Preserving Regional Structural MRI Information in Machine-Learning Brain-age Models Improves Diagnostic Discrimination and Cognitive Prediction in Alzheimer's Disease

Zeeshan Mehmood[1][1] and John-Jon Steyn[2]
[1]Independent researcher, Edinburgh, UK
[2]Independent researcher, Port Louis, Mauritius

zee920786@outlook.com
johnjonsteyn@protonmail.com

Abstract: Alzheimer's disease (AD) involves structural brain change that overlaps with normal ageing but is not reducible to it. Brain-age modelling is an interpretable artificial-intelligence approach in which machine-learning models learn normative age-related patterns from neuroimaging and estimate an individual's brain age. The resulting brain-age gap has been proposed as a marker of disease-related deviation, yet recent evaluations question whether a single scalar gap adds clinical value beyond simple structural measures or higher-dimensional representations. This study tested whether representation choice determines the diagnostic and cognitive utility of brain-age modelling in ADNI structural MRI. Baseline 3T T1-weighted MPRAGE scans from 634 participants were processed with FastSurfer, comprising 171 cognitively normal (CN), 358 mild cognitive impairment (MCI), and 105 AD participants. A CN-only RidgeCV brain-age model was evaluated using subject-level out-of-fold prediction. Three representations were compared within the same leak-safe downstream harness: a scalar whole-brain gap, a 171-region gap vector, and raw FastSurfer anatomical predictors. Sensitivity analyses tested raw, linear, and non-linear age-bias correction, and calibration was assessed alongside discrimination. The regional gap outperformed the scalar gap across all diagnostic and cognitive outcomes. On the primary CN versus MCI+AD contrast, AUC increased from 0.692 with the scalar gap to 0.766 with the regional gap. On CN versus AD, regional-gap AUC reached 0.935. The scalar gap did not outperform hippocampal volume on any outcome, whereas only CN versus AD showed a supported regional advantage over hippocampal volume. Exploratory bias correction had negligible impact on group-level gap distributions. Regional and raw representations also produced lower Brier scores than the demographic baseline and scalar gap. These findings suggest that the clinical limitation of brain-age modelling lies less in the marker itself than in scalar compression of anatomically heterogeneous disease signal. Preserving regional structure retains interpretability while recovering diagnostic and cognitive information lost in a single whole-brain score. For clinical AI, these findings show that representation choice can determine how much disease-relevant information an interpretable brain-age model retains.

Keywords: Artificial intelligence, Machine learning, Alzheimer's disease, Brain age, Structural MRI, Regional morphometry

1. Introduction

Alzheimer's disease is marked by progressive structural brain change that begins before dementia is clinically established and continues through MCI and symptomatic disease. Structural MRI can quantify cortical thinning, hippocampal and medial-temporal atrophy, ventricular enlargement, and broader grey- and white-matter alteration. These disease-related changes overlap with normal ageing without being identical to it. AD therefore creates a methodological problem: clinically relevant neurodegeneration is partly age-like, partly disease-specific, and spatially heterogeneous.

Brain-age modelling is an interpretable machine-learning approach within medical artificial intelligence (AI) for studying this overlap. Trained on neuroimaging from cognitively normal or broadly healthy participants, a model predicts chronological age from structural features. Predicted brain age minus chronological age defines the brain-age gap: positive values indicate older-appearing brain structure, while negative values indicate younger-appearing structure. Early BrainAGE studies reported that increased apparent brain age was associated with MCI, AD, cognitive decline, and conversion from MCI to AD (Gaser et al., 2013; Löwe et al., 2016). Later studies linked brain-age-gap measures to cognition, AD pathology, and longitudinal progression using structural MRI, FDG-PET, or multimodal neuroimaging (Taylor et al., 2022; Millar et al., 2023; Doering et al., 2024; Ly et al., 2024).

The attraction of the brain-age gap is its simplicity. It compresses complex anatomical variation into a single, clinically interpretable deviation from age expectation. However, this compression can make distinct regional patterns indistinguishable to downstream models. AD is not expressed uniformly across the brain, and medial-temporal, hippocampal, ventricular, frontal, temporal, and white-matter features may carry different diagnostic and cognitive information. A scalar gap can indicate that a brain appears older than expected, but it cannot

[1]Zeeshan Mehmood and John-Jon Steyn contributed equally to this work.

identify where that deviation is expressed. Participants with different regional patterns can therefore receive similar scalar values.

