Ensuring Quality of EEG Data Sharing for the Development of Human-Centered Machine Learning and Artificial Intelligence Systems

Aimilia Ntetska[1[0009-0009-4683-1153]], Andreas Miltiadous[2[0000-0003-0675-9088]], Markos G. Tsipouras[1[0000-0002-6757-1698]], Pantelis Angelidis[1[0000-0002-6757-1698]], Nikolaos Giannakeas[2[0000-0002-4278-3301]], Alexandros T. Tzallas[2[0000-0003-1503-8952]] and Katerina D. Tzimourta[1[0000-0001-9640-7005]]

[1] Laboratory of Biomedical Technology and Digital Health, Dept. of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece
[2] Dept. of Informatics and Telecommunications, University of Ioannina, Arta, Greece

Abstract. Machine learning (ML) applied to neurological disorders promises improved diagnosis and discovery of biomarkers but it critically depends on high-quality data. Electroencephalography (EEG) data, being a sensitive biometric recording of brain activity, poses unique ethical, human and social challenges when reused for research. This paper examines these challenges in the context of neurological disorders, focusing on dementia (especially Alzheimer's disease and frontotemporal dementia) and epilepsy. We discuss two recent open-access EEG datasets on dementia and contrast the sparse data landscape for dementia with the abundance of epilepsy EEG datasets. We critically analyze ethical issues, transparency, participant agency, consent granularity and reuse governance – and explore how they influence perceptions of software/AI system quality. Finally, we propose principles for a human-centered, dynamic consent framework tailored to EEG-based ML applications, aiming to align data reuse with ethical and social values.

Keywords: EEG, data sharing, ethical governance, artificial intelligence, consent

1 Introduction

In recent years, the integration of machine learning (ML) with clinical neurophysiology has opened new avenues for early diagnosis and monitoring of neurological disorders. Among the most promising data sources in this domain is electroencephalography (EEG), a non-invasive, low-cost method for capturing real-time brain activity. EEG has proven valuable in detecting patterns associated with conditions such as epilepsy and dementia, including Alzheimer's disease (AD) and frontotemporal dementia (FTD) [1,2]

Despite advances in EEG signal analysis, the progress of ML-based tools in clinical practice remains uneven. A key bottleneck is the limited availability of high-quality, openly accessible EEG datasets, particularly in neurodegenerative disorders. While epilepsy research has long benefited from large public databases (e.g., TUH [3] CHB-

MIT [4]), dementia research has historically lacked similarly structured, well-governed datasets—hindering algorithmic generalization, reproducibility, and comparability across studies 5.

Beyond technical availability, EEG data carry inherent ethical, legal, and social implications [6]. As biometric signals closely tied to personal identity and health status, EEG recordings raise concerns around participant consent, data governance, privacy, and reuse accountability, especially when applied to vulnerable populations such as older adults with cognitive impairment.

This article answers these challenges by comparing two recently published, open-access EEG datasets for Alzheimer's disease and dementia with more established epilepsy datasets, contrasting their quality and governance. In addition to characterizing the structure and safeguards of these dementia datasets, we provide a qualitative framework for assessing EEG data quality from an ethical and governance perspective. These five pillars include explicit consent and ethical approval, licensing openness, preprocessing and standardization, metadata richness, and governance mechanisms. Taking these six popular datasets as a test case, we aim to provide a systematic, human-centered framework for evaluating EEG data transparency and reusability in machine learning research.

2 Background

A wide range of open-access EEG datasets has been published over the past two decades, supporting the development of machine learning models in various neurological domains. While epilepsy has historically been the primary focus of such datasets, only recently high-quality, openly shared EEG databases have become available for dementia research.

The first dataset [7] comprises EEG recordings from 88 participants, including 36 Alzheimer’s disease (AD) patients, 23 frontotemporal dementia (FTD) patients, and 29 cognitively healthy controls. EEG was recorded with a standard 19-electrode scalp montage during resting state (eyes closed), following rigorous quality control [7]. A 19-electrode scalp montage was used to record the EEG while the subjects were in an eyes-closed resting state. Clinical metadata, such as Mini Mental State Exaination (MMSE) scores, was supplied. Following strict quality control, both raw and preprocessed data were shared in BIDS format. Informed consent was acquired, the study was approved by ethics, and the dataset was made available on OpenNeuro under a CC0 license, garnering over 123,000 views and 7,700 downloads.

