or if not yet, will soon. To a large degree, this will be an incremental shift. AI is predictably going to make data mining of the increasingly large datasets more interesting and more informative, especially when those datasets become truly multi-modal across all layers of the biome and include imaging, complex real world data and adjacent data such as the environome. Massive datasets combined with AI analytics will rapidly eclipse current methodologies. Interpreting the results of such datasets for TiTv will also be facilitated by AI which is already today a more effective reader and user of the scientific literature than most of us human scientists.

The goal of target identification for Alzheimer's disease in the age of AI and CI is still to generate improved hypotheses that can lead to compounds that can be tested in human clinical trials (Fig. 3). The percentage of programs that are successful in clinical development are so low that any improvement in the accuracy, speed and variety of targets identified and validated will make a significant impact on the number and variety of clinical development programs. With faster, more comprehensive ways to analyze massive multi-omics datasets there is an opportunity to address the *inverse problem of drug discovery*, complementing approaches focused on characterizing the *forward pathways of disease* (Fig. 1).

Nonetheless, it seems likely that the scientist will also remain an essential part of the target identification process. As well as being a more effective data-miner, the AI combined with the scientist can be a smarter data-miner or user of the data. We have discussed here the ways in which AI could be used beyond data-mining, for example in compound signature matching, in identification of sub-groups of disease for precision intervention, in combining highly disparate types of data. There will be many others.

Experimentation is still required, both to generate signatures to perturb the analytical models, as well as to validate analytically determined hypotheses. Here the scientist is faced with a challenge: if the experiment gives a negative result is that because the model does not replicate the human disease to the same extent as the human-data analytical model, or is it because the human-data analytical model generated an incorrect hypothesis? Still, existing models, even when imperfect, provide the scaffolding for hypothesis generation and falsification, enabling continual learning across *in silico* and *in vitro* domains. The future of AI-driven discovery will be shaped as much by this disciplined cycle of use, evaluation and improvement as by the eventual realization of fully predictive disease models.

Although preclinical models of neurodegeneration remain insufficiently predictive and human validation is still limited, the trajectory is unmistakable. We are moving towards a future in which scientists partner with AI systems, not to replace insight, but to refine it, linking model-trained networks with biological intuition to identify, test and validate therapeutic targets in AD and other neurodegenerative disorders. Humans could not be more "in the loop" as patients, their families and caregivers become the ultimate beneficiaries of this dramatic progress.

9. Glossary

Agentic AI: AI systems capable of autonomously planning, reasoning, and executing sequences of actions toward a defined scientific or analytic goal. In biomedical research, agentic AI refers to "AI co-scientists" that can orchestrate data analysis, hypothesis generation and experiment design through iterative, self-directed workflows.

Artificial Intelligence (AI): Computational systems designed to perform tasks that typically require human intelligence, such as pattern recognition, prediction and reasoning.

Big Data: Extremely large and complex data sets requiring advanced computational tools for storage, integration and analysis.

Cellular Intelligence (CI): The capacity of human-derived cell systems to provide biological grounded insights that complement computational and human analytical approaches reading perturbations through signal transduction cascades, and producing observable or measurable changes in phenotype.

Foundational Model (FM): A large, pre-trained AI model developed on diverse, multimodal data that can be adapted (find-tuned) to specific biomedical tasks.

Generative AI: A class of AI models capable of producing new data, such as text, images, or synthetic omics profiles, based on patterns learned from training datasets.

Genome-Wide Association Study (GWAS): A statistical approach to identify genetic variants associated with disease risk by scanning the entire genome in large populations.

Induced pluripotent step cells (iPSCs): Stem cells reprogrammed from adult comatic cells that can differentiate into multiple cell types enabling disease modeling and drug testing.

Large Language Model (LLM): A neural network trained on massive text corpora to perform language-based reasoning, summarization, and data synthesis; increasingly applied in biomedical research.

Large Reasoning Model (LRM): An AI system integrating structured reasoning and retrieval mechanisms to support hypothesis generation and interpretation across multimodal data.

Multi-omics: Integrated biological data combining multiple "omics" layers such as genetic (DNA), transcriptomics (mRNA), proteomics and metabolomics to provide a holistic view of disease biology.

Population cohort: A large group of individuals followed over time in a research study to assess biological, clinical or genetic factors related to disease risk or progression

Real-World Data (RWD): Data derived from sources outside traditional clinical trials, such as electronic medical records, insurance claims, disease registries, and digital health platforms, increasingly used to complement experimental data.

Self-Supervised Learning: An AI training approach in which models learn patterns or representations from unlabeled data by predicting hidden or missing parts of the input, often used to build large foundation models.

Target Identification and Target Validation (TiTv): The process of discovering, prioritizing and experimentally confirming molecular entities that can be modulated to achieve clinical benefit in disease treatment.

Quantitative Trait Locus (QTL): A genomic region associated with variation in a measurable trait, such as gene expression (eQTL) or protein abundance (pQTL).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT and Claude in order to assist with initial content ideation within human-defined subsections. After using this tool/service, the author(s) reviewed and edited the content heavily and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Gayle Wittenberg: Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Fiona Elwood:** Writing – review & editing, Writing – original draft. **Andrea Houghton:** Writing – review & editing, Writing – original draft. **Tommaso Mansi:** Writing – review & editing, Writing – original draft. **Bart Smets:** Writing – review & editing, Writing – original draft, Visualization. **Simon Lovestone:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gayle Wittenberg reports financial support was provided by Johnson & Johnson Innovative Medicines. Fiona Elwood reports financial support was provided by Johnson & Johnson Innovative Medicines. Andrea Houghton reports financial support was provided by Johnson & Johnson Innovative Medicines. Tommaso Mansi reports financial support was provided by Johnson & Johnson Innovative Medicines. Bart Smets reports financial support was provided by Johnson & Johnson Innovative Medicines. Simon Lovestone reports financial support was provided by Johnson & Johnson Innovative Medicines. Gayle Wittenberg reports a relationship with Johnson & Johnson that includes: employment and equity or stocks. Fiona Elwood reports a relationship with Johnson & Johnson that includes: employment and equity or stocks. Andrea Houghton reports a relationship with Johnson & Johnson that includes: employment and equity or stocks. Tommaso Mansi reports a relationship with Johnson & Johnson that includes: employment and equity or stocks. Bart Smets reports a relationship with Johnson & Johnson that includes: employment and equity or stocks. Simon Lovestone reports a relationship with Johnson & Johnson that includes: employment and equity or stocks. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Noble D. A theory of biological relativity: no privileged level of causation. Interface Focus 2012;2:55–64. https://doi.org/10.1098/rsfs.2011.0067.

[2] Sugimoto H, Ogura H, Arai Y, Iimura Y, Yamanishi Y. Research and development of Donepezil hydrochloride, a new type of acetylcholinesterase inhibitor. Jpn J Pharmacol 2002;89:7–20. https://doi.org/10.1254/jjp.89.7.

[3] Bartus RT. On neurodegenerative diseases, models, and treatment strategies: lessons learned and lessons forgotten a generation following the cholinergic hypothesis. Exp Neurol 2000;163:495–529. https://doi.org/10.1006/exnr.2000.7397.

[4] Tiseo, Rogers, Friedhoff. Pharmacokinetic and pharmacodynamic profile of donepezil HCl following evening administration. Br J Clin Pharmacol 1998;46:13–8. https://doi.org/10.1046/j.1365-2125.1998.0460s1013.x.

[5] Weggen S, Beher D. Molecular consequences of amyloid precursor protein and presenilin mutations causing autosomal-dominant Alzheimer's disease. Alzheimers Res Ther 2012;4:9. https://doi.org/10.1186/alzrt107.

[6] Hsiao K, Chapman P, Nilsen S, Eckman C, Harigaya Y, Younkin S, et al. Correlative memory deficits, aβ elevation, and amyloid plaques in transgenic mice. Science 1996;274:99–103. https://doi.org/10.1126/science.274.5284.99.

[7] Neumann U, Machauer R, Shimshek DR. The β-secretase (BACE) inhibitor NB-360 in preclinical models: from amyloid-β reduction to downstream disease-relevant effects. Br J Pharmacol 2019;176:3435–46. https://doi.org/10.1111/bph.14582.

[8] Barak M, Fedorova V, Pospisilova V, Raska J, Vochyanova S, Sedmik J, et al. Human iPSC-derived neural models for studying Alzheimer's disease: from neural stem cells to cerebral organoids. Stem Cell Rev Rep 2022;18:792–820. https://doi.org/10.1007/s12015-021-10254-3.

[9] Sen T, Thummer RP. CRISPR and iPSCs: recent developments and future perspectives in neurodegenerative disease modelling, research, and therapeutics. Neurotox Res 2022;40:1597–623. https://doi.org/10.1007/s12640-022-00564-w.

[10] Ramos DM, Skarnes WC, Singleton AB, Cookson MR, Ward ME. Tackling neurodegenerative diseases with genomic engineering: a new stem cell initiative from the NIH. Neuron 2021;109:1080–3. https://doi.org/10.1016/j.neuron.2021.03.022.

[11] Quaid K, Xing X, Chen Y-H, Miao Y, Neilson A, Selvamani V, et al. iPSCs and iPSC-derived cells as a model of human genetic and epigenetic variation. Nat Commun 2025;16:1750. https://doi.org/10.1038/s41467-025-56569-4.

[12] Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nat Genet 2015;47:856–60. https://doi.org/10.1038/ng.3314.

[13] Van de Sande B, Lee JS, Mutasa-Gottgens E, Naughton B, Bacon W, Manning J, et al. Applications of single-cell RNA sequencing in drug discovery and development. Nat Rev Drug Discov 2023;22:496–520. https://doi.org/10.1038/s41573-023-00688-4.

[14] Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet 2022;54:412–36. https://doi.org/10.1038/s41588-022-01024-z.

[15] Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet 2019;51:404–13. https://doi.org/10.1038/s41588-018-0311-9.

[16] Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on family history of Alzheimer's disease. Transl Psychiatry 2018;8. https://doi.org/10.1038/s41398-018-0150-6.

[17] Gallagher MD, AS Chen-Plotkin. The Post-GWAS era: from association to function. Am J Hum Genet 2018;102:717–30. https://doi.org/10.1016/j.ajhg.2018.04.002.

[18] Sierksma A, Escott-Price V, De Strooper B. Translating genetic risk of Alzheimer's disease into mechanistic insight and drug targets. Science 2020;370:61–6. https://doi.org/10.1126/science.abb8575.

[19] Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 2016;48:481–7. https://doi.org/10.1038/ng.3538.

[20] Li B, Ritchie MD. From GWAS to gene: transcriptome-wide association studies and other methods to functionally understand GWAS discoveries. Front Genet 2021;12. https://doi.org/10.3389/fgene.2021.713230.

[21] Sun BB, Chiou J, Traylor M, Benner C, Hsu Y-H, Richardson TG, et al. Plasma proteomic associations with genetics and health in the UK Biobank. Nature 2023;622:329–38. https://doi.org/10.1038/s41586-023-06592-6.

[22] Yang C, Farias FHG, Ibanez L, Suhy A, Sadler B, Fernandez MV, et al. Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. Nat Neurosci 2021;24:1302–12. https://doi.org/10.1038/s41593-021-00886-6.

[23] Yang C, Fagan AM, Perrin RJ, Rhinn H, Harari O, Cruchaga C. Mendelian randomization and genetic colocalization infer the effects of the multi-tissue proteome on 211 complex disease-related phenotypes. Genome Med 2022;14:140. https://doi.org/10.1186/s13073-022-01140-9.

[24] Western D, Timsina J, Wang L, Wang C, Yang C, Phillips B, et al. Proteogenomic analysis of human cerebrospinal fluid identifies neurologically relevant regulation and implicates causal proteins for Alzheimer's disease. Nat Genet 2024;56:2672–84. https://doi.org/10.1038/s41588-024-01972-8.

[25] Wingo AP, Liu Y, Gerasimov ES, Gockley J, Logsdon BA, Duong DM, et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. Nat Genet 2021;53:143–6. https://doi.org/10.1038/s41588-020-00773-z.

[26] Hu T, Liu Q, Dai Q, Buchman AS, Bennett DA, Tasaki S, et al. Proteome-wide association studies using summary pQTL data of brain, CSF, and plasma identify 30 risk genes of Alzheimer's disease dementia. Alzheimers Res Ther 2025;17:135. https://doi.org/10.1186/s13195-025-01774-y.

[27] Sun Y, Zhu J, Zhou D, Canchi S, Wu C, Cox NJ, et al. A transcriptome-wide association study of Alzheimer's disease using prediction models of relevant tissues identifies novel candidate susceptibility genes. Genome Med 2021;13:141. https://doi.org/10.1186/s13073-021-00959-y.

[28] Liu S, Cho M, Huang Y-N, Park T, Chaudhuri S, Rosewood TJ, et al. Multi-omics analysis for identifying cell-type-specific and bulk-level druggable targets in Alzheimer's disease. J Transl Med 2025;23:788. https://doi.org/10.1186/s12967-025-06739-1.

[29] Wan Y-W, Al-Ouran R, Mangleburg CG, Perumal TM, Lee TV, Allison K, et al. Meta-analysis of the Alzheimer's Disease Human brain transcriptome and functional dissection in mouse models. Cell Rep 2020;32:107908. https://doi.org/10.1016/j.celrep.2020.107908.

[30] Wang M, Roussos P, McKenzie A, Zhou X, Kajiwara Y, Brennand KJ, et al. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. Genome Med 2016;8:104. https://doi.org/10.1186/s13073-016-0355-3.

[31] Mathys H, Adaikkan C, Gao F, Young JZ, Manet E, Hemberg M, et al. Temporal tracking of microglia activation in neurodegeneration at single-cell resolution. Cell Rep 2017;21:366–80. https://doi.org/10.1016/j.celrep.2017.09.039.

[32] Olah M, Menon V, Habib N, Taga MF, Ma Y, Yung CJ, et al. Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease. Nat Commun 2020;11:6129. https://doi.org/10.1038/s41467-020-19737-2.

[33] Habib N, McCabe C, Medina S, Varshavsky M, Kitsberg D, Dvir-Szternfeld R, et al. Disease-associated astrocytes in Alzheimer's disease and aging. Nat Neurosci 2020;23:701–6. https://doi.org/10.1038/s41593-020-0624-8.

[34] Mathys H, Peng Z, Boix CA, Victor MB, Leary N, Babu S, et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. Cell 2023;186:4365–85. https://doi.org/10.1016/j.cell.2023.08.039. e27.

[35] Swarup V, Chang TS, Duong DM, Dammer EB, Dai J, Lah JJ, et al. Identification of conserved proteomic networks in neurodegenerative dementia. Cell Rep 2020;31:107807. https://doi.org/10.1016/j.celrep.2020.107807.

[36] Johnson ECB, Carter EK, Dammer EB, Duong DM, Gerasimov ES, Liu Y, et al. Large-scale deep multi-layer analysis of Alzheimer's disease brain reveals strong proteomic disease-related changes not observed at the RNA level. Nat Neurosci 2022;25:213–25. https://doi.org/10.1038/s41593-021-00999-y.

[37] Pichet Binette A, Gaiteri C, Wennström M, Kumar A, Hristovska I, Spotorno N, et al. Proteomic changes in Alzheimer's disease associated with progressive Aβ plaque and tau tangle pathologies. Nat Neurosci 2024;27:1880–91. https://doi.org/10.1038/s41593-024-01737-w.

[38] Xu J, Bankov G, Kim M, Wretlind A, Lord J, Green R, et al. Integrated lipidomics and proteomics network analysis highlights lipid and immunity pathways associated with Alzheimer's disease. Transl Neurodegener 2020;9:36. https://doi.org/10.1186/s40035-020-00215-0.

[39] Nativio R, Lan Y, Donahue G, Sidoli S, Berson A, Srinivasan AR, et al. An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. Nat Genet 2020;52:1024–35. https://doi.org/10.1038/s41588-020-0696-0.

[40] Shi L, Xu J, Green R, Wretlind A, Homann J, Buckley NJ, et al. Multiomics profiling of human plasma and cerebrospinal fluid reveals ATN-derived networks and highlights causal links in Alzheimer's disease. Alzheimers Dement J Alzheimers Assoc 2023;19:3350–64. https://doi.org/10.1002/alz.12961.

[41] Imam F, Saloner R, Vogel JW, Krish V, Abdel-Azim G, Ali M, et al. The Global Neurodegeneration Proteomics Consortium: biomarker and drug target discovery for common neurodegenerative diseases and aging. Nat Med 2025. https://doi.org/10.1038/s41591-025-03834-0.

[42] Minikel EV, Karczewski KJ, Martin HC, Cummings BB, Whiffin N, Rhodes D, et al. Evaluating drug targets through human loss-of-function genetic variation. Nature 2020;581:459–64. https://doi.org/10.1038/s41586-020-2267-z.

[43] Carss KJ, Deaton AM, Del Rio-Espinola A, Diogo D, Fielden M, Kulkarni DA, et al. Using human genetics to improve safety assessment of therapeutics. Nat Rev Drug Discov 2023;22:145–62. https://doi.org/10.1038/s41573-022-00561-w.

[44] Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. Nature 2020;581:303–9. https://doi.org/10.1038/s41586-020-2157-4.

[45] Chen J, Wu J, Bai Y, Yang C, Wang J. Recent advances of single-cell RNA sequencing in toxicology research: insight into hepatotoxicity and nephrotoxicity. Curr Opin Toxicol 2024;37:100462. https://doi.org/10.1016/j.cotox.2024.100462.

[46] Li T, Chen X, Tong W. Bridging organ transcriptomics for advancing multiple organ toxicity assessment with a generative AI approach. Npj Digit Med 2024;7:310. https://doi.org/10.1038/s41746-024-01317-z.

[47] Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov 2002;1:727–30. https://doi.org/10.1038/nrd892.

[48] Finan C, Gaulton A, Kruger FA, Lumbers RT, Shah T, Engmann J, et al. The druggable genome and support for target identification and validation in drug development. Sci Transl Med 2017;9:eaag1166. https://doi.org/10.1126/scitranslmed.aag1166.

[49] Schneider M, Radoux CJ, Hercules A, Ochoa D, Dunham I, Zalmas L-P, et al. The PROTACtable genome. Nat Rev Drug Discov 2021;20:789–97. https://doi.org/10.1038/s41573-021-00245-x.

