

The Data Management Crisis Behind AI-Based Brain MRI Diagnosis: Heterogeneity, Governance, and Reproducibility

Tanmoy Debnath
Charles Sturt University
Bathurst, New South Wales, Australia
tdeb Nath@csu.edu.au

Mirosław Narbutt
Technological University Dublin
Dublin, Ireland
mirosław.narbutt@TUDublin.ie

Katarzyna Anna Kazimierczak
University of Bergen
Bergen, Norway
katarzyna.a.kazimierczak@uib.no

ABSTRACT

AI-based brain magnetic resonance imaging (MRI) classifiers face a persistent gap between laboratory performance and clinical deployability. We argue the bottleneck is not model architecture but the absence of principled data management infrastructure. Drawing on our ongoing work in explainable multi-disease MRI classification, we identify five critical data management challenges: heterogeneity, governance misalignment, pipeline irreproducibility, annotation provenance, and demographic bias. To address these issues, we propose a minimal proof-of-concept metadata harmonization toolkit as a practical and tractable starting point for standardization. We further outline a validation protocol that measures field-level mapping coverage, clinician-reviewed preservation of phenotype and severity-score semantics, provenance completeness, and downstream classifier transfer across ADNI and OASIS-3. We invite the data management community to co-design the schema-mapping and query-optimization layers.

VLDB Workshop Reference Format:

Tanmoy Debnath, Mirosław Narbutt, and Katarzyna Anna Kazimierczak. The Data Management Crisis Behind AI-Based Brain MRI Diagnosis: Heterogeneity, Governance, and Reproducibility. VLDB 2026 Workshop: Biomedical Data Management Systems (BioDMS).

1 INTRODUCTION

Brain diseases—encompassing neurodegenerative disorders (Alzheimer’s, Parkinson’s, amyotrophic lateral sclerosis (ALS) and related motor neuron diseases) and primary brain tumors—represent a leading cause of disability-adjusted life years globally. MRI is a central non-invasive modality for brain assessment, yet manual interpretation is time-consuming and subject to substantial inter-observer variability for early-stage presentations. Deep learning approaches, particularly residual convolutional architectures and transfer learning pipelines, have demonstrated strong benchmark performance [4, 8, 24]. However, the gap between laboratory results and clinical deployability has persisted for over a decade [31]. Our ongoing project developing explainable, multi-disease MRI classifiers reveals that the principal bottleneck is not model architecture but the absence of robust supporting data-management

infrastructure. Recent advances in continuous integration for machine learning (ML), data-quality verification, healthcare analytics, cohort discovery, and medical-imaging privacy converge on a common conclusion: reliable clinical AI requires robust infrastructure for data validation, integration, querying, governance, and reproducible execution. This body of work collectively demonstrates that such systems, rather than increasingly complex predictors alone, are essential for trustworthy deployment in clinical settings [5, 6, 18, 19, 21, 32, 34]. We present five problem clusters as concrete open problems for the data systems community.

2 FIVE DATA MANAGEMENT CRISES

2.1 Multi-Source Heterogeneity

Major neuroimaging repositories—including Alzheimer’s Disease Neuroimaging Initiative (ADNI) [17], the Open Access Series of Imaging Studies (OASIS) [22], IXI, and the Brain Tumor Segmentation benchmark (BraTS) [24]—differ substantially in acquisition protocols (field strength, echo time, repetition time, voxel resolution), scanner manufacturer, and patient demographics. Table 1 summarizes five widely used repositories based on public documentation for ADNI, OASIS-3, BraTS, IXI, and UK Biobank.

Figure 1 illustrates how clinical severity metadata is represented differently across ADNI, OASIS-3, and BraTS, highlighting incompatibilities in naming, coding, instrument choice, and recording granularity.