Recent evaluations have sharpened this concern. Dörfel et al. (2026) found that public brain-age packages detected diagnostic group differences in ADNI but had limited clinical utility and did not consistently outperform simple grey-matter measures. Tan et al. (2025) found that a scalar gap performed poorly for AD-related prediction compared with higher-dimensional intermediate representations and direct disease models. Garcia Condado et al. (2025) showed that features useful for chronological age prediction are not necessarily those useful for AD classification. These findings do not show that brain-age models lack disease-relevant signal. They indicate that scalarisation may discard clinically relevant anatomical information.

Regional brain-age modelling tests this possibility directly. Kalc et al. (2024) extended BrainAGE to regional scores, showing that spatially specific estimates can reveal disease-related patterns not visible in a global score. Hwang et al. (2022) likewise showed that typical ageing and AD-related atrophy are anatomically entangled rather than separable by a single global axis. The present study treats regionalisation not as a separate disease classifier, but as an alternative representation of brain-age deviation itself.

This paper evaluates whether preserving regional information in brain-age modelling improves clinical utility in ADNI structural MRI. FastSurfer-derived structural features were compared using three representations that differ in anatomical compression: a scalar whole-brain gap, a regional gap vector, and raw regional features. Models were evaluated using subject-level out-of-fold predictions to reduce leakage. We also tested whether conclusions depended on age-bias correction and reported calibration alongside discrimination. The study asks whether preserving regional age deviations retains more diagnostic and cognitive information than a scalar gap, whether bias correction changes the conclusions, and whether brain-age representations improve calibration as well as discrimination.

2. Related Work

Early BrainAGE studies established apparent brain ageing as a candidate marker of disease-related neurodegeneration. Gaser et al. (2013) reported that higher baseline BrainAGE predicted conversion from MCI to AD and compared favourably with chronological age, cognitive markers, hippocampal measures, and cerebrospinal-fluid markers. Löwe et al. (2016) showed that BrainAGE differed across CN, stable MCI, progressive MCI, and AD groups, and related to cognition and conversion. Later work connected brain-age estimates with cognition, AD pathology, and disease progression across structural MRI, FDG-PET, and multimodal settings (Taylor et al., 2022; Millar et al., 2023; Doering et al., 2024; Ly et al., 2024).

More recent work has questioned the clinical value of the scalar gap. Dörfel et al. (2026) found that brain-age estimates differed by diagnosis but added little over grey-matter volume and showed weak associations with clinically meaningful longitudinal outcomes. Tan et al. (2025) reported that the scalar gap performed poorly for AD-related prediction, while higher-dimensional intermediate representations preserved more useful information. Dular and Špiclin (2024) showed that gap magnitudes vary across model architectures and cohorts, weakening the assumption that a scalar gap is a model-independent biomarker. Garcia Condado et al. (2025) emphasised the tension between age-prediction and disease-discrimination objectives.

Regional and disentangled approaches address this limitation by preserving spatial information. Kalc et al. (2024) extended BrainAGE to regional estimates and showed disease-specific patterns in ADNI and schizophrenia samples. Hwang et al. (2022) showed that typical ageing and AD-related neurodegeneration overlap anatomically and that multivariate modelling improves interpretability. The unresolved issue is whether the scalar gap is an appropriate representation for downstream clinical prediction. Prior work has not jointly compared a scalar gap, regional gap vector, and raw regional features within the same subject-level out-of-fold evaluation while also testing bias-correction sensitivity, calibration, and simple structural baselines.

3. Materials and Methods

3.1 Data and Cohort

Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal multicentre study designed to support biomarker development for Alzheimer's disease and related clinical progression (Petersen et al., 2010). The cohort was restricted to baseline preprocessed 3T T1-weighted MPRAGE scans from CN, MCI, and AD participants, reducing heterogeneity in field strength, sequence type, and visit timing. ADNI early-MCI and late-MCI labels were harmonised into the broader MCI class for diagnostic analyses.

Each image was linked to the nearest same-subject clinical examination within ±90 days of MRI. Images without an eligible visit were excluded. Among eligible records, one scan was retained per participant using a deterministic priority rule based on ADNI standardised-list status, image-clinical date proximity, cognitive-score availability, MPRAGE rank, image size, and image identifier. This produced one row per participant, containing one MRI, one linked clinical visit, and one harmonised diagnostic label.

Selected images were processed with FastSurfer v2.4.2, a deep-learning-based structural MRI segmentation and morphometric pipeline (Henschel et al., 2020). Outputs were parsed into a regional morphometric table containing subcortical, cortical, white-matter parcellation, and global brain-volume measures. The assembled modelling dataset retained participant metadata, diagnosis, cognitive measures, and 1,382 FastSurfer-derived variables before the predictor exclusions described below.