[50] Whitehead KA, Langer R, Anderson DG. Knocking down barriers: advances in siRNA delivery. Nat Rev Drug Discov 2009;8:129–38. https://doi.org/10.1038/nrd2742.

[51] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9. https://doi.org/10.1038/75556.

[52] Consortium The Gene Ontology, SA Aleksander, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The gene ontology knowledgebase in 2023. GENETICS 2023;224:iyad031. https://doi.org/10.1093/genetics/iyad031.

[53] Kanehisa MKEGG. Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.

[54] Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. Bioinforma Oxf Engl 2014;30:523–30. https://doi.org/10.1093/bioinformatics/btt703.

[55] Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 2008;8:37–49. https://doi.org/10.1038/nrc2294.

[56] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.

[57] Jiménez-Jiménez V, Martí-Gómez C, Del Pozo MÁ, Lara-Pezzi E, Sánchez-Cabo F. Bayesian inference of gene expression. In: Helder IN, editor. Bioinformatics. Brisbane (AU): Exon Publications; 2021.

[58] Huang H, Wang Y, Rudin C, Browne EP. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. Commun Biol 2022;5:719. https://doi.org/10.1038/s42003-022-03628-x.

[59] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. Adv. Neural Inf. Process. Syst. 2017;30.

[60] Bommasani R., Hudson D.A., Adeli E., Altman R., Arora S., Arx S von, et al. On the opportunities and risks of foundation models 2022. https://doi.org/10.48550/arXiv.2108.07258.

[61] Gui J, Chen T, Zhang J, Cao Q, Sun Z, Luo H, et al. A survey on self-supervised learning: algorithms, applications, and future trends. IEEE Trans Pattern Anal Mach Intell 2024;46:9052–71. https://doi.org/10.1109/TPAMI.2024.3415112.

[62] Juan Ramon A, Parmar C, Carrasco-Zevallos OM, Csiszer C, Yip SSF, Raciti P, et al. Development and deployment of a histopathology-based deep learning algorithm for patient prescreening in a clinical trial. Nat Commun 2024;15:4690. https://doi.org/10.1038/s41467-024-49153-9.

[63] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Adv. neural inf. process. syst. Adv. neural inf. process. syst, 33. Curran Associates, Inc.; 2020. p. 1877–901.

[64] Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., et al. Learning transferable visual models from natural language supervision 2021. https://doi.org/10.48550/ARXIV.2103.00020.

[65] Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods 2024;21:1470–80. https://doi.org/10.1038/s41592-024-02201-0.

[66] Zeng Y, Xie J, Shangguan N, Wei Z, Li W, Su Y, et al. CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. Nat Commun 2025;16:4679. https://doi.org/10.1038/s41467-025-59926-5.

[67] Avsec Ž., Latysheva N., Cheng J., Novati G., Taylor K.R., Ward T., et al. AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model 2025. https://doi.org/10.1101/2025.06.25.661532.

[68] Cui T., Xu S.-J., Moskalev A., Li S., Mansi T., Prakash M., et al. InfoSEM: a deep generative model with informative priors for gene regulatory network inference 2025. https://doi.org/10.48550/ARXIV.2503.04483.

[69] Snell C., Lee J., Xu K., Kumar A. Scaling LLM test-time compute optimally can be more effective than Scaling model parameters 2024. https://doi.org/10.48550/ARXIV.2408.03314.

[70] Gottweis J., Weng W.-.H., Daryin A., Tu T., Palepu A., Sirkovic P., et al. Towards an AI co-scientist 2025. https://doi.org/10.48550/ARXIV.2502.18864.

[71] Huang K, Zhang S, Wang H, Qu Y, Lu Y, Roohani Y, et al. Biomni: a general-purpose biomedical AI agent. BioRxiv Prepr Serv Biol 2025;2025:05. https://doi.org/10.1101/2025.05.30.656746. 30.656746.

[72] Wang H., He Y., Coelho P.P., Bucci M., Nazir A., Chen B., et al. SpatialAgent: an autonomous AI agent for spatial biology 2025. https://doi.org/10.1101/2025.04.03.646459.

[73] Kaplan J., McCandlish S., Henighan T., Brown T.B., Chess B., Child R., et al. Scaling laws for neural language models 2020. https://doi.org/10.48550/arXiv.2001.08361.

[74] Bourached A, Bonkhoff AK, Schirmer MD, Regenhardt RW, Bretzner M, Hong S, et al. Scaling behaviours of deep learning and linear algorithms for the prediction of stroke severity. Brain Commun 2024;6:fcae007. https://doi.org/10.1093/braincomms/fcae007.

[75] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025;43:1–55. https://doi.org/10.1145/3703155.

[76] Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. Npj Digit Med 2025;8:274. https://doi.org/10.1038/s41746-025-01670-7.

[77] Valmeekam K., Olmo A., Sreedharan S., Kambhampati S. Large language models still can't plan (A Benchmark for LLMs on Planning and Reasoning about Change) n.d.

[78] Shojaee P., Mirzadeh I., Alizadeh K., Horton M., Bengio S., Farajtabar M. The illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity 2025. https://doi.org/10.48550/arXiv.2506.06941.

[79] d'Avila Garcez A, LC Lamb. Neurosymbolic AI: the 3rd wave. Artif Intell Rev 2023;56:12387–406. https://doi.org/10.1007/s10462-023-10448-w.

[80] Berrevoets J, Kacprzyk K, Qian Z, der Schaar M van. Causal deep learning: encouraging impact on real-world problems through causality. Found Trends® Signal Process 2024;18:200–309. https://doi.org/10.1561/2000000123.

[81] Sirugo G, Williams SM, Tishkoff SA. The missing diversity in Human genetic studies. Cell 2019;177:26–31. https://doi.org/10.1016/j.cell.2019.02.048.

[82] Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat Med 2020;26:16–7. https://doi.org/10.1038/s41591-019-0649-2.

[83] Aisen PS, Cummings J, Jack CR, Morris JC, Sperling R, Frölich L, et al. On the path to 2025: understanding the Alzheimer's disease continuum. Alzheimers Res Ther 2017;9:60. https://doi.org/10.1186/s13195-017-0283-5.

[84] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15. https://doi.org/10.1038/s42256-019-0048-x.

[85] Smirnova L, Hartung T. The promise and potential of brain organoids. Adv Healthc Mater 2024;13:2302745. https://doi.org/10.1002/adhm.202302745.

[86] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Med 2015;12:e1001779. https://doi.org/10.1371/journal.pmed.1001779.

[87] The All of us research Program investigators. The "all of us" research program. N Engl J Med 2019;381:668–76. https://doi.org/10.1056/NEJMsr1809937.

[88] Noordijk B, Garcia Gomez ML, ten Tusscher KHWJ, de Ridder D, van Dijk ADJ, Smith RW. The rise of scientific machine learning: a perspective on combining mechanistic modelling with machine learning for systems biology. Front Syst Biol 2024;4. https://doi.org/10.3389/fsysb.2024.1407994.

[89] Procopio A, Cesarelli G, Donisi L, Merola A, Amato F, Cosentino C. Combined mechanistic modeling and machine-learning approaches in systems biology – A systematic literature review. Comput Methods Programs Biomed 2023;240:107681. https://doi.org/10.1016/j.cmpb.2023.107681.

[90] Kamboh MI, Barmada MM, Demirci FY, Minster RL, Carrasquillo MM, Pankratz VS, et al. Genome-wide association analysis of age-at-onset in Alzheimer's disease. Mol Psychiatry 2012;17:1340–6. https://doi.org/10.1038/mp.2011.135.

[91] Thambisetty M, An Y, Kinsey A, Koka D, Saleem M, Güntert A, et al. Plasma clusterin concentration is associated with longitudinal brain atrophy in mild cognitive impairment. NeuroImage 2012;59:212–7. https://doi.org/10.1016/j.neuroimage.2011.07.056.

[92] Cruchaga C, Kauwe JSK, Mayo K, Spiegel N, Bertelsen S, Nowotny P, et al. SNPs associated with cerebrospinal fluid phospho-tau levels influence rate of decline in Alzheimer's disease. PLoS Genet 2010;6:e1001101. https://doi.org/10.1371/journal.pgen.1001101.

[93] Saari TT, Palviainen T, Hiltunen M, Herukka S-K, Kokkola T, Kärkkäinen S, et al. Cross-sectional study of plasma phosphorylated tau 217 in persons without dementia. Alzheimers Dement Amst Neth 2025;17:e70107. https://doi.org/10.1002/dad2.70107.

[94] Oh HS-H, Urey DY, Karlsson L, Zhu Z, Shen Y, Farinas A, et al. A cerebrospinal fluid synaptic protein biomarker for prediction of cognitive resilience versus

decline in Alzheimer's disease. Nat Med 2025;31:1592–603. https://doi.org/10.1038/s41591-025-03565-2.

[95] Ng B, Rowland HA, Wei T, Arunasalam K, Hayes EM, Koychev I, et al. Neurons derived from individual early Alzheimer's disease patients reflect their clinical vulnerability. Brain Commun 2022;4:fcac267. https://doi.org/10.1093/braincomms/fcac267.

[96] Dolan M-J, Therrien M, Jereb S, Kamath T, Gazestani V, Atkeson T, et al. Exposure of iPSC-derived human microglia to brain substrates enables the generation and manipulation of diverse transcriptional states *in vitro*. Nat Immunol 2023;24:1382–90. https://doi.org/10.1038/s41590-023-01558-2.

[97] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 2012;8:e1002375. https://doi.org/10.1371/journal.pcbi.1002375.

[98] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006;313:1929–35. https://doi.org/10.1126/science.1132939.

[99] Bellur O, Kastenmuller G, Requena F, Kaddurah-Daouk RF, Krumsiek J, Arnold M. In silico prioritiziation of drug repositioning candidates for Alzheimer's disease using signature search meta-analysis. Alzheimers Dement 2023;19. https://doi.org/10.1002/alz.076844.

[100] Williams G, Gatt A, Clarke E, Corcoran J, Doherty P, Chambers D, et al. Drug repurposing for Alzheimer's disease based on transcriptional profiling of human iPSC-derived cortical neurons. Transl Psychiatry 2019;9:220. https://doi.org/10.1038/s41398-019-0555-x.

[101] Xu Y, Kong J, Hu P. Computational drug repurposing for Alzheimer's disease using risk genes from GWAS and single-cell RNA sequencing studies. Front Pharmacol 2021;12:617537. https://doi.org/10.3389/fphar.2021.617537.

[102] Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov 2017;16:19–34. https://doi.org/10.1038/nrd.2016.230.

[103] Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1000,000 profiles. Cell 2017;171:1437–52. https://doi.org/10.1016/j.cell.2017.10.049. e17.

[104] Chandrasekaran SN, Cimini BA, Goodale A, Miller L, Kost-Alimova M, Jamali N, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Nat Methods 2024;21:1114–21. https://doi.org/10.1038/s41592-024-02241-6.

[105] Mittal S, Bjørnevik K, Im DS, Flierl A, Dong X, Locascio JJ, et al. β2-Adrenoreceptor is a regulator of the α-synuclein gene driving risk of Parkinson's disease. Science 2017;357:891–8. https://doi.org/10.1126/science.aaf3934.

[106] Giorgianni F, Ernst P, Dell'Aniello S, Suissa S, Renoux C. β 2-agonists and the incidence of Parkinson Disease. Am J Epidemiol 2020;189:801–10. https://doi.org/10.1093/aje/kwaa012.

[107] Searles Nielsen S, Gross A, Camacho-Soto A, Willis AW, Racette BA. β2-adrenoreceptor medications and risk of Parkinson disease. Ann Neurol 2018;84:683–93. https://doi.org/10.1002/ana.25341.

[108] Nevado-Holgado AJ, Ribe E, Thei L, Furlong L, Mayer M-A, Quan J, et al. Genetic and real-world clinical data, combined with empirical validation, nominate Jak-stat signaling as a target for Alzheimer's disease therapeutic development. Cells 2019;8:425. https://doi.org/10.3390/cells8050425.

[109] Cummings J, Zhou Y, Lee G, Zhong K, Fonseca J, Cheng F. Alzheimer's disease drug development pipeline: 2024. Alzheimers Dement Transl Res Clin Interv 2024;10:e12465. https://doi.org/10.1002/trc2.12465.

The Journal of Prevention of Alzheimer's Disease 13 (2026) 100397

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

Solving the ’Goldilocks problem’ in dementia clinical trials with multimodal AI

Andrew E. Welchman [a], Zoe Kourtzi [b,*]

[a] Prodromic Ltd, Milton Hall, Ely Road, Milton, Cambridge CB24 6WZ, UK
[b] Department of Psychology, University of Cambridge, Cambridge CB2 3 EB, UK

ARTICLE INFO

Keywords:
Multimodal AI
Clinical trials
Imaging
Cognition
Alzheimer's disease
Early prediction

ABSTRACT

The development of effective therapeutics for Alzheimer's Disease and related dementias (ADRD) has been hindered by patient heterogeneity and the limitations of current diagnostic tools. New treatments have no chance of working if given to patients who cannot benefit from them. This perspective explores how advances in Artificial Intelligence (AI), particularly multimodal machine learning, can solve the ‘Goldilocks problem’ of identifying patients for inclusion in clinical trials and support precision treatment in real-world healthcare settings. We examine the challenges of patient stratification, grounded by a conceptual framework of identifying each person's stage and subtype of dementia. We review data from several clinical trials of Alzheimer's disease therapeutics, to explore how AI-guided patient stratification can improve trial outcomes, reduce costs and improve recruitment. Further, we discuss the integration of AI into clinical workflows, the importance of model interpretability and generalizability, and ethical imperative to address algorithmic bias. By combining AI with scientific insight, clinical expertise, and patient experience, we argue that intelligent analytics can accelerate the discovery and delivery of new diagnostics and therapeutics, ultimately transforming dementia care and improving outcomes for patients around the globe.

1. Introduction

Developing new therapeutics for Alzheimer's Disease (AD) has been hampered by patient heterogeneity and a lack of sensitive tools to precisely stratify and separate individual patients. Despite scientific progress in promising new drug candidates, clinical trials of potential disease-modifying treatments have been disproportionately unsuccessful [1].

In some cases, patients included in trials were mistakenly believed to have AD, when in fact they did not. In other cases, included patients were too far progressed to benefit from the therapeutic's mechanism of action.

This ‘Goldilocks problem’—finding patients who are neither too early nor too late in disease progression—lies at the heart of the challenge of identifying appropriate patients for (a) inclusion in clinical trials and (b) prescription of therapeutics in real world healthcare settings. Good drugs fail if given to patients who have no possibility of benefitting from them. Ongoing efforts are seeking to use biomarkers to improve accurate identification of dementias (e.g [2,3]). Here we ask, how recent advances in Artificial Intelligence (AI) and machine learning can help. We explore how AI-guided patient stratification can accelerate the development and real-world impact of new therapeutics for AD and related dementias (ADRD). We structure this article around three themes:

(1) Improving the sensitivity of clinical trials to gain a better understanding of the efficacy of a medication for the target population.
(2) Improving the efficiency of clinical trials to accelerate therapeutic development.
(3) Improving the real-world effectiveness of a medication in clinical practice.

We start by framing the problem of patient stratification in dementias conceptually, to expose the complexity of the challenges that need to be solved, and to explore how AI can help.

1.1. Why dementia is a hard problem

Dementia results from a cascade of processes that are not fully

* Corresponding author.
E-mail address: zk240@cam.ac.uk (Z. Kourtzi).

https://doi.org/10.1016/j.tjpad.2025.100397
Received 3 September 2025; Received in revised form 15 October 2025; Accepted 20 October 2025

understood (Fig. 1a). The underlying disease process gives rise to multiple neurological consequences ranging from physical to psychological. To quantify these, the field has adopted a range of measurements (e.g., using Positron Emission Tomography (PET) scans, Magnetic Resonance Imaging (MRI), blood tests, Cerebrospinal Fluid samples, Cognitive tests) that relate (in ways not fully understood) to the underlying pathology.

Further, the 'mapping function'—the relationship between the underlying pathology and its measurable biomarkers—is complex and nonlinear. For instance, a biomarker assay may be insensitive to low concentrations of a protein, while changes in large quantities of the protein may be measurable but not functionally meaningful. Different assay domains (e.g., protein, brain structure, cognitive construct) quantify different aspects of the underlying pathology and are subject to different mapping functions. As a result, the interrelation between different assays (e.g., the correspondence between a cognitive test and measures of the accumulation of amyloid beta (β-Amyloid) in the brain) is highly complex.

To identify patients that could benefit from a potential new treatment, we are faced with the 'inverse problem' of using measurable information about the individual ('patient data' such as demographics, medical history, biomarkers) and working backwards to estimate a given patient's disease status. As current biomarkers lack precision, this gives rise to considerable uncertainty – demonstrated by high rates of misdiagnosis at early stages of dementia [4].

The problem is particularly acute in dementia because of a range of conditions (e.g., AD, Lewy Body Dementia, Frontotemporal Dementia) may appear similar in prodromal phases, and each is likely to have subtypes (e.g., subtypes of AD [5]) and/or co-pathologies that are not yet fully understood. Moreover, because dementias involve a cascade of processes responsible for different pathologies, understanding the timing of the disease is critical to ensure that a drug's mechanism of action (e.g. reduced β-Amyloid synthesis) is appropriate for the patient's current main driver of pathology.

1.2. How to tackle the challenge

We conceptualise the core problem as pinpointing an individual's position (Fig. 1b) in the space of dementia stages / time (*x*-axis) and subtypes (*y*-axis). To do this, we accumulate information across a range of data types (*z*-axis), each of which provides only partial and/or ambiguous signals about the patient's disease status. We can think of the information provided by each data type as a probabilistic cue to the true position of the patient in the stage-by-subtype space, visualised as probability maps.