Metadata schemas in clinical imaging remain inconsistent and often incomplete, with acquisition parameters, scanner calibration records, and clinical phenotype definitions variably encoded or missing altogether. Existing ontology efforts, e.g., NIDM and RadLex, address only a narrow subset of the problem and fall well short of the breadth, maturity, and adoption of the Gene Ontology in genomics, while longitudinal phenotyping and severity-score support remain sparse and non-mandatory. Federated querying, i.e., asking one query across data that remain distributed across institutions, is therefore infeasible without bespoke, manually curated harmonization scripts that neither generalize across studies nor are version-controlled. Prior data-systems work on semantic typing, table embeddings, table-union search, and biomedical identifier mapping demonstrates the infrastructure needed for robust cross-repository discovery [7, 16, 20, 32].

2.2 Governance Misaligned with AI Training

Access conditions remain fragmented: repositories impose differing data use agreements, institutional registration requirements, and embargo periods. Cross-border training exposes unresolved tensions between the General Data Protection Regulation (GDPR)’s

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment, ISSN 2150-8097.

Table 1: Comparison of five major neuroimaging repositories.
Label granularity: scan (S), subject (Su), and cohort (C).

Repository	Access model	BIDS	Labels	Preproc. documented
ADNI	Restricted	Partial	Su	Partial
OASIS-3	Restricted	Partial	Su	Partial
BraTS	Open	No	S	None
IXI	Open	No	None	None
UK Biobank	Tiered	No	Su	Partial

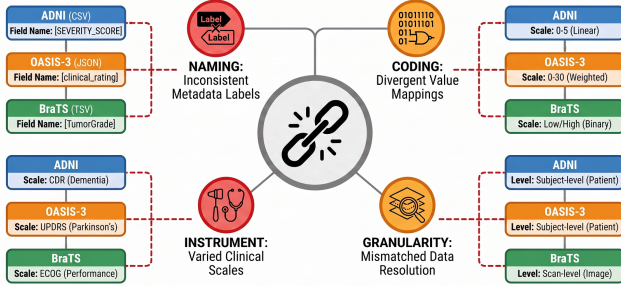


Figure 1: Schema divergence in clinical severity metadata across ADNI, OASIS-3, and BraTS

strict purpose-limitation regime and the Health Insurance Portability and Accountability Act (HIPAA)’s comparatively broader secondary-use allowances, while divergent transfer mechanisms create governance friction that cannot be resolved through administration alone. The resulting gap calls for technical controls that can enforce policy, provenance, and data-use constraints across jurisdictions. Federated benchmarking, biomedical data-discovery interfaces, and privacy-preserving medical-image analysis suggest partial solutions but reveal that governance, cohort selection, and secure computation must be designed together [1, 2, 6, 21, 23].

2.3 Preprocessing Pipeline Irreproducibility

Neuroimaging pipelines (registration, intensity normalization, atlas-based segmentation) are sensitive to software versions (FSL, FreeSurfer, ANTs, etc.) and parameter configurations. The choice of skull-stripping tool alone, e.g., FSL BET, HD-BET, or ANTs, produces measurably different brain masks that propagate divergent volumetric estimates downstream. These pipelines are rarely containerized or publicly deposited, making independent replication practically impossible: performance metrics cannot be unambiguously attributed to architecture versus preprocessing choices versus dataset composition. Continuous integration for ML, automated data-quality verification, validity constraints, and hybrid provenance systems provide concrete building blocks [3, 9, 18, 19, 30].

2.4 Annotation Provenance and Ground Truth

Supervised multi-disease classification requires clinician-verified, scan-level labels. Existing datasets either lack such labels, aggregate multi-rater annotations without recording inter-rater reliability or adjudication protocol, or provide no machine-readable link to the

scoring instruments from which labels derive. Severity scoring requires ordinal labels referenced to validated clinical instruments (e.g., Alberta Stroke Program Early CT Score (ASPECTS) for stroke, Clinical Dementia Rating (CDR) for dementia, and Unified Parkinson’s Disease Rating Scale (UPDRS) for Parkinson’s disease), yet this linkage is rarely maintained in queryable, structured form. For severity and phenotype variables, correctness is not merely syntactic field alignment; mappings should preserve the clinical instrument, scoring range, score directionality, assessment date, rater or adjudication status, and diagnostic context.

