A machine learning model to harmonize volumetric brain MRI data for quantitative neuroradiological assessment of Alzheimer disease

Abstract:

Purpose: To extend a previously developed machine learning algorithm for harmonizing brain volumetric data of individuals undergoing neuroradiological assessment of Alzheimer disease not encountered during model training.

Materials and Methods: Neuroharmony is a recently developed method that uses image quality metrics (IQM) as predictors to remove scanner-related effects in brain-volumetric data using random forest regression. To account for the interactions between Alzheimer disease pathology and IQM during harmonization, the authors developed a multi-class extension of Neuroharmony for individuals with and without cognitive impairment. Cross-validation experiments were performed to benchmark performance against other available strategies using data from 20,864 participants with and without cognitive impairment, spanning 11 prospective and retrospective cohorts and 43 scanners. Evaluation metrics assessed ability to remove scanner-related variations in brain volumes (marker concordance between scanner pairs), while retaining the ability to delineate different diagnostic groups (preserving disease-related signal).

Results: For each strategy, marker concordances between scanners were significantly better ($p < 0.001$) compared to pre-harmonized data. The proposed multi-class model achieved significantly higher concordance (0.75 ± 0.09) than the Neuroharmony model trained on individuals without cognitive impairment (0.70 ± 0.11) and preserved disease-related signal ($\Delta AUC = -0.006 \pm 0.027$) better than the Neuroharmony model trained on individuals with and without cognitive impairment that did not use our proposed extension ($\Delta AUC = -0.091 \pm 0.036$). The marker concordance was better in scanners seen during training (concordance > 0.97) than unseen (concordance < 0.79), independently of cognitive status.

Conclusion: In a large-scale multi-center dataset, our proposed multi-class Neuroharmony model outperformed other available strategies for harmonizing brain volumetric data from unseen scanners in a clinical setting.

Summary Statement:

A novel multi-class Neuroharmony model was developed and evaluated against other approaches for harmonizing volumetric data in a clinical setting using a large, multicenter brain MRI dataset of individuals undergoing neuroradiological assessment of Alzheimer disease.

Key Points:

1. The proposed multi-class Neuroharmony model, trained on 20,864 participants, achieved state-of-the-art performance in harmonizing brain volumetric data from new MRI scanners.
2. The proposed multi-class Neuroharmony model preserved disease-signal on volumetric features better than the other tested approaches.
3. The multi-class Neuroharmony model performed better for harmonizing MRI-derived volumetric data in the clinical setting than other available approaches for harmonizing data from previously unseen scanners.

1 Introduction

Structural MRI scans, such as T1-weighted MRI, are routinely acquired in memory clinics for diagnosing Alzheimer's disease (AD) (1), clinical phenotyping (2), and for differentiating AD from other types of dementias (3). In current clinical practice, radiologists primarily assess global and regional brain atrophy through visual examination of MRI. However, visual examinations are subjective and prone to intra-rater and inter-rater variability. Quantitative imaging markers, such as brain volumetric data, are becoming increasingly popular due to their potential to improve diagnostic confidence (4). Quantitative imaging markers can be used for objective assessment in the radiological workflow either by using automated digital tools based on normative modelling (3) or using latest advances in artificial intelligence, including brain-age estimation (5) and data-driven subtyping (6).

However, differences in MRI acquisition protocols and scanners affect consistency and reproducibility of brain volumetry (7) and are a major impediment for the clinical translation of automated tools. To tackle this problem, many data harmonization tools have emerged in recent years (8). Such algorithms can either harmonize original scans (e.g. DeepHarmony) (9) or derivatives extracted from the scans (e.g. ComBat) (10). Some of these algorithms have been shown to harmonize patient data affected by a neurodegenerative disease (11,12) while preserving disease-related signature. However, such harmonization techniques typically work only for the scanner models they have been trained on, and, in some instances require the same individuals to be scanned with different scanners (13). Harmonizing volumetric data from MRI scanners not encountered during initial model training requires additional training with a substantial number of images from these scanners (14). This poses a challenge for the deployment of such models for clinical use.

Neuroharmony (15) is a recently developed harmonization approach that can harmonize volumetric data from images acquired using new and unseen MRI scanners. The Neuroharmony model is trained to predict the volumetric corrections estimated by ComBat harmonization in the training phase. When trained on large enough samples, the model generalizes well for predictions of harmonized volumes in previously unseen scanners. It works under the assumption that the corrections needed to harmonize data from multiple scanners can be predicted from image quality metrics (IQM) computed from the scans. While the original Neuroharmony study indicates that harmonization works for healthy individuals (15), harmonizing data from patients with neurodegenerative diseases remains an open problem. This is because disease pathology in patients may affect the IQM, and such effects remain unaccounted for in a Neuroharmony model trained on healthy controls.

