



Improving the fairness and robustness of deep learning classification models through bias mitigation and post hoc calibration

Majid Ramedani¹ · Jaya C. Terli¹ · Ole H. Martensen² · Tobias Strauß² · Katharina Buerger^{3,4} · Peter Dechent⁵ · Emrah Düzel^{6,7} · Klaus Fliessbach^{8,9} · Maria Gemenetzi^{10,11} · Wenzel Glanz⁶ · Julian Hellmann-Regen^{11,12,13} · Stefan Hetzer¹⁴ · Enise I. Incesoy^{6,7,15} · Daniel Janowitz^{3,4} · Marie T. Kronmüller⁹ · Christoph Laske^{16,17} · Falk Lüsebrink⁶ · Robert Perneczky^{3,18,19,20} · Oliver Peters^{11,12,13} · Lukas Preis¹² · Josef Priller^{10,11,21,22,23} · Boris-Stephan Rauchmann^{18,24,25} · Ayda Rostamzadeh²⁶ · Anja Schneider^{8,9} · Björn H. Schott^{27,28,29} · Sebastian Sodenkamp^{16,30} · Annika Spottke^{9,31} · Eike J. Spruth^{10,11} · Sophia Stöcklein^{3,32} · Jens Wiltfang^{27,28,33} · Frank Jessen^{9,26,34} · Stefan J. Teipel^{1,35} · Martin Dyrba¹ 

Received: 1 April 2026 / Accepted: 25 September 2026
© The Author(s) 2026

Abstract

Purpose: The goal of this study was to enhance the fairness and clinical applicability of deep learning models for classifying Alzheimer's disease from structural MRI scans. To achieve this, we addressed two key challenges: age-related bias and unreliable probability estimates. Specifically, we investigated in-processing debiasing strategies to achieve age-invariant predictions while preserving diagnostic accuracy. We also systematically evaluated post hoc calibration methods to improve the reliability, interpretability, and clinical usefulness of predicted probabilities across diverse patient populations.

Methods: Structural MRI data from three independent cohorts (ADNI, AIBL, and DELCODE) were preprocessed using a standardized pipeline. Age-related bias was addressed using two in-processing strategies, adversarial training (CF-Net) and GLM-based feature residualization (PMDN), implemented in a shared 3D convolutional neural network (CNN) framework. Additionally, post hoc calibration was performed using four regression methods: logistic, probit, isotonic, and Bayesian binning. We evaluated model performance, fairness, and calibration using classification as well as calibration-specific metrics.

Results: PMDN achieved the best overall performance, with higher balanced accuracy (0.71) and improved calibration (Brier ≈ 0.22), showing substantially reduced variability compared to baseline and CF-Net. PMDN effectively removed age-related bias, achieving near-zero age prediction correlation; CF-Net, however, showed only partial and unstable debiasing. Post hoc calibration further improved probability reliability, with nonparametric methods (isotonic regression and Bayesian binning) yielding the lowest calibration errors (MCE ≈ 0.21 – 0.24).

Conclusion: This study shows that, although deep learning models achieve similar classification performance, they differ significantly in terms of calibration, stability, and fairness. PMDN provides more consistent and effective age bias mitigation than the CF-Net approach, particularly in clinically relevant subgroups. Post hoc calibration improves probability reliability further, with nonparametric methods achieving the best calibration.

Keywords Magnetic resonance imaging · Deep learning · Fairness · Bias detection · Calibration

Introduction

Artificial intelligence (AI) represents a major advancement in modern technology and has become increasingly influential across many aspects of society [1]. Within this domain, deep learning (DL), a subfield of AI inspired by the data-

processing mechanisms of the human brain has emerged as one of the most influential approaches for medical data analysis [2]. Deep learning models can automatically extract complex patterns from large and high-dimensional datasets, demonstrating potential for improving medical image interpretation and disease detection.

Neurodegenerative diseases, particularly Alzheimer's disease (AD), are a leading cause of disability worldwide and are associated with aging [3]. AD primarily manifests as

These authors contributed equally to this work

Extended author information available on the last page of the article

gradual memory loss, cognitive decline, and behavior, which can severely affect daily and social functioning. Advances in magnetic resonance imaging (MRI) have revolutionized the study of the brain by capturing high-resolution images of both its structure and function [4]. MRI provides insights into changes in brain morphology, and this ability makes it an essential tool for identifying early signs of AD [5]. In recent years, DL models that analyze brain MRI images have achieved significant progress in predicting and diagnosing AD, with predictive accuracy that in some cases surpasses expert clinicians [6].