Biomedical mapping resources and cohort-discovery systems show that portable and auditable clinical labels require explicit identifier mappings, curation workflows, and interpretable cohort definitions [5, 16]. Medical-image labeling workflows also need data models that record label creation, reviewer decisions, annotator disagreement, and label reliability [25, 29, 33].

2.5 Ethical Risks from Demographic Bias

Public MRI repositories are not demographically representative: ADNI recruited predominantly non-Hispanic white, college-educated participants [17], and similar skews are documented across major repositories. This systematic under-representation risks AI systems that perpetuate healthcare disparities. Addressing this is a data governance problem—requiring curation-level demographic auditing, dataset documentation, and bias-aware quality checks—as much as an algorithmic one. Recent data-quality work explicitly connects error detection and debugging with fairness and robustness, supporting the view that demographic bias should be monitored as part of the data lifecycle rather than treated as a post-hoc model issue [19].

3 REMAINING GAPS

Brain Imaging Data Structure (BIDS) provides a widely adopted convention for organizing and describing MRI datasets, including structural MRI, resting-state fMRI, and task-fMRI, but clinical disease metadata (severity scores, comorbidities, medication history) remains sparsely covered and non-mandatory. Moreover, repositories such as ADNI and OASIS-3 are often distributed in DICOM or repository-specific formats, while BIDS-compatible versions typically depend on additional conversion and curation workflows. OpenNeuro [27] offers BIDS-validated open sharing but is oriented toward cognitive neuroscience, with no mechanism to link scans to electronic health records. DataLad [14] addresses provenance tracking and version control for large datasets, but it does not resolve clinical metadata harmonization, governance alignment, or cross-repository query semantics. fMRIPrep [12] and sMRIPrep [11] advance preprocessing standardization for functional and structural MRI, respectively, while MRIQC [10] provides automated image-quality assessment; however, these tools do not resolve clinical metadata harmonization, governance alignment, annotation provenance, or cross-repository query semantics.

Federated learning frameworks, e.g., FeTS [28] demonstrate that cross-site training is technically feasible but presuppose harmonized site data—precisely the unsolved problem. Data-systems research offers important component solutions: robust ingestion [32], semantic typing, table embeddings [7], table search [20], biomedical

Table 2: Infrastructure coverage gaps for clinical neuroimaging AI. Coverage is based on public documentation; “Target” denotes planned proof-of-concept functionality.

INFRASTRUCTURE	Clinical Metadata	Provenance & Versioning	Cross-repository Query	Governance & Access	EHR Linkage
BIDS	Sparse	Partial	No	No	No
OpenNeuro	Partial	No	No	Open Only	No
DataLad	No	Yes	No	No	No
fMRIPrep/sMRIPrep	No	Yes	No	No	No
MRIQC	No	Partial	No	No	No
FeTS	No	No	No	Yes	No
Proposed Toolkit	Target	Target	Target	Target	Target

semantic mappings [16], data-quality verification, workflow validity constraints [30], provenance systems [3], cohort discovery [5], disease-centric portals [1], and scalable biomedical indices.

Table 2 positions the proposed toolkit against existing infrastructure and shows that no current system jointly supports the metadata, provenance, query, governance, and linkage capabilities required for multi-disease neuroimaging AI.

4 PROOF-OF-CONCEPT DIRECTION

The challenges above are tractable open research problems in data integration, metadata management, provenance tracking, and federated query optimization—areas of core VLDB and SIGMOD expertise. Five immediate targets are:

- (1) A unified, extensible neuroimaging metadata schema compatible with Fast Healthcare Interoperability Resources (FHIR), Digital Imaging and Communications in Medicine (DICOM), and Brain Imaging Data Structure (BIDS), with mandatory clinical phenotype fields and machine-readable severity-score linkage supported by biomedical semantic mapping resources [13, 15, 16, 26].
- (2) Query languages and optimization strategies for heterogeneous, partially open, multi-site repositories with differing access tiers and schema conventions [5, 7, 20, 32].
- (3) Provenance-aware pipeline management enforcing containerization, data-quality verification, validity constraints, and reproducible re-execution across sites [3, 18, 19, 30].
- (4) Data models for annotation uncertainty, inter-rater disagreement, adjudication history, and identifier-level clinical label provenance in medical image labeling workflows [29, 33].
- (5) Privacy-preserving federated learning infrastructure aligned with biomedical ethics governance, removing the need for pre-harmonized site data [1, 6, 21, 23].