Ntetska et al. [8] expanded on this by publishing a supplementary dataset from the same cohort that recorded EEG during eyes-open photic stimulation, a procedure that probes visual reactivity. It was completely anonymized, ethically approved, and compliant with BIDS, just like the first. When combined, the two datasets enable investigation of both spontaneous and stimulus-driven EEG patterns in AD and FTD. By releasing data from a vulnerable population, the authors set a precedent in transparency and open science for dementia; a field that has historically lagged behind others in data

sharing. The careful oversight and anonymization applied demonstrate that such sensitive data can be shared ethically.

The above datasets stand out because there are no other EEG datasets for dementia publicly available. This scarcity is in contrast to the field of epilepsy, where numerousEEG databases have been openly shared for years. For example, the Temple University Hospital EEG Corpus [3] – specifically its seizure subset (TUSZ) – is the largest open source corpus of its type for epileptic EEG, containing many hours of

seizure recordings from hundreds of patients. Another well-known resource is the CHB-MIT Scalp EEG Database [4], which provides pediatric seizure EEGs from Children's Hospital Boston physionet.org . In Europe, the EPILEPSIAE project [9] established a comprehensive epilepsy EEG database (~275 patients, including long-term recordings and extensive metadata), which is "by far the largest and most comprehensive database for human surface and intracranial EEG data" as of its release[9]. These epilepsy datasets are the main EEG datasets in ML-based epilepsy research, enabling countless studies on seizure detection and prediction[10]. In dementia research, most EEG studies until recently used proprietary data from single labs/hospitals, limiting generalizability and slowing progress. The lack of open dementia EEG data was not just a technical gap but an ethical and social one: without shared data, there is duplication of effort and patients' contributions in one study cannot benefit broader science as readily. The new dementia EEG datasets help bridge this gap, embodying a more open-science approach akin to what epilepsy researchers have long practiced.

3 Methodology

The EEG data utilized and publicly released in the two dementia datasets referenced in this study were collected as part of a clinically approved protocol at the AHEPA University Hospital, Aristotle University of Thessaloniki, Greece. The main goal was to use non-invasive scalp EEG to look into neurophysiological biomarkers of dementia subtypes.

Three diagnostic groups—36 with Alzheimer's disease (AD), 23 with frontotemporal dementia (FTD), and 29 cognitively healthy, age-matched controls—made up the 88 participants who were enrolled. The Department of Neurology handled recruitment, and a multidisciplinary team comprising neurologists and EEG technicians assessed each participant. Standard clinical criteria (such as the Neary criteria for FTD and the NINCDS-ADRDA for AD) were used to make the diagnosis, which was backed up by medical records, neuropsychological testing (such as the MMSE), and neuroimaging when it was available.

3.1 EEG Recording Protocol

Two EEG acquisition protocols were followed:

- **Dataset 1 – Resting-State EEG** [7]: Participants underwent a standard 19-channel scalp EEG using the international 10-20 system. EEG was recorded

during resting state with eyes closed for a duration of 10-15 minutes. The recording environment was controlled to minimize external noise and artifacts.

- **Dataset 2 – Photic Stimulation EEG**[8]**:** The same participants subsequently underwent EEG during intermittent photic stimulation (IPS). Light flashes of varying frequencies, starting from 5 Hz and reaching up to 30 Hz, were presented while participants kept their eyes open. This protocol aimed to capture visually evoked potentials and rhythmic reactivity across different brain regions.

EEG data were recorded using clinical-grade amplifiers with a sampling rate of 500 Hz and impedance kept below 5 kΩ. The signals were bandpass filtered during acquisition (0.5-70 Hz) and visually inspected for quality assurance.