Different types of data provide information in different parts of the space (Fig. 1c). Age, for instance, provides a weak signal (pale pink image intensities) to stage, and does not provide information about

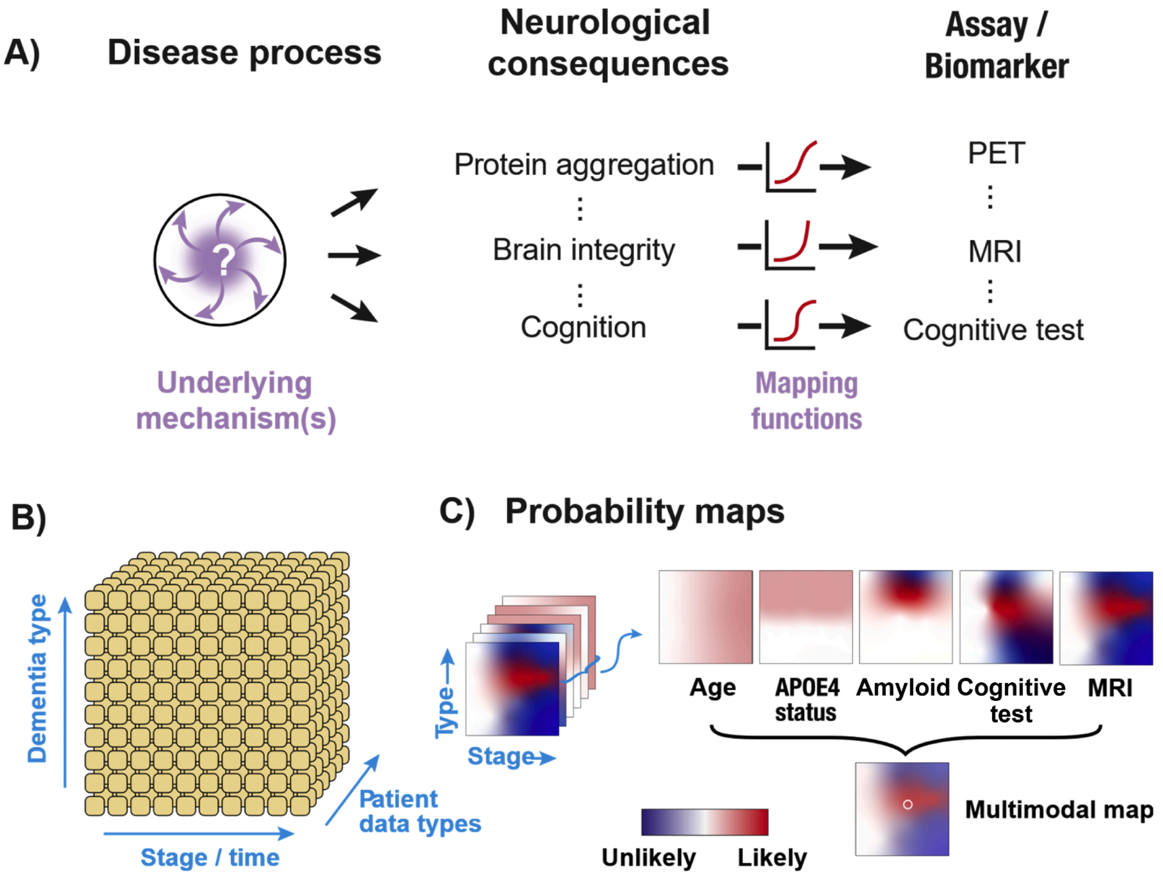


Fig. 1. A) Illustration depicting how an underlying pathology can relate to its measurable biomarkers through a range of different 'mapping functions', shown by different shaped curves. B) A schematic of the conceptual space within which an individual patient can be located – a specific subtype of dementia (y-axis), at a specific stage (x-axis), where data are obtained across a range of different assays and markers (z-axis). C) Illustration of probability maps describing the relationship between information from particular data types in relation to the space of subtypes and disease stages (not empirical data). Different biomarkers provide different degrees of certainty in particular parts of the space shown by variations in the saturation of the colours. White indicates locations where the marker provides no information, saturated coloured regions represent likely (red) or unlikely (blue) disease locations for the patient. Data providing a precise signal to the patient's dementia type and stage would be shown as a small, sharp red point surrounded by dark blue, while a blurred, desaturated region indicates uncertainty about the patient's true location. Combining the diverse data types results in a multimodal map that accumulates all of the signals. Reading out the peak (white circle) indicates the individual's most probable subtype and stage.

subtype. By contrast, Apolipoprotein E allele E4 (ApoE4) status makes some subtypes of dementia more likely but provides no information about a patient's stage. Information about Aβ burden from a PET scan (or blood test) will be weakly predictive at early disease stages, then provide information that can help pinpoint dementia type, with some indication of stage, but thereafter only provide weak information about a patient's stage.

How should we make use of this information? The space of possible biomarkers, subtypes and stages has the potential to be overwhelming. A simple approach for human decision-making is therefore to set thresholds on specific signals, and then sequentially examine a series of markers to see if they are out of range. This is relatively easy to implement (a series of "If… then…" rules that form a decision tree), but can compromise the sensitivity and specificity of clinical decisions. In particular, single biomarkers are unlikely to capture the whole disease process (Fig. 1), so clinical outcomes rarely depend on a single measurement. We therefore need to consider interactions between different signals. Doing this with a decision tree rapidly becomes complex when there are multiple variables (e.g., age, sex, co-pathologies, amyloid, cognition) – creating dense 'branches' to capture all of the possible outcomes for each decision. Moreover, medical information is subject to uncertainty (i.e., it is not perfect): fluctuations in signals and measurement error mean that an erroneous decision from rigid linear cutoffs at an early binary (yes/no) decision stage could lead to a patient being fundamentally misclassified. Finally, fixed thresholds can be particularly problematic for patients whose background characteristics (e.g., ethnicity) are not well represented in the normative samples used to establish threshold values [6]. Together these limitations necessitate a more probabilistic approach that simultaneously combines information from different signals – i.e., a multimodal approach.

Machine learning (ML) methods are inherently well suited to identifying patterns in large, multi-dimensional and multi-modal datasets. They can learn optimal boundaries from data, accommodating multiple predictors in ways that are robust to the uncertainties inherent in clinical data. Given a sufficiently large sample, they will approximate the functions linking measured data and outcomes. In particular, by analysing many patient records composed of different biomarkers and clinical labels, they will learn the probabilistic (data-driven) relationships between markers and outcomes. In this way, the conceptual framework described by Fig. 1 can be made explicit by learning from clinical data. The specific approach taken depends on the data types and the availability and/or certainty of clinical labels (i.e. diagnosis). When clinical labels are available and reliable, associating data with these labels (supervised learning) provides a reliable way to derive a patient's most likely dementia type and clinical stage. In other circumstances (e.g., lack of reliable labels), unsupervised or semi-supervised methods can facilitate the discovery of groupings in the data ('latent classes') that reveal new insights into a patient's specific dementia subtype or clinical stage. This later approach has particular potential to discover new subtypes and/or distinct stages that can, in turn, be related to new biological insights from specific biomarkers (e.g. neuroinflammation, blood, proteomics markers) to improve diagnostic crieria [7].

The viability of these ML approaches has improved dramatically over the past decade thanks to the systematic accumulation of large-scale, high quality data resources. Initiatives such as the AD Neuroimaging Initiative (ADNI) [8], the US National Alzheimer's Coordinate Center (NACC) repository [9], the AD Data Initiative (ADDI) [10], and the Dementias Platform UK (DPUK) [11] have opened up the potential to understand how multimodal markers of ADRD are expressed across large populations of individuals. In particular, carefully curated, multivariate data with clinical labels provides a rich workspace with which to understand the presentation and progression of different forms of dementias using AI methods. How can we leverage these data to improve the search for new treatments?

2. Improving clinical trial sensitivity

The past decade has seen remarkable progress in developing therapeutics that target biological markers associated with AD (e.g., β-Amyloid), but with disappointing results in terms of altering functional (cognitive) symptoms of the disease. Therapies designed to reduce Amyloid production through β-secretase enzyme inhibition (BACE inhibitors) showed efficacy in lowering Amyloid levels within the central nervous system (e.g., Lanabecestat [12], Verubecestat [13,14], Atabecestat [15]), but were not found to positively impact cognition. Other approaches have used Aβ immunotherapies to target the accumulation of Amyloid-beta plaques within the central nervous system. While these have proved effective in engaging their biological targets and reducing β-Amyloid (e.g., Aducanumab [16], Gantenerumab [17]), only two—Lecanemab [18] and Donanemab [19]—slowed down the rate of cognitive decline. What is responsible for this discrepancy between impact on biomarker levels of β-Amyloid and functional symptoms of the disease (cognition)?

There are several logical possibilities: (1) reductions in Aβ may have been insufficient; (2) the chosen biomarkers may be poor surrogates for the disease; (3) cognitive measures (trial endpoints) may lack sensitivity, introducing variation that masks the treatment effect; (4) differences between patients (heterogeneity) may introduce variation that makes it harder to detect true benefits from a treatment – i.e., between-patient variability masks the treatment effect. Here, we focus on how ML approaches help tackle the last of these possibilities by derisking clinical trials and enhancing their efficiency and efficacy.

In a recent study, we leveraged a multimodal ML approach [20] to determine whether precision stratification can reveal a treatment effect in a trial judged to be futile. The study re-examined data from the AMARANTH phase 2/3 trial (ClinicalTrials.gov ID: NCT02245737) of the BACE1 inhibitor Lanabecestat (AZD3293, LY3314814). While the trial had shown successful lowering of β-amyloid, there was no statistically reliable slowing of cognitive decline (the study's primary endpoint) [12]. To determine whether this was due to patient heterogeneity (Fig. 2a), we applied a previously-trained ML model [21–23] to patient data at baseline (i.e., before administration of the drug or placebo). The model used the standard clinical data collected in the trial (structural MRI, florbetapir PET (β-amyloid), ApoE4) to derive an individualised prediction for each patient's future cognitive health. This AI-guided marker predicts progression from early stages of disease (Mild Cognitive Impairment and even pre-symptomatic, Cognitive Normal) to AD more precisely than standard clinical assessments [23] and biomarkers typically used in clinical trials [22]. Thereby, patients were classified as either 'stable' (i.e., unlikely to deteriorate) or 'progressive', where progression was identified as 'slow' or 'rapid'.

Reanalysing the trial data based on stratified subgroups identified an effect on cognitive outcomes for 'slowly progressive' patients. In particular, the stratified analysis revealed a 46 % slowing of cognitive decline as measured by the Clinical Dementia Rating – Sum of Boxes scale (the original primary endpoint for the study) that was specific to patients on a slowly progressive trajectory. Importantly, the model had been trained and optimised on an entirely separate data set (research data from ADNI) with the task of predicting a patient's future dementia trajectory. The fact that this generalised to a new context for a different task (identifying the effectiveness of a drug) suggests the ML model identifies biologically-meaningful clusters of patients. In particular, 'slowly progressive' patients appear to be functionally responsive to a therapeutic that reduces β-amyloid, while rapidly progressive patients do not.

How should we understand 'slowly progressive' and 'rapidly progressive' patient groups in relation to the stage-by-subtype space (Fig. 1b)? Do these statistically-inferred groups relate to biologically meaningful distinctions? It is logically possible that 'slow' vs. 'rapid' form different subtypes of Alzheimer's disease (i.e., different positions along the *y*-axis), and/or that they represent different stages in the

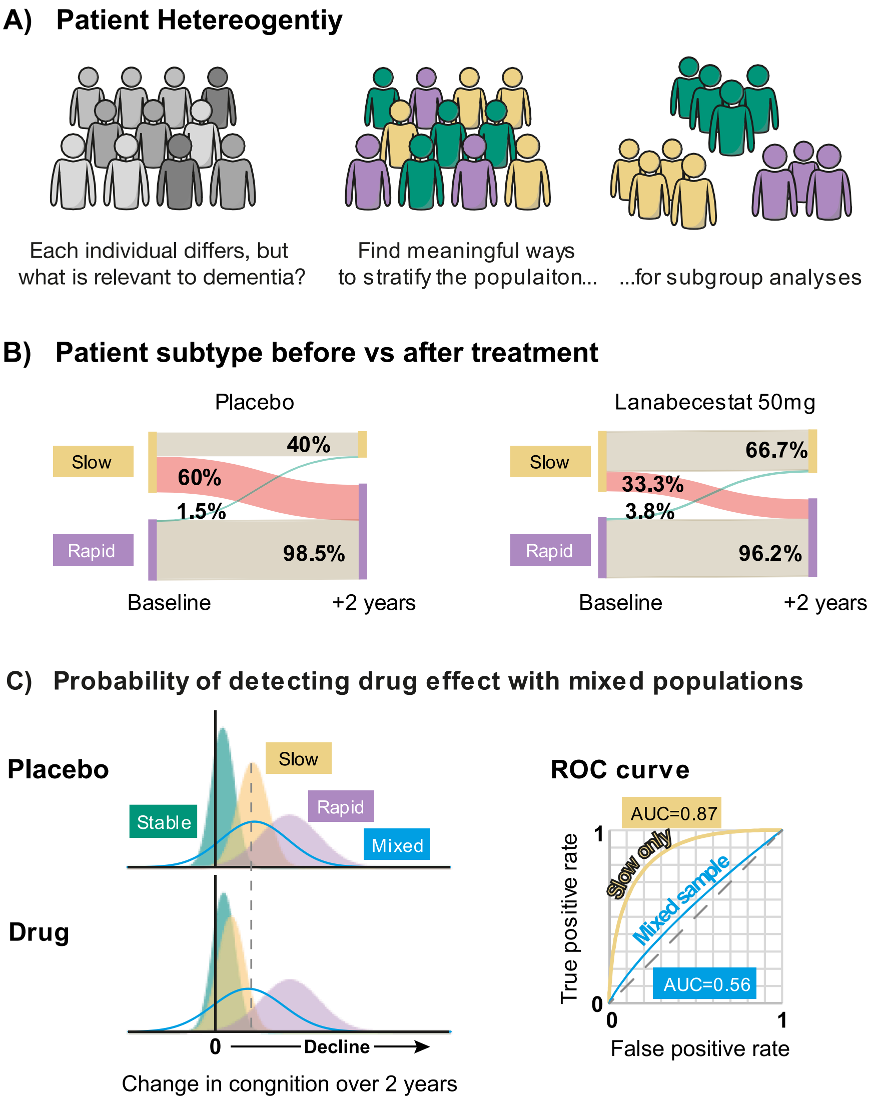


Fig. 2. A) Illustration of the problem of patient heterogeneity. We know individuals differ, but we need to identify the axes along which to segment them into meaningful subtype groupings. B) Alluvial plots reproduced from Vaghari et al [20] illustrating transitions between 'Slow' and 'Rapid' patient classifications at the start (baseline) and end (+2 years) of the AMARANTH trial. The width of the line indicates the proportion of patients in the category. C) Modelling the effects of mixed populations on the ability to detect the effect of a drug that only benefits 'Slow' progressive patients. The left portion of the figure represents probability density for the change in cognition over a two year period. The right portion of the plot shows an ROC curve considering the whole sample (blue curve) or limiting it to comparing the 'Slow' progressive patients (yellow curve). We quantify performance in terms of the Area Under the Curve, where 0.5 represents chance and 1 perfect performance.

progression of the disease (different positions along the *x*-axis). A deeper dive into biomarkers associated with different subtypes could help determine this definitively: for instance, conducting in depth assays on a wide range of multi-omics, cognition, co-pathologies, neuro-inflammation, neuroprotective markers to determine whether biologically distinct mechanisms underlie AI-identified subtypes. Understanding interactions between these factors (as extracted and learned by the AI models) will allow us to determine precisely different subtypes (and potentially discover new ones) that may have different underlying mechanisms. Nevertheless, longitudinal data from a real-world memory clinic [23] and monitoring progression across the duration of the AMARANTH study [20] suggest that slow vs. rapid progressive groups relate to different disease stages. In particular, reclassifying trial participants at the end of the study indicated different probabilities of transitioning from 'slow' to 'rapid' (Fig. 2b). For patients in the placebo group, sixty percent of the patients identified as 'slow' at baseline had transitioned to 'rapid' at the conclusion of the study – i.e., suggestive of a change of dementia stage rather than subtype. For those given the highest dose of Lanabecestat, only a third of patients transitioned from 'slow' to 'rapid' at the conclusion of the study, suggesting that lowering β-Amyloid slows down the progressive nature of AD. This analysis provides an initial indication that prognostic scores derived from multimodal AI models could serve as clinical trial endpoints, delivering multimodal markers that are more sensitive for testing treatment effects than single modality markers alone (e.g. cognitive tests).

Further, this use of longitudinal datasets points to a broader opportunity to develop machine learning approaches for precise patient stratification; that is, modelling longitudinal data to predict individualized trajectories, rather than relying on risk factors at population level or progression rates derived from previous studies. The stratification for the AMARANTH trial was a single 'snapshot' classification approach – using only the data from single time points, rather than modelling the timeseries of clinical measurements. Developing such models would be extremely useful for understanding neurodegenerative conditions (e.g [24]), capturing and predicting disease stages (beyond Aβ and tau deposition [7]), identifying key biomarkers per stage, and would inform both the duration of clinical trials and the timing of interventions.

Finally, in reusing data from a historical trial, our assessment of the use of AI in clinical trials is, by definition, post-hoc. The ML models had not been created when the study was originally designed, so their use did not form part of the statistical analysis plan. Future prospective validation of AI tools (e.g., formally locking the subgroup definition before trial initiation) will be important to build confidence for their widespread adoption. Nevertheless, it is important to note that the ML model's parameters were blind to the trial data (i.e., complete independence of data sets) and that the model was given no information about the outcome of the trial (i.e., only baseline data before treatment). Next, we discuss how the findings can be instructive in the design of future clinical trials.

2.1. Modelling trial sensitivity

From the reanalysis of the AMARANTH trial, it is clear that identifying patient subgroups has the potential to increase the sensitivity of a trial to reveal a therapeutic effect that would otherwise be masked by differences between the subtype or stage of patients within a larger sample. To illustrate a generalised model for this process, consider the simulations presented in Fig. 2c. We model the change in cognition over a two-year period by distributions (probability density functions) for three population subgroups. 'Stable' patients (green) have little change in cognitive function, while 'Slow' (yellow) and 'Rapid' (purple) that involve increasing levels of cognitive decline, where between-patient variability increases (i.e. wider spread) as cognitive decline increases. The level of decline across the whole population (i.e., the mix of different subtypes) is shown by the solid blue curve – that is, the envelope of the whole population.