Minimal proof-of-concept direction. We propose a lightweight community prototype—a *metadata harmonization toolkit*—that (i) ingests repository-specific manifests (e.g., ADNI CSV, OASIS JSON, and BraTS TSV files), (ii) maps fields to a common BIDS-extension schema using embedding-based ontology alignment and

LLM-assisted field disambiguation, and (iii) generates a versioned, queryable SQLite index with provenance links to the source files.

Figure 2 presents the proposed metadata harmonization toolkit, from repository-specific manifest ingestion to ontology-aware field alignment, provenance capture, and construction of a versioned queryable index over immutable source records.

The prototype requires no central storage or duplication of raw imaging data. It will be evaluated on a small cross-repository Alzheimer’s disease cohort constructed from ADNI and OASIS-3. To ensure a falsifiable assessment, evaluation will be conducted at three levels. First, metadata harmonization performance will be quantified by reporting the numbers of source fields ingested, automatically mapped fields, manually validated fields, ambiguous fields, and fields rejected as semantically unsafe. Second, clinical-semantic validity will be assessed by verifying diagnostic labels, severity measures, phenotype definitions, and temporal variables against the original repository data dictionaries and, where feasible, through review by a clinician or neuroimaging domain expert. Third, downstream utility will be evaluated by comparing cohort construction and machine learning model transfer before and after harmonization, reporting cohort overlap, missingness reduction, balanced accuracy, AUROC, calibration, and cross-repository generalization. Figure 3 summarizes the planned validation framework, linking harmonization quality to clinical-semantic validity and downstream utility in cohort construction and classifier transfer.

Table 3 summarizes the proposed evaluation framework and the corresponding metrics used to assess the utility, validity, and auditability of the harmonization toolkit. We invite data management researchers and repository maintainers to contribute to the co-design of the schema-mapping and query-optimization components and to participate in the community-driven evaluation of the framework.

Loss-aware harmonization. Beyond maintaining provenance links, a key design principle of the proposed toolkit is that harmonization should not overwrite or replace source metadata. Instead, harmonized metadata are represented as a queryable view over immutable source records. Each canonical field retains links to the source field, original value, source instrument, unit of measurement, transformation rule, mapping confidence score, reviewer

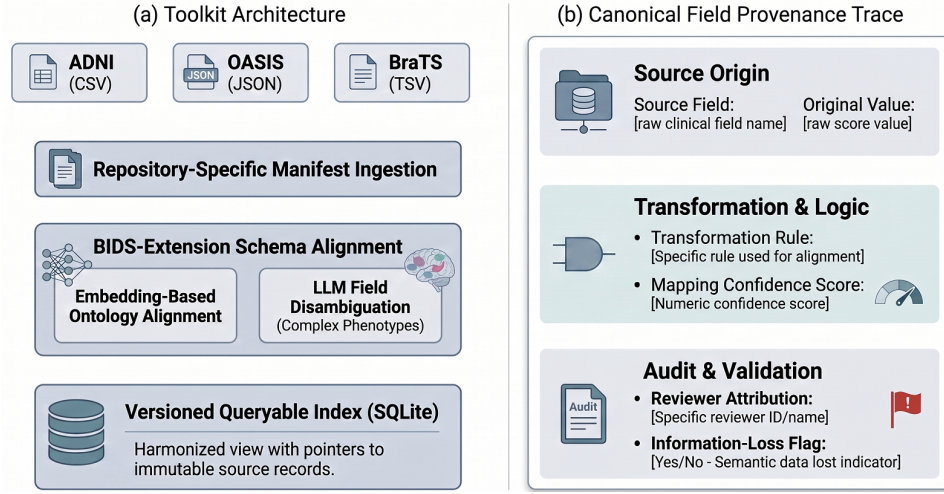


Figure 2: (a) Proposed architecture of the metadata harmonization toolkit. (b) Provenance trace for one canonical field, preserving the source field, original value, transformation rule, mapping confidence, reviewer attribution, and information-loss flag.

