eBRAIN-Health

Public report D5.3

High throughput container workflows

Project number	**101058516**
Project title	**eBRAIN-Health - Actionable Multilevel Health Data**
Submission date	**July 2025**
Authors	**Michael Hanke, Felix Hoffstaedter, Nicolás Nieto (FZJ)**
Dissemination level	**Public (PU)**
Public project website	**https://ebrain-health.eu**

Funded by
the European Union

Table of content

Funded by
the European Union

1. eBRAIN-Health

The Project eBRAIN-Health will deliver a distributed research platform for modeling and simulating complex neurobiological phenomena of human brain function and dysfunction in a data protection compliant environment. It will provide thousands of multilevel virtual brains from patients and healthy human controls for research and innovation. Brain data from multiple sources will be pre-processed. Solving the societal grand challenge of dementia is a big task. Yet it appears feasible in a collective approach. Therefore, we will build an interdisciplinary digital twin for dementia for modeling and simulating complex phenomena at the service of research infrastructure communities. eBRAIN-Health-Cloud will offer end-to-end services for personalized complex brain modeling and simulations in distributed e-infrastructures with data protection by design and by default and simulation-ready human multiscale brain data that range from molecular (genomics, proteomics, metabolomics) and cellular to electrophysiology and imaging to behavioural, clinical, life-style and environmental data as well as data from wearables. Brain data are pre-processed and annotated such that they all relate to a common reference 3D brain space.
eBRAIN-Health-Cloud constitutes a blend of three large-scale research programs: the FET Flagship Human Brain Project with its EBRAINS Research Infrastructure, the EOSC project Virtual Brain Cloud with its Virtual Research Environment for sensitive data and the H2020 project AI-MIND with intelligent tools for dementia risk estimation. The project will have synergies to topics of the Digital Europe Program, such as artificial intelligence, cybersecurity and supercomputing and the Health Data Space.

eBRAIN-Health-Cloud offers a next generation clinical research infrastructure and creates an open yet protected space for groundbreaking digital health innovation by the research infrastructure communities comprising academia and the private sector.

2. eBRAIN-Health consortium

- CHARITE – Universitaetsmedizin Berlin, Germany
- EBRAINS, Belgium
- Forschungszentrum Juelich GmbH, Germany
- Stichting Radboud Universiteit, Netherlands
- Universidad Pompeu Fabra, Spain
- OSLO Universitetssykehus, Norway
- tp21 GMBH, Germany
- Fraunhofer Gesellschaft zur Foerderung der Angewandten Forschung eV, Germany
- INDOC RESEARCH EUROPE gGmbH, Germany
- Universitaet Wien, Austria
- Universidad Complutense de Madrid, Spain
- EODYNE Systems SL, Spain
- ATHENA – Research and Innovation Center, Greece
- University of Oslo, Norway
- Universita degli Studi di Roma la Sapienza, Italy
- Alzheimer Europe, Luxembourg
- Institute National de Recherche en Informatique et Automatique, France
- Centre Hospitalier Universitaire Vaudois, Switzerland
- The University of Edinburgh, United Kingdom

Find the partners on our website

3. Introduction

We present a generalized high-throughput workflow that automatically captures full provenance of data processing in containerized compute environments. The primary focus of this approach is on the further development of the eBRAINs neuroinformatics infrastructure for distributed data management, with a particular emphasis on its interface with high-performance computing environ-ments. A particular consideration is the capability to work with distributed repositories for bringing together compute resources and data. Typically, image processing consists of a series of successive steps: initial conversion from raw scanner formats, within-subject registration, spatial normaliza-tion, confound removal, and processing of measured data into parametric images, among others. In this context full provenance tracking is critical, as each step allows for multiple options and pa-rameters. In light of the necessity for extensive, population-level datasets, scalability and com-pute/storage efficiency also assume a pivotal role in the advancement and refinement of the pipe-line system. This pipeline framework will serves as the base for the development of normative models and allow for training of machine learning models to perform individual prediction tasks.

4. Partners involved

Forschungszentrum Jülich GmbH, Germany

5. Description of work performed

The following work was performed:

1. Software implementation of fully reproducible containerized processing pipelines for structural and functional MRI data including quality assessment.
2. Assessment of the influence of MR image quality on statistical and predictive analysis for generalizable recommendations of data inclusion in robust analysis pipelines

5.1. Reproducible containerized processing pipelines for MRI data

Hence, we propose a bootstrap approach for the reproducible setup of an entire processing workflow for a given dataset with a specific pipeline using a single shell script. This procedure capitalizes on the capabilities of:

1) Datalad as well tested, distributed data management tool.
2) Singularity as reliable software hosting environment.
3) HTCondor and SLURM as powerful processing job scheduling systems.