As deep learning models become increasingly influential, concerns regarding their trustworthiness and reliability have also gained growing attention among radiologists and healthcare professionals [7]. Although these models often achieve high predictive accuracy, several factors limit their safe deployment. In particular, the trustworthiness of DL-based clinical decision support systems is threatened by two key issues.

1. Models trained on real-world medical datasets may inadvertently learn demographic confounders, and it leads to unintended biased predictions across patient subgroups [8]. Such bias can compromise the accuracy of predictions and result in unfair outcomes. These bias and fairness challenges are especially important in automatic classification of AD. Aging is the most significant risk factor for AD, since the process of healthy brain aging naturally leads to structural changes, including gray matter volume reduction, which substantially overlap with AD pathology in many brain regions. This structural overlap means that a disease classifier may inadvertently rely on normal aging patterns rather than disease-specific atrophy, as evidenced by common brain age prediction models that typically estimate older ages for patients with neurodegenerative disease [9].
2. Another challenge concerns the reliability of the probability estimates produced by DL classifiers. In classification tasks, the softmax layer in deep learning models, especially in convolutional neural networks (CNNs), transforms raw output scores into a probability distribution over all classes [10]. Notably, the softmax layer's disease likelihood scores indicate the model's relative confidence instead of true statistical probabilities and lack confidence intervals [11]. Since pretrained softmax outputs often fail to represent actual likelihoods, probability calibration is used to adjust predictions so they align more closely with observed frequencies in an independent calibration dataset.

To address these challenges, this study investigated two complementary strategies for improving the trustworthiness and robustness of DL models in AD classification. First, we

designed and implemented two methodologies, to mitigate age-related confounding in MRI-based features. Second, we evaluated four post hoc probability calibration methods, two parametric and two nonparametric and investigated which calibration method provides the most practical utility for empirical clinical data.

Materials and methods

MRI dataset

In our studies, we employed whole-brain MRI data from different sources. Imaging and clinical data were obtained from three independent cohorts: ADNI2/GO subset of Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>), Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) [12], and German Center for Neurodegenerative Diseases (DZNE) Multicenter Observational Study on Longitudinal Cognitive Impairment and Dementia (DELCODE) [13]. Detailed demographic and MRI acquisition information is summarized in Table 1. The MRI inputs to the classification models consisted of whole-brain, voxel-wise gray matter probability maps derived from the preprocessed T1-weighted MRI scans. These maps were spatially normalized to MNI space and used as input to the networks at the original MNI voxel resolution. Thus, the models operated directly on whole-brain voxel-wise gray matter information rather than on predefined regional volumetric measures or manually extracted regional features. All MRI scans from the three cohorts were processed using the same automated preprocessing pipeline to ensure a consistent processing procedure across datasets. All three studies implemented harmonized MRI acquisition procedures to reduce site-related technical variation prior to preprocessing [14, 15]. Within the modeling framework, age was the only confounding variable explicitly addressed. Notably, for model training, we included mild cognitive impairment (MCI) subjects in the AD group, and subsequently assessed model performances separately or in combination with AD depending on the evaluation study design.

Bias mitigation

To investigate bias mitigation, three classification models were developed using a shared CNN feature extractor architecture to ensure a fair comparison across different debiasing strategies.

CNN architecture

The baseline model is a 3D CNN, illustrated in Fig. 1a, consisting of four convolutional blocks with $3 \times 3 \times 3$ kernels,

Table 1 Demographic and MRI acquisition details of participants across different cohorts

Cohort	Training ADNI2/GO	Evaluation AIBL	DELCODE
N (AD&MCI/CN)	648 (253/395)	606 (158/448)	466 (256/210)
Magnetic field (T)	3	3	3
TR/TE (ms)	7.2/3	2300/2.98	2500/4.37
Voxel size (mm ³)	1 × 1 × 1.2	1 × 1 × 1.2	1 × 1 × 1
Age (y)	74.9 ± 7.6	72.8 ± 6.4	71.4 ± 6.2

Age values are reported as mean ± standard deviation (mean ± std)

batch normalization, ReLU activation, and 2×2×2 max-pooling. A bottleneck module with two 1×1×1 convolutions and a 3×3×3 average pooling layer compresses features into a 4096-dimensional vector (64 channels, 4×4×4 spatial resolution). The fully connected classifier uses dropout (0.5 and 0.3) and L2 regularization (0.02). No bias mitigation is applied, allowing it to serve as a reference baseline for fair comparison. Training converged within 175–235 epochs per cross-validation fold.