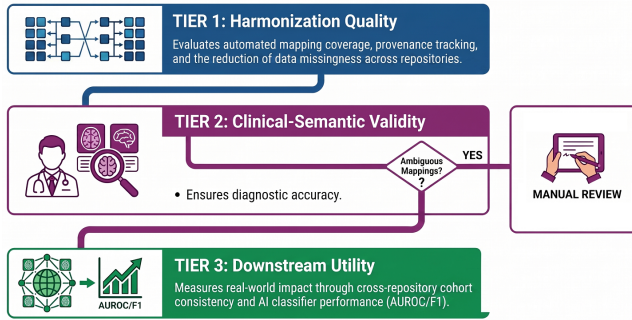


Figure 3: Planned three-tier validation framework.

Table 3: Evaluation dimensions for the toolkit.

Target	Metric	Goal
Schema mapping	% fields mapped automatically/manually/rejected	Demonstrates toolkit coverage beyond a conceptual proposal.
Clinical meaning	Clinician agreement on severity and phenotype mappings	Tests preservation of the source instrument’s meaning.
Provenance	% fields with source and transformation trace	Ensures auditability and reproducibility.
Data quality	Missingness before versus after harmonization	Shows practical value for cohort construction.
Query utility	Cross-repository cohort-query consistency	Demonstrates cross-repository discovery.
Classifier utility	AUROC, F1, balanced accuracy before versus after harmonization	Connects infrastructure to diagnostic utility.

attribution, review status, and an explicit information-loss flag. Ambiguous mappings remain accessible for inspection and querying but are excluded from clinical-label evaluation until manually reviewed. This loss-aware design makes semantic uncertainty and information loss explicit rather than implicit, thereby improving transparency, auditability, and reproducibility. Furthermore, every harmonized value can be traced directly to its repository-specific origin, enabling downstream researchers to verify transformations and assess the impact of harmonization decisions.

5 PROPOSED TALK AND COLLABORATION

This lightning talk will present concrete examples from harmonizing ADNI, OASIS-3, and BraTS for multi-disease classification. It will include preliminary mapping coverage and information-loss tracking outputs, where available, and will conclude with problem statements designed to initiate dialogue on co-developing infrastructure solutions. We seek to establish a working group to develop a metadata harmonization prototype and a community benchmark. This benchmark will consist of a curated cross-repository cohort with harmonized metadata, documented preprocessing provenance, clinical-semantic validation of labels and severity scores, and downstream classifier-transfer checks, released under permissive licenses. Algorithmic progress alone cannot resolve these structural impediments; collaboration with the data management community is essential.

ACKNOWLEDGMENTS

The authors express their gratitude to the anonymous reviewers for their insightful comments and rigorous critique, which improved the quality of this manuscript. Furthermore, we extend our appreciation to the workshop organizers and the Program Committee for their efforts in facilitating the event.

AUTHORS

Dr. Tanmoy Debnath is a Senior Lecturer in Engineering at Charles Sturt University, Australia. His research interests include artificial intelligence, digital health, medical imaging, and data science. From 2024 to 2026, he secured competitive funding for four research projects in AI, e-health, and data science, acting as the principal investigator on two projects. His research has been published in Q1 journals, CRC Press (USA) book chapters, and workshops affiliated with leading CORE A* conferences, including ICDM 2025, SIGMOD 2026, KDD 2026, and VLDB 2026.

Dr. Mirosław Narbutt is a lecturer at Technological University Dublin, Ireland. He holds a PhD in Computer Science from University College Dublin, Ireland, and an MSc in Electronics and Telecommunications Engineering from Poznań University of Technology, Poland. His research interests include real-time multimedia transmission (WebRTC), immersive audio (Ambisonics), deep learning methods in medical diagnostics, and generative audio. Mirosław Narbutt is the inventor of two patented technologies, one of which is used by YouTube/Google.