In this paper, we propose an extension of the Neuroharmony model to account for interactions between disease pathology and IQM to remove scanner-related effects (multi-class model of

Neuroharmony). We systematically compare the performances of the proposed multi-class model in harmonizing data with two other approaches: the original Neuroharmony model trained only on individuals without cognitive impairment (normative model of Neuroharmony) and the original Neuroharmony model trained on individuals with and without cognitive impairment that did not use our proposed multi-class extension (inclusive model of Neuroharmony). We used data from 11 cohorts across three continents for evaluating these approaches. Lastly, we identify key challenges for clinical implementation of the best multicentric harmonization strategy identified in our experiments for enabling quantitative neuroradiological assessment of AD.

2 Materials and Methods

2.1 Study Participants and Data

T1-weighted 3D MRI data of healthy controls (HC), participants with subjective cognitive decline (SCD), mild cognitive impairment (MCI), and AD from 11 prospective and retrospective data cohorts were included in our analysis. The cohorts considered for this study were: Amsterdam Dementia Cohort (ADC) (16), Alzheimer's Disease Neuroimaging Initiative (ADNI) (17), Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) (18), Alzheimer's Repository Without Borders (ARWiBo) (19), European DTI Study on Dementia (EDSD) (20), Hungarian Longitudinal Study of Healthy Brain Aging (HuBA) (21), Italian Alzheimer's Disease Neuroimaging Initiative (I-ADNI) (22), National Alzheimer's Coordination Center (NACC) (23), Open Access Series of Imaging Studies (OASIS, versions 1&2) (24), European Alzheimer's Disease Neuroimaging Initiative (also known as PharmaCOG) (25), and UK Biobank (UKBB) (26). Detailed information about each cohort is summarized in Supplementary Table 1. The clinical diagnoses of participants in these cohorts were made based on international consensus criteria; further details can be found in the respective studies cited above. Each study was approved by the respective institutional ethical committees, with informed consent obtained from each participant.

Minimum inclusion criteria included the availability of a T1-weighted 3D MRI scan along with age, sex, and scanner information, and a clinical diagnosis of either HC, SCD, MCI or AD. All datasets were organized according to the Brain Imaging Data Structure (BIDS) standard (27) to ensure inter-operability and data anonymization. An overview of the scanners used in this study is shown in Table 1 and the scanning parameters are summarized in Supplementary Table 2.

2.2 Image processing

Cortical reconstruction and volumetric segmentation were performed with the cross-sectional pipeline of FreeSurfer v7.1.1 (28) in order to extract volumes of 68 cortical regions in the Desikan-Killiany atlas, 14 subcortical brain regions, as well as total cerebrospinal fluid volume, total gray matter volume and total brain volume with and without ventricles. Supplementary Figure 1 lists all features derived from FreeSurfer. IQM were estimated using MRIQC v0.16.1 (29). Automatic quality control of the FreeSurfer segmentations was performed using the Euler number, where outliers, defined as 1.5×IQR (inter-quartile range) below the first quartile (30), for each scanner were excluded from our experiments.

In order to ensure reproducibility of our results across different computing environments (31), Docker containers for both FreeSurfer and MRIQC were prepared by author NPO, and shared with coauthors

(DA, VV, BW, PB) to process MRI from their local cohort (no images were shared, blinding was not necessary). These authors each have 5–15 years of MRI processing experience. The containers have been made available online to benefit the community (URLs indicated in the "Data and code availability" section).

2.3 Multi-class Neuroharmony model

In the training phase, volumetric data from all individuals in the training set were harmonized using ComBat (10) with empirical Bayes optimization to remove scanner related batch effects. While training, we imposed constraints that preserve the effects of age, sex, and cognitive status. Cognitive status was dichotomized based on the clinical diagnosis as either no cognitive impairment (HC and SCD) or cognitive impairment (MCI and AD). Subsequently, a random forest regressor was trained with MRIQC-derived IQM to predict the corrections needed to harmonize the volumes as predicted by ComBat. Additionally, to preserve disease-related signal during harmonization, we used the synthetic minority oversampling technique (SMOTE) (32) to avoid class imbalance (no cognitive impairment vs cognitive impairment) before training the random forest regressor. This ensured that IQM values with and without neurodegeneration were equally distributed. The use of dichotomized cognitive status instead of clinical diagnosis ensured that in the test phase, a full clinical diagnosis is not required to predict the harmonized volumes. The hyperparameters for the random forest regressor were chosen to be the same as the ones used in the original Neuroharmony paper (15).

