A Novel Paradigm for Neurodegenerative Diseases: Building an AI-Driven Large-Scale Electroencephalography Database Research

Heyuan Xiao[1], Fangzheng Zhou[2], Kunshuan Xie[3], Tao Wang[1,*]

[1] Department of Human Anatomy, Histology and Embryology, Institute of Basic Medical Sciences & School of Basic Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
[2] Department of Computer Science and Technology, School of Computer Engineering and Science, Shanghai University, Shanghai, China
[3] Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, England, UK

***Correspondence:** Tao Wang, Department of Human Anatomy, Histology and Embryology, Institute of Basic Medical Sciences & School of Basic Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China.
E-Mail: wangtao@ibms.pumc.edu.cn

Abstract

The transition to the "Fourth paradigm" of data-driven science offers a major opportunity for Alzheimer's disease (AD) and broader neurodegenerative disease (NDD) research. However, this potential is constrained by a "data void" caused by the long preclinical progression of AD pathology. We argue that the large-scale electroencephalography (EEG), when combined with clinical and biological validation, can provide a scalable functional window onto this gap. Yet, the current EEG data landscape emains fragmented, heterogeneous, and difficult to reuse. This perspective review presents a strategic blueprint to address this deadlock. We first examine existing foundational and AD-focused EEG resources. We then distinguish prospective standardization from retrospective harmonization, arguing that both are required to transform isolated recordings into interoperable and clinically validatable evidence. Furthermore, we analyze how advanced artificial intelligence (AI) architectures, including Transformer-based models, depend on robust data governance and explainable AI (XAI) strategies to avoid "Shortcut Learning" and support biological interpretation. Finally, we propose a collaborative ecosystem designed to ensure Robustness, Interpretability, and Scalability (RIS), linking standardized EEG with longitudinal cohorts, human brain banks, and post-mortem neuropathology. Together, this framework aims to transform EEG from a clinical tool into a scalable engine for discovery in AD and related NDDs.

Keywords: Alzheimer's disease, electroencephalography, deep learning, data standardization, large-scale database, Fourth paradigm

Received: 30 January 2026; **Accepted:** 28 May 2026
DOI: 10.15302/ HB.2026.0004
© The Author(s) 2026. This article is published by Higher Education Press at journal.hep.com.cn.
This is an open access article distributed under the Creative Commons Attribution License 4.0 (CCBY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

An AI-driven research paradigm shift

We are now on the eve of a paradigm shift. This shift is moving us into what is often called "the Fourth paradigm of science", a data-driven paradigm[1]. This new paradigm does not replace the traditional, hypothesis-driven paradigm. Instead, it reorders the scientific process: it prioritizes the large-scale, data-driven discovery over hypothesis-driven research. This transformation is born not from a single breakthrough, but from the unprecedented convergence of three powerful technological waves: scalable cloud infrastructure, advanced AI architectures—represented by the Transformer[2]—for complex pattern discovery, and initiatives for large-scale biological data standardization to ensure signal consistency. Together, these waves provide the foundation for AI-powered tools that can uncover insights at a previously unimaginable scale. However, this wave of innovation has yet to breach the formidable barrier in medical research: the neurodegenerative diseases (NDDs). This barrier is solidified by the "silent phase"—the temporal disconnection between the silent progression of molecular pathology and the eventual appearance of clinical symptoms, particularly within the research of Alzheimer's disease (AD)[3]. The combination of its long latent period and frequent neuropathological comorbidity turns AD from a single disease trajectory into a complex spatiotemporal network. Recent evidence from a Chinese community-based autopsy cohort further supports this mixed-pathology reality[4]. Here lies a paradox: the data-driven paradigm is a core driving paradigm to unlock the spatiotemporal complexity of AD, yet it remains starved of input. This data void creates a translational bottleneck. To empower the paradigm to finally breach the barrier, we require a new source of fuel—a pre-symptomatic biomarker that is scalable and readily detectable[5].