Dr. Katarzyna Kazimierczak received her PhD in Neuroscience from the University of Bergen, Norway, where she is currently a postdoctoral researcher. She specialises in MRI data analysis, including functional (fMRI), structural (sMRI), and magnetic resonance spectroscopy (MRS) using MATLAB, Python, and specialised brain imaging software, with research focused on advancing neuroimaging methods for clinical and cognitive neuroscience.

REFERENCES

- [1] Robert J Allaway, Salvatore La Rosa, Sharad Verma, Lara Mangravite, Justin Guinney, Jaishri Blakeley, Annette Bakker, and Sara JC Gosline. 2019. Engaging a community to enable disease-centric data sharing with the NF Data Portal. *Scientific data* 6, 1 (2019), 319.
- [2] Muhammad Zeerak Awan, Nabeel Ahmed Khan, Petr Strakos, Adil Jhangeer, and Lubomir Riha. 2026. Privacy-preserving federated learning with light-weight attention improved CNNs for automated leukemia detection across distributed medical imaging. *Scientific Reports* (18 Feb. 2026). <https://doi.org/10.1038/s41598-026-40581-9>
- [3] Vasileios Bountris, Lauritz Thamsen, and Ulf Leser. 2025. HyProv: Hybrid provenance management for scientific workflows. *arXiv preprint arXiv:2511.07574* (2025).
- [4] Biagio Breve, Loredana Caruccio, Gianluca Cimino, Stefano Cirillo, Giulia Iuliano, and Giuseppe Polese. 2022. Brain tumors classification from MRI images: A comparative study between different neural networks. In *DMSVIVA 2022 – Proceedings of the 28th International DMS Conference on Visualization and Visual Languages*. 23–30.
- [5] Qingpeng Cai, Kaixin Zheng, H. V. Jagadish, Beng Chin Ooi, and Jonathan W. L. Yip. 2024. CohortNet: Empowering cohort discovery for interpretable healthcare analytics. *Proceedings of the VLDB Endowment* 17, 10 (2024), 2487–2500. <https://doi.org/10.14778/3675034.3675041>
- [6] Stefano Cirillo, Vincenzo Deufemia, Luca Di Biasi, Giuseppe Polese, Gian-domenico Solimando, and Genoveffa Tortora. 2025. CADHE: Privacy-preserving medical image analysis through homomorphic encrypted convolutional networks. In *2025 IEEE International Conference on Big Data*. 7064–7072. <https://doi.org/10.1109/BigData66926.2025.11401707>
- [7] Tianji Cong, Madelon Hulsebos, Zhenjie Sun, Paul Groth, and HV Jagadish. 2023. Observatory: Characterizing embeddings of relational tables. *arXiv preprint arXiv:2310.07736* (2023).
- [8] Tanmoy Debnath and Mirosław Narbutt. 2026. Toward Digital Twin-Ready Brain Tumor Segmentation: A Multi-Pillar Comparison of Transformer and Convolutional Models. In *2026 ACM Knowledge Discovery and Data Mining (KDD) Conference: 25th International Workshop on Data Mining in Bioinformatics (BIOKDD): ACM KDD Conference’26*. Association for Computing Machinery (ACM).
- [9] Xinwei Du, Shijun Li, Xiaojun Wang, Yuan Shen, and Tingwei Quan. 2026. A high-resolution dataset of mouse brain vasculature for deep learning-based reconstruction. *Frontiers in Neuroinformatics* 20 (2026), 1809341. <https://doi.org/10.3389/fninf.2026.1809341>
- [10] Oscar Esteban, Daniel Birman, Marie Schaefer, Oluwasanmi O. Koyejo, Russell A. Poldrack, and Krzysztof J. Gorgolewski. 2017. MRIQC: Advancing the Automatic Prediction of Image Quality in MRI from Unseen Sites. *PLOS ONE* 12, 9 (2017), e0184661. <https://doi.org/10.1371/journal.pone.0184661>
- [11] Oscar Esteban, Christopher J. Markiewicz, Ross Blair, Russell A. Poldrack, and Krzysztof J. Gorgolewski. 2023. sMRIprep: Structural MRI PREProcessing workflows. <https://doi.org/10.5281/zenodo.8253909>
- [12] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. 2019. fMRIprep: a robust preprocessing pipeline for functional MRI. *Nature methods* 16, 1 (2019), 111–116.
