# A machine-learning model to harmonize brain volumetric data for quantitative neuro-radiological assessment of Alzheimer's disease

**Authors:**

Damiano Archetti[1,*], Vikram Venkatraghavan[2,3,*], Béla Weiss[4,5], Pierrick Bourgeat[6], Tibor Auer[4,7], Zoltán Vidnyánszky[4], Stanley Durrleman[8], Wiesje M. van der Flier[2,3,9], Frederik Barkhof[10,11,12], Daniel C. Alexander[12], Andre Altmann[12], Alberto Redolfi[1], Betty M. Tijms[2,3], Neil P. Oxtoby[12], for the Australian Imaging Biomarkers and Lifestyle Study[†], for the Alzheimer's Disease Neuroimaging Initiative[‡], and for the E-DADS Consortium

1 Laboratory of Neuroinformatics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy;

2 Alzheimer Center Amsterdam, Neurology, Vrije Universiteit, Amsterdam UMC, location VUmc, Amsterdam, the Netherlands;

3. Amsterdam Neuroscience, Neurodegeneration, Amsterdam, the Netherlands;

4 Brain Imaging Centre, Research Centre for Natural Sciences, Budapest, Hungary;

5 Biomatics and Applied Artificial Intelligence Institute, John von Neumann Faculty of Informatics, Óbuda University, Budapest, Hungary

6 The Australian e-Health Research Centre, CSIRO Health and Biosecurity, Brisbane, Queensland, Australia;

7 School of Psychology, University of Surrey, Guildford, United Kingdom;

8 Sorbonne Université, Institut du Cerveau - Paris Brain Institute – ICM, CNRS, Inria, Inserm, AP-HP, Hôpital Pitié-Salpêtrière, Paris, France;

9 Department of Epidemiology and Data Science, Vrije Universiteit, Amsterdam UMC, location VUmc, the Netherlands;

10 Department of Radiology & Nuclear Medicine, Amsterdam UMC, Vrije Universiteit, the Netherlands;

11 Queen Square Institute of Neurology, University College London, United Kingdom;

12 UCL Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering and Department of Computer Science, University College London, United Kingdom;

*Indicates shared first co-authorship, with the first co-authors being listed alphabetically.

† Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle (AIBL) Study. Unless named, the AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at https://aibl.org.au

‡ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

**Corresponding author:**

Damiano Archetti (OrcID: 0000-0003-0818-7685)

E-mail: darchetti@fatebenefratelli.eu

Phone: +39 0303501502

Address: Via Pilastroni 4, Brescia, 25125, Italy

**Contributors' contacts**

Vikram Venkatraghavan (OrcID: 0000-0001-9759-0462): v.venkatraghavan@amsterdamumc.nl

Béla Weiss (OrcID: 0000-0003-1031-0283): weiss.bela@ttk.hu

Pierrick Bourgeat (OrcID: 0000-0002-2605-4766): pierrick.bourgeat@csiro.au

Tibor Auer (OrcID: 0000-0001-5153-1424): t.auer@surrey.ac.uk

Zoltán Vidnyánszky (OrcID: 0000-0003-3914-3087): vidnyanszky.zoltan@ttk.hu

Stanley Durrleman (OrcID: 0000-0002-9450-6920): stanley.durrleman@inria.fr

Wiesje M. van der Flier (OrcID: 0000-0001-8766-6224): wm.vdflier@amsterdamumc.nl

Frederik Barkhof (OrcID: 0000-0003-3543-3706): f.barkhof@amsterdamumc.nl

Daniel C Alexander (OrcID: 0000-0003-2439-350X): d.alexander@ucl.ac.uk

Andre Altmann (OrcID: 0000-0002-9265-2393): a.altmann@ucl.ac.uk

Alberto Redolfi (OrcID: 0000-0002-4145-9059): aredolfi@fatebenefratelli.eu

Betty M. Tijms (OrcID: 0000-0002-2612-1797): b.tijms@amsterdamumc.nl

Neil P Oxtoby (OrcID: 0000-0003-0203-3909): n.oxtoby@ucl.ac.uk

**Abstract:**

**Background**: Structural MRI plays a pivotal role in the radiological workup for assessing neurodegeneration. Scanner-related differences hinder quantitative neuroradiological assessment of Alzheimer's disease (QNAD). This study aims to train a machine-learning model to harmonize brain volumetric data of patients not encountered during model training.

**Method**: 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 AD-pathology and IQM during harmonization, we developed a multi-class extension of Neuroharmony. We performed cross-validation experiments to benchmark performance against existing approaches using data from 20,864 participants comprising cognitively unimpaired (CU) and impaired (CI) individuals, spanning 11 cohorts and 43 scanners. Evaluation metrics assessed ability to remove scanner-related variations in brain volumes (biomarker concordance), while retaining the ability to delineate different diagnostic groups (preserving disease-related signal).

**Results**: For each strategy, biomarker concordances between scanners were significantly better ($p < 10^{-6}$) compared to pre-harmonized data. The proposed multi-class model achieved significantly higher concordance than the Neuroharmony model trained on CU individuals (CI: $p < 10^{-6}$, CU: $p = 0.02$) and preserved disease-related signal better than the Neuroharmony model trained on all individuals without our proposed extension ($\Delta AUC = -0.09$). The biomarker concordance was better in scanners seen during training (concordance $> 97\%$) than unseen (concordance $< 79\%$), independent of cognitive status.