This critical demand has spurred the revival of electroencephalography (EEG). The NIA-AA ATN framework has redefined AD around Amyloid, Tau, and Neurodegeneration, providing a biological anchor beyond symptom-based disease[6]. Yet this framework leaves the functional dynamics of AD insufficiently captured. While historically viewed as a classical tool, EEG is uniquely positioned to address this gap[7]. Advances in modern machine learning have substantially expanded its analytical potential, transforming it from a simple chart into a high-dimensional data stream[8,9]. Its characteristics are perfectly suited to the new data-driven paradigm: unlike metabolic or structural neuroimaging (PET/MRI), it offers millisecond-level temporal resolution to capture real-time neural dynamics; unlike invasive fluid biopsies, it is non-invasive and cost-effective[10]. However, EEG also has clear limitations, including low spatial resolution, weak sensitivity to deep neural sources, and vulnerability to artefacts. Therefore, EEG should not be viewed as a replacement for molecular, imaging or neuropathological markers, but as a functional window onto network-level dysfunction. Thus, EEG serves not merely as a measurement tool, but as a critical bridge connecting the computational discovery with clinical-biological validation. For AD, this bridge is particularly important. Data-driven models can search for latent electrophysiological signatures across the silent and prodromal phases, while hypothesis-driven research remains essential for testing their clinical and biological validity against cognitive trajectories, established biomarkers, neuroimaging findings and post-mortem pathology. However, a significant barrier prevents the full realisation of this potential. At present, EEG data remain trapped in fragmented and heterogeneous silos. Without comprehensive standardization, spanning from acquisition protocols, data formats, clinical metadata to privacy governance, raw recordings cannot be transformed into trustworthy, clinically validatable evidence.

This perspective review aims to provide a blueprint for breaking the current deadlock. We first map the landscape of EEG databases, ranging from foundational archives in sleep, epilepsy, and emotion research to emerging AD collections. Their value is immense. However, they remain largely isolated and cannot be readily integrated due to the absence of a common standard. We argue that the field requires a dual strategy: prospective standardization to prevent further irreversible information loss in the future EEG cohorts, and retrospective harmonisation to maximise the value of existing heterogeneous datasets. This is not merely a regulatory formality, but a technical and biological necessity for reproducibility, interoperability

and clinical validation. It is driven by the critical need to transform technical model outputs into trustworthy, clinically validatable evidence. Finally, drawing on direct experience with the comprehensive Standardized Operational Protocol (SOP) for the China Human Brain Bank Consortium (CHBBC)[11,12], we propose a concrete initiative to build this future: a framework for shared data to accelerate collaborative discovery in NDDs, particularly in AD.

The EEG database landscape: the potential in fragments

Our blueprint for the AI-driven AD research begins with the raw material: data. The field is not starting from nothing. Pioneering efforts have yielded vast repositories, creating an apparent wealth of information. This chapter maps this complex terrain.

Our survey begins with the representative foundational EEG archives and platforms established in classic domains, including sleep, epilepsy, and emotion research (summarized in Table 1). These resources demonstrate both the feasibility of large-scale data collection and the practical diversity of existing sharing models, thereby providing valuable experience for developing future databases.

As Table 1 illustrates, repositories such as Temple University Hospital (TUH) EEG Corpus have achieved an unprecedented scale, providing the volume necessary for developing robust deep learning algorithms. Concurrently, platforms such as PhysioNet and OpenNeuro have been instrumental in fostering a culture of open data sharing.

Many of these archives are well curated, widely cited, and have supported decades of productive research. However, they were constructed for domain-specific research questions rather than for cross-disease integration or large-scale translational application. Table 1 highlights several dimensions of heterogeneity that directly affect cross-dataset integration:

- **Sampling rate:** explicitly reported sampling rates range from 100 Hz to 1,000 Hz, while several repositories further depend on dataset- or device-specific settings.
- **Channel configuration:** channel counts vary from low-density or device-dependent recordings to 118-channel configurations, and several archives do not enforce a uniform montage.
- **File format:** data formats include .edf, .bdf, .cnt, .mat, .txt, .pkl, and BIDS-compatible structures, creating substantial preprocessing barriers before joint analysis.
- **Accessibility and reusability:** beyond technical heterogeneity, several resources are affected by unstable links, incomplete documentation, or platform-dependent access restrictions, further limiting practical reuse.

These differences mean that direct integration of otherwise valuable datasets can introduce acquisition-, format-, and access-related confounders that compromise scientific rigour. The field is therefore left in a paradoxical situation: possessing a wealth of data that it cannot collectively mine.

The experience with foundational databases exposes a universal challenge. The issue of data fragmentation becomes even more acute when we turn our attention to the specific and urgent challenge of AD.