3.2 Data Preprocessing and Format

All EEG recordings underwent post-hoc preprocessing that included:

- Artifact rejection (e.g., eye movements, muscle noise)
- Independent Component Analysis (ICA)
- Band-specific filtering for common EEG rhythms

Both datasets were curated and formatted in compliance with the Brain Imaging Data Structure (BIDS) standard for EEG, ensuring compatibility and reusability by the global research community. Data are available in both raw and preprocessed versions.

3.3 Ethical Approval and Informed Consent

The entire study protocol — including both EEG recording procedures and open data release plan — was approved by:

a. The Scientific Committee of the AHEPA University Hospital
b. The Administrative Board of the hospital

All participants (or legal representatives) provided written informed consent for participation and open public dissemination of anonymized data. Consent processes were compliant with the Declaration of Helsinki, GDPR guidelines, and national ethical standards.

This strong ethical and institutional framework created participant trust and enabled the efficient global dissemination of these datasets. The end-to-end process—ensuring transparency, traceability, and ethical integrity along the data lifecycle—is illustrated in Figure 1.

The process begins with Institutional Approval, where scientific and administrative boards sign off on the EEG protocol and data-sharing plan. This is followed by Participant Consent, where there is explicit anonymized open-access release under GDPR and Declaration of Helsinki. The EEG signals then undergo Preprocessing and Standardization (e.g., artifact rejection, ICA, BIDS formatting). Researchers then prepare Extensive Metadata and Documentation for reproducibility. Finally, the chosen data are made available on an Open-Access EEG Repository (e.g., OpenNeuro) under a permissive license (e.g., CC0).

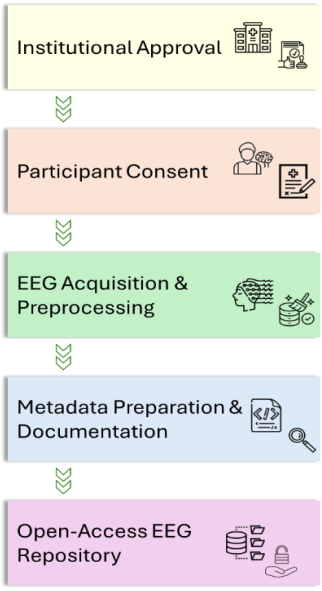

Fig. 1. Workflow for the ethical collection, curation, and public dissemination of EEG data in neurological research.

4 Discussion

The open-access dementia EEG data sets studied in this paper are not only technical successes at data sharing but also models of ethical integration throughout the data life cycle. Institutional clearance, consent for reuse, and CC0-licensed, BIDS-compliant releases increase trust and usefulness particularly in very vulnerable communities, such as those with cognitive impairment. However, challenges in neurophysiological data sharing persist. Traditional consent models tend to be either overly broad, risking inadequate participant understanding, or overly narrow, limiting future, unforeseen research.

To address this, granular consent has gained traction. It enables participants to permit certain data uses (e.g., neurological, psychiatric, or AI research) while excluding others, and to define acceptable data recipients (e.g., academic-only or ethics-approved projects). Yet, implementing such flexibility remains difficult. Withdrawal of consent following data release is practically difficult, and most ethics committees prefer fixed study objectives. Open-ended sharing tends to need specific advance consent [11]. For retrospective data sets, absence of advance sharing consent can bar release except where participants are traced and re-contacted [11]. This highlights the need for prospective and potentially dynamic consent models [11].

To investigate how these ethical aspects affect the practical quality and reusability of EEG data for machine learning, we compared six commonly used datasets. Based on this, we defined five fundamental pillars of EEG data quality:

a) Ethical approval and explicit consent
b) License/openness
c) Pre-processing and standardization
d) Metadata richness
e) Governance and traceability

Table 1 benchmarks each dataset against these criteria. This comparison does not aim to assign scores or judge scientific merit, but rather to descriptively assess how EEG datasets align with five key pillars derived from recurring norms in open neuroscience and neuroethics, emphasizing ethical transparency, reusability, and governance. Data were collected from publicly available documentation, emphasizing structural variation such as adherence to BIDS, licensing, metadata richness, and governance instruments.