- [13] Krzysztof J. Gorgolewski, Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, Satrajit S. Ghosh, Tristan Glatard, Yaroslav O. Halchenko, Daniel A. Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, B. Nolan Nichols, Thomas E. Nichols, John Pellman, Jean-Baptiste Poline, Ariel Rokem, Gunnar Schaefer, Vanessa Sochat, William Triplett, Jessica A. Turner, Gaël Varoquaux, and Russell A. Poldrack. 2016. The Brain Imaging Data Structure, a Format for Organizing and Describing Outputs of Neuroimaging Experiments. *Scientific Data* 3 (2016), 160044. <https://doi.org/10.1038/sdata.2016.44>
- [14] Yaroslav O. Halchenko, Kyle Meyer, Björn Poldrack, Debanjum Solanky, Adina Wagner, Jason Gors, Dave MacFarlane, Dorian Pustina, Vanessa Sochat, Satrajit S. Ghosh, et al. 2021. DataLad: distributed system for joint management of code, data, and their relationship. *Journal of Open Source Software* 6, 63 (2021), 3262. <https://doi.org/10.21105/joss.03262>
- [15] HL7 International. 2023. FHIR Release 5.0.0: Fast Healthcare Interoperability Resources. <https://hl7.org/fhir/R5/> FHIR v5.0.0, generated March 26, 2023.
- [16] Charles Tapley Hoyt, Klas Karis, and Benjamin M. Gyori. 2025. Assembly and reasoning over semantic mappings at scale for biomedical data integration. *Bioinformatics* 41, 11 (2025), btaf542. <https://doi.org/10.1093/bioinformatics/btaf542>
- [17] Shaveta Kanoria, Dallas P Veitch, Melanie J Miller, Paul S Aisen, Laurel A Beckett, Robert C Green, Danielle J Harvey, Clifford R Jack Jr, William Jagust, Edward B Lee, et al. 2026. Clinical impact of the Alzheimer’s Disease Neuroimaging Initiative: A review of studies using ADNI data (2023 to June 2025). *Alzheimer’s & Dementia* 22, 4 (2026), e71353.
- [18] Bojan Karlas, Matteo Interlandi, Cedric Renggli, Wentao Wu, Ce Zhang, Deepak Maddix I. Babu, Jordan Edwards, Chelsea Lauren, Andy Xu, and Markus Weimer. 2020. Building continuous integration services for machine learning. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2407–2415. <https://doi.org/10.1145/3394486.3403290>
- [19] Bojan Karlas, Babak Salimi, and Sebastian Schelter. 2025. Navigating data errors in machine learning pipelines: Identify, debug, and learn. In *Companion of the 2025 International Conference on Management of Data*. 895–898. <https://doi.org/10.1145/3722212.3725636>
- [20] Aamod Khatiwada, Grace Fan, Roe Shraga, Zhengjie Chen, Wolfgang Gatterbauer, Renée J. Miller, and Mirek Riedewald. 2023. SANTOS: Relationship-based semantic table union search. In *Proceedings of the 2023 ACM SIGMOD International Conference on Management of Data*. <https://doi.org/10.1145/3588689>
- [21] Eugenie Lai, Gerardo Vitagliano, Zainab Zhang, Shrainik Sudhir, Omid Chabira, Andrew Zeng, Aleksander A. Zabreyko, Chunwei Li, Frank Kossmann, Jialin Ding, Jie Chen, Manolis Markakis, Michael Russo, Weijie Wang, Ziniu Wu, Michael J. Cafarella, Lydia Cao, Samuel Madden, and Tim Kraska. 2025. KramaBench: A benchmark for AI systems on data-to-insight pipelines over data lakes. *arXiv preprint arXiv:2506.06541* (2025).
- [22] Pamela J. LaMontagne, Tammie L. S. Benzing, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. 2019. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv* (2019). <https://doi.org/10.1101/2019.12.13.19014902>
- [23] Dustin Lange, Shuying Gao, Pamela Sui, Andrew Money, Patrick Misner, Marinka Zitnik, and Nils Gehlenborg. 2025. A generative AI system for biomedical data discovery with grammar-based visualizations. *IEEE VIS Workshop on GenAI, Agents, and the Future of VIS*. <https://doi.org/10.48550/arXiv.2509.16454>
- [24] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. 2015. The multimodal brain tumour image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging* 34, 10 (2015), 1993–2024. <https://doi.org/10.1109/TMI.2014.2377694>
- [25] Mirosław Narbutt and Tanmoy Debnath. 2026. Compact Metrics, Catastrophic Failures: Trustworthy Benchmarking of Vision-Language Foundation Models in 3D Medical Segmentation. In *The 1st International Workshop on Sustainable and Efficient Language, Vision, and Action Models (SELVA) at ACL 2026*.