**Conclusion**: In a large-scale multi-center dataset, our proposed multi-class Neuroharmony model outperformed other strategies available for harmonizing brain-volumetric data in a clinical setting. This paves the way for enabling QNAD in the future.

## 1 Introduction

Structural MRI such as T1-weighted (T1w) sequences 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 atrophy through visual examination of MRI. However, visual examinations are subjective and prone to intra-rater and inter-rater variability. Quantitative imaging biomarkers such as brain volumetric data are becoming increasingly popular because of their potential to improve diagnostic confidence[4]. Quantitative imaging biomarkers 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 such as brain-age estimation[5], or 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 to automated tools. To tackle this problem, many harmonization tools have emerged in recent years[8]. Such algorithms can either harmonize original scans[9] or derivatives extracted from the scans[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 subjects to be scanned with multiple different scanners[13]. Harmonizing volumetric data from MRI scanners not encountered during initial model training needs additional training with a substantial number of images from these scanners[14]. This poses a challenge for the deployment of such methods for clinical use.

Neuroharmony[15] is a recently developed harmonization approach that can harmonize volumetric data from new and unseen MRI 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 experiments indicate that harmonization works for healthy controls, 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 Neuroharmony to account for interaction 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 cognitively unimpaired (CU) individuals (normative model of Neuroharmony) and the original Neuroharmony model trained on

cognitively unimpaired as well as cognitively impaired (CI) individuals without 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 Alzheimer's disease (QNAD).

## 2 Materials and Methods

### *2.1 Participants and scanner characteristics*

T1w MRI data of healthy controls (CN), participants with subjective cognitive decline (SCD), mild cognitive impairment (MCI), and Alzheimer's disease (AD) from 11 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 Bio-bank (UKBB)[26]. Detailed information about each cohort is summarized in Supplementary Table 1.

Minimum inclusion criteria included the availability of a T1w MRI scan along with age, sex, scanner information, and a clinical diagnosis of either CN, SCD, MCI or AD. All datasets were organized according to the BIDS standard[27] to ensure inter-operability and data anonymization. An overview of the scanners used in this study is shown in Table 1.

### *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 CSF volume, total gray matter volume and total brain volume with and without ventricles. All features derived from FreeSurfer are listed in Supplementary Figure 1. 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 and 1.5×IQR above the third quartile[30] per scanner were excluded from our experiments.

In order to ensure reproducibility of our results across different computing environments[31], Docker containers for both FreeSurfer (https://github.com/E-DADS/freesurfer) and MRIQC (https://github.com/E-DADS/mriqc) have been made available online.

### *2.3 Multi-class Neuroharmony model*

The volumetric data from all individuals in the training set were harmonized using ComBat harmonization[10] with empirical Bayes optimization, for removing scanner related batch effects while

preserving the effects of age, sex, and cognitive status. The cognitive status was dichotomized based on the clinical diagnosis as either CU (CN and SCD) or CI (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. To ensure the regressor learns to predict the corrections needed to harmonize accurately in the presence of AD pathology, we used synthetic minority oversampling technique (SMOTE) for data augmentation[32] before training the random forest regressor. This ensured that IQM values with and without neurodegeneration were equally distributed by removing data imbalance between CI and CU individuals. The use of dichotomized cognitive status instead of clinical diagnosis ensured that in the test phase, a full clinical diagnosis is not required as an input 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 to Neuroharmony was compared with two other machine-learning based harmonization strategies that are generalizable to external datasets:

*2.4.1 Normative model:* In the training phase, volumetric data from only the CU individuals 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 both CU and CI individuals were used for training.

### *2.5 Measures for model evaluations*

We use two evaluation metrics to assess the ability of the harmonization strategies to remove scanner-related variations in brain-volumetric data between each pair of scanners (biomarker concordance), while retaining the ability to delineate diagnostic groups (preserving disease-related signal).

Firstly, to assess if the volumetric data are harmonized, we compared the distributions of each volumetric measure for each pair of scanners. 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 CU individuals. The Kolmogorov-Smirnov (KS) test was used for comparing these distributions with the null hypothesis that the distributions between any two pair of scanners were the same. For statistical validity we excluded scanners with fewer than 10 participants of the same diagnostic group from this evaluation. A measure for evaluating harmonization was defined as the

182 percentage of such comparisons across brain regions for each scanner pair, where distributions were
183 *not* statistically different from each other ($p \geq 0.05$) after correcting for multiple testing via false
184 discovery rate (FDR). To provide a reference measure for biomarker concordance, we also computed
185 this measure in the non-harmonized data.

186 Secondly, to assess if disease-related signal was preserved in the volumetric measures, we used area
187 under the receiver operating characteristic (ROC) curve (AUC) as an auxiliary evaluation measure. The
188 ROC curve for distinguishing CN participants from AD patients was computed independently for each
189 volumetric measure. A reference measure for AUC was also computed for the non-harmonized
190 dataset.

191 *2.6 Cross-validation experiments*

192 We performed two experiments in a cross-validation framework. Experiment 1 assessed concordance
193 of the three harmonization strategies, by performing cross-validation at the scanner-level. Experiment
194 2 performed cross-validation at the participant level, using the best-performing scanner-level
195 harmonization models.