Unlike in foundational domains such as sleep, epilepsy, and emotion research, the large-scale, standardized public EEG databases specifically designed for AD are exceedingly rare. The majority of studies, typically involving only dozens of participants, do not release their raw data to the public. The few studies that have made their data available are summarized in Table 2. Consequently, the data landscape for AD is even more challenging than for foundational research: available datasets are not only highly heterogeneous but also severely limited in quantity.

The evidence in Table 2 solidifies this conclusion. To make this imbalance explicit, Figure 1 compares the participant numbers of publicly available AD-focused EEG datasets with the TUHEEG Corpus as a foundational benchmark. The comparison shows that most public AD-focused EEG datasets remain at the scale of dozens to hundreds of participants. Taken together, the foundational archives (Table 1) and the emerging

Table 1: Representative foundational EEG archives and platforms worldwide.

Database	Study focus	Participants	Records	Sampling rate	Channels	File formats	Source
TUH EEG Corpus	Epilepsy: 78% Stroke: 12% Concuss: 0.3%	14,987	26,846	250 Hz	24–36	.edf	isip.piconepress.com/projects/tuh_eeg/index.shtml
PhysioNet	Depends on dataset (including sleep disorder and auditory research)	Depends on dataset From a few (e.g., 8, Sleep-EDF) to hundreds (e.g., 607, I-Care v2.1)	Depends on dataset From a few (e.g., 8, Sleep-EDF) to tens of thousands (e.g., 80,809, I-Care v2.1)	Depends on dataset 100–256 Hz	Depends on dataset	Depends on dataset	physionet.org/
OpenNeuro	Depends on dataset (275 EEG datasets, until 14/March/2025, including epilepsy, sleep disorders, etc.)	Depends on dataset 1–100 +	Depends on dataset 5–10,000 +	Depends on dataset	Depends on dataset	Follow the FAIR principles and BIDS	openneuro.org/
MindBigData	EEG signals captured with the stimulus of seeing a digit (from 0 to 9) and thinking about it	1 (David Vivancos)	1,207,293 (version 1.03, 2014)	Depends on Device 128–512 Hz	Depends on Device 1–14	.txt	mindbigdata.com/opendb/index.html
DEAP	Human affective states	32	1280	512 Hz	32	.bdf	eecs.qmul.ac.uk/mmv/datasets/deap/
SEED series	Affective EEG and multimodal emotion recognition; extended to vigilance, fatigue, visual perception and sleep deprivation	Depends on dataset 8–40	Depends on dataset 34–80	Depends on dataset 200 Hz (down-sampled, 1000 Hz raw)	Depends on dataset 18–62	.cnt & .edf (raw) .mat (processed)	bcmi.sjtu.edu.cn/home/seed/index.html
FACED	Affective EEG (9 emotions)	123	3444	250 Hz, 1000 Hz	32	.bdf, .mat & .pkl	doi.org/10.7303/syn50614194
BCI Competitions	Motor imagery EEG data	Depends on dataset < 10 each	Depends on dataset	Depends on dataset 250–1000 Hz	Depends on dataset 3–118	.mat	bbci.de/competition/
MAHNOB-HCI tagging database	Implicit tagging of multimedia content	30	120	Not specified	32	.bdf	ieeexplore.ieee.org/document/5975141mah-nob-db.eu/hci-tagging/media/uploads/manual.pdf

Table 2: Summary of AD-focused EEG databases.

Database	Study focus	Participants	Records	Sampling rate	Channels	File formats #	Source
CAUEEG	Dementia, MCI (Mild Cognitive Impairment), Normal, FTD (Frontotemporal dementia), Parkinson's Syndrome, NPH (Normal Pressure Hydrocephalus), TGA (Transient Global Amnesia)	1155	1379	200 Hz	19	.edf	github.com/ipis-mjkim/caueeg-dataset
A dataset of EEG recording from: AD, FTD and Healthy participants	Resting stage closed-eyes recordings Open-eyes recording in a photo-stimulation settings	88 (AD: 36; FTD: 23; healthy: 29)	88 (closed-eyes) 88 (open-eyes)	500 Hz	19	**.set**	Closed-eyes: openneuro.org/datasets/ds004504/versions/1.0.8 Open-eyes: openneuro.org/datasets/ds006036/versions/1.0.5
Data from: Computational Methods of EEG Signals Analysis for Alzheimer's Disease Classification	AD (eye-opened and eye-closed)	184 (12 eye-opened healthy elderly; 12 eye-closed healthy elderly; 80 eye-opened probable AD; 80 eye-closed probable AD)	390	128 Hz	19	.txt	osf.io/2v5md/overview
40 Hz Auditory Entrainment	Auditory stimulation on Normal, MCI and AD	35 (17 females)	35 (.fdt) 35 (.set)	250 Hz	19	**.fdt & .set**	openneuro.org/datasets/ds005048/versions/1.0.1/
Brain EEG Recording during Olfactory Stimulation in MCI & AD Patients	Olfactory oddball perception task on AD, MCI, and healthy participants	35 (57.14% females; AD: 13, MCI: 7, healthy: 15)	35	200 Hz (down-sampled, 2 kHz raw)	4	.mat	data.mendeley.com/datasets/sgzbgwjfkr/5
EEG Signals from Normal and MCI Cases	Closed-eyes resting status MCI and normal	27 (16 normal and 11 MCI)	27	256 Hz	19	.edf	misp.mui.ac.ir/en/eeg-data-0