Consider what happens under administration of a drug that is only effective in reducing decline for the 'Slow' patient subtype (bottom portion of the figure). While the difference in the distributions of decline for patients on vs. off the drug is clear for 'Slow' patients (contrast the position of yellow distributions between top and bottom), at the whole population level (blue curves) it is much less obvious. We quantify using the Area Under the Curve (AUC) in a Receiver Operating Characteristic (ROC) analysis, a standard method for a diagnostic test (Fig. 2c). If a trial includes only 'Slow' patients, the treatment produces clear difference between the measured samples (AUC=0.87) while this is much harder to detect for a population with equal proportions of 'Stable', 'Slow' and 'Rapid' patients (AUC=0.56). In a clinical trial sample, the mix of underlying subpopulations is typically unknown – introducing random variability into the trial and making it harder to detect a true positive. In the case of the AMARANTH trial, the patient selection process resulted in a trial sample that subsequent analysis revealed was composed of approximately one third of 'slowly progressive' patients, two thirds 'rapidly progressive' and only a handful of 'stable' patients.

It is important to understand that the choice of patients with slowly progressive dementia is illustrative. Lanabecestat has a therapeutic action related to β-amyloid, making it suitable for patients in earlier stages of AD. A therapeutic targeting tauopathy stages may be more suited to rapidly progressive patients. Moreover, the labels 'stable', 'slow' and 'rapid' themselves are categorical descriptors derived by learning from disease trajectories in patient cohorts used for model training. Underpinning the labels is a continuous prognostic score that we derived from a multimodal AI model and binned into different groups based on normative data (considering data over 3-years). However, like any metric, there is inherent uncertainty in the measurements, and a small difference in prognostic score can lead to a change in a patient's categorisation. Scores therefore represent an estimate of the patient's state, subject to uncertainty in the measurements and the performance of the ML model. This uncertainty can be estimated using statistical techniques [25] and potentially reduced as biomarker precision improves to the point at which AI enables the integration of signals to precisely pinpoint an individual to within the stage-by-subtype space (Fig. 1c).

2.2. AI generalizability

Understanding how well results from a clinical trial generalize to new settings and populations is critical for any study. The clinical and scientific communities have accumulated knowledge to evaluate how well a particular study sample (e.g., a group of patients recruited from medical centres in North America and Europe) models broader populations (e.g., patients from around the globe) by considering a range of biological and demographic factors. While far from perfect, such human intuitions into which differences 'matter' for generalization underpin regulatory frameworks and commercial contracting.

Introducing AI tools into patient for selection can complicate understanding of how well a study result will generalize, particularly if the AI model is an 'opaque (black) box' model whose rationale for selecting specific individuals is not transparent. In particular, an AI model could learn a reliable association between multiple data features and a particular clinical presentation that works well for a specific set of training and test data. However, if we cannot map the AI engine's operation to our intuitive understanding of the features used by the model it will be hard to evaluate how well the model will generalize to new data sets, in different contexts, geographies and for patients with different demographic backgrounds.

When developing the ML engine used for reanalysis of the ANMARANTH trial, we used two principal ways to address the challenge of AI generalizability. First, we adopted a metric learning approach that is "interpretable-by-design", meaning that the model's decision-making process (i.e., the features used by the model and their weights) can be traced and understood transparently, so the operation of the AI tool can be fully understood. Second, we tested generalisation performance by evaluating a family of trained ML engines on different data sets, demonstrating reliable stratification performance for data obtained in different research studies, as well as real-world memory clinics, from North America, Europe and Asia [21–23]. This provides reassurance that using this ML engine for clinical stratification is robust to different populations and contexts, while further work is ongoing to evaluate this family of models on data from non-AD dementias and underrepresented patient groups.

3. Improving clinical trial efficiency

We have seen how AI-guided patient selection can increase the sensitivity with which therapeutic efficacy is measured. This has direct implications for the efficiency of clinical trials: if we have a more sensitive measure, we need fewer patients to assess the therapeutic to a standard statistical significance threshold (e.g., $\alpha < 0.05$). In this section, we consider how AI methods can be used to optimise human participation in clinical trials (Fig. 3a).

We start with trial design considerations related to sample size. We conducted statistical power analyses on the results from the AMARANTH trial with- and without- AI-guided patient stratification [20]. The results (Table 1) show dramatic reductions in the number of patients that need to be included within a trial for 90 % statistical power to a chosen level of statistical significance.

Reducing the sample size has obvious potential to reduce the costs of a trial. The per patient costs of clinical trials for drug development in the central nervous system are around $40,000 [26]. Depending on trial design, using stratified patient samples translates into potential cost savings of $54–$126 million per trial (Table 1), underscoring the financial impact of AI-guided stratification. Higher statistical power provides the opportunity for more agile trial designs. For instance, testing a broader range of dosing regimen or using adaptive trial designs [27,28] that enable optimization of trial parameters based on interim results.

Repowering a trial in this way requires a high degree of confidence

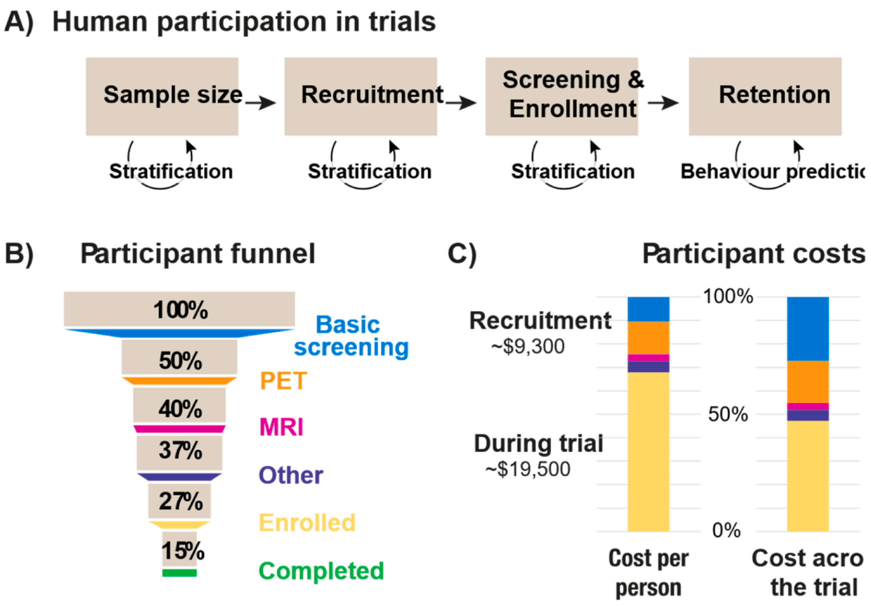


Fig. 3. A) Overview on human participation in clinical trials and how AI can help. B) Patient recruitment screening flow modelled on two AD trials [32]. Of 100 people considered for a trial, only 15 on average complete the full protocol. C) Estimated costs related to procedures in the trial (based on US estimates). Basic screening and cognitive testing estimated at $3056, PET $4000, MRI $850, Other (e.g. genetics, blood) $1400. Costs during the trial encompass multiple MRIs, PET, cognitive tests, bloods / CSF, participant compensation and travel.

Table 1
Power analyses based on results from the AMARANTH trial showing the number of participants required to test lanabecestat 50 mg vs. placebo with 90 % statistical power (1-β) at different thresholds for statistical significance (α). The table contrasts the original sampling technique to an AI-stratified approach, producing sample sizes around 90 % smaller.

	α = 0.05	α = 0.01	α = 0.001
Original	1524	2396	3520
Stratified	164	234	328
Reduction	89 %	90 %	91 %

that a particular group of patients has the right profile for a specific therapeutic. This may be justified by earlier phases of clinical development, but, in many circumstances, there may be uncertainty about which subgroups would benefit most from the treatment. In these cases, a trial could be designed to sample different subgroups in a pre-planned way. For instance, using stratified randomisation [29] to ensure that there is a balance of subgroup types across the different experimental conditions, with a pre-specified definition of the statistical analysis plan for subgroups. Adaptive enrichment might also be deployed, using an interim analysis (overseen by an independent data monitoring committee) to enrich the sample for the group that appears to benefit most from the treatment (e.g., response-adaptive randomization) and/or stopping enrolment for non-responding subgroup(s) (e.g. enrichment using futility rules) [30,31]. Finally, multimodal markers derived from AI-guided tools could provide interim trial endpoints that may be more sensitive than single markers for identifying treatment responders.

Implementing patient stratification could encompass the entire recruitment funnel from initial screening to the point of enrolment. We analysed recruitment across major trials of AD therapeutics [12,14, 17–19,32] involving around 50,000 patients. This showed a screen out rate of 72.9 % - i.e., for every 100 people screened, only 27 of them are enrolled into an Alzheimer's trial.

Fig. 3b breaks down different screening stages in the clinical trial process for AD, modelled on the EMERGE (NCT02484547) and ENGAGE (NCT02477800) trials [32]. In Fig. 3c we further estimate costs (using US prices) based on these trial designs in terms of the raw cost per participant, and the proportionate cost across the whole trial. It is apparent that while the largest cost at the single patient level is incurred during the trial, total expenditure is greater during the recruitment and screening phases. Indeed, more cost can be incurred on the three quarters of patients who do not make it into the trial than those who stay in the trial for two years. Trials would be more cost efficient if more of the patients entering the top of the funnel ended up being recruited to the trial – i.e., getting it right first time. How might AI help?

The stratification models we considered for the AMARANTH trial relied on MRI, PET and genetic information to support prediction and patient stratification. These specialist data are expensive and most suited to making a final decision on participant enrolment based on the highest possible precision in stratification. However, predictions can be derived from less-invasive, lower cost data (e.g. blood tests, cognitive tests, electronic healthcare records). While the features used by a model affect its accuracy [21], the ability to perform a first pass at the top of the funnel that is more accurate than standard screening methods has significant potential to enhance the lower portions of the enrolment funnel.

Implementing stratification at the point of recruitment (i.e., via ethically consented analysis of data contained within electronic health records [33]) and/or at the point of basic screening can reduce patient burden, lower costs, and speed up trial recruitment. In terms of trial efficiency, this will lower the proportion of resources expended on participants who are not enrolled in the study. Pre-screening using AI models to identify likely candidates from electronic health records (prior to formal screening), would increase the probability that a given patient is enrolled in the study before the individual has any active contact with the study and any new tests are run.

Recent advances in blood-based biomarkers are opening new opportunities to lower patient burden and refine inclusion within clinical trials [3]. The plasma-derived assay pTau217, for instance, provides a means of detecting elevated amyloid pathology using a much less costly or invasive measure than PET imaging. Current trial protocols envisage its use as a preliminary measure that is confirmed by PET imaging (e.g., [34]) and recent FDA approval [35] means it may be used without confirmatory PET. Further, recent advances in proteomics and large-scale datasets (e.g. GNPC [36]) hold great promise for the discovery of precision biomarkers and new ADRD subtypes. However, blood markers or proteomics provide single sources of information that should be statistically combined with other markers (e.g. MRI scans, genetics, neuroinflammation, cognition) to provide the best basis to pinpoint an individual's disease subtype and stage (Fig. 1). AI-guided tools synthesising blood and proteomics markers can match patients with different pathology profiles and at different progression stages to the right targets. This has strong potential to facilitate the design of multi-arm trials that test multiple targets against the same placebo group, accelerating and enhancing the efficiency of clinical trials.

Beyond recruitment, trial efficiency is also affected by participant retention in trials. Unfortunately, not every participant that enrols into a clinical trial is able to complete it. A range of factors influence drop out including mortality and morbidity, adverse events, patient burden, mental health comorbidities and worsening of disease state [37,38]. While some of these factors are outside the control of study investigators, AI models designed to predict behaviour and identify risk of dropout have potential to be used for proactive monitoring and early intervention to facilitate continued engagement in a trial. This can be particularly important for underrepresented populations [39] where support for patients and caregivers can reduce drop out from trials. Finally, interpretable AI models allow us to determine key combinations of predictive markers for patient stratification, allowing smarter and more efficient selection of data types to be collected at different clinical trials stages, accelerating trials and enhancing retention rates.

4. Improving therapeutic effectiveness

Successfully executing a clinical trial to show the efficacy of a new therapeutic provides the foundation for regulatory approval. However, it does not guarantee real-world success or reimbursement. In this section we consider how AI approaches can support precision treatment for the best use of therapeutics, as well as key considerations in evaluating AI tools when used in wider populations.

4.1. Adoption in clinical practice

The challenge of accurate diagnoses in dementia has been a key contributing factor to the lack of success in therapeutic development over the past two decades [1]. Yet clinical trials represent a high resource environment that use costly biomarkers. What are the prospects for accurate diagnosis in real world healthcare that is typically less well-resourced than a clinical trial?

A range of AI methods have been explored to enhance dementia diagnosis, with a particular focus on AD and the interpretation of imaging data [40,41]. Such systems could be useful adjuncts to the interpretation of radiological data, but as outlined above (Fig. 1), any one diagnostic marker will give an incomplete picture of the patient's disease state. Multimodal approaches [22] are critical, particularly when used for differential diagnosis to identify subtypes of dementia and their overlap [42].

How could multimodal models be useful when different types of data are available in different clinical settings? We can conceptualise the process as a "Russian Nesting Doll (Matryoshka)" family of models: constructed using the same architecture and functional goal, but using different input features that influence the precision and specificity of the predictions that are produced. We have seen that a model trained with PET, MRI and ApoE4 information can precisely separate patients in the context of a clinical trial [20]. We conceptualise this as the 'core' of the Russian Doll – providing us with the most tightly defined sense of where a patient sits in the space of dementia stage and subtype. Within a secondary care setting, typical data that a clinician has access to would be MRI, a cognitive test, and demographic information. This situates the patient, although less precisely than the core model. Finally, within primary care, models that integrate demographics and cognitive data would provide more information to the clinician than a memory test alone. This approach can benefit from using a 'privileged information' [43] framework where models are trained using richer data than are available at test time, allowing for robust predictions even within more limited inputs. For instance, training a model on MRI, cognitive data and demographics, and applying it in a setting where only cognitive data and demographics are available. Integrating developments in scalable, remotely-collected data through mobile technology (e.g., RADAR-AD consortium [44]) has the potential to enrich AI models and democratize diagnosis in community settings. While we have example instantiations of these models [21], further work is needed to provide full validation across a family of models and build tools that bridge drug discovery with adoption of therapeutics in healthcare.

Designing and implementing models that are compatible with existing clinical workflows is key for advancing dementia therapeutics. The 'best' model is not necessarily the one with highest performance, but rather the one that has most chance of improving clinical pathway decisions that can range from deciding that the patient needs an onward referral to making a refined choice between two specific medicines (i.e., its clinical utility). Ultimately, the AI models that produce most clinical impact may be those that are (a) most easily integrated within Electronic Healthcare Record systems and (b) robust to real world healthcare data that often contains missing or degraded data. Multimodal approaches are inherently more robust than single markers, and statistical approaches based on probabilistic imputation of missing data can further support real world deployment in healthcare [45–47].

Finally, it is worth considering the computing resource requirements associated with running AI models. Large AI models can necessitate large data flows and extensive computation. As AI usage becomes ubiquitous, these considerations may reduce, but currently low-bandwidth data networks and/or lack of dedicated IT hardware in clinical settings can introduce barriers to adoption. These barriers may be particularly acute in lower- and middle-income country settings [48,49].

4.2. Interpretability

Many AI algorithms operate as 'opaque (black) boxes' that make their operations difficult to understand. Deep neural networks can involve many millions of free parameters, making the operation of the system hard to understand. This lack of transparency is a challenge when the system may be informing clinical decision making and treatment planning [50]. Interpretability varies according to the types of machine learning used in a solution, and solutions can be made 'interpretable by design' [42,51]. Where this is not possible, techniques can be used to infer or visualise how the system is operating to make solutions more interpretable. For instance, reverse engineering features to test their clinical relevance (e.g., relationships between biomarkers and cognitive decline), and/or implementing methods based on the integration of concepts [52], attention [53] and logic [54]. As outlined above, interpretability considerations are related to the ability of humans to assess the likelihood of AI model generalizability.

4.3. Algorithmic bias

It is critical to understand the potential for AI models to entrench or even widen existing health inequalities. When AI models are trained on specific data sets, their parameters are tuned to the properties of those data. Clinical trials and research cohorts generally overrepresent majority populations [55,56] creating the potential for models to perform poorly when presented with data from underrepresented groups. This is particularly critical when building models for ADRD given known increased risks in specific racial/ethnic groups [57,58]. Further, when AI models have been trained using diagnostic labels derived by clinicians, there is potential for AI to amplify biases in human decision-making through the automated use of AI tools at scale. AI methods can be used to detect algorithmic bias (e.g [59]), however rigorous testing of generalization to minority populations, and improving the diversity of the underlying datasets used to train models are critical to longer term efforts. Real-world data sets (as opposed to research cohorts) are advantageous in this regard as they tend to be more broadly representative of the underlying population.

Finally, it is important to consider that models should not be static: clinical standards evolve, data types are refined, and models being applied in real world practice may becoming increasingly divergent from the populations on which they were trained. It is critical therefore to ensure continued model relevance. Interpretability can help with this, but specific procedures can be used to quantify drift (i.e., differences between training and test data sets) [60] so that models can be retrained or recalibrated once deployed. Addressing bias, maintaining model relevance, and improving dataset diversity are all critical to ensuring equitable access to AI-enhanced dementia care.

4.4. Regulatory considerations

As AI tools move from research into clinical practice, regulatory oversight becomes essential. Adopting AI stratification within healthcare is likely to require regulatory approval as a diagnostic technology – either as a companion to a new medicine or as a diagnostic algorithm that would support clinicians in treatment planning. Regulators around the globe are currently grappling with the right way to balance the risks and opportunities of AI for patient benefit. Much of this involves extending thinking about patient safety, the efficacy of solutions, risk stratification, and software lifecycle management that has been part of Software as a Medical Device (SaMD) for two decades (e.g., FDA guidance published in 2005). Specific additional elements relate to the adoption of Good Machine Learning Practice principles outlined jointly between US, UK and Canadian regulators [61]. These are designed to ensure that models are robust, well validated, secure, transparent and that performance is actively monitored once products are introduced into the market (i.e., post-market surveillance). Where AI algorithms can update themselves, or produce highly variable outputs (e.g., natural language), regulators require performance monitoring plans to ensure ongoing safety and effectiveness as AI models evolve under real-world usage.