196 Experiment 1: To investigate the generalizability of the model to unseen scanners (not included in the
197 training set), we performed 5-fold cross-validation across the 43 available scanners, where in each
198 fold 80% of the scanners were used for training the models, and the remaining 20% of the scanners
199 were used for evaluation. In this experiment, to evaluate the bias introduced by using single-scanner
200 data from the large UKBB cohort, we repeated this experiment for increasing portions of UKBB
201 participants such that when the UKBB data is included in the training data the proportions included
202 were: 10%; 33%; 67%; 100%. However, when UKBB cohort data is used in the test set, we always used
203 100% of the cohort. The non-parametric McNemar Chi-square test was used to compare accuracies
204 across harmonization strategies.

205 Experiment 2: In this experimental setup, the test set consisted of participants from the scanners
206 which were also present in the training set. We selected the two best performing models from
207 Experiment 1 and performed a stratified 5-fold cross-validation across participants, stratified based
208 on the dichotomized cognitive status. For this, the proportion of the UKBB participants included was
209 also decided based on Experiment 1. To provide a reference measure, we compared the accuracies
210 obtained with the corresponding accuracies obtained in Experiment 1.

## 3 Results

### *3.1 Participants*

Table 2 shows descriptive statistics for the combined study sample used in our experiments, which consisted of QC-passed volumetric data from 20,864 participants (53.3% female) from 43 scanners across 11 cohorts. Figure 1 shows age distributions by scanner and cognitive group.

### *3.2 Model evaluation*

Figure 2 shows the first result of Experiment 1: biomarker concordance under cross-validation, independently for each diagnostic group and with increasing proportions of the UKBB dataset. Reference concordance for non-harmonized dataset are also shown for each diagnostic group for comparison. As expected, concordance for each harmonization strategy were significantly higher than the non-harmonized data for all the diagnostic groups ($p < 10^{-6}$). 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 ($p < 10^{-6}$). For diagnostic groups of CN and SCD, the concordance of the multi-class model was significantly higher than the normative model when the proportion of UKBB subjects included was 100% ($p_{CN} = 0.01, p_{SCD} = 0.02$). The inclusive model's concordance for CN and SCD subjects was statistically similar to that of the normative model ($p > 0.05$).

Figure 3 shows the second result of Experiment 1: CN vs AD AUC 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 imbalance in the number of CN and AD participants in the different scanners. The AUCs of the normative model and multi-class model are slightly lower than non-harmonized volumes ($\Delta AUC = -0.01$, paired t-test $p = 10^{-5}$). However, the multi-class model significantly outperformed the inclusive model ($\Delta AUC = -0.09$, paired t-test $p < 10^{-10}$), indicating relative loss of disease related signal when using the inclusive model harmonization strategy.

### *3.3 Harmonization in seen vs unseen MRI scanners*

Figure 4 shows the results of Experiment 2: biomarker concordance as a function of scanners seen vs unseen scanners during model training, for the normative model and the multi-class model. Supplementary Figure 1 shows the same for each brain volume individually. Biomarker concordance of the multi-class model was significantly higher than the normative model for unseen scanners, based on McNemar test for all diagnostic categories ($p_{CN} = 0.01$, $p_{SCD} = 0.02, p_{MCI} < 10^{-6}, p_{AD} < 10^{-6}$). For seen scanners, the multi-class model harmonization strategy significantly outperformed

the normative model for the diagnostic groups of CN, MCI, and AD ($p_{CN} < 10^{-6}, p_{MCI} < 10^{-6}, p_{AD} < 10^{-6}$), but significantly underperformed for SCD ($p_{SCD} = 0.02$). Harmonization accuracy using the multi-class model in a seen scanner (accuracy > 97%) was better for all diagnostic groups than in unseen scanners (accuracy < 79%).

## 4 Discussion

We introduced a novel extension to the Neuroharmony harmonization model to train a generalizable machine learning model for harmonizing multicentric brain-volumetric data for quantitative neuro-radiological assessment of Alzheimer's disease. The data for these evaluation experiments were derived from T1w MRIs acquired with 43 different scanners from 20,864 participants spanning 11 cohorts. The newly introduced multi-class model would be helpful in harmonizing volumetric data while using automated methods in clinics and research where there could be data from new scanners not included in training. The trained model has been made openly available at https://www.neugrid2.eu/index.php/edads_harmonization/ while the code to train the model has been made available in: https://github.com/e-dads/Multiclass-Neuroharmony/.

Our experiments showed that the multi-class model, that accounts for the interaction between disease pathology and image quality metrics to remove scanner-related effects, significantly improved harmonization accuracy for patients in unseen scanners, as compared to normative modelling. For seen scanners, it improved the harmonization accuracy for all diagnostic groups except SCD, potentially due to the low sample size of the SCD group (n=1,376). Additionally, we showed that the multi-class model of Neuroharmony preserves disease-related signature during harmonization.

Harmonization of biomarker data from unseen scanners remains a challenge: biomarker concordance for both normative and multi-class models in unseen scanners was lower than obtained for 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 which uses a smaller dataset for training (normative model). However, we think this setting is a realistic and fair comparison, because normative modelling always discards data from CI individuals. Through our modifications to the Neuroharmony model, we provided a way to include both CI and CU individuals in the training data, which our experiments show improves harmonization in both seen and unseen scanners while preserving disease-related signature.