After processing and quality control, the final analytic cohort contained 634 participants: 171 CN, 358 MCI, and 105 AD. CN participants had a mean age of 73.0 years (SD 6.3), MCI participants 70.9 (SD 7.5), and AD participants 74.0 (SD 7.9). The impaired groups were not systematically older than CN participants, reducing the likelihood that downstream brain-age differences were driven by age imbalance. The CN group included 88 women and 83 men, the MCI group 170 women and 188 men, and the AD group 50 women and 55 men. Mean education was 16.6 years in CN, 16.2 in MCI, and 15.8 in AD.

Table 1: Demographic characteristics of the final analytic cohort. Values are shown as mean (SD) for continuous variables and n (%) for sex. CN = cognitively normal; MCI = mild cognitive impairment; AD = Alzheimer's disease

Diagnosis	N	Age, mean (SD)	Female, n (%)	Male, n (%)	Education, mean (SD)
CN	171	73.0 (6.3)	88 (51.5%)	83 (48.5%)	16.6 (2.5)
MCI	358	70.9 (7.5)	170 (47.5%)	188 (52.5%)	16.2 (2.7)
AD	105	74.0 (7.9)	50 (47.6%)	55 (52.4%)	15.8 (2.6)

3.2 Quality Control

The modelling dataset contained no harmonised manual raw-image quality-control flag. FastSurfer produces a whole-brain surface reconstruction defect count, SurfaceHoles, which serves as an automated proxy for reconstruction quality, analogous to the Euler number approach described by Rosen et al. (2018). Subjects with whole-brain SurfaceHoles values exceeding the median plus three scaled median absolute deviations were excluded before modelling, following Leys et al. (2013). The threshold was approximately 49 defects, excluding 54 of 688 subjects after one processing failure (RID 4377). Exclusion rates were 6.0% for CN, 7.5% for MCI, and 11.8% for AD. SurfaceHoles columns were removed from the predictor set because defect count reflects reconstruction quality rather than anatomy.

This procedure provides an automated proxy for reconstruction quality, but it is not raw-image QC; motion artefacts or acquisition noise may therefore remain undetected. Because no manual raw-image QC label was available in the modelling data, SurfaceHoles provided the strongest automated quality signal available.

3.3 Brain-age Model

The brain-age model was trained exclusively on CN participants using RidgeCV. The global scalar model predicted chronological age from the FastSurfer structural predictor set after removal of SurfaceHoles columns. Five-fold GroupKFold cross-validation used participant IDs as groups, ensuring that no participant contributed to both training and test sets. CN participants received out-of-fold predicted brain ages. MCI and AD participants were scored using the average of fold-specific models, so the brain-age model had no exposure to either impaired diagnostic group during training.

Three leakage rules were enforced. First, StandardScaler was fitted on the CN training fold only and applied to held-out and non-CN subjects without refitting. Second, RidgeCV alpha was tuned by inner cross-validation within each CN training fold only. Third, exploratory bias-correction steps were fitted on the CN training fold only and applied without refitting.

3.4 Representations

Three representations were evaluated. The scalar gap was predicted brain age minus chronological age, producing one value per participant. The regional gap vector used a separate brain-age model for each of 171 brain regions, producing 171 per-region gap scores per participant. The raw feature matrix comprised 1,379

FastSurfer anatomical predictors directly, without a brain-age modelling step. These representations span progressively greater structural detail, allowing direct comparison of retained information.

3.5 Bias-correction and Evaluation

Brain-age gaps can exhibit age-related bias, where older participants are under-predicted and younger participants over-predicted. Bias-correction sensitivity was assessed separately using the exploratory Ridge (α=1.0) brain-age model. Three variants were tested: raw gap, linear residual correction using age, and non-linear residual correction using age and age squared. Within each five-fold CN GroupKFold split, correction models were fitted on the training fold raw gaps only and applied to held-out CN and non-CN subjects without refitting.

The primary evaluation tested incremental value over a demographics-only baseline comprising age, sex, and education. Two simple structural comparators were also evaluated: bilateral hippocampal volume and head-size-normalised grey matter volume, each added to the same demographic baseline. All downstream models were evaluated out-of-fold using shared five-fold StratifiedKFold splits with shuffling and random seed 42; cognitive folds were stratified by diagnostic status. Every model used a leak-safe Pipeline combining StandardScaler and the downstream estimator. Diagnostic outcomes used LogisticRegressionCV with five-fold internal cross-validation; AUC and Brier score were reported. The primary diagnostic contrast was CN versus MCI+AD; CN versus AD was secondary. Cognitive outcomes used RidgeCV with five-fold internal cross-validation; incremental R^2 over the demographic baseline was reported for MMSE, CDRSB, ADAS, and FAQ.