The three ready to use pipelines implemented to date deal with the following basic MRI processing workflows:

1. Quality assessment of MRI data for quality control (QC)
2. Structural preprocessing for anatomical analysis and inference
3. Functional preprocessing for activity and connectivity analyses

1. The **QC-workflow** is designed to inform and optimize subsequent statistical analyses of MRI datasets using a combination of established quality assessment tools: CAT12 and MRIQC. It provides a

Funded by the European Union

multitude of image quality metrics (IQMs) that are available for the whole dataset including flags for potential outliers with low image quality. Data is presented in machine readable CSV tables, as well as in an interactive HTML format provided by MRIQC to browse individual subjects IQMs.

2. **CAT12** - the Computational Anatomy Toolbox (Gaser et at., 2024) performs full reparation of structural MRI data for volumetric and cortical surface based analyses. CAT12 is an extension to SPM12 in Matlab/Octave used here as compiled standalone version in a Singularity container. The toolbox covers diverse morphometric analysis methods such as Voxel-based morphometry (VBM), surface-based morphometry (SBM) and label- or region-based morphometry (RBM). The CAT-workflow provides morphometric derivatives for subsequent statistical and predictive modeling of brain anatomy.

3. **fMRIprep** flexibly prepares all commonly acquired flavors of functional MRI data including optional preparation of related structural data for surface based workflows with Freesurfer. fMRIPrep is a flexible pipeline for preprocessing functional magnetic resonance imaging (fMRI) data. It’s designed to be robust to variations in scan acquisition protocols and relying on the Brain Imaging Data Structure (BIDS) convention requires minimal user input to providing easily interpretable and comprehensive reporting in addition to analysis ready derivatives.

5.2. Influence of MR image quality on statistical and predictive analysis

Image quality is known for influencing the inferences drawn from analyses based on MRI data of any acquisition method, such as structural (i.e. T1 weighted contrast), functional (i.e. EPI T2*) or diffusion weighted. Image quality of MRI brain scans is strongly influenced by within scanner head motion and the resulting image artifacts alter derived measures like brain volume, cortical thickness, functional activation or connectivity as well as diffusivity or structural connectivity.

Here, we developed a quality control (QC) workflow consisting of an image quality assessment pipeline gathering a full battery of quality measurements for image-wise as well as sample-wise quality assessment. Quality data is made centrally accessible alongside each sample's raw data to simplify sample compilation based on image quality and inform downstream analyses.

Currently, there is little general consensus on criteria for image inclusion or exclusion for most statistical analysis on any type of MRI data apart from omitting images with severe distortions or noise and missing data. To address this problem, we additionally conducted a first assessment of the impact of MRI image quality on statistical and predictive analysis of brain volume.

Automated QC methods provide scalable and, in particular, objective solutions for processing large datasets, reducing reliance on time-intensive manual inspections and enhancing the consistency of data evaluation. These methods are vital for maintaining the integrity of studies that rely on structural MRI scans, enabling more reliable findings and improving reproducibility.

6. Results

6.1. Framework for fully reproducible data processing pipelines for MRI data

The framework applied here implements the FAIRly big workflow (Wagner, et al., 2022), which is based on existing free and open-source tools like git, git-annex, and DataLad (Halchenko, et al., 2021). Software containerization via Singularity and the integration with common scientific compute job scheduling systems (HTCondor & SLURM) enables fully reproducible computation in highly efficient computational environments. The workflow consists in the following three consecutive steps:

1. Automated setup of all data processing prerequisites including input dataset, software pipeline and submission scripts to trigger job scheduling in High-Throughput and High-Performance Computing environments (HTC & HPC).
2. Data processing with full provenance tracking in Datalad that captures machine readable and re-executable run records for each and every compute job, which is also associated to every file produced by the workflow.
3. Consolidation of produced data derivatives in a ready to use, cloneable Datalad dataset that provides granular file level availability including access control mechanisms for data privacy preservation.