The harmonization performance obtained with the normative model in our experiments was lower than reported in the original Neuroharmony paper[15]. We think this is 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 future clinical implementation.

Some limitations of the original Neuroharmony model[15] also 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 T1w 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 biomarker concordance across scanner-pairs is a surrogate measure to measure consistency in the absence of a ground-truth. To overcome this limitation, it would be useful to measure the harmonization performance for participants scanned with multiple scanners, in the future.

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 machine learning 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[34] 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 data inclusion from under-represented countries.

Challenges in the clinical implementation of the harmonization strategy: while the multi-class model outperformed the normative model in terms of harmonization accuracy, 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 CI from CN/SCD can be done with high accuracy using MRI[35]. To avoid a circular dependency between the two tasks, we think that developing multi-task machine learning models to jointly harmonize and predict cognitive status is an important avenue of future work.

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). Future work validating harmonization approaches for patients with other types of dementia is crucial for eventual clinical implementation.

In summary, we have generalized the Neuroharmony model to harmonize imaging biomarker data from multisite studies while retaining disease signal that could otherwise be removed by the harmonization procedure. Demonstrating on brain MRI biomarker data from the Alzheimer's disease spectrum, our new method outperforms others on both seen and unseen scanners, making it more

313 suitable for clinical applications related to cognitive decline, such as memory clinics and clinical trials
314 of new interventions for neurodegenerative diseases.

**Funding**

This study was supported by the Early Detection of Alzheimer's Disease Subtypes (E-DADS) project, an EU Joint Programme — Neurodegenerative Disease Research (JPND) project (see www.jpnd.eu). The project is supported under the aegis of JPND through the following funding organizations: United Kingdom, Medical Research Council (MR/T046422/1); Netherlands, ZonMW (733051106); France, Agence Nationale de la Recherche (ANR-19-JPW2–000); Italy, Italian Ministry of Health (MoH); Australia, National Health & Medical Research Council (1191535); Hungary, National Research, Development and Innovation Office (2019–2.1.7-ERA-NET-2020–00008).

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-5353. F.B. is supported by the NIHR Biomedical Research Centre at UCLH. B.W. and Z.V. were supported by Project no. RRF-2.3.1-21-2022-00015, which has been implemented with the support provided by the European Union. B.W. was supported by the Consolidator Researcher program of the Óbuda University.

N.P.O. is supported by a UKRI Future Leaders Fellowship (UK Medical Research Council MR/S03546X/1).

**Acknowledgments**

The authors would like to thank the developers of the original Neuroharmony algorithm for making the code available, and all the research participants and their families for donating their data for scientific research.

Data collection and sharing for ADNI was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by

the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education , and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The AIBL study (https://aibl.org.au) is a consortium between Austin Health, CSIRO, Edith Cowan University, the Florey Institute (The University of Melbourne), and the National Ageing Research Institute. The study has received partial financial support from the Alzheimer's Association (US), the Alzheimer's Drug Discovery Foundation, an Anonymous foundation, the Science and Industry Endowment Fund, the Dementia Collaborative Research Centres, the Victorian Government's Operational Infrastructure Support program, the Australian Alzheimer's Research Foundation, the National Health and Medical Research Council (NHMRC), and The Yulgilbar Foundation. Numerous commercial interactions have supported data collection and analyses. In-kind support has also been provided by Sir Charles Gairdner Hospital, Cogstate Ltd, Hollywood Private Hospital, The University of Melbourne, and St Vincent's Hospital. The AIBL team wishes to thank all clinicians who referred patients with AD and/or MCI to the study. We also thank all those who took part as subjects in the study for their commitment and dedication to helping advance research into the early detection and causation of AD. We thank all the investigators within the AIBL who contributed to the design and implementation of the resource and/or provided data but did not actively participate in the development, analysis, interpretation or writing of this current study.

The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary

Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

## Conflicts of interest

F.B. is on the steering committee or Data Safety Monitoring Board member for Biogen, Merck, ATRI/ACTC and Prothena. F.B. has been a consultant for Roche, Celltrion, Rewind Therapeutics, Merck, IXICO, Jansen, Combinostics and has research agreements with Merck, Biogen, GE Healthcare, Roche. F.B. and D.C.A. are co-founders and shareholders of Queen Square Analytics Ltd. N.P.O. is a consultant for Queen Square Analytics Ltd.

## Data and code availability

ADC data can be made available to academic researchers upon reasonable request;

ADNI and AIBL data are stored at the Laboratory of Neuroimaging at the University of Southern California and are available to the general scientific community for download: http://adni.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 containers for FreeSurfer and MriQC are available on the E-DADS 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;

Harmonization model files are available for all researchers on the NeuGRID2 platform: https://www.neugrid2.eu/index.php/edads_harmonization.