Incremental and head-to-head metric differences were evaluated using paired 95% percentile-bootstrap confidence intervals with 9,999 subject-level resamples of the out-of-fold predictions. A confidence interval excluding zero was taken as evidence of a difference. Sensitivity to QC threshold was tested by re-running the full pipeline at four SurfaceHoles levels: no exclusion (n=688), 97.5th percentile (approximately 79 defects), 95th percentile (approximately 59), and the primary median plus 3 scaled MADs threshold (approximately 49; n=634). At each level, data were reloaded, the threshold applied, the CN-only brain-age model retrained, gap representations recomputed, and the evaluation harness re-run with identical fold seeds and bootstrap settings.

4. Results

4.1 Brain-age Model Performance

Under the CN-only out-of-fold scheme, the RidgeCV brain-age model achieved MAE 3.58 years and R^2 0.459 on CN holdout participants. This improved on the exploratory fixed-alpha Ridge model (α=1.0), which achieved MAE 4.11 years and R^2 0.358. The tuned alpha was 1,000 in every fold, indicating that alpha 1.0 was under-regularised in this high-dimensional setting. RidgeCV was therefore used for the primary analyses, and the out-of-fold CN prediction pattern is shown in Figure 1.

The mean out-of-fold CN gap was approximately +0.09 years, indicating that the CN reference group was centred close to zero. Elevated gaps in MCI or AD are therefore unlikely to be explained by a systematic offset in the CN reference model. The demographics-only baseline showed near-chance diagnostic discrimination, with AUC 0.535 on CN versus MCI+AD, and negative R^2 on MMSE. Age, sex, and education carried little diagnostic or cognitive signal in this cohort, making simple chronological-age imbalance an unlikely explanation for the subsequent brain-age results. The scalar gap distributions by diagnosis are shown in Figure 2.

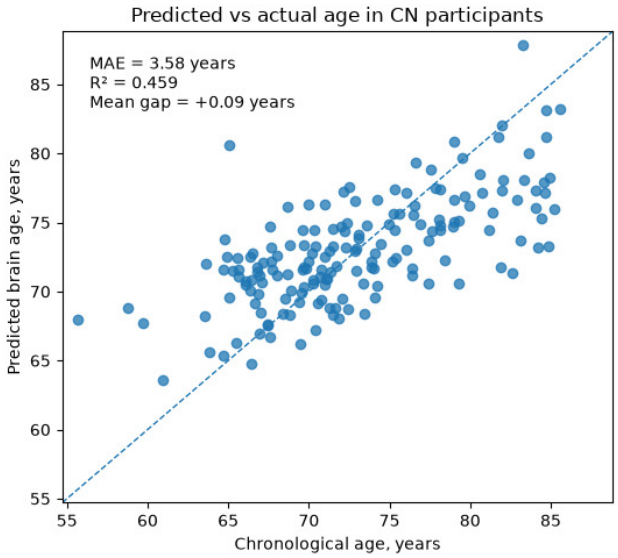

Figure 1: Predicted brain age versus actual age in cognitively normal participants. Scatter plot of out-of-fold RidgeCV predicted brain age against chronological age for cognitively normal participants (n=171). The diagonal reference line represents perfect prediction. The model achieved MAE 3.58 years and R^2 0.459, with mean CN gap approximately +0.09 years

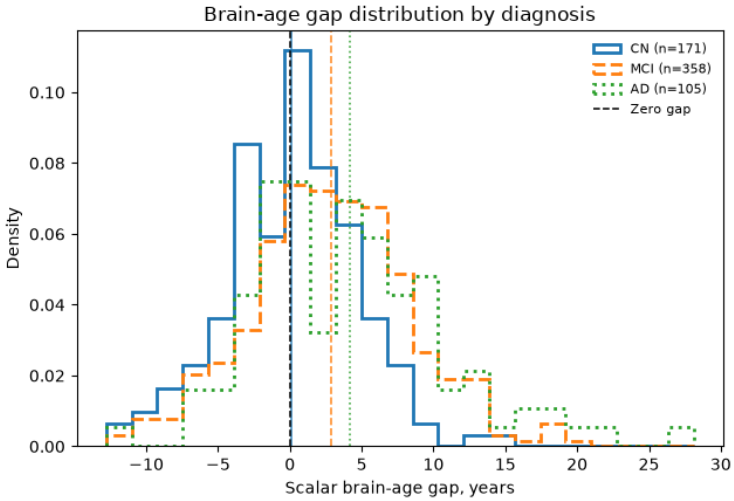

Figure 2: Brain-age gap distribution by diagnosis group. Histogram of scalar brain-age gap scores across CN (n=171), MCI (n=358), and AD (n=105) participants. The distribution shifts rightward from CN, centred near zero, through MCI to AD, consistent with increased apparent brain ageing across the disease spectrum. The vertical dashed line marks zero gap