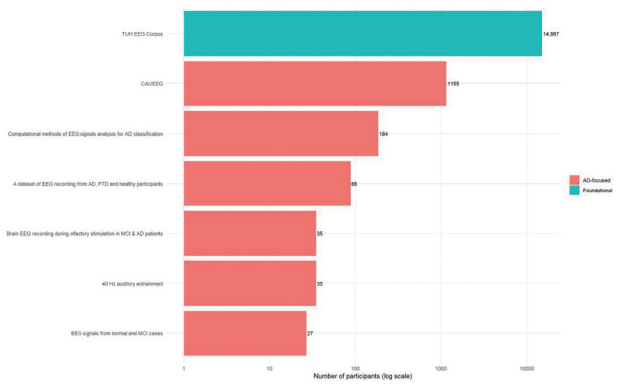

Figure 1: Sample-size disparity between AD-focused public EEG datasets and foundational EEG resources. Note: EEG, electroencephalography.

AD datasets (Table 2) form a fragmented landscape of potential. This landscape of valuable but disconnected data is fundamentally incapable of powering the large-scale, data-driven research we envision. This is the central bottleneck. The solution to this paralysis is standardization. In the following chapter, we will propose a concrete framework to implement it.

Standardization: the cornerstone of the new paradigm

The fragmentation of existing data does not merely cause inconvenience; it leads research astray. Integrating heterogeneous data, particularly by merging EEG records with different sampling rates, filter settings, or electrode montages, is scientifically unsound. It inevitably introduces profound biases that can obscure true biological signals or generate spurious findings[13–15]. While modern large model architectures demonstrate increasing technical resilience to heterogeneous inputs, relying on such inconsistency for clinical discovery remains a fundamental risk. Establishing robust, clinically reproducible biomarkers demands a harmonised baseline, regardless of the model's capacity. Feeding these models a patchwork of inconsistent data guarantees "garbage in, garbage out" rendering any resulting biomarker unreliable for medical applications[13,16]. Furthermore, this lack of a common standard fuels the reproducibility crisis in neuroscience, including NDD research[16], and impedes the development of automated, high-throughput analysis pipelines, trapping researchers in a Sisyphean cycle of writing bespoke scripts[17].

Here, we distinguish between prospective standardization and retrospective harmonisation. Prospective standardization refers to predefined protocols applied before data generation, including clinical metadata, acquisition parameters, ethical governance, and file organisation. Its purpose is to prevent irreversible information loss. Retrospective harmonization, by contrast, refers to post-acquisition strategies that improve the interoperability of already collected heterogeneous datasets, including BIDS conversion, signal normalisation, metadata mapping, domain adaptation, and, federated learning. These approaches complement, rather than replace, one another.

For future AD-oriented EEG cohorts, the core instrument of prospective standardization is a comprehensive SOP. To address the full spectrum of challenges from compliance to analysis, this SOP should include four levels of content: **(1) Ethical and procedural standardization**: establishing strict protocols for data desensitisation, privacy protection, consent for longitudinal follow-up, and compliance governance to ensure the legality of large-scale, multi-centre sharing; **(2) Clinical standardization**: using Common Data Elements (CDEs)[18] for demographic information, cognitive scales, diagnostic criteria, medication status, comorbidities, biomarker status, and longitudinal outcomes; **(3) Acquisition standardization**: defining unified sampling rates, filters, reference schemes, electrode montages, impedance criteria, recording duration, and task status, such as eyes-closed resting-state EEG and selected task-related paradigms, to maximise the retention of original information while minimising irreversible operations; **(4) Format standardization**: mandating the community-wide adoption of the Brain Imaging Data Structure (BIDS)[19,20], which transforms isolated files into a unified, machine-readable, analysis-ready dataset. For newly collected EEG, BIDS should be embedded prospectively into the SOP; for legacy datasets, BIDS conversion provides a practical route for retrospective harmonisation.