## References

1. Johnson KA, Fox NC, Sperling RA, Klunk WE. Brain imaging in Alzheimer disease. *Cold Spring Harb Perspect Med*. Apr 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*. Nov 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*. May 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*. Sep 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*. Oct 15 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*. Oct 15 2018;9(1):4273. doi:10.1038/s41467-018-05892-0
7. Liu S, Hou B, Zhang Y, Lin T, Fan X, You H, Feng F. Inter-scanner reproducibility of brain volumetry: influence of automated brain segmentation software. *BMC Neurosci*. Sep 4 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*. Apr 1 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*. Dec 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*. Feb 15 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*. Jan 2023;270(1):446-459. doi:10.1007/s00415-022-11387-2
12. Zhou HH, Singh V, Johnson SC, Wahba G, Alzheimer's Disease Neuroimaging I. Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. *Proc Natl Acad Sci U S A*. Feb 13 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*. Oct 15 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*. Apr 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*. Aug 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*. Jul 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*. Jan 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). *medRxiv*. 2023:2023.11.09.23298159. doi:10.1101/2023.11.09.23298159
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*. Jul-Sep 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*. Sep 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*. Jun 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*. Mar 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*. Jun 21 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*. Jan 31 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*. Aug 15 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*. Jun 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:DOI 10.1613/jair.953
33. Chen H, Young A, Oxtoby NP, Barkhof F, Alexander DC, Altmann A, investigators A. Transferability of Alzheimer's disease progression subtypes to an independent population cohort. *Neuroimage*. May 1 2023;271:120005. doi:10.1016/j.neuroimage.2023.120005
34. Adams HHH, Evans TE, Terzikhan N. The Uncovering Neurodegenerative Insights Through Ethnic Diversity consortium. *Lancet Neurol*. Oct 2019;18(10):915. doi:10.1016/S1474-4422(19)30324-2

507 35. Bron EE, Klein S, Papma JM, et al. Cross-cohort generalizability of deep and conventional
508 machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *Neuroimage Clin*.
509 2021;31:102712. doi:10.1016/j.nicl.2021.102712

510

511 **Figures and Tables**

| Manufacturer | Scanner Model | Magnetic Field (T) | Number of scans (Female %) |
|---|---|---|---|
| Canon | Titan | 3.0 | 581 (43.4%) |
| GE | Discovery MR750 | 3.0 | 662 (43.8%) |
| | Discovery MR750w | 3.0 | 24 (33.3%) |
| | Genesis Signa | 1.5 | 9 (66.7%) |
| | Signa Excite | 1.5 | 378 (47.8%) |
| | Signa PET/MR | 3.0 | 31 (48.3%) |
| | Signa HDx | 1.5 | 29 (34.5%) |
| | Signa HDx | 3.0 | 99 (44.4%) |
| | Signa HDxt | 1.5 | 486 (46.3%) |
| | Signa HDxt | 3.0 | 998 (46.4%) |
| | Signa Premier | 3.0 | 14 (42.9%) |
| Philips | Achieva | 1.5 | 11 (36.3%) |
| | Achieva | 3.0 | 295 (60.6%) |
| | Achieva dStream | 3.0 | 23 (56.5%) |
| | Eclipse | 1.5 | 41 (68.3%) |
| | Gemini | 3.0 | 526 (59.3%) |
| | Gyroscan NT | 1.0 | 195 (65.1%) |
| | Ingenia | 3.0 | 51 (35.3%) |
| | Ingenuity | 3.0 | 637 (46.7%) |
| | Intera | 1.0 | 436 (63.1%) |
| | Intera | 1.5 | 61 (31.1%) |
| | Intera | 3.0 | 54 (50.0%) |
| | Intera Achieva | 1.5 | 5 (20.0%) |
| | Intera Gyroscan | 1.5 | 27 (40.7%) |
| Siemens | Allegra | 3.0 | 84 (59.5%) |
| | Avanto | 1.5 | 303 (49.5%) |
| | Biograph | 3.0 | 5 (0.0%) |
| | Espree | 1.5 | 7 (42.8%) |
| | Magnetom Expert | 1.0 | 813 (48.8%) |
| | Magnetom Impact | 1.0 | 9 (77.8%) |

| | | | |
|---|---|---|---|
| | Magnetom Vida | 3.0 | 20 (30.0%) |
| | Magnetom Vision | 1.5 | 27 (74.1%) |
| | Prisma | 3.0 | 253 (51.8%) |
| | Prima fit | 3.0 | 216 (56.5%) |
| | RCNS | 3.0 | 116 (58.6%) |
| | Skyra | 3.0 | 11181 (54.9%) |
| | Sonata | 1.5 | 415 (45.5%) |
| | Sonata Vision | 1.5 | 5 (60%) |
| | Symphony | 1.5 | 156 (57.7%) |
| | Trio | 3.0 | 62 (66.2%) |
| | Trio Tim | 3.0 | 837 (54.6%) |
| | Verio | 3.0 | 323 (57.6%) |
| | Vision | 1.5 | 359 (64.9%) |