In considering the Regulatory approach, there are unresolved questions about the relationships between AI tools used as part of a clinical trial, and those required once a new medicine has been approved as safe and effective. In particular, if the success of the trial depends in part on the use of an AI-guided stratification tool, does the tool become essential for prescribing the therapeutic in healthcare settings? Here, we see significant potential for "Russian doll" families of models that are related, but use lower-cost less-invasive types of data from those typically collected within a clinical trial (e.g. blood tests and cognition instead of PET scans). This could result in the use of simple, interpretable models that aid clinicians to assign patients to the right treatment and can be easily related to quantities already known and trusted by Regulators and clinicians.

Ultimately Regulatory decisions will be guided by evidence that the use of a particular tool is safe and effective. Because clinical trials already collect multimodal data, including lower cost data, there is potential to simultaneously validate the use of 'gold standard' stratification tools alongside tools that have practical use in real world clinical settings. This will necessitate comparing patient inclusion/exclusion criteria across different models, and relating these to outcomes in the placebo and treatment groups. Similar considerations are needed for thinking about AI-guided patient selection at different stages of the recruitment funnel, particularly to ensure that patients with particular demographics are not being systematically excluded from trials.

5. Conclusion

The search for effective treatments for Alzheimer's and related dementias has long been hindered by patient heterogeneity and the lack of sensitive tools for stratification. In this review, we have explored how advances in AI—particularly those leveraging multimodal data—can enhance both the development and deployment of new therapeutics.

Multimodal approaches enable more precise identification of a patient's disease subtype and stage, matching patients based on their pathology profile and stage to targets. This has strong potential to transform clinical trial outcomes by increasing statistical power, improve operational efficiency through adaptive design of multi-arm trials, and pave the way to precision medicine and combination therapies in ADRD. These benefits extend beyond clinical trial settings, opening up pathways to more targeted healthcare in real-world settings.

However, realizing the full potential of AI in dementia care requires that we address key challenges: ensuring model transparency, fairness, and generalizability across diverse populations of patients requiring dementia care, and making tools compatible with clinical workflows and resource constraints.

AI alone will not solve the complex challenge of ADRD. But, when integrated with scientific insight, clinical expertise, and lived experience of patients and caregivers, intelligent analytics can accelerate the discovery and delivery of diagnostics and therapeutics—ultimately transforming dementia care and improving outcomes for individuals and their families worldwide.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zoe Kourtzi, Andrew Welchman report a relationship with Prodromic that includes: board membership.

Acknowledgments

This work was supported by grants to ZK from the Royal Society (INF\R2\202107), Wellcome Trust (221633/Z/20/Z), Alzheimer's Research UK (ARUK-PG2023B-010), RAi UK (EP/Y009800/1). For the purpose of open access, the authors have applied for a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

[1] Kim CK, et al. Alzheimer's Disease: key insights from two decades of clinical trial failures. J Alzheimers Dis 2022;87:83–100.

[2] Blennow K, Zetterberg H. Biomarkers for Alzheimer's disease: current status and prospects for the future. J Intern Med 2018;284:643–63. https://doi.org/10.1111/joim.12816. Preprint at.

[3] Frisoni GB, et al. New landscape of the diagnosis of Alzheimer's disease. Lancet 2025;406:1389–407. https://doi.org/10.1016/S0140-6736(25)01294-2. Preprint at.

[4] Gaugler JE, et al. Characteristics of patients misdiagnosed with Alzheimer's disease and their medication use: an analysis of the NACC-UDS database. BMC Geriatr 2013;13:137.

[5] Ferreira D, Nordberg A, Westman E. Biological subtypes of Alzheimer disease: a systematic review and meta-analysis. Neurology 2020;94:436–48.

[6] Molina-Henry DP, et al. Racial and ethnic differences in plasma biomarker eligibility for a preclinical Alzheimer's disease trial. Alzheimers Dement 2024;20:3827–38.

[7] Jack CR, et al. Revised criteria for diagnosis and staging of Alzheimer's disease: alzheimer's association workgroup. Alzheimer 19s Dement 2024;20:5143–69.

[8] adni.loni.usc.edu.

[9] Besser, L. et al. Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. www.alzheimerjournal.com (2018).

[10] McHugh CP, Clement MHS, Phatak M. AD Workbench: transforming Alzheimer's research with secure, global, and collaborative data sharing and analysis. Alzheimer 19s Dement 2025;21.

[11] Bauermeister S, et al. The dementias platform UK (DPUK) Data Portal. Eur J Epidemiol 2020;35:601–11.

[12] Wessels AM, et al. Efficacy and safety of lanabecestat for treatment of early and mild Alzheimer disease: the AMARANTH and DAYBREAK-ALZ randomized clinical trials. JAMA Neurol 2020;77:199–209.

[13] Egan MF, et al. Randomized trial of verubecestat for mild-to-moderate Alzheimer's disease. N Engl J Med 2018;378:1691–703.

[14] Egan MF, et al. Randomized trial of verubecestat for prodromal Alzheimer's disease. N Engl J Med 2019;380:1408–20.

[15] Novak G, et al. Long-term safety and tolerability of atabecestat (JNJ-54861911), an oral BACE1 inhibitor, in early Alzheimer's disease spectrum patients: a randomized, double-blind, placebo-controlled study and a two-period extension study. Alzheimers Res Ther 2020;12:58.
[16] Sevigny J, et al. The antibody aducanumab reduces aβ plaques in Alzheimer's disease. Nature 2016;537:50–6.
[17] Bateman RJ, et al. Two phase 3 trials of Gantenerumab in early Alzheimer's disease. N Engl J Med 2023;389:1862–76.
[18] van Dyck CH, et al. Lecanemab in early Alzheimer's Disease. N Engl J Med 2023; 388:9–21.
[19] Sims JR, et al. Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. JAMA 2023;330:512–27.
[20] Vaghari D, et al. AI-guided patient stratification improves outcomes and efficiency in the AMARANTH Alzheimer's Disease clinical trial. Nat Commun 2025;16:6244.
[21] Giorgio J, Landau SM, Jagust WJ, Tino P, Kourtzi Z. Modelling prognostic trajectories of cognitive decline due to Alzheimer's disease. Neuroimage Clin 2020; 26.
[22] Giorgio J, et al. A robust and interpretable machine learning approach using multimodal biological data to predict future pathological tau accumulation. Nat Commun 2022;13.
[23] Lee LY, et al. Robust and interpretable AI-guided marker for early dementia prediction in real-world clinical settings. EClinicalMedicine 2024;74.
[24] Salvadó G, et al. Disease staging of Alzheimer's disease using a CSF-based biomarker model. Nat Aging 2024;4:694–708.
[25] Seoni S, et al. Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023). Comput Biol Med 2023;165. https://doi.org/10.1016/j.compbiomed.2023.107441. Preprint at.
[26] Moore TJ, Heyward J, Anderson G, Alexander GC. Variation in the estimated costs of pivotal clinical benefit trials supporting the US approval of new therapeutic agents, 2015-2017: a cross-sectional study. BMJ Open 2020;10.
[27] Oikonomou EK, et al. An explainable machine learning-based phenomapping strategy for adaptive predictive enrichment in randomized clinical trials. NPJ Digit Med 2023;6.
[28] Fountzilas E, Tsimberidou AM, Vo HH, Kurzrock R. Clinical trial design in the era of precision medicine. Genome Med 2022;14:101.
[29] Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. J Clin Epidemiol 1999;52.
[30] Tu, Y. & Renfro, L.A. Latest developments in "adaptive enrichment" clinical trial designs in oncology. Ther Innov Regul Sci vol. 58 1201–13 Preprint at https://doi.org/10.1007/s43441-024-00698-3 (2024).
[31] Robertson DS, Lee KM, López-Kolkovska BC, Villar SS. Response-adaptive randomization in clinical trials: from myths to practical considerations. Stat Sci 2023;38:185–208.
[32] Budd Haeberlein S, et al. Two randomized phase 3 studies of Aducanumab in early Alzheimer's disease. J Prev Alzheimers Dis 2022;9:197–210.
[33] Boustani M, et al. Passive digital signature for early identification of Alzheimer's disease and related dementia. J Am Geriatr Soc 2020;68:511–8.
[34] Rafii MS, et al. The AHEAD 3-45 study: design of a prevention trial for Alzheimer's disease. Alzheimers Dement 2023;19:1227–33.
[35] FDA. FDA clears first blood test used in diagnosing Alzheimer's disease. https://www.fda.gov/news-events/press-announcements/fda-clears-first-blood-test-used-diagnosing-alzheimers-disease (2025).
[36] Imam F, et al. The Global Neurodegeneration Proteomics Consortium: biomarker and drug target discovery for common neurodegenerative diseases and aging. Nat Med 2025;31:2556–66.
[37] Burke SL, et al. Factors influencing attrition in 35 Alzheimer's disease centers across the USA: a longitudinal examination of the National Alzheimer's coordinating center's uniform data set. Aging Clin Exp Res 2019;31:1283–97.
[38] Raman R, et al. Pre-randomization predictors of study discontinuation in a preclinical Alzheimer's disease randomized controlled trial. J Prev Alzheimers Dis 2024;11:874–80.
[39] Gilmore-Bykovskyi AL, et al. Recruitment and retention of underrepresented populations in Alzheimer's disease research: a systematic review. Alzheimers Dement 2019;5:751–70.
[40] Myszczynska MA, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol 2020;16:440–56.
[41] Martin SA, Townend FJ, Barkhof F, Cole JH. Interpretable machine learning for dementia: a systematic review. Alzheimers Dement 2023;19:2135–49.
[42] Xue C, et al. AI-based differential diagnosis of dementia etiologies on multimodal data. Nat Med 2024;30:2977–89.
[43] Vapnik V, Izmailov R, Gammerman A, Vovk V. Learning using privileged information: similarity control and knowledge transfer. J Mach Learn Res 2015;16: 2023–49.
[44] https://www.radar-ad.org.
[45] Giorgio J, et al. A robust harmonization approach for cognitive data from multiple aging and dementia cohorts. Alzheimer 19s Dement: Diagn Assess Dis Monit 2023; 15.
[46] Zhang F, et al. Recent methodological advances in federated learning for healthcare. Patterns 2024;5. https://doi.org/10.1016/j.patter.2024.101006. Preprint at.
[47] Shadbahr T, et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. Commun Med 2023;3:139.
[48] López DM, Rico-Olarte C, Blobel B, Hullin C. Challenges and solutions for transforming health ecosystems in low- and middle-income countries through artificial intelligence. Front Med 2022;9:958097.
[49] Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. NPJ Digit Med 2022;5.
[50] Ravi D, et al. Deep learning for health informatics. IEEE J Biomed Health Inf 2017; 21:4–21.
[51] Allen GI, Gan L, Zheng L. Interpretable machine learning for discovery: statistical challenges and opportunities. Annu Rev Stat Appl 2024;11:97–121.
[52] Wei Koh P, et al. Concept bottleneck models. In: Proceedings of the 37th International Conference on Machine Learning; 2020. p. 5338–48 (Proceedings of Machine Learning Research 119).
[53] Vaswani A, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17) 6000–6010; 2017.
[54] Ciravegna G, et al. Logic Explained Networks. Artif Intell 2023;314:103822.
[55] Bibbins-Domingo K, Helman A. Improving representation in clinical trials and research. Washington, D.C.: National Academies Press; 2022. https://doi.org/10.17226/26479.
[56] Alarcón Garavito GA, et al. Enablers and barriers of clinical trial participation in adult patients from minority ethnic groups: a systematic review. Trials 2025;26:65.
[57] Shiekh SI, et al. Ethnic differences in dementia risk: a systematic review and meta-analysis. J Alzheimers Dis 2021;80:337–55.
[58] Chen C, Zissimopoulos JM. Racial and ethnic differences in trends in dementia prevalence and risk factors in the United States. Alzheimers Dement 2018;4: 510–20.
[59] Smith, J., Holder, A., Kamaleswaran, R. & Xie, Y. Detecting algorithmic bias in medical-AI models using trees. ArXiv 2312.02959 http://arxiv.org/abs/2312.02959 (2024).
[60] Abroshan, M. et al. Safe AI for health and beyond – Monitoring to transform a health service. (2023).
[61] Good Machine Learning Practice for Medical Device Development: Guiding Principles. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (2021).

Glossary of Terms

Clinical Label: Diagnostic category assigned to a patient (e.g., MCI, AD).
Machine Learning (ML): Algorithms that learn patterns from data to make predictions/ decisions.
Metric Learning: ML approach that separates similar patients that are spatially close to each other by learning discriminable prototypes.
Probability Density Function (PDF): A curve describing how likely different outcome values are within a population.
SaMD (Software as a Medical Device): Software intended for medical purposes.
Supervised Learning: ML algorithms trained on labeled data.
Un- /Semi-Supervised Learning: ML algorithms that learns from unlabeled or a mix of labeled and unlabeled data.

The Journal of Prevention of Alzheimer's Disease 13 (2026) 100396

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

AI-augmented frameworks for enhancing Alzheimer's disease clinical trials: A memory clinic perspective

Francesco K. Yigamawano [a,b], Aubrey R. Odom [a], Chonghua Xue [a,c], Hemant K. Pandey [d], Vijaya B. Kolachalama [a,e,f,*]

[a] Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA
[b] Department of Biomedical Engineering, University of South Carolina, Columbia, SC, USA
[c] Department of Electrical & Computer Engineering, Boston University, Boston, MA, USA
[d] Brain & Spine Center of Arizona, Chandler, AZ, USA
[e] Department of Computer Science, Boston University, Boston, MA, USA
[f] Faculty of Computing & Data Sciences, Boston University, Boston, MA, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Clinical trials
Language models

ABSTRACT

Alzheimer's disease (AD) clinical trials continue to face major hurdles in patient identification, resulting in delayed timelines, underpowered studies, and escalating costs. This perspective explores these challenges through the lens of a memory clinic, where hundreds of cases often translate into only a handful of enrollments. We highlight the potential of artificial intelligence (AI) to address this gap by powering chatbots for awareness and pre-screening, decision support tools for case identification, and algorithms for matching patients to trial-specific criteria, automating and streamlining the recruitment process. We also examine critical considerations in developing such AI-driven tools, including data standardization, privacy protections, and ethical safeguards. With thoughtful implementation, these innovations could accelerate more inclusive and efficient AD trials, ultimately bringing therapies to patients faster.

1. Introduction

Consider a representative memory clinic tasked with managing a consistent influx of patients presenting with early cognitive impairments, including memory deficits or disorientation. Although assessing numerous potential participants annually, such clinics typically secure enrollment for only a limited number in Alzheimer's disease (AD) clinical trials. These recruitment difficulties arise from multifaceted practical barriers (Fig. 1): patients and their families often attribute mild symptoms to normative aging processes, thereby postponing medical consultation and contributing to diagnostic delays [1]; overburdened neurologists and other providers are constrained by time limitations, impeding thorough manual examination of electronic health records (EHRs) to determine eligibility amid broader system resource constraints; diagnostic obstacles, such as the requirement for costly biomarker evaluations, discourage patient referrals; and rigorous trial protocols, mandating precise cognitive thresholds or genetic indicators, contribute to elevated screen-failure rates [2]. These impediments reflect wider systemic challenges in AD therapeutic advancement, wherein a large majority of qualified candidates remain unreferred or uninvolved in trials [3], culminating in investigations that exhibit protracted enrollment periods, extended durations, and heightened expenditures relative to other medical domains.

Artificial intelligence (AI) provides a practical, actionable path forward by automating and enhancing key steps in the trial recruitment process [4–7]. In the memory clinic scenario, AI tools based on large language models (LLMs) and other analytical tools have the potential to integrate diverse data sources, EHRs, neuroimaging, genetics, and even digital biomarkers from apps, to identify and match patients efficiently. This perspective uses the memory clinic example to dissect these challenges and outline a vision for AI integration, guiding readers toward implementing more precise, inclusive, and accelerated AD trial recruitment to advance therapeutic breakthroughs. To facilitate user comprehension, we present a collection of technical terms with brief explanations in Table 1.

* Corresponding author at: 72 E. Concord St, Evans 636, Boston, MA 02118.
E-mail address: vkola@bu.edu (V.B. Kolachalama).

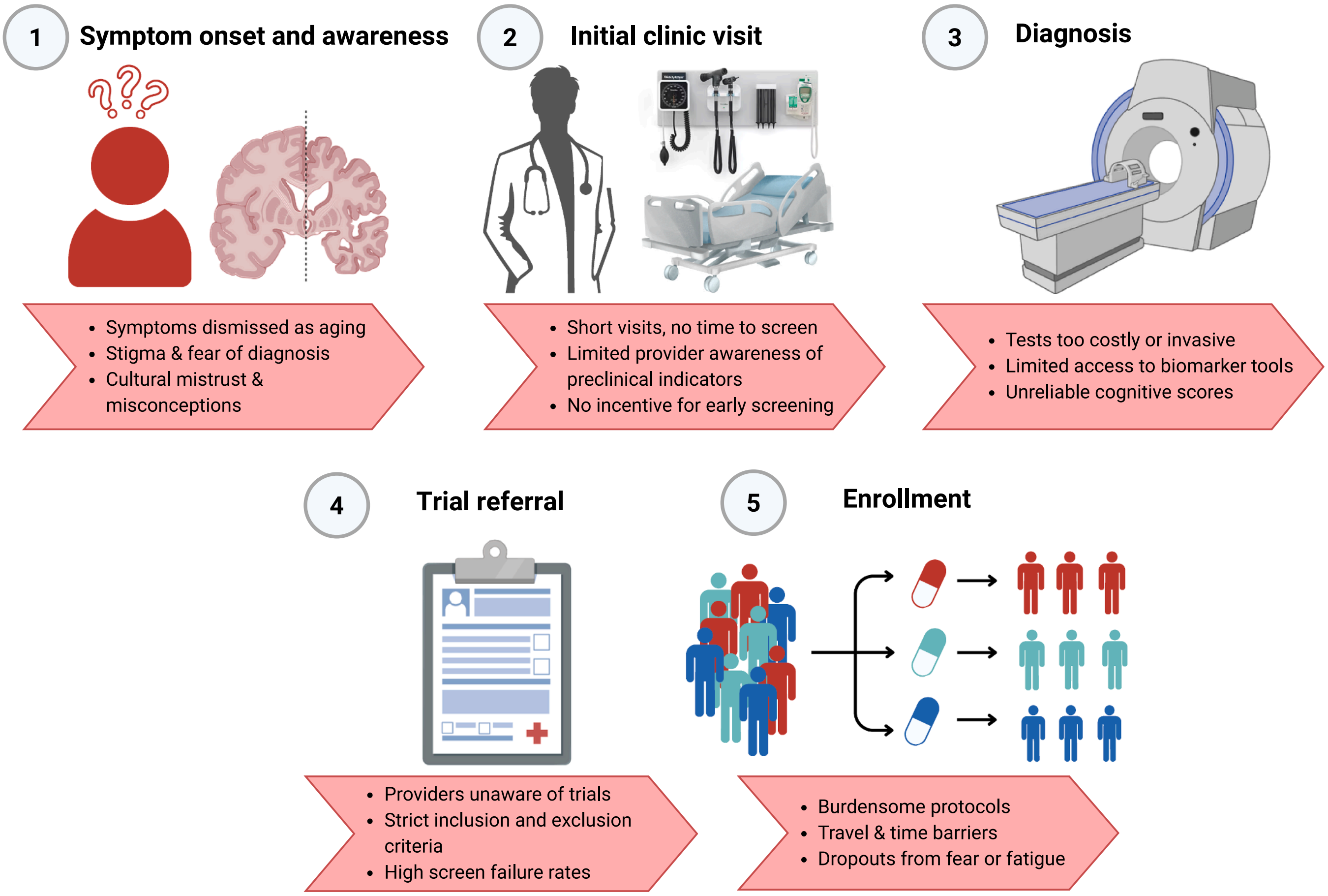
https://doi.org/10.1016/j.tjpad.2025.100396
Received 4 September 2025; Received in revised form 15 October 2025; Accepted 20 October 2025

2. Systemic barriers to effective clinical trial participant identification in AD

AD clinical trials face interconnected, system-wide barriers that accumulate across the patient journey, from community perceptions to data infrastructure, resulting in enrollment shortfalls (Fig. 1). In our representative memory clinic, these barriers manifest daily: despite seeing hundreds of patients with cognitive concerns annually, only a small fraction might be referred to trials, with even fewer enrolling due to delays, dropouts, and mismatches. Addressing these barriers is crucial for accelerating therapeutic development and ensuring trial populations reflect the diverse demographics affected by AD [8].