Miltiadous et al. [12] and Ntetska et al.'s [8] datasets on dementia fully meet all pillars in that they combine twin hospital-board approval with written consent, CC0 licensing, BIDS-compliant curation (raw and cleaned), complete clinical metadata coverage, and open, versioned repositories. Each of these are non-uniform in compliance compared to epilepsy datasets: CHB-MIT, though BIDS-conversion and open-licensed, lacks rich metadata; TUH/TUSZ has rich annotation but is bound by a Data Use Agreement; EPILEPSIAE has clinical and imaging depth but remains contractually bounded and proprietary; the older Bonn dataset lacks licensing, standardization, and metadata. While they all have contributed to machine learning based on EEG, these discrepancies highlight the need for harmonized, ethically sound standards. Compared to general frameworks such as FAIR or GDPR, which are non-specific for neurophysiological s, the five-pillar model enhances them by addressing domain-specific problems such as consent granularity, annotation rigor, and licensing clarity to support human-centered governance and actionable practice that translate ethical ideals into explicit EEG data curation.

Table 1. Comparative assessment of EEG datasets across five quality pillars (ethics, licensing, standardization, metadata, governance).

Dataset (*year*)	Institutional ethics & consent	Licence/ Openness	Pre-processing & standardizationn	Metadata richness	Governance / traceability
Miltadous *et al.,* (2023) [7]	✓ Scientific & administrative hospital boards; written informed consent	✓ CC0 public release on OpenNeuro	✓ Raw & cleaned files, full BIDS package	✓ Demo-graphics, MMSE, clinical notes	✓ Public logs; permanent DOI; versioning
Ntetska *et al.,* (2025) [8]	✓ Scientific & administrative hospital boards; written informed consent	✓ CC0 public release on OpenNeuro	✓ Raw & cleaned files, full BIDS package	✓ Mirrors first dataset; photic-stim parameters	✓ Public logs; permanent DOI; versioning
TUH / TUSZ [3]	✓ IRB approval; clinical consent	‣ Data-Use Agreement (free but request-based)	‣ EDF; no native BIDS	✓ Extensive seizure annotations	‣ Access logs; DUA enforcement

Dataset (*year*)	Institutional ethics & consent	Licence/ Openness	Pre-processing & standardizationn	Metadata richness	Governance / traceability
CHB-MIT [4]	✓ IRB Boston Children's; parental consent	✓ Open Data Commons Attribution	✓ Community BIDS conversion available	‣ Basic demographics, limited clinical info	‣ Public download; no granular consent
EPILEPSIAE [9]	✓ Multi-centre ethics approvals	✗ Closed; access only by signed contract	✗ Proprietary structure; not BIDS	✓ Rich clinical, imaging & seizure meta	‣ Steering-committee gate-keeping
Bonn [13]	✓ Local IRB	✗ No explicit licence	✗ Raw ASCII; no standardi-sation	✗ No demographics / clinical	✗ No access logs or governance

Overall, this framework offers a methodical approach to the evaluation of EEG datasets used in machine learning pipelines for their ethical readiness. Though applied initially here to dementia and epilepsy data, it can readily be extended to other neurophysiological modalities (e.g., MEG, fNIRS) and diseases where data sensitivity is significant. Integrating ethical and governance-aware assessments upfront duringthe data life cycle may improve not only public trust and transparency, but also downstream fairness and resilience of AI systems trained on them.

5 Conclusion

Reusing EEG data in machine learning applications for neurological disorders holds great promise, but also raises important ethical, legal, and governance-related concerns. In this paper, we examined two recently published open-access dementia EEG datasets as exemplary cases that combine technical rigor with ethical transparency. We further proposed and applied a five-pillar framework for evaluating EEG dataset quality from a human-centered perspective, focusing on consent, licensing, standardization, metadata richness, and governance. This framework, presented as a proof-of-concept, offers a structured approach for assessing ethical readiness in neurophysiological data reuse. By benchmarking dementia and epilepsy datasets through this lens, we highlighted critical disparities and opportunities for improvement. Embedding such ethics-aware assessments into dataset design and sharing practices can strengthen both scientific reproducibility and public trust, laying a stronger foundation for equitable and responsible AI in neuroscience.