4.2 Regional Versus Scalar Brain-age gap

All three evaluated representations added value over demographics on the primary diagnostic contrast. The scalar gap achieved AUC 0.692, corresponding to ΔAUC +0.157 with 95% CI [+0.102, +0.211]. The regional gap achieved AUC 0.766, corresponding to ΔAUC +0.232 with 95% CI [+0.177, +0.286]. Raw anatomical features achieved AUC 0.740, corresponding to ΔAUC +0.205 with 95% CI [+0.140, +0.269]. On CN versus AD, discrimination was higher for all three representations: scalar AUC 0.842, regional AUC 0.935, and raw-feature AUC 0.921. The paired comparison directly tested whether preserving regional brain-age information improved performance relative to the scalar gap. On the primary diagnosis contrast, the regional gap exceeded the scalar gap by +0.075 AUC units, with 95% CI [+0.042, +0.109]. The same pattern held across cognitive outcomes. Regional-gap incremental R^2 exceeded scalar-gap incremental R^2 for MMSE (0.309 versus 0.216), CDRSB (0.320 versus 0.232), ADAS (0.292 versus 0.171), and FAQ (0.195 versus 0.147). The regional-minus-scalar difference

was smallest for FAQ (+0.048 incremental R^2), while all four cognitive outcomes numerically favoured the regional gap.

Comparison with the simple structural comparators qualified this finding. The bilateral hippocampal volume comparator, added to demographics, achieved AUC 0.727 on CN versus MCI+AD. The scalar gap did not outperform the hippocampal comparator on any diagnostic or cognitive outcome. The regional gap showed a narrower advantage. On CN versus MCI+AD, the paired confidence interval for the regional-minus-hippocampal AUC difference included zero. On CN versus AD, the regional gap achieved AUC 0.935 compared with 0.881 for the hippocampal comparator, with a paired difference +0.053 and 95% CI [+0.015, +0.094]. The regional gap therefore consistently outperformed the scalar representation, whereas evidence that it outperformed the hippocampal comparator was limited to CN versus AD.

4.3 Bias-correction Sensitivity

In the exploratory Ridge (α=1.0) analysis, the three bias-correction variants produced nearly identical gap distributions. Across variants, the CN mean gap was 0.099 years, and the standard deviation differed by less than 0.01 years between raw and corrected gaps. The MCI+AD mean gap was 3.127 years under the raw variant, 3.125 under linear correction, and 3.125 under non-linear correction. The corrections therefore had negligible effects on the group-level gap means and standard deviations. Linear and quadratic correction therefore changed the observed gap distributions only minimally in this exploratory analysis. This sensitivity analysis concerned the gap distributions themselves; downstream diagnostic, cognitive, and calibration models were not independently re-evaluated under each correction variant. The finding is therefore limited to the stability of the gap distributions in this cohort and modelling design.

4.4 Calibration

Calibration assesses whether predicted probabilities are trustworthy, not merely whether cases are ranked correctly. The demographics-only baseline produced a Brier score of 0.2507 on CN versus MCI+AD. Scores were lower for the scalar gap (0.2227), regional gap (0.1924), and raw features (0.1865). Hippocampal volume produced 0.2119, and normalised grey matter volume 0.2377. Regional and raw representations produced the lowest Brier scores, roughly a quarter lower than the demographics baseline and numerically below the hippocampal comparator. These differences were not subjected to paired significance testing. Calibration curves for the primary contrast are shown in Figure 3.