6.1.1. Implementation

The High-Throughput Container Workflow implementation (https://cerebra.fz-juelich.de/inm7/HTContainerWorkflow) is based on DataLad, git-annex, and git. Additionally, the utilized FAIRly big workflow (Wagner, et al., 2022) requires the following prerequisites:

- **BIDS formatted input data as Datalad dataset:** The well established BIDS represents a standardized way of organizing neuroimaging and behavioral data. All processing pipelines require the input data to be BIDS conform Datalad datasets in cloneable data repositories accessible within the used computational environment.
- **Singularity container environment**: Singularity containers provide operating-system-level virtualization for reproducibility of scientific computing in HTC and HPC systems. Image processing pipelines are provided via publicly available Singularity containers to guarantee reliable transferability of the software environment.
- **Datalad-container extension**: This packages seamlessly and integrates container technology into the data management workflow to document, track and manage the use of software containers through Datalad.
- **Compute job scheduling with HTCondor and SLURM**: Scheduling systems are built for distributed parallelization of computationally intensive tasks to manage workloads on a dedicated cluster of computers. Every HTC and HPC system relies on effective compute job distribution, which is often provided by HTCondor or SLURM.

For each workflow, a template bootstrap script is provided to automatically set up all necessary parts for processing a whole MRI dataset in ephemeral clones. Only after successful completion of a compute job in the pipeline, results are pushed to a Datalad special remote, from where the processed data can be cloned as part of the generated dataset.

6.1.2. Stages of pipeline execution

The computation of results files is executed in 3 stages:

1. **Dataset preparation**: All prerequisites are automatically setup of for data processing, including the input dataset, the software pipeline, and the submission scripts that trigger job scheduling in high-throughput and high-performance computing environments (HTC and HPC). The bootstrap template script is tailored to a given pipeline's (I) input data, (II) storage setup for saving the pipeline’s output, and (III) available job scheduling system for efficient, parallelized data processing. When executed, the bootstrap script creates an empty dataset that includes all the necessary scripts for processing the data, as well as links to the input dataset, the software containers, and the dataset repository, which will gather all the processed data derivatives.

2. **Job submission**: Submit compute jobs for processing the full dataset with provenance tracking in Datalad. This captures machine-readable, re-executable run records for every computed job associated with each derivative file produced by the workflow. Executing the prepared job submission script for the available HTC/HPC environment triggers the maximum parallel processing of the entire input dataset. The pipeline setup guarantees that data transfer to the desired remote location will only occur if the data processing is fully successful.

3. **Dataset consolidation:** As parallel data processing is only possible for independent job execution, it is necessary to consolidate the resulting data into a final dataset. Running the initially prepared merge script ensures that all derivatives are available in one place as ready-to-use, cloneable Datalad dataset including access control mechanisms for data privacy preservation.

6.1.3. High-throughput (HTC) and high-performance computing (HPC)

A typical workflow for data preprocessing and quality control of a brain imaging cohort dataset requires substantial computational resources. Per participant processing times on fast machines can be several hours, depending on the complexity of the pipeline. Only HPC or HTC systems can manage such loads within reasonable time frames. At the same time, these often complex and non-standard environments pose a threat to the reproducibility of the computational aspects of scientific study outcomes. Access to these resources is often not universally possible, internal system configuration properties limit the portability of optimally tuned processing pipelines, and the lifetime of these compute environments is substantially more limited than the relevance of scientific results. To overcome this HPC/HTC systems specific limitation, the FAIRly big workflow separates the compute job itself from the computational environment running the job so that outcomes of highly specific HPC/HTC-based data processing provide portable computational reproducibility.

6.1.4. Pipeline properties

QC-pipeline

The quality control pipeline runs CAT12 and MRIQC within the FAIRly big workflow. This platform-independent QC pipeline relies on BIDS for MRI input data and produces a range of quality measures for each input image in any given modality. Furthermore, sample specific group statistics are provided in a condensed form to enable informed decisions on analysis specific data inclusion and exclusion.

bootstrapQC: https://cerebra.fz-juelich.de/f.hoffstaedter/bootstrapQC

CAT-pipeline

This structural image processing pipeline CAT12 computes derivatives for diverse morphometric analyses such as Voxel-based morphometry (VBM), surface-based morphometry (SBM), deformation-based morphometry (DBM), and label- or region-based morphometry (RBM).

bootstrapCAT: https://cerebra.fz-juelich.de/f.hoffstaedter/bootstrapCAT

fMRIprep-pipeline

Analysis-ready data derivatives are provided by applying minimal necessary procedures of image homogenization, movement correction, slice-timing and alignment into the same-subject’s T1w space or in some standard space alongside a multitude of time-series confounds for data-specific denoising.
bootstrapfMRIprep: https://cerebra.fz-juelich.de/f.hoffstaedter/bootstrapfMRIprep

6.1.5. Application of the High-throughput container workflow

Funded by the European Union

To illustrate the outcome of the developed High throughput container workflows, all three pipelines were run on the following publicly available dataset from OpenNeuro: https://openneuro.org/datasets/ds005479. The dataset contains structural and functional MRI data of 37 subjects and showcases the outcome of parallel processing of the dataset using HTCondor on a medium sized high-throughput compute cluster at Research Centre Jülich.