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

| | Participants (processed/ considered after removing outliers) | Age [years] † | Sex (F/M)† | Diagnosis (CN / SCD / MCI / AD) † | Unique scanners † |
|---|---|---|---|---|---|
| ADC | 4,086 / 3,722 | 63.9 ± 9.2 | 1,717 / 2,005 | 0 / 1,355 / 805 /1562 | 12 |
| ADNI | 2,044 / 1,830 | 72.2 ± 7.06 | 889 / 941 | 687 / 0 / 851 / 292 | 27 |
| AIBL | 557 / 524 | 72.7 ± 6.5 | 299 / 225 | 388 / 0 / 83 / 53 | 3 |
| ARWiBo | 913 / 831 | 56.3 ± 16.2 | 529 / 302 | 603 / 16 / 116 / 96 | 7 |
| EDSD | 416 / 384 | 70.4 ± 7.3 | 197 / 187 | 143 / 0 / 119 / 122 | 8 |
| HuBA | 121 / 116 | 62.4 ± 6.9 | 68 / 48 | 116 / 0 / 0 / 0 | 1 |
| I-ADNI | 179 / 172 | 72.2 ± 8.0 | 106 / 66 | 2 / 5 / 35 / 130 | 4 |
| NACC | 1,861 / 1,731 | 71.9 ± 9.8 | 910 / 821 | 0 / 0 / 949 / 782 | 22 |
| OASIS | 373 / 359 | 73.2 ± 10.7 | 233 / 126 | 211 / 0 / 111 / 37 | 1 |
| PharmaCog | 141 / 137 | 69.0 ± 7.3 | 80 / 57 | 0 / 0 / 137 / 0 | 7 |
| UKBB | 12,259 / 11,058 | 63.5 ± 7.6 | 6,083 / 4,975 | 11,058 / 0 / 0 / 0 | 1 |
| Total | 22,950 / 20,864 | 65.3 ± 9.4 | 11,111 / 9,753 | 13,208 / 1,376 / 3,206 / 3,074 | 43 |

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

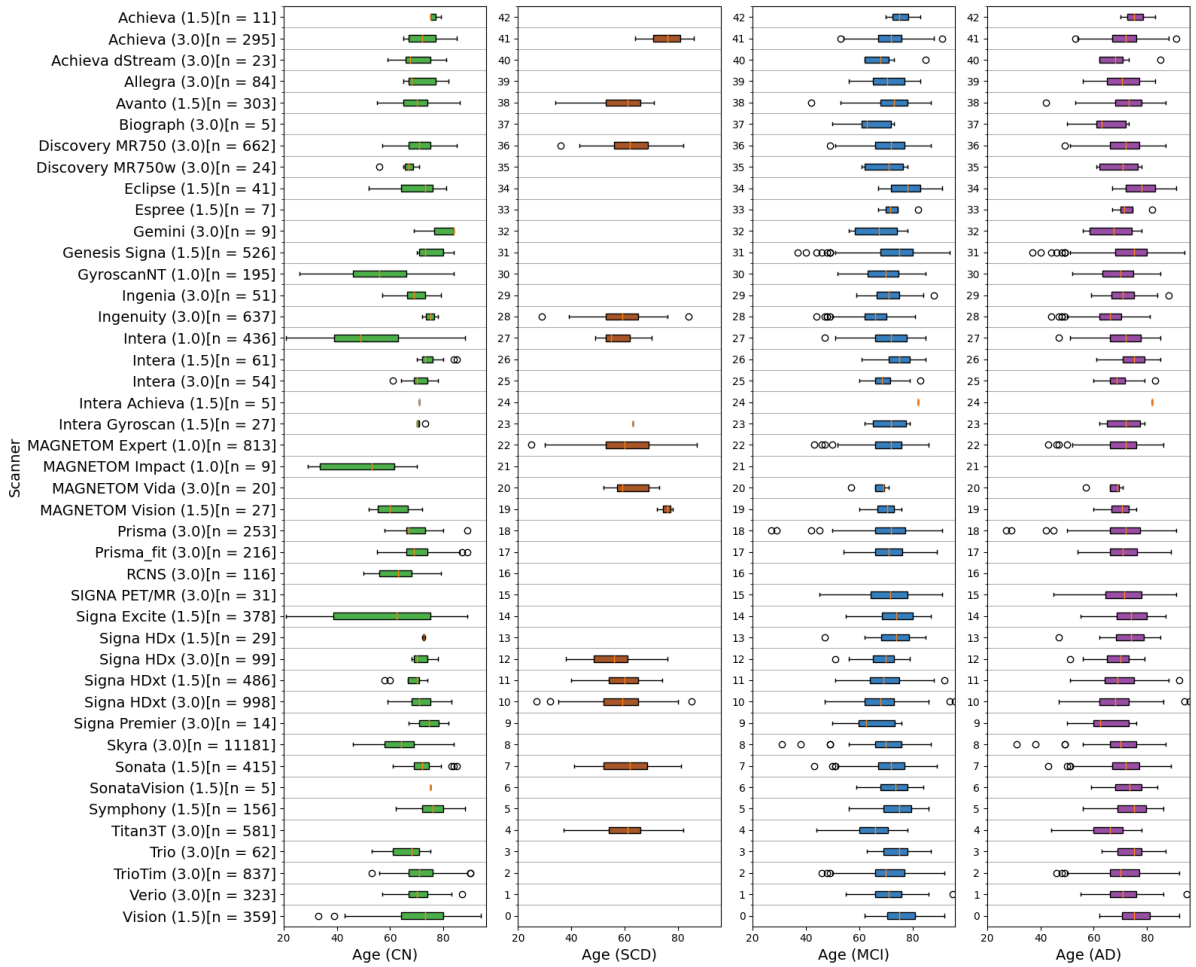


**Figure 1**: Diagnosis-wise distributions of age for each scanner in the training cohort. Abbreviations: CN: cognitively normal; SCD: subjective cognitive decline; MCI: mild cognitive impairment; AD: Alzheimer's Disease.