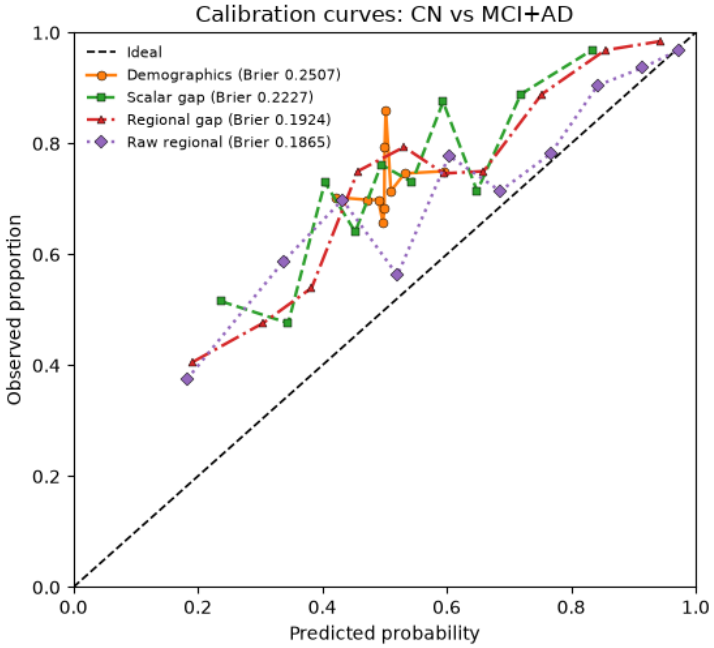

Figure 3: Calibration curves for the CN versus MCI+AD diagnostic model. Calibration curves show predicted probability against observed positive-class proportion for the demographics baseline and the three main representations. Adding the scalar gap improved calibration over the demographics baseline, while the regional gap and raw features produced the lowest Brier scores and curves closest to the ideal diagonal

4.5 Quality-control Sensitivity

The sensitivity analysis showed that the regional advantage was not an artefact of the SurfaceHoles threshold. Re-running the entire pipeline at no exclusion, 97.5th percentile, 95th percentile, and median plus 3 scaled MADs reproduced the same qualitative result. The regional-minus-scalar AUC difference on CN versus MCI+AD excluded zero at every threshold: +0.067 with no QC, +0.053 at 97.5th percentile, +0.078 at 95th percentile, and +0.075 at the primary threshold. On cognitive outcomes, regional-gap ΔR^2 exceeded scalar-gap ΔR^2 on MMSE and CDRSB at all four thresholds.

The hippocampal comparison remained specific. On CN versus MCI+AD, the regional gap's advantage over hippocampal volume stayed within sampling error at all thresholds. On CN versus AD, the regional gap exceeded hippocampal volume at every threshold, with paired AUC differences of +0.049 with no QC, +0.056 at 97.5th percentile, +0.045 at 95th percentile, and +0.053 at the primary threshold. Thus, the regional-versus-scalar advantage was QC-robust, while superiority over the hippocampal comparator remained specific to CN versus AD.

5. Discussion

The central finding is that representation affects how much clinically useful signal the brain-age gap retains. The per-region gap significantly outperformed the scalar gap across all diagnostic and cognitive outcomes tested, while the scalar gap failed to exceed hippocampal volume on any outcome. Scalar compression discards spatial information relevant to diagnosis and cognition, yielding an interpretable but weaker marker.

This result reframes recent sceptical evaluations. The present study provides a mechanism for their findings: the scalar gap is weak because aggregation discards anatomical information, not because brain-age modelling lacks disease signal. This is consistent with Dörfel et al. (2026), who found brain-age packages added little beyond grey-matter volume; Tan et al. (2025), who found latent representations outperformed the scalar gap; and Garcia Condado et al. (2025), who found tension between age-prediction and disease-discrimination objectives. The regional gap preserves the conceptual framing of deviation from age expectation while retaining the spatial structure that scalarisation removes.

The results also address whether brain-age signal merely recaptures chronological age. In this cohort, age, sex, and education produced near-chance diagnostic discrimination and poor cognitive prediction. MCI participants were not older than CN participants, and AD participants were only marginally older. The observed gap signal therefore reflects deviation from a CN-trained structural ageing trajectory rather than merely chronological age masquerading as a biomarker.

The hippocampal comparison is deliberately qualified. The scalar gap did not beat hippocampal volume, aligning with prior scepticism toward scalar brain-age biomarkers. Only CN versus AD showed a supported regional advantage over hippocampal volume. On CN versus MCI+AD, which includes a heterogeneous MCI group, the advantage remained within sampling error. The correct claim is therefore not that regional brain age supersedes established structural markers, but that regionalising the gap recovers information that scalar compression loses and can outperform hippocampal volume when diagnostic separation is strong.

Several limitations remain. First, quality control used FastSurfer SurfaceHoles as a reconstruction-quality proxy rather than manual raw-image QC. This is defensible because no harmonised raw-image QC label was available, but it cannot guarantee removal of motion artefacts or acquisition noise. AD participants were excluded at a higher rate than CN or MCI participants, which may introduce differential selection and limit generalisability.

Second, the analysis was cross-sectional and used baseline scans only. Longitudinal regional-gap trajectories may be more sensitive to disease progression. Third, all data came from ADNI, an AD-enriched research cohort that is not fully population-representative. External validation in cohorts such as OASIS or UK Biobank is required before claiming generalisability. Fourth, representations were not feature-matched: the scalar gap has one feature, the regional gap 171, and the raw representation 1,379. Out-of-fold evaluation reduces overfitting risk, and raw features did not consistently outperform the regional gap, but the asymmetry remains a limitation. Fifth, ADAS and FAQ were available only for subsets of participants. Finally, the two-stage cross-validation design computed gap features out-of-fold before downstream cross-validation. Each gap encodes only the subject's own deviation from the CN-trained model and contains no diagnostic-label information, so this does not create label leakage, but a fully nested single-stage pipeline would be stronger.