Community and healthcare system barriers. In the memory clinic setting, early AD symptoms such as subtle memory lapses or difficulty with daily tasks, are frequently misattributed to normal aging by patients and families [9], leading to delayed recognition and underreporting, particularly in preclinical or mild cognitive impairment (MCI) stages. First, many patients present at a much later stage in the disease continuum, often after significant cognitive decline has already occurred [1]. At that point, they may no longer meet eligibility criteria for early-intervention trials, which increasingly focus on prodromal or preclinical AD [3,8]. This likely postpones consultations, with many patients only seeking help when symptoms become severe, narrowing the window for trial eligibility. Stigma and fear surrounding an AD diagnosis further exacerbate this [10–12], as individuals worry about impacts on independence, employment, or insurability, often minimizing or concealing symptoms to avoid social or professional repercussions. For instance, in underserved communities served by the clinic, cultural misconceptions or historical mistrust of medical research can reduce initial visits [13,14], limiting the potential participant pool from the outset.

Within the healthcare system, clinicians report insufficient time and resources to thoroughly discuss AD with patients [15], especially those without overt symptoms, amid packed schedules. Structural disincentives, such as limited reimbursement for cognitive screenings or referrals, further discourage proactive involvement [16]. Many neurologists lack deep familiarity with preclinical AD indicators, contributing to delayed diagnoses and missed referral opportunities. The perceived absence of effective disease-modifying therapies diminishes the incentive for early screening among both providers and patients. Second, there is a lack of sophisticated investigational tools in routine clinical settings to help stratify patients accurately along the amyloid versus tau pathology axis. While plasma biomarkers are emerging as valuable tools, they are not yet widely adopted or reimbursed, and positron emission tomography (PET) or cerebrospinal fluid (CSF) analyses remain cost-prohibitive or inaccessible. This makes precise phenotyping difficult, thereby limiting appropriate trial assignments and resulting in referral drop-offs, higher costs, and identification of only a fraction of eligible patients. Third, the variability in cognitive testing, both in administration and patient performance, makes it challenging to predict with confidence whether a patient will meet cognitive thresholds for trial inclusion. Factors such as education, cultural background,



1 Symptom onset and awareness

- Symptoms dismissed as aging
- Stigma & fear of diagnosis
- Cultural mistrust & misconceptions

2 Initial clinic visit

- Short visits, no time to screen
- Limited provider awareness of preclinical indicators
- No incentive for early screening

3 Diagnosis

- Tests too costly or invasive
- Limited access to biomarker tools
- Unreliable cognitive scores

4 Trial referral

- Providers unaware of trials
- Strict inclusion and exclusion criteria
- High screen failure rates

5 Enrollment

- Burdensome protocols
- Travel & time barriers
- Dropouts from fear or fatigue

Fig. 1. Challenges in the Alzheimer's disease patient journey to clinical trial enrollment. This infographic outlines the progression of Alzheimer's disease (AD) patients from symptom onset to trial enrollment, highlighting barriers at each step. Symptom onset and awareness involve dismissing symptoms as aging, stigma/fear of diagnosis, and cultural mistrust. Initial clinic visit faces short durations without screening time, low provider awareness of preclinical indicators, and no early screening incentives. Diagnosis encounters costly/invasive tests, limited biomarker access, and unreliable cognitive scores. Trial referral includes provider unawareness of trials, strict eligibility criteria, and high screen failures. Enrollment is impeded by burdensome protocols, travel/time barriers, and dropouts from fear or fatigue.

Table 1
Glossary of technical terms.

Application programming interface (API)	A set of rules and communication standards that allow different software systems to communicate and share data.
Clinical decision support (CDS) system	A technology system that provides healthcare professionals with evidence-based knowledge and patient-specific recommendations to enhance clinical decisions and improve patient care.
Data fragmentation	The scattering of data across multiple disconnected systems, applications, and locations, which makes it difficult to manage, analyze, and integrate data effectively.
Data infrastructure	Systems and resources that enable the collection, storage, management, integration, processing, and accessibility of data. This includes hardware, software, standards, and governance policies.
Differential privacy	A method of adding carefully calibrated noise to the output of an algorithm to protect the privacy of individual data points, ensuring that the presence or absence of any single person's data does not significantly affect the algorithm.
Fairness-constrained modeling	Integration of mathematical fairness criteria directly into the training process of an AI model.
Federated learning	Computational approach where models are trained in a decentralized way without sharing source data.
Fast Healthcare Interoperability Resources (FHIR)	A standard for exchanging electronic health information that makes it easier for different healthcare systems to share data securely and efficiently.
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.
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.
Regex	Regex, short for "regular expression," is a sequence of characters that defines a search pattern.
Retrieval grounding	Connecting a large language model to a verifiable, external knowledge base to generate more accurate, relevant, and trustworthy responses.
Scalability	The ability of a system to handle increased workload efficiently, without prohibitive cost or waiting times.
Structure-code crosschecks	Structure-code crosschecks verify that a program's code matches its intended architectural design to ensure consistency and reduce errors.

testing environment, and examiner differences can significantly impact results, often contributing to high screen-failure rates.

Clinical trial ecosystem barriers. Most physicians remain unaware of ongoing AD trials or lack the detailed knowledge to refer patients effectively [17,18], treating trials as an afterthought rather than a viable care option. In the memory clinic, this translates to ad hoc referrals, with neurologists relying on sporadic emails from sponsors rather than integrated systems, resulting in missed matches for a large fraction of suitable patients. Stringent inclusion and exclusion criteria often demanding biomarker positivity (e.g., amyloid or tau confirmation) and specific cognitive thresholds, exacerbate issues, leading to high screen-failure rates. Exclusion criteria frequently outnumber inclusions, such as barring patients with comorbidities common in older adults, compounding recruitment challenges. Even post-enrollment, participants may withdraw due to burdensome protocols (e.g., frequent visits or invasive monitoring), perceived risks, or logistical hurdles like transportation, further underpowering studies and straining clinic resources.

Data and operational barriers. Data fragmentation and operational inefficiencies pose additional hurdles in the clinic environment. Patient information is often siloed in disparate, unstructured formats across EHR systems, making it labor-intensive to identify candidates efficiently. While standards like Health Level Seven (HL7) and Fast Healthcare Interoperability Resources (FHIR) exist to facilitate integration, inconsistent adoption and varying versions hinder seamless data harmonization, as evidenced in efforts to align sources like the National Alzheimer's Coordinating Center (NACC) or the Alzheimer's Disease Neuroimaging Initiative (ADNI). Current practices rely on manual EHR reviews, which are time-consuming and error-prone; a single neurologist might spend hours weekly scanning records yet overlook key details. Furthermore, AD trials disproportionately enroll participants who are more educated, engaged, and research-positive, leading to underrepresentation of racial and ethnic minorities and underserved populations [19–22]. In the memory clinic, this bias means trials may not capture the full spectrum of AD, with minorities comprising only a small fraction of participants despite higher disease prevalence, limiting generalizability and perpetuating health disparities. Given these inefficiencies, there is a need for automated, scalable solutions to match patients against complex criteria in real-time.

3. Role of AI in enhancing trial readiness

Building upon the barriers outlined in the memory clinic context, AI emerges as a useful tool to enhance trial readiness by directly addressing these challenges through automation, precision, and scalability (Fig. 2). Numerous AI-based solutions for clinical trial recruitment are available and have shown promise in enhancing efficiency, yet the primary hurdles remain their adoption and integration into clinical workflows, influenced by factors such as implementation barriers, lack of uniform standards, and the need for clinician literacy [4,23,24]. In our representative memory clinic, serving approximately 500-1000 patients with cognitive concerns annually, such tools can transform recruitment from a manual, inefficient process yielding only 10-20 enrollments per year to a streamlined system identifying a much larger set of viable candidates, reducing timelines and costs. Below is a phased vision tailored to the clinic, leveraging tools like LLMs and predictive analytics:

Community-level AI for awareness building and pre-screening.

To expand the patient funnel and combat stigma, the clinic could deploy accessible AI tools on its website or app, such as LLM-based chatbots (e.g., built on open-source frameworks like Hugging Face models or commercial platforms) [25–29].

1. Patients or families input symptoms via text or voice; the AI analyzes responses to provide personalized education (e.g., "`The frequent forgetfulness you described may indicate early cognitive concerns rather than just typical aging. Here's a comparison based on common patterns. Disclaimer: This is not a diagnosis. Consult a doctor.`").
2. Integrated LLMs scan local social media or forums (with anonymized data) to detect misconceptions, enabling targeted outreach campaigns (e.g., emails or ads to high-risk demographics like those over 65 in underserved areas). To mitigate risks of LLMs generating content misaligned with clinical intentions, we recommend incorporating human oversight, such as clinician review of LLM outputs, to ensure alignment and accuracy.
3. Anonymized pre-screening via digital biomarkers such as speech analysis apps where users read prompts, and AI detects hesitations or patterns indicative of decline using models trained on neuropsychological data, stratifies risk levels (low/medium/high) while employing federated learning and differential privacy to safeguard identities. High-risk users receive gentle nudges (e.g., "`Schedule a visit for further evaluation.`") with links to clinic appointments or trial info.

Regarding feasibility and timing, while LLMs are mature and widely available with proven applications in healthcare, their immediate

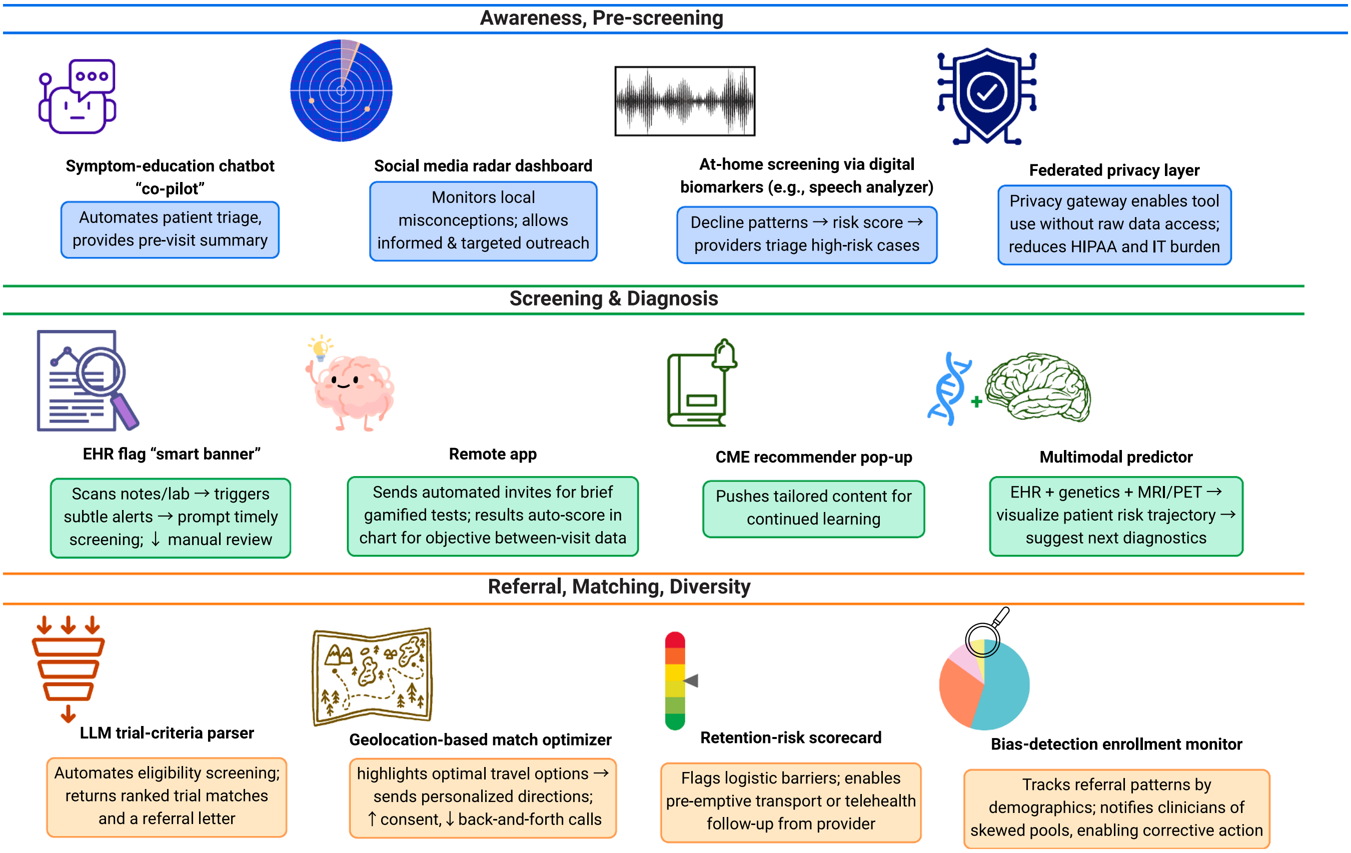

Awareness, Pre-screening
Symptom-education chatbot "co-pilot"
Automates patient triage, provides pre-visit summary
Social media radar dashboard
Monitors local misconceptions; allows informed & targeted outreach
At-home screening via digital biomarkers (e.g., speech analyzer)
Decline patterns → risk score → providers triage high-risk cases
Federated privacy layer
Privacy gateway enables tool use without raw data access; reduces HIPAA and IT burden
Screening & Diagnosis
EHR flag "smart banner"
Scans notes/lab → triggers subtle alerts → prompt timely screening; ↓ manual review
Remote app
Sends automated invites for brief gamified tests; results auto-score in chart for objective between-visit data
CME recommender pop-up
Pushes tailored content for continued learning
Multimodal predictor
EHR + genetics + MRI/PET → visualize patient risk trajectory → suggest next diagnostics
Referral, Matching, Diversity
LLM trial-criteria parser
Automates eligibility screening; returns ranked trial matches and a referral letter
Geolocation-based match optimizer
highlights optimal travel options → sends personalized directions; ↑ consent, ↓ back-and-forth calls
Retention-risk scorecard
Flags logistic barriers; enables pre-emptive transport or telehealth follow-up from provider
Bias-detection enrollment monitor
Tracks referral patterns by demographics; notifies clinicians of skewed pools, enabling corrective action

Fig. 2. AI-driven solutions to overcome barriers in the Alzheimer's disease patient journey for clinical trials. This diagram showcases AI-powered tools across three stages to streamline the Alzheimer's disease (AD) patient journey toward clinical trials. In "Awareness, Pre-screening," it features a symptom-education chatbot for triage and summaries, a social media dashboard for targeted outreach, an at-home biomarker screening tool for risk detection, and a privacy layer for secure data handling. The "Screening & Diagnosis" section includes an EHR alert banner for subtle detections, a remote gamified testing app for objective data, a CME recommender for provider education, and a multimodal predictor for risk visualization using EHR, genetics, and imaging. Finally, "Referral, Matching, Diversity" highlights an AI parser for trial eligibility and referrals, a geolocation optimizer for travel and consent, a retention-risk scorer for proactive support, and a bias monitor to ensure diverse enrollment.

implementation for chatbots and text-based pre-screening requires caution, particularly for vulnerable populations such as older adults with cognitive impairment, due to concerns including biases, privacy risks, and technical limitations in handling interactions with this group [30]. Speech-derived acoustic and linguistic markers, while promising and supported by ongoing research [31–38], are at a moderate maturity level, requiring additional clinical validation, and could be feasibly integrated as datasets expand and standardization improves [39]. Social media mining faces greater variability in maturity due to privacy constraints, data quality issues, and ethical challenges in health research, with effective deployment potentially taking a few years as tools, regulations, and frameworks advance [40–42].

Healthcare system-level AI for screening and workflow optimization.

Within the clinic, AI integrates into EHR systems (e.g., via APIs from vendors like Epic or Cerner, enhanced with tools from IQVIA) to alleviate resource strains and diagnostic gaps. Rather than deploying separate bots for each of the ICD-coded diseases, a common foundational chatbot with modular, disease-specific extensions (e.g., for AD-focused pre-screening) can provide efficient, scalable support across conditions.