2.4 Model comparisons

The performance of the proposed multi-class extension of Neuroharmony was compared with two other harmonization strategies that are generalizable to external datasets:

2.4.1 Normative model: In the training phase, volumetric data from only individuals without cognitive impairment were harmonized using ComBat harmonization using the aforementioned strategy, while preserving the effects of age and sex. Subsequently, a random forest regressor was trained to predict the corrections needed to harmonize the volumes, as predicted by ComBat using MRIQC-derived IQM.

2.4.2 *Inclusive model*: The training strategy remained the same as for the normative model, but volumetric data of individuals with and without cognitive impairment were used.

2.5 Measures for model evaluation

We used two measures for model evaluation to assess how well each method removes unwanted scanner-related noise while retaining disease-related signal. First, we define "marker concordance" (details below) as a statistical measure of similarity between brain-volumetric data from different scanners. Increased marker concordance after harmonization shows that a method successfully reduces scanner-related variance (see "Statistical Analysis" section for details). Second, we used classification performance (HC vs AD) to assess the amount of disease-related signal. The best performing harmonization model will return the best classification performance.

We used area under the receiver operating characteristic (ROC) curve (AUC) to quantify the amount of disease-related signal that is retained in the volumetric measures after harmonization. The ROC curve for distinguishing healthy control participants from participants with AD was computed independently for each volumetric measure with logistic regression. A reference measure for AUC was also computed for the non-harmonized data.

2.6 Cross-validation experiments

We performed two experiments in a cross-validation framework. Experiment 1 assessed concordance of the three harmonization strategies, by performing cross-validation at the scanner-level. Experiment 2 performed cross-validation at the participant level, using the best-performing scanner-level harmonization models.

Experiment 1: To investigate the generalizability of the model to unseen scanners (not included in the training set), we performed 5-fold cross-validation across the 43 available scanners. In each fold, 80% of the scanners were used for training the models, and the remaining 20% of the scanners were used for evaluation. To evaluate the bias introduced by using single-scanner data from the large UKBB cohort, we repeated this experiment for increasing portions of UKBB participants such that when the UKBB data were included in the training data, the proportions included were: 10%; 33%; 67%; 100%. However, in the cross-validation folds when UKBB cohort data were not used for training, we always used 100% of the cohort.

To investigate if this approach can be used for harmonizing cortical thickness measures, we selected the two best performing approaches from the above analysis and repeated our experiment on cortical thickness measures obtained from 68 brain regions defined by the Desikan-Killiany atlas.

Experiment 2: We selected the two best performing models from Experiment 1 and performed a stratified 5-fold cross-validation across participants, stratified based on the dichotomized cognitive status. Differently from Experiment 1, the scanner was not used to define folds for cross validation in Experiment 2 in order to test the generalizability to new participants in seen scanners as opposed to unseen scanners tested in Experiment 1. For this, the proportion of the UKBB participants included was also decided based on Experiment 1. To provide a reference measure, we compared the accuracies obtained with the corresponding accuracies obtained in Experiment 1.

2.7 Statistical Analysis

To compute marker concordance, we compared the distributions of each volumetric measure for each pair of scanners by means of the Kolmogorov-Smirnov (KS) test with the null hypothesis that the distributions between any pair of scanners were the same. This was done independently within each diagnostic group, and after correcting for the confounding effects of age and sex by regressing out their effects estimated in individuals without cognitive impairment. Marker concordance was calculated as the proportion of such comparisons where there was no evidence that distributions were different between each pair of scanners across all brain regions after controlling for multiple testing via false discovery rate (FDR ≥ 0.05) based on the p-values of Kolmogorov-Smirnov tests. For statistical validity, we excluded scanners with fewer than 10 participants of the same diagnostic group from this evaluation.

AUCs of classification tasks for each harmonization strategy were compared with the AUCs in the case of non-harmonized data separately for each feature by means of a DeLong test.

The non-parametric McNemar Chi-square test was used to compare concordances across harmonization strategies. To control for multiple hypothesis testing, resulting p-values were used to estimate the FDR.

2.8 Data and code availability

ADC data can be made available to academic researchers upon reasonable request; ADNI and AIBL data are managed by the Laboratory of Neuroimaging at the University of Southern California and are available to the general scientific community for download (http://ida.loni.usc.edu/); ArWiBO, EDSD, I-ADNI, OASIS and PharmaCog data are available for all researchers on the NeuGRID2 platform (https://www.neugrid2.eu/ https://doi.org/10.17616/R31NJN1E); HuBA data can be made available upon reasonable request; NACC data is available through the National Alzheimer's Coordinating Center platform (https://naccdata.org/); UKBB data is available at the UK Biobank platform (https://www.ukbiobank.ac.uk/);

Docker container source code for FreeSurfer and MriQC is available on GitHub (https://github.com/E-DADS/freesurfer, https://github.com/E-DADS/mriqc); Multi-class Neuroharmony harmonization algorithm is available on GitHub (https://github.com/88vikram/Multiclass-Neuroharmony);Trained model files forHarmonization using Multi-class Neuroharmony are available for all researchers on the NeuGRID2 platform (https://www.neugrid2.eu/index.php/edads_harmonization).