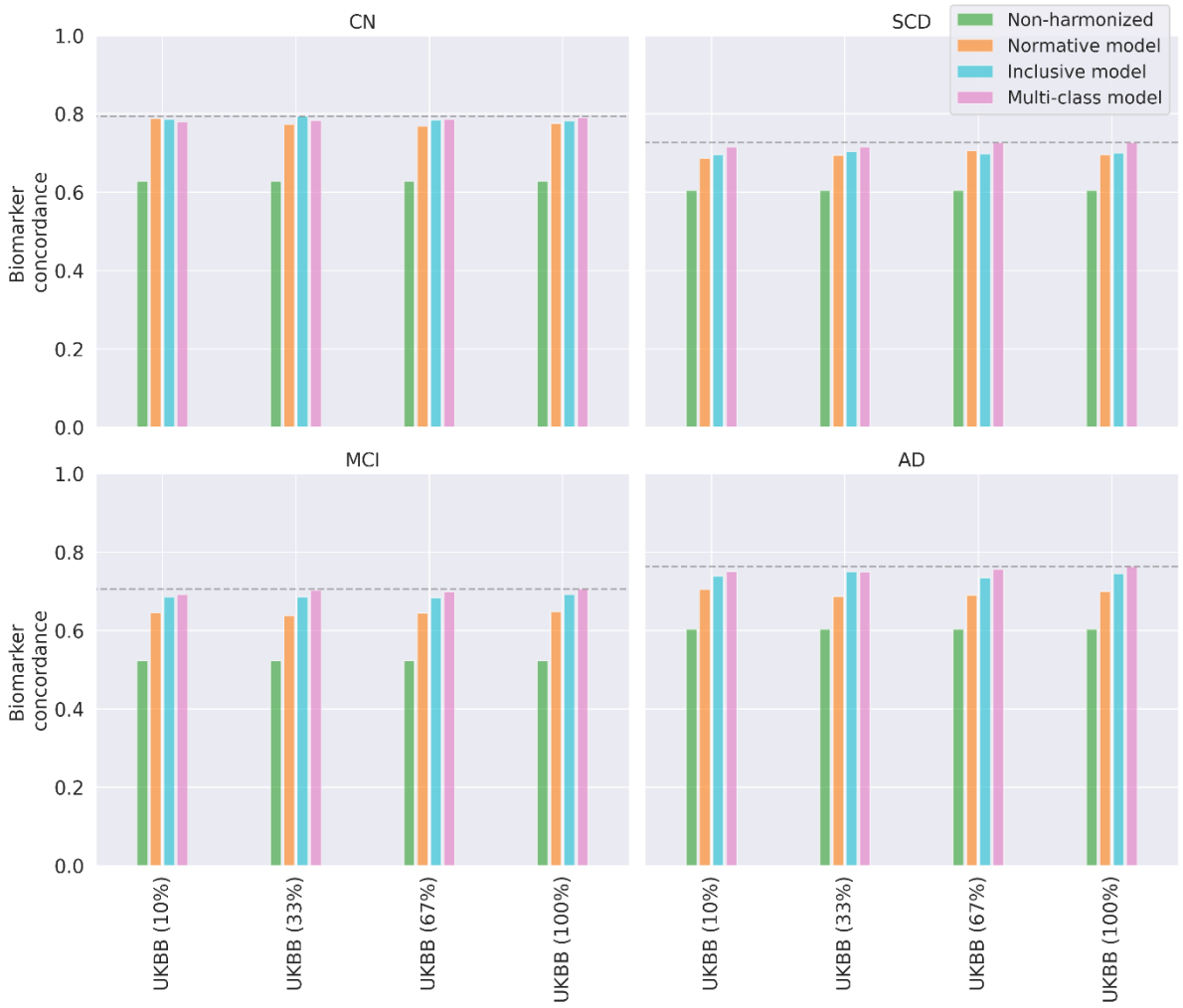


**Figure 2:** Experiment 1: Biomarker concordance for brain volumes on unseen scanners using different harmonization strategies. Concordance for non-harmonized data has also been shown here as a reference measure for comparison. Abbreviations: CN: cognitively normal; SCD: subjective cognitive decline; MCI: mild cognitive impairment; AD: Alzheimer's Disease.