QC-pipeline: https://cerebra.fz-juelich.de/f.hoffstaedter/ds005479_QCworkflow

CAT-pipeline: https://cerebra.fz-juelich.de/f.hoffstaedter/ds005479_cat12.9

fMRIprep-pipeline: https://cerebra.fz-juelich.de/f.hoffstaedter/ds005479_fmriprep24.1.1_workflow

6.1.6. Impact on personal data processing

Computational derivatives computed from personal data, such as brain imaging data, generally have to be considered personal data too. Using the FAIRly big workflow to compute such derivatives greatly reduces the complexity associated with the pipeline setup and hence with the task of personal data protection because all data derivatives are kept in storage systems suitable for personal data hosting. To illustrate the minimally necessary information to be recorded in a dataset with derivative data, we can consider a typical configuration when working with DataLad datasets. A top-level (super)dataset is populated with computed derivative data. This dataset links two other DataLad datasets: (1) the source data (i.e., as acquired from an instrument), and (2) a dataset providing a containerized computational environment (i.e., a docker or singularity container image), and all custom software implementations necessary to compute a data derivative. These two datasets are linked, rather than the associated files being included in the derivative dataset directly, because they are typically maintained/curated separately, and can be reused in other contexts than this particular computation too. When linking DataLad datasets, only a sub-dataset (download) URL and a version identifier of the sub-dataset are included in the super-dataset.

6.2. Influence of MR image quality on statistical and predictive analysis

We explored the impact of MRI image quality on both statistical and predictive analyses, to gain insights into the influence of data quality on commonly used MRI analysis methods. Recent work was generalized to broader analysis and additional datasets (Hoffstaedter et al., 2024).

6.2.1. Implementation

In this work, gray matter volume was extracted as features (3747) from 4013 T1w MRI images from five publicly available dataset to train a sex/gender classification model as exemplar prediction problem. The weighted image quality rating (IQR), derived from CAT12, was used as metric of interest. Each included dataset exhibited a different IQR distribution, while being balanced for each sex/gender class (**Figure 1**). The features were adjusted for brain size via Total Intracranial Volume (TIV) using a linear confound removal approach. We proposed a sampling strategy to create subsamples of relatively high and low image quality, respectively, by maximizing or minimizing the IQR of the sample, while maintaining the same age and sex/gender distributions. For this sampling strategy, a number of age bins are selected to control for the critical influence of ageing. In our experiments a more liberal control with 3 age bins and a strict control with 10 age bins were applied. Using 3 age bins allows to retain more data with the cost that the image quality is more similar between subsamples. Conversely, using 10 age bins allows for more distinct IQR between low and high quality subsamples with the cost of retaining less participants. All experiments were performed within sample to account for site specific variance both in features and in IQR. Additional analyses were performed on pooled data for generalization. The code to generate and reproduce all results is publicly available at (https://github.com/N-Nieto/QC).