The next steps are external validation, longitudinal modelling of regional gap trajectories, disease-aware finetuning that preserves regional interpretability, and incorporation of raw-image QC where original scans are available.

6. Conclusion

This study tested whether representation choice affects the diagnostic and cognitive utility of the brain-age gap in Alzheimer's disease. Using ADNI baseline structural MRI processed with FastSurfer, we compared a scalar whole-brain gap, a 171-region gap vector, and raw regional features within a leak-safe out-of-fold pipeline. The per-region gap outperformed the scalar gap across all six primary outcomes; QC sensitivity analyses reproduced the regional advantage on the diagnostic contrasts and on MMSE and CDRSB. The scalar gap did not outperform hippocampal volume on any outcome. Only CN versus AD showed a supported regional advantage over hippocampal volume, with a modest but stable margin.

Exploratory bias correction had negligible impact on the observed gap distributions. All three representations had lower Brier scores than demographics, with regional and raw representations lowest. These findings support a precise conclusion: prior scepticism toward scalar brain-age-gap biomarkers is justified, but it should not be generalised to the brain-age framework. The limitation lies in scalar compression of anatomically heterogeneous disease signal. Preserving regional structure retains the interpretability of age-deviation modelling while recovering clinically relevant information that a single whole-brain score discards.

Ethics Declaration: Ethical clearance for the original collection of imaging and clinical data was obtained by the Alzheimer's Disease Neuroimaging Initiative (ADNI) through the relevant institutional review boards at participating sites, and informed consent was obtained from participants by ADNI. The present study used de-identified secondary ADNI data. Access to ADNI data was obtained through the ADNI/LONI data-access approval process, and the analysis was conducted in accordance with the applicable ADNI data-use conditions. No new human participants were recruited, and no new primary data were collected for this study.

AI Declaration: AI tools were used during the preparation of this paper to support outlining, language refinement, editing for clarity, and reduction of word count. AI tools were not used to generate the empirical results, alter the data, run the statistical analyses, or make independent scientific interpretations. All AI-assisted text was reviewed, edited, and verified by the authors, who take full responsibility for the accuracy, integrity, and final content of the submitted paper.

Data availability Statement: Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and are available to approved users through the ADNI/LONI data-access process. Raw ADNI imaging and clinical data are not redistributed with this manuscript. Analysis scripts will be made available separately; restricted ADNI data and participant-level files will not be redistributed.

References

Doering, E., Antonopoulos, G., Hoenig, M., van Eimeren, T., Daamen, M., Boecker, H., Jessen, F., Düzel, E., Eickhoff, S., Patil, K., Drzezga, A., Alzheimer's Disease Neuroimaging Initiative and DELCODE Study Group (2024) 'MRI or 18F-FDG PET for brain age gap estimation: links to cognition, pathology, and Alzheimer disease progression', *Journal of Nuclear Medicine*, 65(1), pp. 147–155. doi: 10.2967/jnumed.123.265931.

Dörfel, R.P., Ozenne, B., Ganz, M., Svensson, J.E., Plavén-Sigray, P. and Alzheimer's Disease Neuroimaging Initiative (ADNI) (2026) 'Prediction of brain age using structural magnetic resonance imaging: a comparison of clinical utility of publicly available software packages', *EBioMedicine*, 123, 106094. doi: 10.1016/j.ebiom.2025.106094.

Dular, L., Špiclin, Ž., Alzheimer's Disease Neuroimaging Initiative and Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing (2024) 'Analysis of brain age gap across subject cohorts and prediction model architectures', *Biomedicines*, 12(9), 2139. doi: 10.3390/biomedicines12092139.

Garcia Condado, J., Verdugo Recuero, I., Tellaetxe-Elorriaga, I., Birkenbihl, C., Carrigan, M., Diez, I., Buckley, R.F., Erramuzpe, A. and Cortes, J.M. (2025) 'Aging as an active player in Alzheimer's disease classification: insights from feature selection in BrainAge models', *NeuroImage*, 322, 121548. doi: 10.1016/j.neuroimage.2025.121548.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H. and Alzheimer's Disease Neuroimaging Initiative (2013) 'BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease', *PLoS ONE*, 8(6), e67346. doi: 10.1371/journal.pone.0067346.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B. and Reuter, M. (2020) 'FastSurfer: a fast and accurate deep learning based neuroimaging pipeline', *NeuroImage*, 219, 117012. doi: 10.1016/j.neuroimage.2020.117012.