Acknowledgments. This research is implemented within the framework of the National Recovery and Resilience Plan "Greece 2.0" with funding from the European Union – NextGenerationEU through the program "SUB1.1. Research Excellence Partnerships ", grant number YP3TA-0559562, project "Dynamic CONSENT management in complex data workfows: techniques and tools" - CONSENT.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Lal, U.; Chikkankod, A.V.; Longo, L. A Comparative Study on Feature Extraction Techniques for the Discrimination of Frontotemporal Dementia and Alzheimer's Disease with Electroencephalography in Resting-State Adults. *Brain Sci* **2024**, *14*, doi:10.3390/brainsci14040335.
2. Joshi, V.; Nanavati, N. A Review of EEG Signal Analysis for Diagnosis of Neurological Disorders Using Machine Learning. *J Biomed Photonics Eng* 2021, *7*.
3. Shah, V.; von Weltin, E.; Lopez, S.; et al. The Temple University Hospital Seizure Detection Corpus. *Front Neuroinform* **2018**, *12*, doi:10.3389/fninf.2018.00083.
4. Goldberger, A.L.; Amaral, L.A.N.; Glass, ; Leon; Hausdorff, J.M.; Plamen, ; Ivanov, C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; et al. *PhysioBank, PhysioToolkit, and PhysioNet Components of a New Research Resource for Complex Physiologic Signals*; 2000;
5. Cassani, R.; Estarellas, M.; San-Martin, R.; Fraga, F.J.; Falk, T.H. Systematic Review on Resting-State EEG for Alzheimer's Disease Diagnosis and Progression Assessment. *Dis Markers* 2018, *2018*.
6. Amadio, J.; Bi, G.Q.; Boshears, P.F.; Carter, A.; et al. Neuroethics Questions to Guide Ethical Research in the International Brain Initiatives. *Neuron* 2018, *100*, 19–36.
7. Miltiadous, A.; Tzimourta, K.D.; Afrantou, T.; et al. A Dataset of Scalp EEG Recordings of Alzheimer's Disease, Frontotemporal Dementia and Healthy Subjects from Routine EEG. *Data (Basel)* **2023**, *8*, doi:10.3390/data8060095.
8. Ntetska, A.; Miltiadous, A.; Tsipouras, M.G.; et al. A Complementary Dataset of Scalp EEG Recordings Featuring Participants with Alzheimer's Disease, Frontotemporal Dementia, and Healthy Controls, Obtained from Photostimulation EEG. *Data (Basel)* **2025**, *10*, 64, doi:10.3390/data10050064.
9. Ihle, M.; Feldwisch-Drentrup, H.; Teixeira, C.A.; Witon, A.; Schelter, B.; Timmer, J.; Schulze-Bonhage, A. EPILEPSIAE - A European Epilepsy Database. *Comput Methods Programs Biomed* **2012**, *106*, 127–138, doi:10.1016/j.cmpb.2010.08.011.
10. Miltiadous, A.; Tzimourta, K.D.; Giannakeas, N.; et al. Machine Learning Algorithms for Epilepsy Detection Based on Published EEG Databases: A Systematic Review. *IEEE Access* **2023**, *11*, 564–594, doi:10.1109/ACCESS.2022.3232563.
11. Poline, J.B.; Breeze, J.L.; Ghosh, S.; et al. Data Sharing in Neuroimaging Research. *Front Neuroinform* 2012, *6*.
12. Miltiadous, A.; Tzimourta, K.D.; Giannakeas, N .; et al. Diagnostics Alzheimer's Disease and Frontotemporal Dementia: A Robust Classification Method of EEG Signals and a Comparison of Validation Methods. **2021**, *11*, 1437, doi:10.3390/diagnostics.
13. Andrzejak, R.G.; Lehnertz, K.; Mormann, F.; et al. Indications of Nonlinear Deterministic and Finite-Dimensional Structures in Time Series of Brain Electrical Activity: Dependence on Recording Region and Brain State. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **2001**, *64*, 8, doi:10.1103/PhysRevE.64.061907.