Funded by the European Union

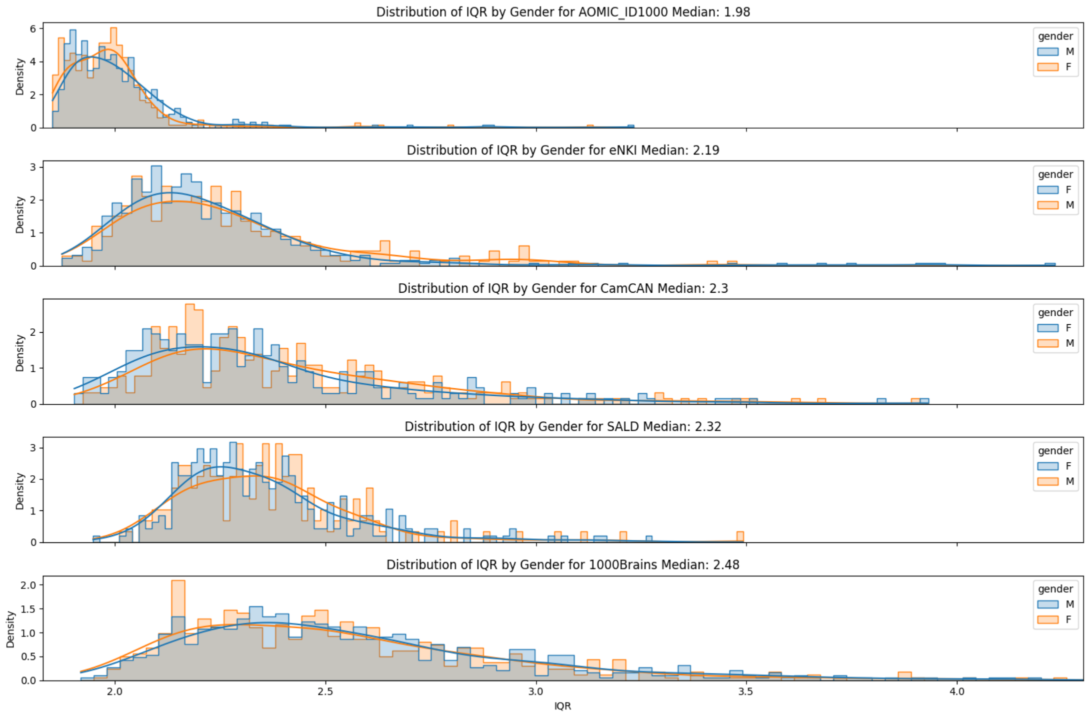

Figure 1. IQR distribution from the five publicly available datasets. Lower IQR represents better image quality. Different datasets present different overall distribution of IQR, but similar for each sex/gender class within dataset.

6.2.2. Results

For classical univariate statistical approaches we perform a t-test comparing the features (GMV value of each voxel) of box sex/gender. As 3747 features were tested, we applied a Bonferroni correction for multiple testing. We found that poorer-quality data drastically reduces sensitivity to detect group differences, especially in studies with smaller sample sizes (**Figure 2**). On the other hand, samples with higher-quality images revealed more robust and consistent findings, emphasizing the critical role of QC in enhancing the validity of such analyses.

Funded by the European Union

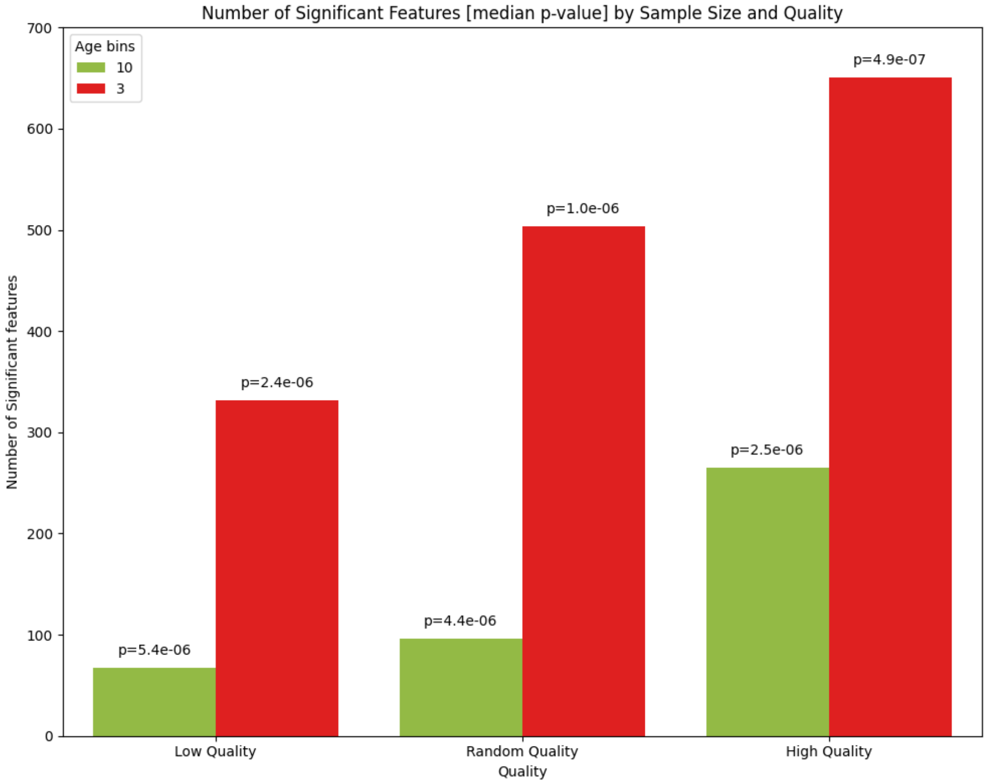

Figure 2. Number of features from which the feature distribution from both sex/genders are significantly different with respect to the quality sampling strategy. The number of significantly different features were calculated using the pooled data for low, random and high images quality for subsamples with 10 age bins (n=1528) and for subsamples with 3 age bins (n=2060).