1. During intake or visits, clinical decision support (CDS) tools use LLMs to scan unstructured notes, labs, and medication histories, flagging at-risk patients in real-time (e.g., "`Word repetition in notes suggests MCI; recommend MoCA screening`) [43].
2. Virtual assistants or chatbots pre-collect data (e.g., administering remote cognitive tests like digital memory assessments) and educate patients [44,45], freeing clinicians a few hours weekly for direct care.
3. To address diagnostic shortages, AI analyzes low-cost digital biomarkers (e.g., typing patterns from app-based tasks or eye movements via webcam) for non-invasive risk scores [46–48], prioritizing patients for advanced tests and reducing unnecessary procedures.
4. Predictive models integrate diverse data types (EHRs, genetics, neuroimaging) to forecast progression, tailoring physician education via medically focused LLMs [25], or personalized continuing medical education modules (e.g., "`Based on your query history, review this update on preclinical AD detection`"). This not only motivates screening by linking to emerging treatments but also counters perceived futility.

Clinical trial referral and matching.

For seamless referrals, AI platforms embed into the clinic's EHR for end-to-end matching.

- Post-screening, the system extracts patient data (age, biomarkers, medication use, neuroimaging reports, cognition scores) and uses LLMs to parse trial criteria from databases like ClinicalTrials.gov.
- It generates matches with probabilistic scores (e.g., "`92% eligibility for AMARANTH trial [NCT identifier], biomarker-positive MCI cohort, 15 miles away`").
- Real-time notifications alert physicians during visits, with LLM assistants answering queries (e.g., "`What are exclusion details?`") 24/7.

- Geolocation and predictive analytics prioritize local trials, forecast retention risks (e.g., based on travel distance or comorbidities), and suggest outreach to sponsors for clinics with high eligible pools. To enhance diversity, bias-detection algorithms flag underrepresented groups (e.g., by zip code or ethnicity) and recommend targeted pre-screening campaigns. Case studies, including AI-stratified trials [49], demonstrated reductions in recruitment time, lower screen failures, and cost savings, while improving inclusivity.

When evaluating LLMs versus other AI approaches in healthcare, they perform best for language-focused tasks [50–52]. These include developing patient- and clinician-facing chatbots, where retrieval grounding and rule-based guardrails enhance safety. LLMs also excel at converting unstructured text, such as EHR notes, into structured fields through parsing and entity extraction, with validation using rule-based dictionaries, regex, and structured-code crosschecks. Additionally, LLMs aid in interpreting narrative eligibility criteria for normalization, often paired with rules and ontologies for formalization. However, LLMs have limitations in areas like calibrated risk prediction [53,54], fairness-constrained modeling, and regulated decision support, where probabilistic and auditable models are preferred. Instead, traditional ML is superior for tabular risk scoring, reproducible cohort definitions, and deterministic pre-screening, with optional LLM assistance for filling missing text values [55–57]. Traditional ML also manages compliance-critical workflows effectively and excels in bias detection and monitoring, limiting LLMs to textual explanations. For clinical decision support prompts, combining LLM summarization with rules and validated models enforces thresholds. Multi-source, diverse data pipelines can integrate LLM-based text extraction with non-LLM models to improve prediction and question-answering, incorporating rule checks as guardrails for reliability and compliance [58].

4. Technical and clinical considerations for implementation

Implementing an AI solution requires careful consideration of technical feasibility, clinical integration, and human oversight to ensure they augment rather than supplant professional judgment. While AI tools like LLMs and predictive models offer probabilistic outputs (e.g., eligibility scores or risk predictions) to boost efficiency, their deployment must prioritize seamless workflow fit, data security, and validation to avoid errors or biases that could undermine trust or outcomes. For instance, in a clinic handling 500-1000 cognitive cases yearly, starting with pilot integrations, such as embedding AI tools in EHRs, can yield faster identifications, but only if addressed through structured steps. A recent study showed that an AI-powered system coined as Automated Clinical Trial Eligibility Screener (ACTES) reduced patient screening time by 34 % compared to manual processes, while improving the numbers of subjects screened, approached, and enrolled by 14.7 %, 11.1 %, and 11.1 %, respectively [59]. These efficiencies not only expedite identifications but also alleviate clinician burden in resource-constrained memory clinics, where manual processes often yield low enrollment rates.

From a technical standpoint, successful AI implementation in memory clinics requires a foundational assessment of infrastructure readiness. Many clinics operate with legacy EHR systems that are incompatible with modern AI tools, necessitating a shift toward interoperable standards such as FHIR to enable standardized data exchange. This includes converting unstructured clinical notes into structured formats (i.e., JSON), using cloud-based platforms, though such adoption can involve costs (e.g., storage fees of $0.25-0.50 per GB per month), and accessibility challenges for smaller clinics, including integration complexities and the need for specialized IT expertise. For a broad audience including smaller clinics with limited AD expertise, resource pooling across multiple sites (e.g., via medical center consortia) can distribute costs and expertise. Beyond EHR standardization, AD-specific tuning may require an additional 3-6 months, involving fine-tuning models on AD-relevant datasets (e.g., for biomarker integration), customizing APIs for databases like ClinicalTrials.gov, and conducting iterative pilot tests to align with evolving regulations. Establishing such interoperability allows AI models to process diverse data types including EHRs, imaging, and biomarkers, without fragmentation. A typical implementation roadmap begins with auditing existing systems for API compatibility and, where necessary, using middleware or open-source NLP tools (e.g., spaCy) to bridge integration gaps. Models are then trained or fine-tuned on de-identified clinic data, often employing federated learning approaches to preserve patient privacy while ensuring robustness across demographic variations. Real-time inference can be supported through edge computing e.g., using on-site servers to analyze speech biomarkers during clinic visits, while more computationally intensive tasks, such as trial matching, can leverage cloud resources. Technical challenges, such as incomplete or noisy data, can be addressed by ensemble approaches that combine LLMs for text parsing with computer vision for imaging analysis, achieving high accuracy in eligibility assessments in pilot settings. Continuous model refinement ensures adaptability to evolving trial criteria, and scalability testing (e.g., simulating high query volumes) is essential to ensure operational reliability in busy clinical environments.

For AI to be effective in memory clinics, it must integrate seamlessly into clinical workflows, enhancing patient-centered care. This requires tailoring tools to provider needs. The first step involves conducting user training sessions (e.g., workshops) focused on interpreting AI outputs such as "`75% MCI risk, recommend trial NCT456,`" with an emphasis on the probabilistic and supportive nature of these outputs to avoid over-reliance. The next step entails piloting the system on a subset of patients, validating AI recommendations against manual review, and tracking key metrics such as false positives, which can be reduced through iterative tuning, along with clinician satisfaction. The subsequent step integrates feedback loops in which clinicians can override or annotate AI suggestions, feeding into model refinement for continuous improvement. A major concern is bias amplification. For example, models trained on non-diverse datasets may under-identify eligible patients from minority groups. This can be addressed through dataset audits and the application of debiasing techniques. Ethical considerations include obtaining informed consent (e.g., "`This tool analyzes your data to suggest relevant trials`") and monitoring impacts on clinical workflow, where initial increases in validation time are often offset by long-term gains, such as a reduction in administrative burden.

5. Data standardization and interoperability

Effective AI deployment for AD trial recruitment hinges on overcoming data fragmentation, as models based on LLMs require structured, consistent inputs to accurately parse eligibility from sources such as EHRs, lab results, and demographics. However, the healthcare data landscape remains plagued by inconsistencies, posing barriers to AI systems that must integrate heterogeneous information for patient matching. For a mid-sized clinic managing a few hundred cognitive cases annually, this means siloed, unstructured records lead to inefficient manual harmonization, delaying identifications and inflating errors in trial referrals. Below, we outline practical data challenges in this setting and a vision for standardization, leveraging tools like FHIR and privacy-preserving techniques to enable AI-driven efficiency.

Data issues in the clinic amplify recruitment hurdles, mirroring broader AD research obstacles. For instance, patient data scatters across incompatible systems, e.g., legacy EHRs with free-text notes varying in syntax, alongside disparate lab or imaging files, making it difficult to aggregate for AI analysis, often requiring hours of manual review per case and overlooking a good fraction of eligibility criteria. In the context of AD, this is exacerbated by diverse data inputs like longitudinal narratives on cognitive decline, leading to misinterpretations in assessments. Moreover, inconsistent adoption of standards (e.g., multiple HL7 versions) hinders data exchange with external trial databases or collaborators, such as NACC datasets not natively FHIR-formatted, causing delays in matching patients to biomarker-positive trials and contributing to referral inefficiencies. Strict regulations limit data sharing for AI training, with siloed records restricting access to high-quality datasets; techniques like annotation add complexity, while noisy or biased data (e.g., underrepresenting minorities) reduces model accuracy in diverse clinic populations. EHRs blend structured data with clinical language, challenging LLMs to process lengthy documents or eligibility criteria in natural language, resulting in oversight of subtle AD indicators and prolonging recruitment by weeks. These challenges culminate in underpowered trials, with clinics like ours contributing fewer participants due to data silos, underscoring the need for standardized pipelines.

To transform fragmented data into AI-powered assets, the clinic can implement interoperability frameworks. Here's a practical vision:

- Begin by auditing EHRs for fragmentation, then adopt FHIR as a core standard to convert heterogeneous sources into unified formats (e.g., transforming free-text notes and labs into FHIR resources via APIs). Step 1: Use cloud platforms to ingest and normalize data from multiple systems, creating a centralized repository where AD-specific elements (e.g., cognitive scores, biomarkers) are structured for AI querying. Step 2: Employ LLMs for automated harmonization, parsing unstructured eligibility criteria into operationalizable rules, reducing manual effort and enabling real-time trial matching.
- Integrate legacy and modern systems through FHIR-compliant pipelines (e.g., JSON conversions for inputs like NACC data). Step 1: Implement hybrid integrations to bridge HL7 gaps, allowing seamless exchange with trial sponsors or registries. Step 2: Use AI to flag inconsistencies (e.g., semantic variations in clinical notes), automating corrections and improving accuracy for analyses like combining EHRs with neuroimaging. This fosters collaborative research and enhances clinic contributions to diverse trials.
- To address data silos and confidentiality, apply federated learning and differential privacy for model training without centralizing sensitive information. Step 1: Train AI on distributed datasets (e.g., across clinic sites) using federated approaches, adding noise via differential privacy to protect identities while maintaining utility for AD detection. Step 2: Validate with de-identified clinic data, ensuring compliance with HIPAA/GDPR and reducing bias through diverse annotations.
- Leverage advanced AI architectures for complex data. Step 1: Use LLMs to process complex texts and integrate with structured elements, e.g., extracting AD risk from narrative histories. Step 2: Invest in quality datasets via collaborations, fine-tuning models to handle clinic-specific variations and cutting eligibility oversights [49].

6. Ethical and regulatory considerations

Integrating AI into AD trial recruitment workflows such as using LLMs for patient matching or digital biomarkers for pre-screening, introduces ethical and regulatory challenges that must be navigated to protect patient well-being, ensure equity, and sustain trust in healthcare. A robust framework is essential for responsible deployment, balancing innovation with safeguards like informed consent and bias mitigation. For a mid-sized clinic handling 500-1000 cognitive cases annually, these issues can manifest in daily operations, potentially eroding patient participation if unaddressed, while compliance with regulations like HIPAA and GDPR adds operational layers. Below, we outline practical challenges in this clinic context and a step-by-step vision for ethical integration, drawing on privacy-preserving techniques and explainable AI (XAI) to foster trust and inclusivity.

Ethical and regulatory hurdles in the clinic significantly amplify recruitment barriers, contributing to hesitancy among both patients and providers. The collection of sensitive data such as genomics, EHRs, and behavioral metrics, for use in AI tools raises privacy concerns under HIPAA and GDPR. Beyond HIPAA and GDPR, evolving AI-specific regulatory frameworks, such as the FDA's guidance on AI/ML-based software as a medical device (SaMD) [60–63], and the European Union AI Act [64–67], must be integrated; these classify AI tools by risk (e.g., Class I minimal risk for low-stakes chatbots, which may not require rigorous RCTs, but still necessitate local validation to mitigate risks like hallucinations) and mandate requirements for transparency, clinical validation, robustness, and ongoing post-market monitoring to ensure safety and efficacy. As discussed by Shuren et al., for MCI/AD [68], SaMD like digital therapeutics are regulated via a risk-based approach, with no clearances yet for AD treatment, emphasizing breakthrough programs and benefit-risk assessments in diverse populations. Health systems should implement safeguards such as rule-based constraints, human oversight for outputs, and pilot testing to prevent misinformation, particularly given the potential for patient over-reliance with serious consequences.

Patients, particularly those from certain subgroups, are often less trusting of AI systems, which can deter participation, especially since retracting data once incorporated into AI models is difficult [69]. Furthermore, the "black box" nature of many AI algorithms obscures decision-making, making it challenging for clinicians to explain why a patient was flagged for a trial. This undermines informed consent and fosters "therapeutic misconception [70]," where patients overestimate the certainty of AI predictions especially in preclinical AD assessments, posing autonomy risks for individuals, particularly those with lower health literacy [71]. Over-reliance on AI may also erode trust in clinicians, with some patients perceiving them as less competent, potentially leading to "de-skilling" and a diminished emphasis on human skills like empathy [72–75]. In some cases, this can increase clinician workload for validating AI outputs, negating promised efficiencies and contributing higher participant withdrawal rates. Given that only a few health systems have the capacity for ongoing monitoring, responsibility should involve multidisciplinary teams or external partners, potentially utilizing a CMS Registry to track safety, performance, and risks post-installation, ensuring neurologists are not solely burdened. Finally, AI-generated pre-symptomatic predictions echo ethical dilemmas seen in Huntington's disease testing [76,77], where disclosing risk in the absence of treatments can provoke anxiety or stigma. This not only threatens patients' right not to know but also raises the specter of discrimination in insurability or employment, deterring engagement of at-risk individuals who fear unequal access or social repercussions. Additionally, generative AI's tendency to fabricate human-like experiences (e.g., claiming personal anecdotes) poses deception risks [78], especially for vulnerable AD patients; safeguards such as disclaimers, prompt restrictions, and family oversight are essential to prevent exploitation [79].

These challenges, if unaddressed, risk perpetuating disparities and regulatory violations, underscoring the need for proactive, ethics-driven frameworks. Misaligned financial incentives must be countered by independent vetting to prevent recruitment into potentially unsafe trials; chatbots should transparently discuss high failure rates and historical risks, programmed by multidisciplinary teams including patient advocates and lawyers for balance. Clinics can mitigate these issues by adopting a structured approach that emphasizes AI as a tool for augmentation anchored in human oversight. First, a secure data ecosystem should be established using techniques like federated learning and differential privacy to enable AI training without transmitting identifiable data. This includes auditing data flows for HIPAA/GDPR compliance, encrypting digital biomarkers, and collaborating on de-identified datasets, practices that can reduce breach risks and foster trust via transparent, opt-in data policies. Second, integrating explainable AI ensures that clinicians can communicate how and why decisions are made, for instance, using saliency maps to show why certain speech features suggest cognitive impairment. In parallel, patient consent must be redesigned with clarity and tailored literacy, supported by clinician training modules that reduce therapeutic misconceptions and align with consent standards for analogous conditions like HD. Consent should occur at multiple steps: for non-standard assessments (e.g., digital biomarkers or speech analysis) that may inform models beyond immediate care; an explicit opt-in for trial eligibility screening, with disclosures on procedures specific to recruitment to avoid therapeutic misconception; and a separate consent for aggregating de-identified data with others for research purposes. Third, AI should be explicitly framed as an assistive tool that preserves human strengths like empathy and judgment. Hybrid workflows should allow clinicians to override AI in some cases, with regular staff feedback and soft-skills training to counter overreliance, strategies that may reduce patient withdrawals. Finally, clinics must adapt protocols for preclinical predictions by offering patients the option to decline results, ensuring access to psychosocial support, and partnering with ethicists to protect against stigma and discrimination.

7. Future directions

AI's role in AD trial recruitment promises significant advancements, but realizing this requires strategic investments in technology, policy, and partnerships to scale beyond current limitations. For our mid-sized clinic, future tools could evolve from basic matching to proactive, personalized systems that predict trial success and address diversity gaps, potentially doubling enrollments while halving timelines. Below, we outline emerging challenges and a step-by-step vision for future AI integration, emphasizing multimodal enhancements, regulatory evolution, and collaborative ecosystems to drive inclusive AD therapeutic progress.

As AI matures, new hurdles will arise in clinic settings, building on existing barriers. For example, current models may struggle with real-time multimodal data (e.g., integrating live speech analysis with EHRs), risking biases in underrepresented groups and reducing accuracy as patient volumes grow. In diverse clinics, this could perpetuate under-enrollment of minorities, limiting trial generalizability. Evolving guidelines (e.g., FDA's AI frameworks) may lag innovations like generative AI, complicating validation and increasing compliance burdens without clear sandboxes for testing. Without longitudinal data on AI's effects (e.g., on patient trust or outcomes), clinics risk unintended harms like increased anxiety from predictive tools, potentially raising withdrawals. Siloed advancements hinder global sharing, with clinics missing out on large-scale datasets, slowing progress in AD-specific AI.

To overcome these challenges, clinics can adopt phased, stakeholder-engaged innovations that support scalable and ethical AI integration. Advanced models that fuse LLMs, computer vision, and sensor data can enable precise, real-time participant matching by analyzing modalities such as speech, gait, and EHRs. Embedding bias-detection algorithms trained on diverse datasets helps correct underrepresentation and has been shown to improve minority enrollment [80]. Edge-AI devices, such as wearables for continuous biomarker monitoring, allow proactive risk flagging during clinical visits [81–84]. Participation in regulatory sandboxes (e.g., FDA or EMA) provides a safe environment to test AI protocols, while generative AI enables virtual trial simulations through synthetic patient profiles with ethical oversight to minimize risks like hallucination. Longitudinal studies tracking AI's clinical and psychosocial effects (e.g., annual patient trust surveys) inform iterative refinements. Ongoing input from ethicists supports responsible tool development, while adaptive clinician training ensures AI augments empathy-driven care, potentially increasing trial retention. Engagement in data-sharing consortia such as ADNI, NACC or the Critical Path Institute helps standardize tools and access high-quality datasets. Community-focused applications, including VR tools to destigmatize AD, can extend outreach to underserved populations, and clinic-led policy advocacy for inclusive trial incentives may accelerate global progress through unified benchmarks.