This call for a comprehensive SOP is not merely theoretical. A powerful precedent already exists. The CHBBC has successfully implemented a multi-centre SOP for the complex logistics of human tissue collection[11,12,21,22]. Crucially, this infrastructure is designed to manage

dual repositories: physical biological specimens and their corresponding digital assets (e.g., clinical records, digitised pathology, and omics). Therefore, incorporating digital EEG into this archive is not a reconstruction but a logical and seamless extension. Rather than claiming a completed resource, we propose a staged initiative to link ante-mortem functional EEG with post-mortem neuropathology. Longitudinal brain-donation cohorts have already demonstrated the feasibility of connecting ante-mortem clinical data with neuropathological endpoints, including BDR in the UK[23], ROSMAP in the United States[24], and BANC-PUMC cohort in China[25]. Importantly, isolated precedents such as Arizona Study of Aging and Neurodegenerative Disorders and Brain and Body Donation Program (AZSAND/BBDP) have further shown that premortem resting-state EEG can be analyzed in autopsy-confirmed NDD cohorts[26]. However, the systematic incorporation of standardized EEG into such living-to-post-mortem frameworks remains rare and underdeveloped.

In short, prospective standardization provides the trustworthy baseline for future EEG cohorts, while retrospective harmonization maximises the usability of existing heterogeneous datasets. Together, they build the foundation that the new paradigm scientifically demands.

Analytical evolution: from manual feature extraction to AI-driven discovery

The standardized, large-scale data infrastructure established in the previous chapter provides the essential "clean fuel". However, this fuel requires more powerful analytical engines to convert it into discovery. This chapter charts the evolution of these engines, transitioning from traditional methods to data-driven approaches enabled by this new paradigm.

Historically, EEG analysis depended on visual inspection and quantitative methods such as spectral analysis. These approaches were foundational, translating the raw signals into interpretable features based on prior hypotheses. This process, while essential for interpretability, is limited by its design, risking the oversight of complex, non-linear patterns[27]. The field evolved towards more advanced feature engineering, using techniques such as functional connectivity and microstate analysis. This knowledge-driven approach peaked in systems such as the dementia screen proposed by Huang et al.[28]. By manually crafting novel non-linear features, they achieved remarkable classification accuracy (AUC > 0.91). This success proves the immense value hidden in EEG signals, but it also reveals the fundamental limitation: the discovery of features remains a manual, hypothesis-bound process.

This leads to end-to-end (E2E) deep learning. A common view frames this as a complete replacement for manual feature extraction. This perspective is inaccurate. E2E learning does not bypass human knowledge; it abstracts it. Our decades of domain expertise are not discarded. Instead, they are encoded into the model's architecture as inductive biases[29]. While the Transformer architecture currently spearheads this revolution, it operates within a broader architectural constellation, where each design mirrors a classical neuroscientific hypothesis. Just as Convolutional Neural Networks (CNNs) enforce local receptive fields to capture time-frequency patterns, and Graph Neural Networks (GNNs) model the brain's topology to mimic functional connectivity, the Transformer's attention mechanisms[2] is a direct implementation of the hypothesis that long-range temporal dependencies are critical in EEG signals[30]. Thus, the E2E model is not an unbiased "black box" but a powerful tool that scales up our human-defined priors.

However, this power is bound by an inescapable reality regarding data. It is true that modern foundation models demonstrate the engineering resilience to ingest heterogeneous inputs—handling inconsistent length or sampling rates via technical adaptations. However, clinical discovery demands more than engineering feasibility; it demands biological validity. Relying on inconsistent data risks training models that learn the "noise" of the acquisition protocol rather than the "signal" of the pathology. Therefore, simply feeding data—following the "Scaling Laws"[31] —is insufficient if the data source acts as a confounder; rigorous harmonisation of processing pipelines remains a non-negotiable imperative. This data dependency reinforces the central argument of this review: the analytical methods

have evolved to a point where our fragmented data landscape (chapter 2) is the primary bottleneck.