CF-Net

The adversarial confounder-free network (CF-Net) is an in-model debiasing approach that learns feature representations invariant to confounding variables through adversarial training. As illustrated in Fig. 1b, the architecture consists of three components: a feature extractor (FE), a classifier (P), and a confounder predictor (CP). The FE generates a features vector, which is used by the classifier to predict the target label Y , while the CP attempts to predict the confounder C from the same features [16].

The model is trained using a min–max objective:

$$\min_{\theta_{fe}, \theta_c} \max_{\theta_{cp}} [L_c(\mathbf{X}, Y) - \lambda L_{cp}(\mathbf{X}^p, C^p)] \quad (1)$$

where L_c is the classification loss and L_{cp} is the adversarial confounder loss. To enforce statistical independence between the learned features and confounders, the adversarial loss is defined using the squared Pearson correlation, $L_{cp} = \sum_{k=1}^k \text{corr}^2(C_k, \hat{C}_k)$. Minimizing this loss drives the correlation toward zero, encouraging mean independence between features and confounders.

Training proceeds in three phases. In Phase 1, the shared CNN encoder and classifier are trained jointly to establish robust features vector. In Phase 2, the CP is activated and updated while the encoder and classifier remain frozen. In Phase 3, two S3 switches are engaged, enabling a final joint training loop where the encoder and CP interact to refine vector. The classifier (with dropout 0.5/0.3) predicts diagnosis, and the CP (tanh activations) predicts age. Learning rates follow an 8:1 ratio ($\eta_{FE/C} = 1 \times 10^{-4}$, $\eta_{CP} = 8 \times 10^{-4}$) to

balance the capacities of the classifier (6.3M parameters) and CP (0.8M parameters). Training converged within 245–276 epochs per cross-validation fold, with regressor loss curves monitored to ensure stable adversarial equilibrium.

PMDN

The penalty-based metadata normalization (PMDN) is an in-model strategy that achieves invariant feature representations by residualizing confounder effects directly in the network's feature space. PMDN treats the confounding effect as a general linear model (GLM), modeling feature vectors as $\mathbf{f} - \mathbf{M}\boldsymbol{\beta} = \mathbf{r}$, where $\mathbf{M}$ contains metadata, $\boldsymbol{\beta}$ are coefficients, and $\mathbf{r}$ is the residual, confounder-free feature vector (Fig. 1c) [17].

In the original MDN formulation, $\boldsymbol{\beta}$ is computed analytically via least-squares over the entire training set [18]. However, in deep learning, optimization is performed batch-wise, which causes the $\boldsymbol{\beta}$ estimates to vary across batches and introduces instability. PMDN addresses this by formulating a differentiable, bilevel optimization problem that jointly updates network weights $\mathbf{W}$ and deconfounding parameters $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}, \mathbf{W}} L(\psi(\varphi(\mathbf{X}) - \mathbf{M}\boldsymbol{\beta}), Y) + \lambda L^*(\varphi(\mathbf{X}); \mathbf{M}), \quad (2)$$

where L is the classification loss, L^* penalizes feature correlation with metadata, and λ balances the trade-off. Alternating updates of $\mathbf{W}$ and $\boldsymbol{\beta}$ ensure stable convergence of $\boldsymbol{\beta}$ while learning task-relevant, residualized features.

At inference, only confounding components are subtracted. PMDN is applied to the 4,096-dimensional encoder features, with $\boldsymbol{\beta}$ as a 3×4,096 matrix regressing features onto metadata. Learning rates are set to $\eta_{\mathbf{W}} = 5 \times 10^{-5}$, $\eta_{\boldsymbol{\beta}} = 2.5 \times 10^{-3}$, with a penalty weight $\lambda = 0.9$.