3 Results

3.1 Participants

Table 2 shows descriptive statistics for the combined study sample used in our experiments, which consisted of volumetric data that passed quality control from 20,864 participants (53.3% female, 11,111/20,864) from 43 scanners across 11 cohorts. A total of 2,086 individuals were excluded based on low Euler number. Figure 1 shows age distributions by scanner and cognitive group.

3.2 Model evaluation

Figure 2 shows the first result of Experiment 1: marker concordance under cross-validation, independently for each diagnostic group and with increasing proportions of the UKBB dataset. Reference concordances for non-harmonized data are also shown for each diagnostic group for comparison. As expected, concordances for each harmonization strategy were significantly higher than the non-harmonized data for all the diagnostic groups ($FDR < 0.001; p < 0.001$). The use of the inclusive and multi-class models significantly improved the concordance with respect to the normative model for the diagnostic categories of MCI and AD ($FDR < 0.001; p < 0.001$). For diagnostic groups of HC and SCD, the concordance of the multi-class model was significantly higher than the normative model with 100% of UKBB included ($11{,}058/11{,}058$) ($FDR_{HC} = 0.01; p_{HC} = 0.009, FDR_{SCD} = 0.02; p_{SCD} = 0.02$). There was no evidence of a difference in concordance for HC and participants with SCD between the inclusive model and normative model ($FDR = 0.23; p = 0.21$).

Figure 3 shows the second result of Experiment 1: the AUCs for classifying HC versus participants with AD, which were computed independently for each brain regional volume in the test set. Removing scanner-related differences decreased AUC for all harmonization approaches, potentially due to the significant imbalance ($p < 10^{-4}$) in the number of HC and participants with AD in the different scanners (Supplementary table 3). For the normative model, the AUC was significantly lower than the pre-harmonized data for 45 volumetric features ($\Delta AUC = -0.013 \pm 0.023$). For the inclusive model, the AUC was significantly lower than the pre-harmonized data for 82 features ($\Delta AUC = -0.091 \pm 0.036$). For the multi-class model, the AUC was significantly lower than pre-harmonized AUC for 40 features ($\Delta AUC = -0.006 \pm 0.027$), indicating relative loss of disease-related signal when using the inclusive model harmonization strategy. Across all brain regions, the best AUCs were achieved for the amygdalae and hippocampi in all harmonization scenarios (Supplementary Figure 1).

Based on marker concordance and AUC, the two best models were the normative and the multi-class Neuroharmony models, when trained with 100% UKBB data. For harmonizing cortical thickness measures, our proposed multi-class model achieved significantly higher marker concordance than the normative model (Supplementary Figure 2).

3.3 Harmonization in seen vs unseen MRI scanners

Figure 4 shows the results of Experiment 2: marker concordance for seen vs unseen scanners during model training for both the normative model and multi-class model. Supplementary Figure 3 shows these results for each brain volume individually. Marker concordance of the multi-class model was significantly higher than the normative model for unseen scanners for all diagnostic categories ($FDR_{HC} = 0.01; p_{HC} = 0.009$, $FDR_{SCD} = 0.02;\ p_{SCD} = 0.02, FDR_{MCI} < 0.001;\ p_{MCI} < 0.001, FDR_{AD} < 0.001;\ p_{AD} < 0.001$). For seen scanners, the multi-class model harmonization strategy significantly outperformed the normative model for the diagnostic groups of HC, MCI, and AD ($FDR < 0.001;\ p < 0.001$), but significantly underperformed for SCD ($FDR \quad = 0.02; p = 0.02$). Marker concordance using the multi-class model in a seen scanner ($concordance > 0.97$) was better for all diagnostic groups than in unseen scanners ($concordance < 0.79$).

4 Discussion

We introduced a novel extension of the Neuroharmony harmonization model (15) to train a generalizable machine learning model for harmonizing multicentric brain volumetric data for quantitative assessment of AD. The data for these evaluation experiments were derived from T1-weighted 3D MRIs acquired with 43 different scanners from 20,864 participants spanning 11 cohorts. Our experiments showed that the multi-class model, which accounts for the interaction between disease pathology and IQM to remove scanner-related effects, significantly improved marker concordance between scanner pairs for participants in unseen scanners as compared to normative modelling for all diagnostic groups $(FDR_{HC} = 0.01,\ FDR_{SCD} = 0.02, FDR_{MCI} < 0.001\ FDR_{AD} < 0.001)$. For seen scanners, it improved the marker concordance for all diagnostic groups except SCD, potentially due to the lower sample size of the SCD group or uncertainty in the etiology of this diagnostic category. Additionally, we showed that the multi-class model of Neuroharmony preserves disease-related signal during harmonization better than the other tested approaches that represent state-of-the-art methodologies. The newly introduced multi-class model would be helpful in harmonizing volumetric data while using automated tools in clinics and research where there could be data from new scanners not included in training.