To deploy these next-generation discovery engines, we must first build the foundation from chapter 3. However, this synergy between data and models is not automatic, and significant challenges remain. First, robustness is not guaranteed. A model trained on one database can learn site-specific artefacts, especially in the low signal-to-noise ratio EEG data, risking "Shortcut Learning" that compromises clinical validity across centres[32]. Besides, the issue of interpretability is paramount[33]. An E2E model may discover predictive patterns that are physiologically opaque, requiring a return to rigorous, hypothesis-driven methods for validation[34]. These challenges—robustness and interpretability—cannot be solved by a better algorithm alone. Furthermore, they highlight a third deficit: the lack of scalable infrastructure to validate and deploy these models across diverse centres. These systemic problems demand a new ecosystem for collaborative discovery. Building this predictable and reliable blueprint is the task we turn to next.

Future blueprint: a new ecosystem for brain research

The transition from manual analysis to AI-driven discovery faces a "triad of deficits": Robustness, Interpretability, and Scalability (RIS). As outlined in chapter 4, these are not merely technical footnotes but systemic barriers. They cannot be overcome by isolated algorithmic improvements; they demand a new infrastructure for analysis and a new ecosystem for collaboration. This final chapter proposes a blueprint for this future, built directly upon the three technological waves we introduced at the outset.

The foundation of this solution is the technical infrastructure, and this is where scalable cloud computing becomes the engine of discovery. We envision an open, modular platform that moves beyond simple data storage[35]. This vision is grounded in reality; successful precedents such as the CHBBC's existing portal have already demonstrated the feasibility of centralizing resources on a national scale. It is precisely this unified architecture that addresses the first challenge: robustness (R). It allows for large-scale, federated model training and validation across diverse, multi-centre datasets, ensuring that biomarkers are generalizable and not mere site-specific artefacts[36]. By requiring models to be trained, stress-tested, and externally validated across centres, the platform turns multi-centre heterogeneity from an uncontrolled confounder into an explicit benchmark for robustness.

This cloud platform is also the stage for our new analytical engines. Here, the Transformer architecture (chapter 4) and prospective standardization (chapter 3) converge. The platform hosts advanced E2E models, including Transformer-based EEG models, while the standardized data provides the "clean fuel" they require[37]. This synergy directly addresses the second challenge: interpretability (I). For EEG, practical explainable artificial intelligence (XAI) should be organized around neurophysiological dimensions rather than generic model explanations alone, using methods such as saliency analysis, occlusion testing, or attention/relevance visualization[38,39]. First, temporal attribution can identify which recording windows contribute most to a prediction. Second, spectral attribution can test whether the model relies on biologically plausible frequency components, such as slowing-related change in AD. Third, spatial or connectivity-based attribution can examine whether relevant channels, regions, or network patterns are consistent with known neurodegenerative involvement. Embedding these XAI outputs into the cloud platform would allow different centres to compare, reproduce, and challenge the same model explanations. By validating these explanatory outputs against standardized data, cognitive trajectories, neuroimaging markers, and post-mortem pathology, we can democratize access to advanced computation and collectively translate opaque patterns into physiological understanding.

With this infrastructure and its powerful analytics in place, the platform addresses the final challenge: scalability (S). This is not just about data volume, but about longitudinal depth, and biological dimensionality. This platform becomes the hub for a new collaborative ecosystem. Its true power is unlocked when it integrates data far beyond resting-state EEG. The highest priority is to create synergy with Human Brain Banks,

allowing ante-mortem functional dynamics to be correlated with post-mortem neuropathology. Equally important is integration with large-scale prospective cohorts. Retrospective harmonization can maximize the value of existing EEG datasets, but it cannot create the missing silent-phase observations by itself. These observations require prospectively standardized EEG embedded into longitudinal studies that capture cognition, imaging, genomics, and multi-omics. Such cohorts would allow us to track the transition from silent pathology to clinical decline and explore the complex "gene-environment-brain" interactions. This ecosystem must also capture real-world data through wearable devices, moving from clinical snapshots to the longitudinal monitoring required for true personalisation.

This technical and scientific blueprint, however, cannot succeed without a corresponding cultural shift. The technical capacity for scalability is useless without the communal will to share. We must champion a move towards Open Science, grounded in the FAIR Principles (Findable, Accessible, Interoperable, and Reusable). This openness must be built on a foundation of trust. To sustain this ecosystem, the privacy of participants is non-negotiable, requiring state-of-the-art techniques like differential privacy and strong ethical governance. Only by securing trust can we ensure the data flow necessary for a truly robust and scalable science.