Post hoc calibration

For calibration, we used a CNN baseline model from section [CNN architecture](#). The output layer employed a softmax activation function to convert class scores into likelihood

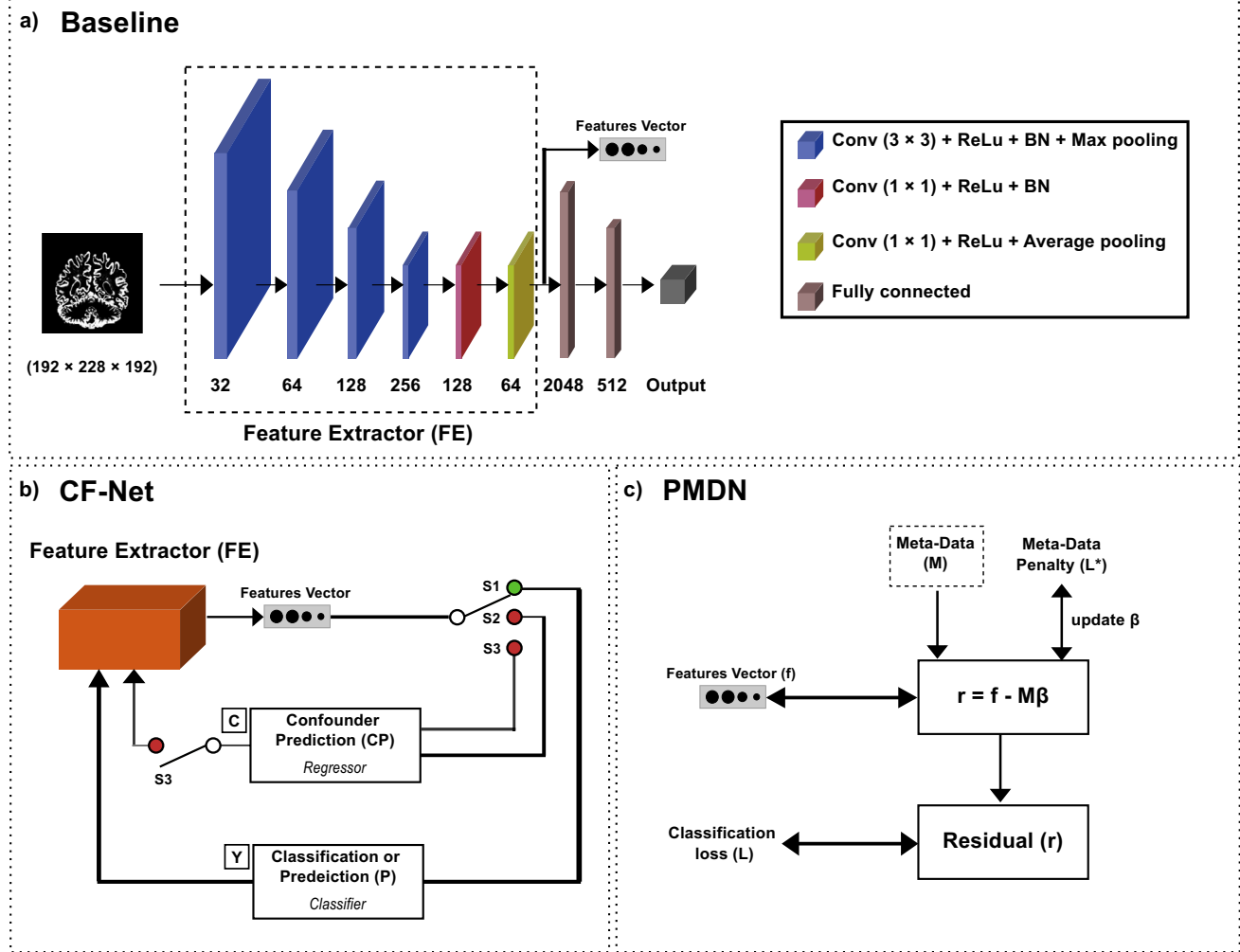


Fig. 1 Overview of model architectures. **a** Baseline: CNN extracts features from brain images. **b** CF-Net: Training occurs in three phases. Phase I is dedicated to prediction module, Phase 2 activates the con-

founder prediction module while freezing others, and Phase 3 enables interaction via S3 switches for final joint training. **c** PMDN: Metadata is residualized from features with a penalty

probabilities. To improve probability estimates, established calibration methods were applied to the CNN outputs [19].

Parametric calibration methods

We employed two parametric calibration methods: logistic regression and probit regression. These methods transform the raw model output scores X into well-calibrated probability estimates of the true labels T .

Logistic regression applies Bayes' theorem to estimate the posterior probability $\Pr(T = 1 | X = x) = \sigma(a)$, where σ is the logistic function, and a depends on the class-conditional probabilities $\Pr(X = x | T = t)$ for $t \in \{0, 1\}$. Assuming Gaussian class-conditional distributions $X | T = t \sim \mathcal{N}(\mu_t, \Sigma)$, the transformation reduces to an affine function $a = wx + b$.