In the predictive modeling approach, logistic regression was applied for classifying sex/gender. Results demonstrate that machine learning seems relatively less sensitive to image quality variations, particularly when overall image quality is at a good to acceptable level (**Figure 3**). Increasing sample size in these predictive analyses showed a clear benefit for improving model performance in those datasets with smaller sampler sizes, underscoring the complementary importance of both data quality and quantity.

Funded by the European Union

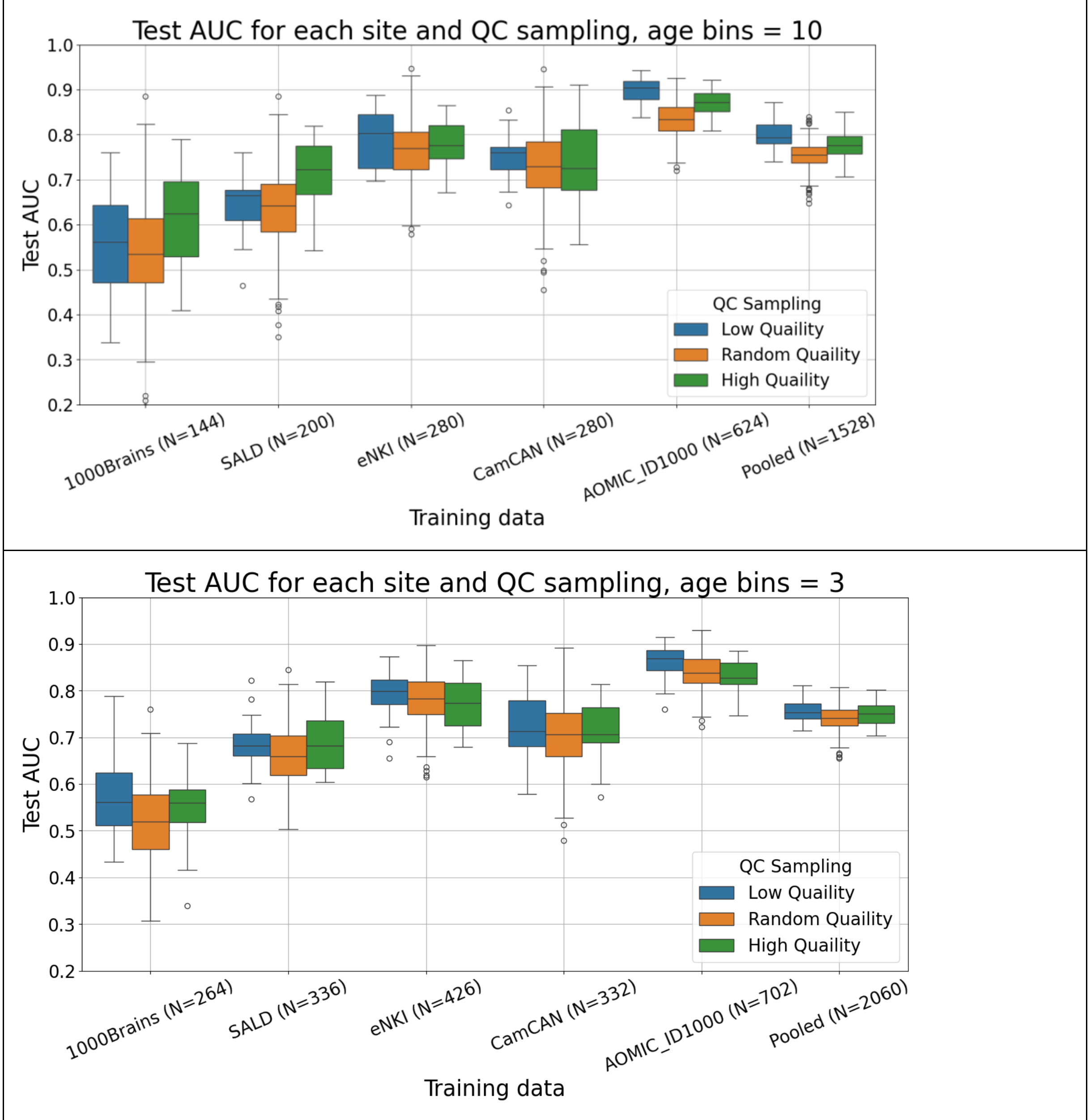

Figure 3. Area under the Receiver Operating Characteristic Curve (AUC) for sex/gender classification. A) AUC for subsamples with 10 age bins (n=1528) for low, random and high-quality images and overall for the pooled data. B) AUC for bigger subsamples with 3 age bins (n=2060) for each subsample and pooled data.