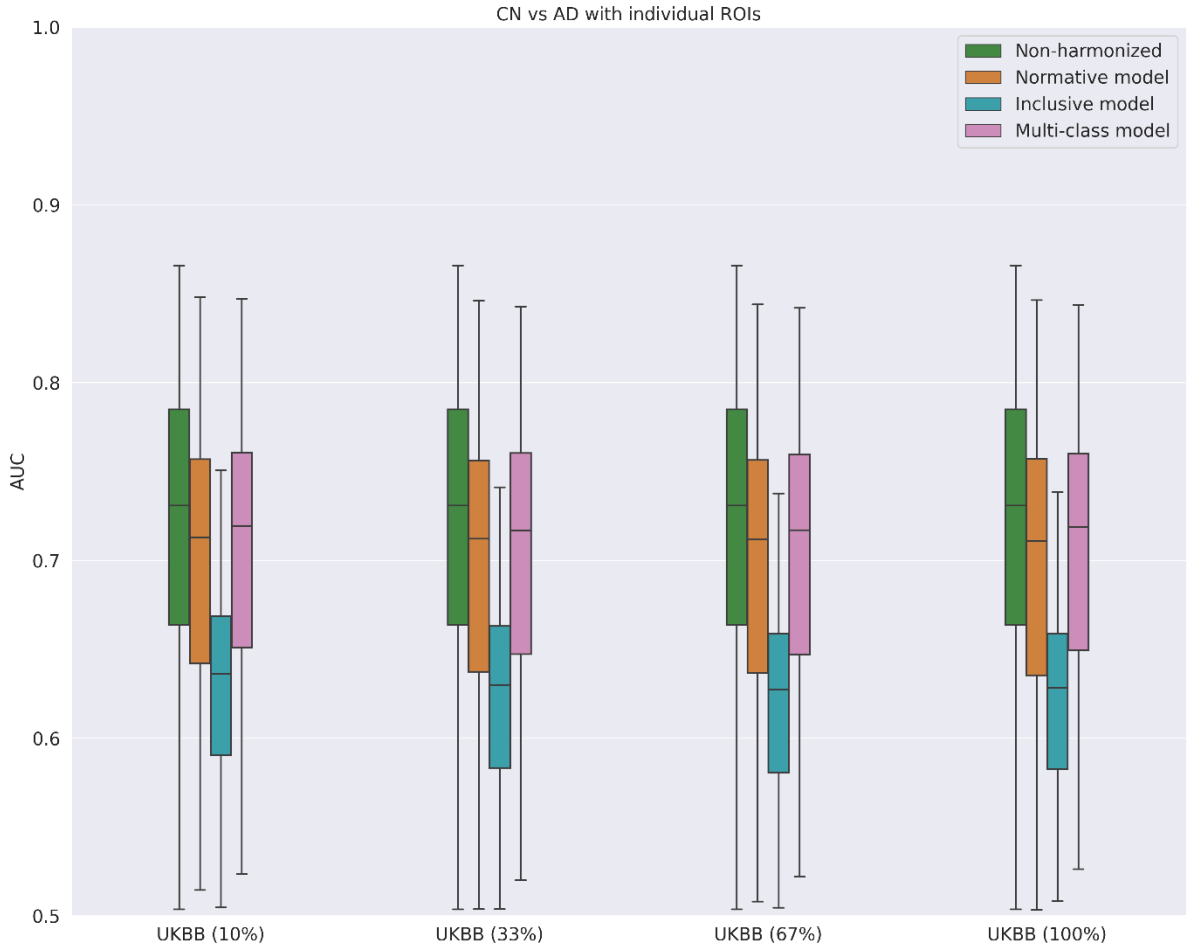


**Figure 3:** Experiment 1: Boxplots of AUCs for distinguishing CN participants from AD patients in the test set based on the 86 brain ROIs considered before and after harmonization. Abbreviations: CN: cognitively normal; AD: Alzheimer's Disease.

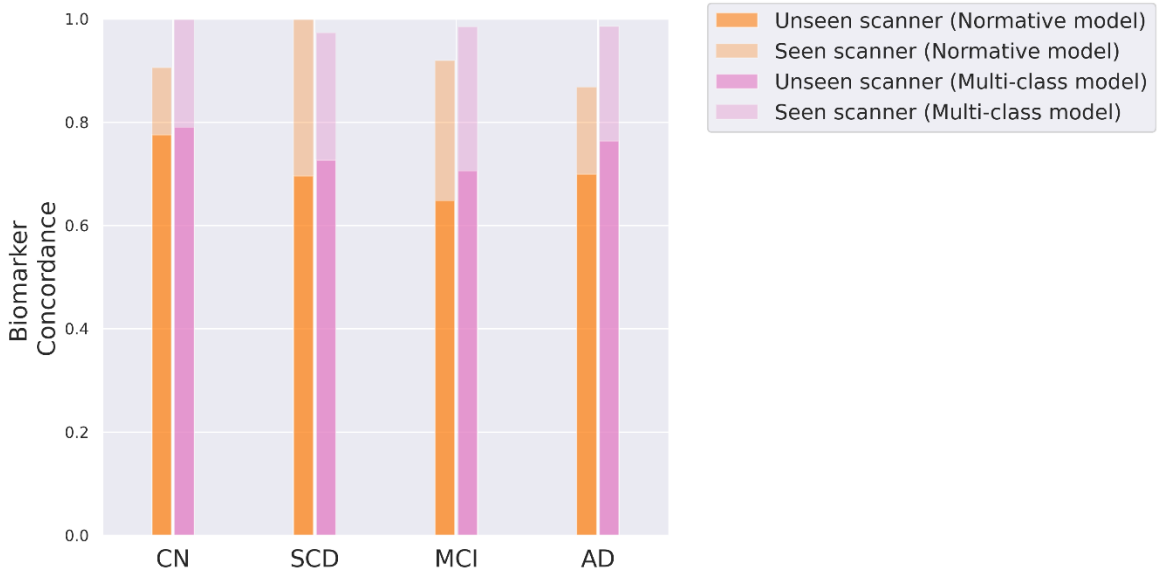


**Figure 4:** Experiment 2: Accuracy of harmonization for harmonizing brain volumes on seen versus unseen scanners using normative model and multi-class model. Abbreviations: CN: cognitively normal; SCD: subjective cognitive decline; MCI: mild cognitive impairment; AD: Alzheimer's Disease.