However, we note that the AUC was slightly reduced compared to non-harmonized data for some brain regions, implying that Multi-class Neuroharmony can remove some disease-related signal in the presence of diagnostic class imbalance across scanners. Future work should explore model-based mechanisms for disentangling such associations to preserve disease-related signal.

Harmonization of marker data from unseen scanners remains a challenge: marker concordance for both normative and multi-class models in unseen scanners was lower than in seen scanners. While this leaves scope for further methodological improvements to harmonization strategies for unseen scanners, it would also be useful to investigate if the achieved harmonization performance is sufficient for the generalizability of machine learning approaches such as classification, subtyping (33), and brain aging.

The different number of participants used to train the respective models could potentially bias the results against the model that uses a smaller dataset for training (normative model). However, we believe that this setting is a realistic and fair comparison because normative modelling always discards data from individuals with cognitive impairment. Through our modifications to the Neuroharmony model, we provided a way to include individuals with and without cognitive impairment in the training data, and our experiments showed improved harmonization in both seen and unseen scanners while preserving disease-related signal.

The harmonization performance obtained with the normative model in our experiments was lower than reported in the original Neuroharmony paper (15). This may be due to removal of sex and age variability in the original Neuroharmony method. We preserved these effects, retaining this biological variability, which we would argue is important for both research studies and clinical implementation.

Challenges in the clinical implementation of the harmonization strategy: while the multi-class model outperformed the normative model in terms of marker concordance, the implementation of the model in memory clinics might require additional work to include cognitive status of a patient during regular radiological workup. Machine learning models could potentially be used to overcome this limitation, as it has been shown in recent studies that classifying cognitive impairment from healthy control or SCD can be done with high accuracy using MRI (34). To avoid a circular dependency between the two tasks, developing multi-task machine learning models to jointly harmonize and predict cognitive status is an important avenue of future work. Also, for broader employment in memory clinics, the harmonization algorithm should be validated on other segmentation algorithms beyond Freesurfer.

While the current work was focused on the AD spectrum, we expect that our new method will be valuable for impaired cognition in general (e.g.: vascular dementia, frontotemporal dementia, dementia with Lewy bodies). We expect the approach to also be applicable for patients with psychiatric disorders, but further work would be needed for patients with other neurological conditions — especially those where the brain is affected by large lesions and other major structural modifications.

Some limitations of the original Neuroharmony model (15) apply to this work as well. The harmonization performance for an individual in the test-set depends on the contrast-to-noise ratio in the T1-weighted 3D MRI and the pipeline cannot guarantee effective harmonization if the ratio is outside the range seen in our training data, and might lead to incorrect harmonization. Secondly, the harmonization performance based on marker concordance across scanner-pairs is a surrogate measure to measure consistency in the absence of a ground-truth. A potential limitation of this study is the lack of a study to assess within-participant variability across scanners, i.e., where a group of participants including all diagnostic classes are scanned across multiple scanners. This would allow for evaluation of the ability of the model to remove scanner effects at the individual level, but such a study would face considerable ethical issues related to repeatedly scanning patients. Another potential issue of the present work is the definition of marker concordance, which may not be statistically robust as it implies that the failure of rejection of the null hypothesis (i.e. failure to state that marker distributions are significantly different) corresponds to the null hypothesis being true (i.e.:

marker distributions are similar), consistency of future works may benefit from more apt definitions of marker concordance. An important limitation of this study, as with most research studies in this field, is that the imaging data used predominantly came from the developed Western countries of the EU, US, UK, and Australia. A more generalizable and inclusive model for harmonization would require data from nations in South-America, Asia, and Africa. This would include low field-strength scanners that are predominantly used in these regions, as well as more diverse biological variation in the training data. Large global consortia such as the UNITED consortium (35) could potentially help in getting access to such diverse neuroimaging data. Further developing Neuroharmony for distributed or federated learning for harmonizing imaging data can also facilitate inclusion from under-represented countries.

In summary, we have generalized the Neuroharmony model to harmonize Freesurfer-based MRI marker data from multiple scanners and sites while retaining disease signal that could otherwise be removed by the harmonization procedure. When evaluated on brain MRI marker data from participants along the AD spectrum, our new model outperformed the other approaches we tested on both seen and unseen scanners. Further validation using different processing pipelines and evaluation criteria would be essential for clinical use of the model in applications related to cognitive decline, such as memory clinics and clinical trials of new interventions for neurodegenerative diseases.

Funding

Anonymized for review.