Probit regression is similar but uses the inverse normal link function instead of the logit, $a = \Phi^{-1}(p)$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution.

Nonparametric calibration methods

We used two nonparametric calibration methods: Bayesian binning into quantiles (BBQ) and isotonic regression. These methods do not assume a specific parametric form and rely on the data to estimate well-calibrated probabilities.

BBQ divides the model outputs X into B bins based on empirical quantiles, ensuring that bins are evenly populated and avoiding empty or sparse bins. Within each bin, probabilities are estimated using a Bayesian approach based on the beta distribution. The final calibration function is obtained by

averaging models with varying numbers of bins, with B constrained between a minimum and maximum value to ensure stability. BBQ is entirely data-driven and may produce non-monotonic calibration functions.

Isotonic regression fits a nondecreasing, piecewise-constant function to map raw scores to calibrated probabilities, minimizing the mean squared error (MSE) between predicted and true labels. Using the pool-adjacent-violators algorithm (PAVA), it achieves an optimal $O(n)$ solution.

Evaluation criteria

To evaluate the performance of the proposed strategies, we computed the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Based on these quantities, several evaluation metrics (sensitivity, specificity, accuracy, F1-score, Brier score, ECE, and MCE) were calculated to assess both classification performance and probability calibration [20]. In addition to these metrics, we evaluated the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC).

Brier Score: The Brier score measures the accuracy of probabilistic predictions and reflects both classification and calibration performance. It is defined as the mean squared difference between predicted probabilities and the true class labels, where lower values indicate better calibration.

ECE & MCE: For the post hoc calibration experiments, we further considered calibration-specific metrics. To isolate calibration quality from overall classification accuracy, we also computed the expected calibration error (ECE) and the maximum calibration error (MCE). These metrics measure, respectively, the mean and maximum absolute deviation between predicted confidence and observed accuracy across m quantile bins of the predicted probabilities.

Results

Classification performance

After training three models on ADNI data to evaluate bias mitigation methods, we applied them to an external cohort (AIBL + DELCODE), as described in Section [MRI dataset](#). Table 2 presents diagnostic performance aggregated across five cross-validation folds, reported as mean $\pm$ standard deviation for each model on the test dataset.

PMDN achieved the highest balanced accuracy (0.71 ± 0.01), outperforming CF-Net (0.70 ± 0.06) and the baseline model (0.66 ± 0.11). Notably, PMDN exhibited remarkably low variability ($SD < 0.02$), approximately ten times lower than the baseline ($SD \approx 0.11$) and five times lower than CF-Net ($SD \approx 0.06$), suggesting more consistent learning dynamics across folds. PMDN showed better probability

calibration across the test dataset, achieving the lowest Brier scores (0.22 ± 0.01), whereas the baseline model exhibited worse calibration (0.32 ± 0.16). AUC values were comparable across all models, with CF-Net and PMDN performing similarly on the test data (0.82 ± 0.02 and 0.82 ± 0.01 , respectively).

To evaluate the statistical significance of differences between models at the individual subject level, paired t -tests were performed on prediction errors ($|\text{true_label} - \text{predicted_probability}|$). CF-Net achieved significantly lower prediction error than the baseline model (mean difference = 0.029 , $p = 0.01$), whereas PMDN showed significantly higher prediction error than the baseline across both test configurations (mean difference = -0.038 , $p < 0.001$).

Age bias quantification

To assess whether debiasing reduces age-dependent predictions, we examined the association between chronological age and predicted AD logits on the age-balanced test set. Linear regression was applied within each diagnostic group, and correlation coefficients (r) with p -values were computed. Near-zero correlations ($r \approx 0$, $p \gg 0.05$) indicate age-independent predictions, while significant positive correlations suggest the model retains residual age-related signals rather than disease-specific features.

Table 3 shows that the baseline model (CN: $r = 0.16 \pm 0.03$) exhibits a consistent positive association between age and predicted logits, indicating that, in the absence of debiasing, higher AD risk is systematically assigned to older cognitively normal individuals. CF-Net (CN: $r = 0.15 \pm 0.08$) demonstrates a comparable mean correlation to the baseline model, but with substantially greater variability, suggesting only partial and unstable mitigation of age-related bias through adversarial training. In contrast, PMDN reduces this association to near zero (CN: $r = -0.07 \pm 0.06$).