- [26] National Electrical Manufacturers Association. [n.d.]. NEMA PS3 / ISO 12052: Digital Imaging and Communications in Medicine (DICOM) Standard. <https://www.dicomstandard.org/> Available free online.
- [27] OpenNeuro Team. 2026. OpenNeuro: An open archive for analysis and sharing of BRAIN Initiative data. <https://openneuro.org>. Accessed 6 July 2026.
- [28] Sarthak Pati, Ujjwal Baid, Maximilian Zenk, Brandon Edwards, Micah Sheller, G. Anthony Reina, Patrick Foley, Alexey Gruzdev, Jason Martin, Shadi Albarqouni, et al. 2022. Federated learning enables big data for rare cancer boundary detection. *Nature Communications* 13 (2022), 7346. <https://doi.org/10.1038/s41467-022-33407-5>
- [29] Tim Radsch, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Schreck, A. Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. 2023. Labelling Instructions Matter in Biomedical Image Analysis. *Nature Machine Intelligence* 5, 3 (2023), 273–283. <https://doi.org/10.1038/s42256-023-00625-5>
- [30] Florian Schintke, Khalid Belhajjame, Ninon De Mecquenem, David Frantz, Vanessa Emanuela Guarino, Marcus Hilbrich, Fabian Lehmann, Paolo Missier, Robert Sattler, Jürgen A. Sparka, Daniel T. Speckhard, Hendrik Stolte, Anh Duy Vu, and Ulf Leser. 2024. Validity constraints for data analysis workflows. *Future Generation Computer Systems* 157 (2024), 82–97. <https://doi.org/10.1016/j.future.2024.03.037>
- [31] Sahar Soltanieh, Farzad Khalvati, and E. Ann Yeh. 2026. Federated Learning in Neurology: Bridging Data Privacy and Artificial Intelligence for Brain Health. *Seminars in Neurology* 46, 1 (2026), 38–48. <https://doi.org/10.1055/a-2769-6752> Epub 2025 Dec 29.
- [32] Gerardo Vitagliano, Mazhar Hameed, Lan Jiang, Lars Reisener, Edward Wu, and Felix Naumann. 2023. Pollock: A data loading benchmark. *Proceedings of the VLDB Endowment* 16, 8 (2023), 1870–1882. <https://doi.org/10.14778/3594512.3594518>
- [33] Feng Yang, Ghada Zamzmi, Sandeep Angara, Sivaramakrishnan Rajaraman, André Aquilina, Zhiyun Xue, Stefan Jaeger, Emmanouil Papagiannakis, and Sameer K. Antani. 2023. Assessing Inter-Annotator Agreement for Medical Image Segmentation. *IEEE Access* 11 (2023), 21300–21312. <https://doi.org/10.1109/ACCESS.2023.3249759>
- [34] Kaixin Zheng, Shaofeng Cai, Horng Ruey Chua, Melanie Herschel, Meihui Zhang, and Beng Chin Ooi. 2022. DyHealth: Making neural networks dynamic for effective healthcare analytics. *Proceedings of the VLDB Endowment* 15, 12 (2022), 3445–3458. <https://doi.org/10.14778/3554821.3554835>