To assess whether structural brain differences are associated with variations in MRI image quality, we performed a linear regression analysis predicting quality scores from the extracted features. We trained a linear regression model to predict image quality based on gray matter volume. Since this analysis aimed to evaluate general associations rather than optimize predictive performance, no sampling or stratification was applied, and all available samples were used for each dataset. Although the model achieved a low mean absolute error, the predictions were primarily clustered around the mean of the distribution, suggesting limited sensitivity to extreme quality values. In order to provide additional validation of the information carried by the gray matter features regarding image quality, a permutation test was applied. In this test, the features were randomly permuted 100 times and a linear regressor was trained to predict the IQR using the shuffled features to compare the real and a baselineMAE (**Figure 4**). While significant difference were found between real and permutation based model performance for all sites ($p < 0.05$), the real MAE of image quality prediction suggests largely the prediction of the mean image quality of each sample.

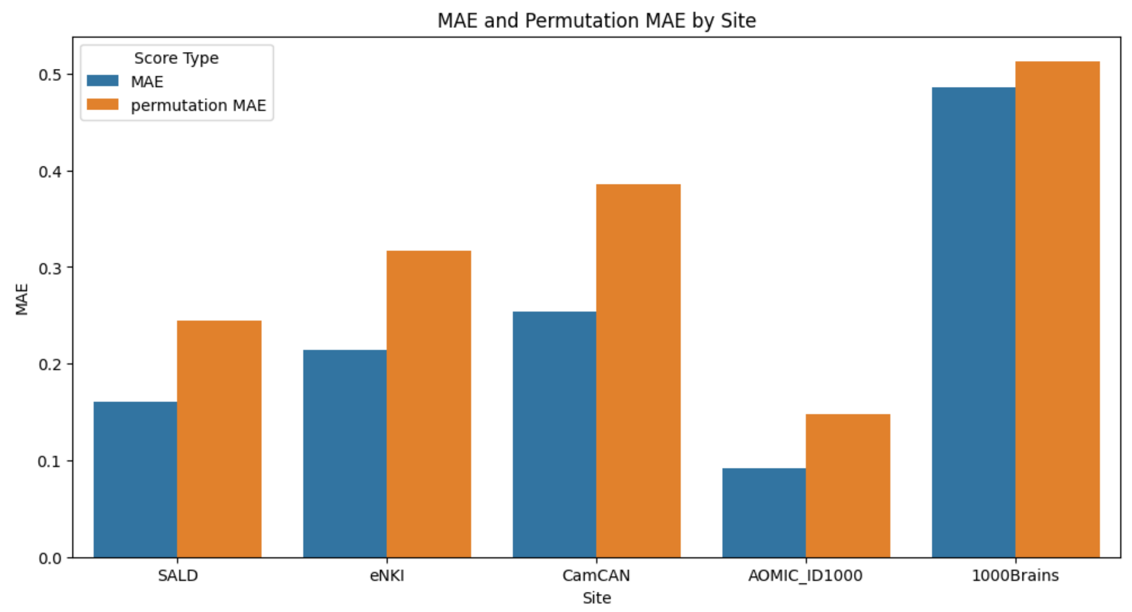

Figure 4. Comparison of Mean Absolute Error (MAE) and Permutation MAE across the datasets (SALD, eNKI, CamCAN, AOMIC_ID1000, 1000Brains). The bars represent the model's prediction accuracy (MAE) and the robustness of the model assessed via permutation testing (Permutation MAE). Lower values indicate better performance.

6.2.3. Impact

The quality control (QC) of neuroimaging data, as measured by the IQR, demonstrates substantial variability across different acquisition sites, including datasets such as SALD, CamCAN, eNKI, 1000Brains, and AOMIC. This variability likely reflects differences in scanning protocols, hardware, or population characteristics, emphasizing the challenges of implementing consistent QC standards in multi-site studies. Such variability has a pronounced and consistent impact on traditional statistical analyses, particularly in large cohorts where QC differences can systematically bias results. This effect underscores the critical need for QC pipelines when employing conventional statistical approaches.

In contrast, machine learning (ML) methods appear more robust to these QC variations, without exhibiting clear performance trends tied to data quality. This resilience may stem from ML's inherent capacity to handle heterogeneous data effectively. The noise introduced by QC variability could function similarly to explicit regularization techniques, preventing overfitting and encouraging the discovery of more generalizable patterns. Thus, while statistical methods are vulnerable to QC fluctuations, ML models may inadvertently benefit from the heterogeneity, turning potential noise into a useful feature for robust learning.