Acknowledgments

Anonymized for review.

Conflicts of interest

Anonymized for review.

References

1. Johnson KA, Fox NC, Sperling RA, Klunk WE. Brain imaging in Alzheimer disease. Cold Spring Harb Perspect Med 2012;2(4):a006213. doi: 10.1101/cshperspect.a006213
2. Ossenkoppele R, Cohn-Sheehy BI, La Joie R, et al. Atrophy patterns in early clinical stages across distinct phenotypes of Alzheimer's disease. Hum Brain Mapp 2015;36(11):4421-37. doi: 10.1002/hbm.22927
3. Hedderich DM, Dieckmeyer M, Andrisan T, et al. Normative brain volume reports may improve differential diagnosis of dementing neurodegenerative diseases in clinical practice. Eur Radiol 2020;30(5):2821-2829. doi: 10.1007/s00330-019-06602-0
4. Goodkin O, Pemberton H, Vos SB, et al. The quantitative neuroradiology initiative framework: application to dementia. Br J Radiol 2019;92(1101):20190365. doi: 10.1259/bjr.20190365
5. Wang J, Knol MJ, Tiulpin A, et al. Gray Matter Age Prediction as a Biomarker for Risk of Dementia. Proc Natl Acad Sci U S A 2019;116(42):21213-21218. doi: 10.1073/pnas.1902376116
6. Young AL, Marinescu RV, Oxtoby NP, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nat Commun 2018;9(1):4273. doi: 10.1038/s41467-018-05892-0
7. Liu S, Hou B, Zhang Y, et al.. Inter-scanner reproducibility of brain volumetry: influence of automated brain segmentation software. BMC Neurosci 2020;21(1):35. doi: 10.1186/s12868-020-00585-1
8. Gebre RK, Senjem ML, Raghavan S, et al. Cross-scanner harmonization methods for structural MRI may need further work: A comparison study. Neuroimage 2023;269:119912. doi: 10.1016/j.neuroimage.2023.119912
9. Dewey BE, Zhao C, Reinhold JC, et al. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magn Reson Imaging 2019;64:160-170. doi: 10.1016/j.mri.2019.05.041
10. Fortin JP, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 2018;167:104-120. doi: 10.1016/j.neuroimage.2017.11.024
11. Pagani E, Storelli L, Pantano P, et al. Multicenter data harmonization for regional brain atrophy and application in multiple sclerosis. J Neurol 2023;270(1):446-459. doi: 10.1007/s00415-022-11387-2
12. Zhou HH, Singh V, Johnson SC, Wahba G. Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. Proc Natl Acad Sci U S A 2018;115(7):1481-1486. doi: 10.1073/pnas.1719747115
13. Potvin O, Chouinard I, Dieumegarde L, et al. The Canadian Dementia Imaging Protocol: Harmonization validity for morphometry measurements. Neuroimage Clin 2019;24:101943. doi: 10.1016/j.nicl.2019.101943
14. Kia SM, Huijsdens H, Rutherford S, et al. Closing the life-cycle of normative modeling using federated hierarchical Bayesian regression. PLoS One 2022;17(12):e0278776. doi: 10.1371/journal.pone.0278776
15. Garcia-Dias R, Scarpazza C, Baecker L, et al. Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. Neuroimage 2020;220:117127. doi: 10.1016/j.neuroimage.2020.117127
16. van der Flier WM, Pijnenburg YA, Prins N, et al. Optimizing patient care and research: the Amsterdam Dementia Cohort. J Alzheimers Dis 2014;41(1):313-27. doi: 10.3233/JAD-132306
17. Jack CR, Jr., Bernstein MA, Fox NC, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 2008;27(4):685-91. doi: 10.1002/jmri.21049
18. Ellis KA, Bush AI, Darby D, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal

study of Alzheimer's disease. Int Psychogeriatr 2009;21(4):672-87. doi: 10.1017/S1041610209009405
19. Frisoni GB, Prestia A, Zanetti O, et al. Markers of Alzheimer's disease in a population attending a memory clinic. Alzheimers Dement 2009;5(4):307-17. doi: 10.1016/j.jalz.2009.04.1235
20. Brueggen K, Grothe MJ, Dyrba M, et al. The European DTI Study on Dementia - A multicenter DTI and MRI study on Alzheimer's disease and Mild Cognitive Impairment. Neuroimage 2017;144(Pt B):305-308. doi: 10.1016/j.neuroimage.2016.03.067
21. Bankó ÉM, Weiss B, Hevesi I, et al. Study protocol of the Hungarian Longitudinal Study of Healthy Brain Aging (HuBA). Ideggyogy Sz 2024;77(1-2):51-59. doi: 10.18071/isz.77.0051
22. Cavedo E, Redolfi A, Angeloni F, et al. The Italian Alzheimer's Disease Neuroimaging Initiative (I-ADNI): validation of structural MR imaging. J Alzheimers Dis 2014;40(4):941-52. doi: 10.3233/JAD-132666
23. Beekly DL, Ramos EM, Lee WW, et al. The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. Alzheimer Dis Assoc Disord 2007;21(3):249-58. doi: 10.1097/WAD.0b013e318142774e
24. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci 2007;19(9):1498-507. doi: 10.1162/jocn.2007.19.9.1498
25. Galluzzi S, Marizzoni M, Babiloni C, et al. Clinical and biomarker profiling of prodromal Alzheimer's disease in workpackage 5 of the Innovative Medicines Initiative PharmaCog project: a 'European ADNI study'. J Intern Med 2016;279(6):576-91. doi: 10.1111/joim.12482
26. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12(3):e1001779. doi: 10.1371/journal.pmed.1001779
27. 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. doi: 10.1038/sdata.2016.44
28. Fischl B, Salat DH, Busa E, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33(3):341-55. doi: 10.1016/s0896-6273(02)00569-x
29. Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. PLoS One 2017;12(9):e0184661. doi: 10.1371/journal.pone.0184661
30. Monereo-Sanchez J, de Jong JJA, Drenthen GS, et al. Quality control strategies for brain MRI segmentation and parcellation: Practical approaches and recommendations - insights from the Maastricht study. Neuroimage 2021;237:118174. doi: 10.1016/j.neuroimage.2021.118174
31. Matelsky J, Kiar G, Johnson E, Rivera C, Toma M, Gray-Roncal W. Container-Based Clinical Solutions for Portable and Reproducible Image Analysis. J Digit Imaging 2018;31(3):315-320. doi: 10.1007/s10278-018-0089-4
32. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. doi: 10.1613/jair.953
33. Chen H, Young A, Oxtoby NP, Barkhof F, Alexander DC, Altmann A. Transferability of Alzheimer's disease progression subtypes to an independent population cohort. Neuroimage 2023;271:120005. doi: 10.1016/j.neuroimage.2023.120005
34. Bron EE, Klein S, Papma JM, et al. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. Neuroimage Clin 2021;31:102712. doi: 10.1016/j.nicl.2021.102712

35. Adams HHH, Evans TE, Terzikhan N. The Uncovering Neurodegenerative Insights Through Ethnic Diversity consortium. Lancet Neurol 2019;18(10):915. doi: 10.1016/S1474-4422(19)30324-2

Figures and Tables

Manufacturer	Scanner Model	Magnetic Field (T)	Number of scans n (Female), n (Male)
Canon	Titan	3.0	252, 329
GE	Discovery MR750	3.0	290, 372
	Discovery MR750w	3.0	8, 16
	Genesis Signa	1.5	6, 3
	Signa Excite	1.5	181, 197
	Signa PET/MR	3.0	15, 16
	Signa HDx	1.5	10, 19
	Signa HDx	3.0	44, 55
	Signa HDxt	1.5	225, 261
	Signa HDxt	3.0	463, 535
	Signa Premier	3.0	6, 8
Philips	Achieva	1.5	4, 7
	Achieva	3.0	179, 116
	Achieva dStream	3.0	13, 10
	Eclipse	1.5	28, 13
	Gemini	3.0	312, 214
	Gyroscan NT	1.0	127, 68
	Ingenia	3.0	18, 33
	Ingenuity	3.0	298, 339
	Intera	1.0	275, 161
	Intera	1.5	19, 42
	Intera	3.0	27, 27
	Intera Achieva	1.5	1, 4
	Intera Gyroscan	1.5	11, 16
Siemens	Allegra	3.0	50, 34
	Avanto	1.5	150, 153
	Biograph	3.0	0, 5
	Espree	1.5	3, 4
	Magnetom Expert	1.0	397, 416
	Magnetom Impact	1.0	7, 2
	Magnetom Vida	3.0	6, 14

	Magnetom Vision	1.5	20, 7
	Prisma	3.0	131, 122
	Prima fit	3.0	122, 84
	RCNS	3.0	68, 48
	Skyra	3.0	6146, 5035
	Sonata	1.5	189, 226
	Sonata Vision	1.5	3, 2
	Symphony	1.5	90, 66
	Trio	3.0	41, 21
	Trio Tim	3.0	457, 380
	Verio	3.0	186, 137
	Vision	1.5	233, 126

Table 1: Scanners considered in this study and their characteristics.