8. Conclusion

The memory clinic scenario illustrates the challenges in AD clinical trial recruitment, from community stigma and resource constraints to data silos and ethical dilemmas, that collectively hinder therapeutic advancement, often resulting in a small number of enrollments annually despite evaluating several potential participants. AI offers a practical solution by automating identification, enhancing matching precision, and promoting inclusivity through tools like LLMs integrated into everyday workflows. By addressing barriers step-by-step, from community pre-screening chatbots that boost early awareness to EHR-embedded CDS systems reducing screen failures and interoperable platforms ensuring data readiness, AI can accelerate enrollment, cut costs and ensure diverse representation, ultimately alleviating the global AD burden affecting millions. Ethical deployment, with privacy safeguards and human oversight, is paramount to building trust, while future directions in multimodal AI and collaborations promise even greater efficiencies. Through coordinated adoption in clinics worldwide, AI-augmented strategies can pave the way for faster breakthroughs in AD treatments, turning persistent obstacles into broader access to timely, inclusive AD therapies.

Funding

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

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: 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.

References

[1] Bradford A, Kunik ME, Schulz P, Williams SP, Singh H. Missed and delayed diagnosis of dementia in primary care. Alzheimer Dis Assoc Disord. 2009;23:306–14.

[2] Langbaum JB, Zissimopoulos J, Au R, Bose N, Edgar CJ, Ehrenberg E, Fillit H, Hill CV, Hughes L, Irizarry M, Kremen S, Lakdawalla D, Lynn N, Malzbender K, Maruyama T, Massett HA, Patel D, Peneva D, Reiman EM, Romero K, Routledge C, Weiner MW, Weninger S, Aisen PS. Recommendations to address key recruitment challenges of Alzheimer's disease clinical trials. Alzheimer's Dement. 2022;19:696–707.

[3] Watson JL, Ryan L, Silverberg N, Cahan V, Bernard MA. Obstacles and opportunities In Alzheimer's clinical trial recruitment. Health Aff 2014;33:574–9.

[4] Lu X, Yang C, Liang L, Hu G, Zhong Z, Jiang Z. Artificial intelligence for optimizing recruitment and retention in clinical trials: a scoping review. J Am Med Inform Assoc. 2024;31:2749–59.

[5] Beattie J, Neufeld S, Yang D, Chukwuma C, Gul A, Desai N, Jiang S, Dohopolski M. Utilizing large language models for enhanced clinical trial matching: a study on automation in patient screening. Cureus 2024.

[6] Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci 2019;40:577–91.

[7] Lu X, Chen M, Lu Z, Shi X, Liang L. Artificial intelligence tools for optimising recruitment and retention in clinical trials: a scoping review protocol. BMJ Open 2024:14.

[8] Grill JD, Karlawish J. Addressing the challenges to successful recruitment and retention in Alzheimer's disease clinical trials. Alzheimers Res Ther 2010;2.

[9] Arenaza-Urquijo EM, Vemuri P. Resistance vs resilience to Alzheimer disease. Neurology 2018;90:695–703.

[10] Herrmann LK, Welter E, Leverenz J, Lerner AJ, Udelson N, Kanetsky C, Sajatovic M. A systematic review of dementia-related stigma research: can we move the stigma dial? Am J Geriatr Psychiatry 2018;26:316–31.

[11] Werner P, Heinik J. Stigma by association and Alzheimer's disease. Aging Ment Health 2008;12:92–9.

[12] Rosin ER, Blasco D, Pilozzi AR, Yang LH, Huang X. A narrative review of Alzheimer's Disease stigma. J Alzheimer's Dis. 2020;78:515–28.

[13] Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved 2010;21:879–97.

[14] Hamel LM, Penner LA, Albrecht TL, Heath E, Gwede CK, Eggly S. Barriers to clinical trial enrollment in racial and Ethnic Minority patients with cancer. Cancer Control 2016;23:327–37.

[15] Wollney EN, Armstrong MJ, Bedenfield N, Rosselli M, Curiel-Cid RE, Kitaigorodsky M, Levy X, Bylund CL. Barriers and best practices in disclosing a dementia diagnosis: a clinician interview study. Health v Insights 2022:15.

[16] Okpalauwaekwe U, Franks H, Kuo Y-F, Raji MA, Passy E, Tzeng H-M. What helps or hinders annual wellness visits for detection and management of cognitive impairment among older adults? A scoping review guided by the consolidated framework for implementation research. Nurs Rep 2025:15.

[17] Galvin JE, Meuser TM, Boise L, Connell CM. Predictors of physician referral for patient recruitment to Alzheimer disease clinical trials. Alzheimer Dis Assoc Disord. 2009;23:352–6.

[18] Park L, Kouhanim C, Lee S, Mendoza Z, Patrick K, Gertsik L, Aguilar C, Gullaba D, Semenova S, Jhee S. Implementing a memory Clinic model to facilitate recruitment into early phase clinical trials for mild cognitive impairment and Alzheimer's disease. J Prev Alzheimers Dis 2019;6:135–8.

[19] Shaw AR, Perales-Puchalt J, Moore T, Weatherspoon P, Robinson M, Hill CV, Vidoni ED. Recruitment of older African Americans in Alzheimer's Disease clinical trials using a community education approach. J Prev Alzheimers Dis 2022;9:672–8.

[20] Dabiri S, Raman R, Grooms J, Molina-Henry D. Examining the role of community engagement in enhancing the participation of racial and ethnic minoritized communities in Alzheimer's Disease clinical trials; a rapid review. J Prev Alzheimers Dis 2024;11:1647–72.

[21] Franzen S, Smith JE, van den Berg E, Rivera Mindt M, van Bruchem-Visser RL, Abner EL, Schneider LS, Prins ND, Babulal GM, Papma JM. Diversity in Alzheimer's disease drug trials: the importance of eligibility criteria. Alzheimer's Dement. 2021;18:810–23.

[22] Raman R, Aisen P, Carillo MC, Detke M, Grill JD, Okonkwo OC, Rivera-Mindt M, Sabbagh M, Vellas B, Weiner M, Sperling R. Tackling a major deficiency of diversity in Alzheimer's disease therapeutic trials: an CTAD Task Force report. J Prev Alzheimers Dis 2022;9:388–92.

[23] Wenderott K, Krups J, Weigl M, Wooldridge AR. Facilitators and barriers to implementing AI in routine medical imaging: systematic review and qualitative analysis. J Med Internet Res 2025:27.

[24] Hu Z, Hu R, Yau O, Teng M, Wang P, Hu G, Singla R. Tempering expectations on the medical artificial intelligence revolution: the medical trainee viewpoint. JMIR Med Inf. 2022:10.

[25] Jia S, Bit S, Searls E, Lauber MV, Fan P, Wang WM, Claus LA, Jasodanand VH, Veerapaneni D, Au R, Kolachalama VB. PodGPT: an audio-augmented large language model for research and education. NPJ Biomed Innov 2025;2:26.

[26] Nievas M, Basu A, Wang Y, Singh H. Distilling large language models for matching patients to clinical trials. J Am Med Inf Assoc 2024;31:1953–63.

[27] Jin Q, Wang Z, Floudas CS, Chen F, Gong C, Bracken-Clarke D, Xue E, Yang Y, Sun J, Lu Z. Matching patients to clinical trials with large language models. Nat Commun 2024;15:9074.

[28] Rybinski M, Kusa W, Karimi S, Hanbury A. Learning to match patients to clinical trials using large language models. J Biomed Inf. 2024:159.

[29] Chen H, Li X, He X, Chen A, McGill J, Webber EC, Xu H, Liu M, Bian J. Enhancing patient-trial matching with large language models: a scoping review of emerging applications and approaches. JCO Clin Cancer Inform. 2025.

[30] Abadir PM, Battle A, Walston JD, Chellappa R, Lipsitz LA. Enhancing care for older adults and dementia patients with large language models: proceedings of the National Institute on Aging—Artificial Intelligence & Technology Collaboratory for Aging Research Symposium. J Gerontol A: Biol Sci Med Sci. 2024:79.

[31] Karjadi C, Xue C, Cordella C, Kiran S, Paschalidis IC, Au R, Kolachalama VB. Fusion of low-level descriptors of digital voice recordings for dementia assessment. J Alzheimers Dis 2023;96:507–14.

[32] 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.

[33] Amini S, Hao B, Zhang L, Song M, Gupta A, Karjadi C, Kolachalama VB, Au R, Paschalidis IC. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. Alzheimer's Dement. 2022;19:946–55.

[34] Tavabi N, Stück D, Signorini A, Karjadi C, Al Hanai T, Sandoval M, Lemke C, Glass J, Hardy S, Lavallee M, Wasserman B, Ang TFA, Nowak CM, Kainkaryam R, Foschini L, Au R. Cognitive digital biomarkers from automated transcription of spoken language. J Prev Alzheimers Dis 2022;9:791–800.

[35] Kiyoshige E, Ogata S, Kwon N, Nakaoku Y, Hayashi C, Blaylock N, Brueckner R, Subramanian V, Joseph Oconnell H, Yoshikawa Y, Teramoto K, Nakatsuka K, Saito S, Ihara M, Takegami M, Nishimura K. Developing and testing AI-based voice biomarker models to detect cognitive impairment among community dwelling adults: a cross-sectional study in Japan. Lancet Reg Health - West Pac. 2025:59.

[36] Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine. 2020. p. 28.

[37] Fraser KC, Meltzer JA, Rudzicz F, Garrard P. Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimer's Dis. 2015;49:407–22.

[38] Guo Z, Ling Z, Li Y, Peña-Casanova J. Detecting Alzheimer's disease from continuous speech using language models. J Alzheimer's Dis. 2019;70:1163–74.

[39] Qi W, Zhu X, Wang B, Shi Y, Dong C, Shen S, Li J, Zhang K, He Y, Zhao M, Yao S, Dong Y, Shen H, Kang J, Lu X, Jiang G, Boots LMM, Fu H, Pan L, Chen H, Yan Z, Xing G, Cao S. Alzheimer's disease digital biomarkers multidimensional landscape and AI model scoping review. NPJ Digit Med. 2025;8.

[40] Ford E, Shepherd S, Jones K, Hassan L. Toward an ethical framework for the text mining of Social Media for health research: a systematic review. Front Digit Health 2021;2.

[41] Azizi M, Jamali AA, Spiteri RJ. Identifying X (Formerly Twitter) posts relevant to dementia and COVID-19: machine learning approach. JMIR Form Res. 2024;8.

[42] Alharthi NH, Alanazi EM, Liu X. Awareness level of Huntington disease: comprehensive analysis of tweets during Huntington Disease Awareness month. Comput Methods Progr Biomed Update 2023:4.

[43] Du X, Novoa-Laurentiev J, Plasek JM, Chuang Y-W, Wang L, Marshall GA, Mueller SK, Chang F, Datta S, Paek H, Lin B, Wei Q, Wang X, Wang J, Ding H, Manion FJ, Du J, Bates DW, Zhou L. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. eBioMedicine 2024:109.

[44] Serafimovska A, Swavley K, Zhang Qian Ao A, Challinor KL, Florio T. Cognitive status assessment of older adults - test administration by conversational artificial intelligence (AI) chatbot: proof-of-concept investigation. J Clin Exp Neuropsychol 2025;47:472–84.

[45] Hasan WU, Zaman KT, Wang X, Li J, Xie B, Tao C. Empowering Alzheimer's caregivers with conversational AI: a novel approach for enhanced communication and personalized support. NPJ Biomed Innov. 2024:1.

[46] Alfalahi H, Khandoker AH, Chowdhury N, Iakovakis D, Dias SB, Chaudhuri KR, Hadjileontiadis LJ. Diagnostic accuracy of keystroke dynamics as digital biomarkers for fine motor decline in neuropsychiatric disorders: a systematic review and meta-analysis. Sci Rep 2022:12.

[47] Tripathi S, Acien A, Holmes AA, Arroyo-Gallego T, Giancardo L. Generalizing Parkinson's disease detection using keystroke dynamics: a self-supervised approach. J Am Med Inform Assoc. 2024;31:1239–46.

[48] Li Q, Yan J, Ye J, Lv H, Zhang X, Tu Z, Li Y, Guo Q. Construction of a prediction model for Alzheimer's disease using an AI-driven eye-tracking task on mobile devices. Aging Clin Exp Res 2024:37.

[49] Vaghari D, Mohankumar G, Tan K, Lowe A, Shering C, Tino P, Kourtzi Z. AI-guided patient stratification improves outcomes and efficiency in the AMARANTH Alzheimer's Disease clinical trial. Nat Commun 2025:16.

[50] Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med. 2025;8.

[51] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, Hou L, Clark K, Pfohl SR, Cole-Lewis H, Neal D, Rashid QM, Schaekermann M, Wang A, Dash D, Chen JH, Shah NH, Lachgar S, Mansfield PA, Prakash S, Green B, Dominowska E, Agüera y Arcas B, Tomašev N, Liu Y, Wong R, Semturs C, Mahdavi SS, Barral JK, Webster DR, Corrado GS, Matias Y, Azizi S, Karthikesalingam A, Natarajan V. Toward expert-level medical question answering with large language models. Nat, Med 2025;31:943–50.

[52] Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med. 2023:6.

[53] Brown KE, Yan C, Li Z, Zhang X, Collins BX, Chen Y, Clayton EW, Kantarcioglu M, Vorobeychik Y, Malin BA. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. J Am Med Inform Assoc. 2025;32:811–22.

[54] Liu M, Ning Y, Teixayavong S, Liu X, Mertens M, Shang Y, Li X, Miao D, Liao J, Xu J, Ting DSW, Cheng LT-E, Ong JCL, Teo ZL, Tan TF, RaviChandran N, Wang F, Celi LA, Ong MEH, Liu N. A scoping review and evidence gap analysis of clinical AI fairness. NPJ Digit Med. 2025;8.

[55] Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, Makadia R, Jin P, Shang N, Kang T, Weng C. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc. 2019;26:294–305.

[56] Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. Deep neural networks and tabular data: a survey. IEEE Trans Neural Netw Learn Syst 2024;35: 7499–519.

[57] Lopez I, Swaminathan A, Vedula K, Narayanan S, Nateghi Haredasht F, Ma SP, Liang AS, Tate S, Maddali M, Gallo RJ, Shah NH, Chen JH. Clinical entity augmented retrieval for clinical information extraction. NPJ Digit Med. 2025;8.

[58] Adam H, Lin J, Lin J, Keenan H, Wilson A, Ghassemi M. Clinical information extraction with large language models: a case study on organ procurement. AMIA Annu Symp Proc 2024;2024:115–23.

[59] Ni Y, Bermudez M, Kennebeck S, Liddy-Hicks S, Dexheimer J. A real-time automated patient screening system for clinical trials eligibility in an emergency department: design and evaluation. JMIR Med Inf. 2019:7.

[60] Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. NPJ Digit Med. 2020:3.

[61] Yang S-R, Chien J-T, Lee C-Y. Advancements in clinical evaluation and regulatory frameworks for AI-driven software as a medical device (SaMD). IEEE Open J Eng Med Biol. 2025;6:147–51.

[62] Singh V, Cheng S, Kwan AC, Ebinger J. United States Food and Drug Administration regulation of clinical software in the era of artificial intelligence and machine learning. Mayo Clin Proc.: Digit Health 2025:3.

[63] Muehlematter UJ, Bluethgen C, Vokinger KN. FDA-cleared artificial intelligence and machine learning-based medical devices and their 510(k) predicate networks. Lancet Digit Health 2023;5:e618–26.

[64] Cancela-Outeda C. The EU's AI act: a framework for collaborative governance. Internet Things 2024:27.

[65] Kusche I. Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk. J Risk Res 2024:1–14.

[66] Laux J, Wachter S, Mittelstadt B. Trustworthy artificial intelligence and the European Union AI act: on the conflation of trustworthiness and acceptability of risk. Regul Gov. 2023;18:3–32.

[67] Pham B-C, Davies SR. What problems is the AI act solving? Technological solutionism, fundamental rights, and trustworthiness in European AI policy. Crit Policy Stud. 2024;19:318–36.

[68] Shuren J, Doraiswamy PM. Digital Therapeutics for MCI and Alzheimer's disease: a regulatory perspective - highlights from the Clinical Trials on Alzheimer's Disease conference (CTAD). J Prev Alzheimers Dis 2022;9:236–40.

[69] Shah R, Robertson C, Woods A, Bergstrand K, Findley J, Balser C, Slepian MJ. Diverse patients' attitudes towards artificial intelligence (AI) in diagnosis. PLOS Digit Health 2023;2.

[70] Wilkins JM, Forester BP. Informed consent, therapeutic misconception, and clinical trials for Alzheimer's disease. Int J Geriatr Psychiat 2020;35:430–5.

[71] Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In Artif Intell Healthc. 2020:295–336.

[72] Reis M, Reis F, Kunde W. Public perception of physicians who use artificial intelligence. JAMA Netw Open 2025;8.

[73] Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19: 121–7.

[74] Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. J Saf Sci Resil. 2024;5:460–9.

[75] Natali C, Marconi L, Dias Duran LD, Cabitza F. AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. Artif Intell Rev 2025:58.

[76] Wusthoff C. The dilemma of confidentiality in Huntington disease. JAMA: J Am Med Assoc. 2003;290:1219–20.

[77] Kromberg JGR, Wessels T-M. Ethical issues and Huntington's disease. S Afr Med J. 2013:103.

[78] Williamson SM, Prybutok V. The era of artificial intelligence deception: unraveling the complexities of false realities and emerging threats of misinformation. Information 2024:15.

[79] Maddox T, Babski D, Embi P, Gerhart J, Goldsack J, Parikh R, Sarich T, Krishnan S, Elliott A. Gener Artif Intell Health Med. 2025.

[80] Chang CY, Yuan J, Ding S, Tan Q, Zhang K, Jiang X, Hu X, Zou N. Towards fair patient-trial matching via patient-criterion level fairness constraint. AMIA Annu Symp Proc 2023;2023:884–93.

[81] Mahajan A, Heydari K, Powell D. Wearable AI to enhance patient safety and clinical decision-making. NPJ Digit Med. 2025;8.

[82] Huang G, Chen X, Liao C. AI-driven wearable bioelectronics in Digital Healthcare. Biosensor 2025:15.

[83] Kovur P, Kovur KM, Rayat DY, Wishart DS. POC sensor systems and artificial intelligence—where we are now and where we are going? Biosensor 2025:15.

[84] Shajari S, Kuruvinashetti K, Komeili A, Sundararaj U. The emergence of AI-based wearable sensors for digital health technology: a review. Sensors 2023;23.