Finally, while it is feasible to predict IQR from image-derived features, these predictions generally do not outperform simple mean-based benchmarks. This observation suggests that QC may not be linearly encoded in standard features. Together, these findings highlight the differential sensitivity of statistical and ML methods to QC variability, with important implications for the design and interpretation of large-scale neuroimaging studies and aligns with the growing body of evidence supporting the efficacy of automated QC metrics, such as IQR measures, for facilitating large-scale image analyses without labor-intensive, full-scale visual inspection of MRI data.

Funded by the European Union

7. Conclusion, next steps

A generalized, high-throughput, containerized workflow was developed to capture the full provenance of data processing applicable to HTC and HPC systems. The proposed bootstrap approach facilitates the reproducible configuration of an entire processing workflow for a specified dataset with a particular pipeline via a singular shell script. To date, three ready-to-use pipelines have been implemented, each designed to serve standard MRI processing needs: MRI data quality assessment for quality control, structural preprocessing for anatomical analysis and inference and functional preprocessing for activity and connectivity analyses. These well tested pipelines represent standard MRI data processing in a scaleable and easy to use format, ready for extraction of image features and standard downstream analyses and model building. The variability in MRI image quality was found to differentially affect classical statistical and machine learning methods, with important implications for the design and interpretation of large-scale neuroimaging studies.

7.1. Feature additions for containerized processing pipelines

Support for command templates is being considered for adoption by the EBRAINS computing platform. Import/export compatibility would enable direct submission of compute jobs to the EBRAINS platform.

7.2. QC integration on pipelines

The next step is to extend the impact assessment of image quality to larger datasets, like Dallas Lifespan Brain Study (DLBS) and HCP to improve understanding of the interplay between image quality, sample size, and analytical outcomes. The goal of this next phase is to develop best practices for using QC metrics in neuroimaging research. With a focus on automated methods and scalability, this phase will contribute to the broader goal of enhancing the reliability and reproducibility of MRI-based studies by further integrating QC into large-scale processing pipelines.

8. References

Esteban, O., Birman, D., Schaer, M., Koyejo, O.O., Poldrack, R.A. & Gorgolewski, K.J. (2017). MRIQC: Advancing the Automatic Prediction of Image Quality in MRI from Unseen Sites; PLOS ONE 12(9):e0184661. https://doi.org/10.1371/journal.pone.0184661

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S.S., Wright, J., Durnez, J., Poldrack, R.A. & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nature methods, 16(1), 111-116. https://doi.org/10.1038/s41592-018-0235-4

Gaser C, Dahnke R, Thompson PM, Kurth F, Luders E & The Alzheimer's Disease Neuroimaging Initiative. CAT: a computational anatomy toolbox for the analysis of structural MRI data. Gigascience. 2024 Jan 2;13:giae049. https://doi.org/10.1093/gigascience/giae049

Halchenko, Y. O., Meyer, K., Poldrack, B., Solanky, D. S., Wagner, A. S., Gors, J., MacFarlane, D., Pustina, D., Sochat, V., Ghosh, S. S., Mönch, C., Markiewicz, C. J., Waite, L., Shlyakhter, I., de la Vega, A., , Hayashi, S., Häusler, C. O., Poline, J.-B., Kadelka, T., Skytén, K., Jarecka, D., Kennedy, D., Strauss, T., Cieslak, M., Vavra, P., Ioanas, H.-I., Schneider, R., Pflüger, M., Haxby, J. V., Eickhoff, S. B. & Hanke, M. (2021). DataLad: distributed system for joint management of code, data, and their relationship. Journal of Open Source Software, 6(63), 3262. https://doi.org/10.21105/joss.03262

Hoffstaedter, F., Nieto, N., Eickhoff, S. B., & Patil, K. R. (2024). The impact of MRI image quality on statistical and predictive analysis on voxel based morphology. *arXiv preprint arXiv:2411.01268*. https://doi.org/10.48550/arXiv.2411.01268

Wagner, A. S., Waite, L. K., Wierzba, M., Hoffstaedter, F. Waite, A. Q., Poldrack, B., Eickhoff, S. B. & Hanke, M. (2022) FAIRly big: A framework for computationally reproducible processing of large-scale data. Scientific Data, 9, 80. https://doi.org/10.1038/s41597-022-01163-2

Disclaimer

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101058516. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or other granting authorities. Neither the European Union nor other granting authorities can be held responsible for them.