Data Cohort	Participants (processed/ considered after removing outliers)	Age (mean ± standard deviation) [years] †	Sex (F (%) /M (%))†	Diagnosis (HC (%) / SCD (%) / MCI (%) / AD (%)) †	Unique scanners †
ADC	4,086 / 3,722	63.9 ± 9.2	1,717 (46.1%) / 2,005 (53.9%)	0 (0%) / 1,355 (36.4%) / 805 (21.6%) / 1562 (42.0%)	12
ADNI	2,044 / 1,830	72.2 ± 7.06	889 (48.6%) / 941 (51.4%)	687 (37.5%) / 0 (0%) / 851 (46.5%) / 292 (16.0%)	27
AIBL	557 / 524	72.7 ± 6.5	299 (57.1%) / 225 (42.9%)	388 (74.0%) / 0 (0%) / 83 (15.8%) / 53 (10.1%)	3
ARWiBo	913 / 831	56.3 ± 16.2	529 (63.7%) / 302 (36.3%)	603 (72.6%) / 16 (1.9%) / 116 (14.0%) / 96 (11.6%)	7
EDSD	416 / 384	70.4 ± 7.3	197 (51.3%) / 187 (48.7%)	143 (37.2%) / 0 (0%) / 119 (31.0%) / 122 (31.8%)	8
HuBA	121 / 116	62.4 ± 6.9	68 (58.6%) / 48 (41.4%)	116 (100%) / 0 (0%) / 0 (0%) / 0 (0%)	1
I-ADNI	179 / 172	72.2 ± 8.0	106 (61.6%) / 66 (38.4%)	2 (1.2%) / 5 (2.9%) / 35 (20.3%) / 130 (75.6%)	4
NACC	1,861 / 1,731	71.9 ± 9.8	910 (52.6%) / 821 (47.4%)	0 (0%) / 0 (0%) / 949 (54.8%) / 782 (45.2%)	22
OASIS	373 / 359	73.2 ± 10.7	233 (64.9%) / 126 (35.1%)	211 (58.8%) / 0 (0%) / 111 (30.9%) / 37 (10.3%)	1
PharmaCog	141 / 137	69.0 ± 7.3	80 (58.4%) / 57 (41.6%)	0 (0%) / 0 (0%) / 137 (100%) / 0 (0%)	7
UKBB	12,259 / 11,058	63.5 ± 7.6	6,083 (55.0%) / 4,975 (45.0%)	11,058 (100%) / 0 (0%) / 0 (0%) / 0 (0%)	1
Total	22,950 / 20,864	65.3 ± 9.4	11,111 (53.3%) / 9,753 (46.7%)	13,208 (63.3%) / 1,376 (6.6%) / 3,206 (15.4%) / 3,074 (14.7%)	43

Table 2: Participant Demographics. † Values indicated in the column are calculated after removing the outliers, as described in section 2.2. Abbreviations: HC: healthy control; SCD: subjective cognitive decline; MCI: mild cognitive impairment; AD: Alzheimer's Disease.

Figure 1: Age distributions by diagnosis for each scanner in the training cohort. Abbreviations: HC: healthy control; SCD: subjective cognitive decline; MCI: mild cognitive impairment; AD: Alzheimer's Disease.

Figure 2: Experiment 1: Marker concordance for brain volumes on unseen scanners using different harmonization strategies. Concordance for non-harmonized data has also been shown here as a reference measure for comparison. In each diagnostic class, colored stars on top of bars indicate statistically significant differences (FDR<0.05) between the model where the bar is located and the model indicated by the color of the star. Boxes represent individuals between the first and third quartile, black lines inside the boxes represent the medians, whiskers represent individuals above the third quartile and below the first quartile and circles indicate concordance outliers. Abbreviations: HC: healthy control; SCD: subjective cognitive decline; MCI: mild cognitive impairment; AD: Alzheimer's Disease; UKBB: UK Bio-Bank.

Figure 3: Experiment 1: Boxplots of AUCs for distinguishing CN participants from participants with AD in the test set based on the 86 brain ROIs considered before and after harmonization. Boxes represent individuals between the first and third quartile, black lines inside the boxes represent the medians and whiskers represent individuals above the third quartile and below the first quartile. Abbreviations: AUC = area under the receiver operating characteristic curve; HC: healthy control; AD: Alzheimer's Disease; UKBB: UK Bio-Bank.

Figure 4: Experiment 2: Marker concordance for brain volumes on unseen versus seen scanners using normative model and multi-class model. For each diagnostic class, crosses on top of bars indicate statistically significant differences (FDR<0.05) between marker concordances of normative and multi-class model in seen scanners, whereas stars on top of bars indicate statistically significant differences between marker concordances of normative and multi-class model in unseen scanners. Boxes represent individuals between the first and third quartile, black lines inside the boxes represent the medians, whiskers represent individuals above the third quartile and below the first quartile and circles indicate concordance outliers. Abbreviations: HC: healthy control; SCD: subjective cognitive decline; MCI: mild cognitive impairment; AD: Alzheimer's Disease.