The paradigm shift we described in chapter 1 is not a spectator sport. It requires building. The blueprint outlined here—a synergistic platform designed to ensure robustness, clarify interpretability, and enable scalability (RIS)—is the necessary next step. It is not merely a technical roadmap for a database; it is a call for a new, collaborative method of scientific discovery. Although AD serves as the index disease in this perspective, mixed pathology makes the proposed framework relevant to NDDs more broadly. By solving the RIS challenges, we can move from fragmented EEG records to clinically and biologically grounded understanding across the ageing brain.

Author contributions

Heyuan Xiao conceptualised the review, conducted the literature search, and drafted the original manuscript. Fangzheng Zhou and Kunshuan Xie provided critical revisions and refined the conceptual framework. Tao Wang supervised the project, acquired funding, and critically reviewed the manuscript. All authors have read and approved the final version of the manuscript.

Acknowledgements

This work was supported by grants from: STI2030-Major Projects-2022ZD0211800, Task 3 #2022ZD021180003.

Data availability statement

No datasets were generated or analyzed for this work.

Ethics statement

Not applicable.

AI statement

The authors used Gemini 3.1 Pro as a language-editing tool to improve grammar, clarity, and readability. The AI tool was not used for data analysis, data generation, image processing, or scientific content creation. All outputs were carefully reviewed and verified by the authors, who take full responsibility for the final content.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Hey AJG, editor. The fourth paradigm: data-intensive scientific discovery. Redmond, Washington: *Microsoft Research*; 2009. 251
2. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: advances in neural information processing systems [Internet]. *Curran Associates, Inc.*; 2017 [cited 2025 Oct 19]. Available from the website of proceedings.
3. Dubois B, Hampel H, Feldman HH, et al. Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimers Dement*. 2016;12(3):292-323.
4. Wu W, Wang X, Xu Y, Ma C, Qian X, Qiu W. Neuropathological comorbidity, genetics and cognition in a Chinese community-based autopsy cohort. *Brain*. 2025; 148(7):2481-2492.
5. Counts SE, Ikonomovic MD, Mercado N, Vega IE, Mufson EJ. Biomarkers for the early detection and progression of Alzheimer's disease. *Neurotherapeutics*. 2017;14(1): 35-53.