Hwang, G., Abdulkadir, A., Erus, G., Habes, M., Pomponio, R., Shou, H., Doshi, J., Mamourian, E., Rashid, T., Bilgel, M., Fan, Y., Sotiras, A., Srinivasan, D., Morris, J.C., Albert, M.S., Bryan, N.R., Resnick, S.M., Nasrallah, I.M., Davatzikos, C., Wolk, D.A., iSTAGING Consortium and Alzheimer's Disease Neuroimaging Initiative (ADNI) (2022) 'Disentangling Alzheimer's

disease neurodegeneration from typical brain ageing using machine learning', *Brain Communications*, 4(3), fcac117. doi: 10.1093/braincomms/fcac117.

Kalc, P., Dahnke, R., Hoffstaedter, F., Gaser, C. and Alzheimer's Disease Neuroimaging Initiative (2024) 'BrainAGE: revisited and reframed machine learning workflow', *Human Brain Mapping*, 45(3), e26632. doi: 10.1002/hbm.26632.

Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L. (2013) 'Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median', Journal of Experimental Social Psychology, 49(4), pp. 764-766. doi: 10.1016/j.jesp.2013.03.013

Löwe, L.C., Gaser, C., Franke, K. and Alzheimer's Disease Neuroimaging Initiative (2016) 'The effect of the APOE genotype on individual BrainAGE in normal aging, mild cognitive impairment, and Alzheimer's disease', *PLoS ONE*, 11(7), e0157514. doi: 10.1371/journal.pone.0157514.

Ly, M., Yu, G., Son, S.J., Pascoal, T., Karim, H.T. and Alzheimer's Disease Neuroimaging Initiative (2024) 'Longitudinal accelerated brain age in mild cognitive impairment and Alzheimer's disease', *Frontiers in Aging Neuroscience*, 16, 1433426. doi: 10.3389/fnagi.2024.1433426.

Millar, P.R., Gordon, B.A., Luckett, P.H., Benzinger, T.L.S., Cruchaga, C., Fagan, A.M., Hassenstab, J.J., Perrin, R.J., Schindler, S.E., Allegri, R.F., Day, G.S., Farlow, M.R., Mori, H., Nübling, G., The Dominantly Inherited Alzheimer Network, Bateman, R.J., Morris, J.C. and Ances, B.M. (2023) 'Multimodal brain age estimates relate to Alzheimer disease biomarkers and cognition in early stages: a cross-sectional observational study', *eLife*, 12, e81869. doi: 10.7554/eLife.81869.

Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C.R. Jr, Jagust, W.J., Shaw, L.M., Toga, A.W., Trojanowski, J.Q. and Weiner, M.W. (2010) 'Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization', *Neurology*, 74(3), pp. 201–209. doi: 10.1212/WNL.0b013e3181cb3e25.

Rosen, A.F.G., Roalf, D.R., Ruparel, K., Blake, J., Seelaus, K., Villa, L.P., Ciric, R., Cook, P.A., Davatzikos, C., Elliott, M.A., Garcia de La Garza, A., Gennatas, E.D., Quarmley, M., Schmitt, J.E., Shinohara, R.T., Tisdall, M.D., Craddock, R.C., Gur, R.E., Gur, R.C. and Satterthwaite, T.D. (2018) 'Quantitative assessment of structural image quality', NeuroImage, 169, pp. 407-418. doi: 10.1016/j.neuroimage.2017.12.059

Tan, T.W.K., Nguyen, K.N., Zhang, C., Kong, R., Cheng, S.F., Ji, F., Chong, J.S.X., Chong, E.J.Y., Venketasubramanian, N., Orban, C., Chee, M.W.L., Chen, C., Zhou, J.H., Yeo, B.T.T., Alzheimer's Disease Neuroimaging Initiative and Australian Imaging Biomarkers and Lifestyle Study of Aging (2025) 'Mind the gap: does brain age improve Alzheimer's disease prediction?', *Human Brain Mapping*, 46(12), e70276. doi: 10.1002/hbm.70276.

Taylor, A., Zhang, F., Niu, X., Heywood, A., Stocks, J., Feng, G., Popuri, K., Beg, M.F. and Wang, L. (2022) 'Investigating the temporal pattern of neuroimaging-based brain age estimation as a biomarker for Alzheimer's disease related neurodegeneration', *NeuroImage*, 263, 119621. doi: 10.1016/j.neuroimage.2022.119621.