6. Jack CR, Bennett DA, Blennow K, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement*. 2018;14(4):535-562.
7. Simmatis LE, Russo EE, Steininger T, et al. EEG biomarkers for Alzheimer's disease: a novel automated pipeline for detecting and monitoring disease progression. *J Alzheimers Dis*. 2025;108(1_suppl):S94-S104.
8. Michail Marinis, Eleni Vrochidou, George A. Papakostas. Recent developments in electroencephalography with deep learning models. In: advances in bioelectromagnetism [Internet]. *Academic Press*; 2026 [cited 2026 Jan 30]. p.177-93.
9. Smailovic U, Jelic V. Neurophysiological markers of Alzheimer's disease: quantitative EEG approach. *Neurol Ther*. 2019;8(Suppl 2):37-55.
10. Nunez PL, Srinivasan R. Electric fields of the brain [Internet]. *Oxford University Press*; 2006 [cited 2026 Jan 30]. Available from OUP.
11. Xue W, Juanli W, Naili W, et al. Standardized operational protocol for the China Human Brain Bank Consortium. *Human Brain*. 2022 Oct 24;1(1).
12. Chen Z, Wang X, Wu J, et al. Standardized operational protocol of human brain banking for amyotrophic lateral sclerosis. *Human Brain*. 2023 Feb 25;2(1):1.
13. Liu J, Cui P. Data heterogeneity modeling for trustworthy machine learning. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 [Internet]. *Toronto ON Canada: ACM*; 2025 [cited 2025 Oct 19]. p. 6086-6095.
14. Cote MP, Lubowitz JH, Rossi MJ, Brand JC. Reviews pooling heterogeneous, low-evidence, high-bias data result in incorrect conclusions: but heterogeneity is an opportunity to explore. *Arthroscopy*. 2018;34(12):3126-3128.
15. Kim B, Kim H, Kim K, Kim S, Kim J. Learning not to learn: training deep neural networks with biased data. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. *Long Beach, CA, USA: IEEE*; 2019 [cited 2025 Oct 19]. p. 9004-9012. Available from the website of ieeexplore.
16. Troller-Renfree SV, Morales S, Buzzell GA, Sandre A. Heterogeneity in pediatric resting EEG data processing and analysis: a state of the field. *Psychophysiology*. 2025;62(2):e14733.
17. Putrama IM, Martinek P. Heterogeneous data integration: challenges and opportunities. Data Brief. 2024;56:110853.
18. Hao X, Abeysinghe R, Zheng F, Schulz PE, Cui L. Mapping of Alzheimer's disease related data elements and the NIH common data elements. *BMC Med Inform Decis Mak*. 2024;24(Suppl 3):103.
19. Gorgolewski KJ, Auer T, Calhoun VD, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data*. 2016;3:160044.
20. Pernet CR, Appelhoff S, Gorgolewski KJ, et al. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci Data*. 2019;6(1):103.
21. Liu F, Yin XS, Cong C, Wang Y, Ma C. Human brain banking as a convergence platform of neuroscience and neuropsychiatric research. *Human Brain*. 2022 Oct 24; 1(1):46-62.
22. Qiu W, Zhang H, Bao A, et al. Standardized operational protocol for human brain banking in China. *Neurosci Bull*. 2019;35(2):270-276.
23. Francis PT, Costello H, Hayes GM. Brains for dementia research: evolution in a longitudinal brain donation cohort to maximize current and future value. *J Alzheimers Dis*. 2018;66(4):1635-1644.
24. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious orders study and rush memory and aging project. *J Alzheimers Dis*. 2018;64 (s1):S161-S189.
25. Han F, Qiu WY, Zhou LX, et al. The brain aging national cohort-PUMC: study design and baseline characteristics. *Human Brain*. 2023 Mar 25;2(2).
26. Choi A, Zhang N, Adler CH, et al. Resting-state EEG predicts cognitive decline in a neuropathologically diagnosed longitudinal community autopsied cohort. *Parkinsonism Relat Disord*. 2024;128:107120.
27. Alqudah A, Moussavi Z. A review of deep learning for biomedical signals: current applications, advancements, future prospects, interpretation, and challenges. *Comput Mater Continua*. 2025;83(3):3753-841.
28. Huang NE, Yuan WS, Yuan F, et al. A fast screen system for dementia based on EEG signals. *Human Brain*. 2025 Jun 28;4(1):1.
29. Battaglia PW, Hamrick JB, Bapst V, et al. Relational inductive biases, deep learning, and graph networks [Internet]. arXiv; 2018 [cited 2026 Jan 30]. Available from the website of arxiv.
30. Wan Z, Li M, Liu S, Huang J, Tan H, Duan W. EEGformer: a transformer-based brain activity classification method using EEG signal. *Front Neurosci*. 2023;17:1148855.
31. Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models [Internet]. arXiv; 2020 [cited 2026 Jan 30]. Available from the website of arxiv.
32. Geirhos R, Jacobsen JH, Michaelis C, et al. Shortcut learning in deep neural networks. *Nat Mach Intell*. 2020 Nov;2(11):665-73.
33. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models [Internet]. arXiv; 2017 [cited 2026 Jan 30]. Available from the website of arxiv.
34. Kriegeskorte N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu Rev Vis Sci*. 2015;1:417-446.
35. Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. *Nat Neurosci*. 2014;17(11): 1510-1517.
36. Zhu SY, Tang W, Xie YT, Xie LZ. A comprehensive

review of federated learning for multi-center medical data. Biomedical Engineering Communications [Internet]. 2025 Feb 21 [cited 2026 May 24];4(2). Available from the website of tmrjournals.

37. Acharya M, Deo RC, Tao X, et al. Deep learning techniques for automated Alzheimer's and mild cognitive impairment disease using EEG signals: a comprehensive review of the last decade (2013-2024). Comput Methods *Programs Biomed*. 2025;259:108506.
38. Khan W, Khan MS, Qasem SN, et al. An explainable and efficient deep learning framework for EEG-based diagnosis of Alzheimer's disease and fronto-temporal dementia. *Front Med* (*Lausanne*). 2025;12: 1590201.
39. Sujatha Ravindran A, Contreras-Vidal J. An empirical comparison of deep learning explainability approaches for EEG using simulated ground truth. *Sci Rep*. 2023; 13(1):17709.