The MCI subgroup showed the clearest difference between debiasing strategies. The baseline model exhibited a substantial age logits correlation (MCI: $r = 0.19 \pm 0.07$), indicating persistent age dependence in this clinically critical category, where age-related atrophy and pathological changes overlap. CF-Net demonstrated an even stronger correlation (MCI: $r = 0.20 \pm 0.05$), suggesting that adversarial training was ineffective at mitigating age bias in this population. In contrast, PMDN achieved a near-zero correlation in the MCI subgroup (MCI: $r = 0.06 \pm 0.03$) reflecting successful removal of age-confounded variance.

In the AD subgroup, however, age-related associations persisted across all three models. The baseline model showed a positive correlation between age and predicted logits ($r = 0.15 \pm 0.07$), which remained essentially unchanged with CF-Net ($r = 0.17 \pm 0.07$) and PMDN ($r = 0.15 \pm 0.04$). Thus, unlike the CN and MCI groups, PMDN did not substantially

Table 2 Models performance on external dataset for CN versus AD&MCI

Model	bAcc	AUC	Sensitivity	Specificity	F1-score	Brier
Baseline	0.66±0.11	0.81±0.02	0.83±0.17	0.49±0.37	0.51±0.28	0.32±0.16
CF-Net	0.70±0.06	0.82±0.02	0.78±0.16	0.61±0.26	0.63±0.12	0.26±0.08
PMDN	0.71±0.01	0.82±0.01	0.84±0.02	0.58±0.02	0.68±0.01	0.22±0.01

Values are reported as mean ± standard deviation (mean ± std). bAcc stands for balanced accuracy
 Bold values indicate the best performance for each respective metric

Table 3 Age-related dependency of outputs from different CNN models on the external test set

Model	Group	<i>r</i>	<i>p</i> -range
Baseline	AD	0.15±0.07	0.004–0.546
	CN	0.16±0.03	< 0.001–0.015
	MCI	0.19±0.07	< 0.001–0.246
CF-Net	AD	0.17±0.07	0.001–0.167
	CN	0.15±0.08	< 0.001–0.389
	MCI	0.20±0.05	< 0.001–0.023
PMDN	AD	0.15±0.04	0.014–0.256
	CN	−0.07±0.06	0.008–0.727
	MCI	0.06±0.03	0.109–0.874

r: Pearson correlation. CN: Cognitively Normal; MCI: Mild Cognitive Impairment; AD: Alzheimer's Disease
 Bold values indicate substantial decorrelation

reduce the mean age-related dependency of predictions in the AD group. However, the lower standard deviation observed with PMDN indicates more consistent age-related behavior across the evaluation folds.

Figure 2 illustrates these relationships. Each subplot presents scatter points color-coded by diagnostic group (CN: black, MCI: gold, AD: red) with fitted regression lines and corresponding correlation statistics. PMDN demonstrates a nearly flat regression line for CN subjects and a minimal age gradient for MCI ($r = 0.06$), in contrast to the positive slopes observed for MCI subjects in baseline and CF-Net ($r = 0.19$, $r = 0.20$, respectively), where older individuals are assigned higher AD-risk logits regardless of true diagnosis, consistent with PMDN predictions being less driven by chronological age.

Post hoc calibration

Table 4 presents the performance of four calibration techniques applied to the test folds aggregated across five cross-validation folds, alongside the uncalibrated CNN outputs for reference. Overall classification performance, measured by accuracy and F1-score, remained largely stable across all methods (accuracy ≈ 0.77 , F1 ≈ 0.73).

Calibration-specific metrics highlight more substantial differences. The raw CNN outputs exhibited the largest calibration errors (ECE = 0.28, MCE = 0.64). Parametric

methods, including logistic and probit regression, produced moderate improvements (ECE ≈ 0.13 , MCE ≈ 0.29 , for both) with smooth, constrained calibration curves (Fig. 3b). Non-parametric approaches, specifically isotonic regression and the BBQ method, achieved the lowest calibration errors (ECE ≈ 0.07 –0.08, MCE ≈ 0.19 –0.20), and near-good alignment in calibration plots (Fig. 3a). Notably, isotonic also achieved the lowest log-loss (0.50) and Brier score (0.16), while maintaining the same F1-score (0.73), highlighting its benefit for reliable probability estimates.

Discussion

In this paper, we presented two approaches to enhance fairness and robustness of deep learning models for Alzheimer's disease through bias mitigation and post hoc calibration. Age-related bias was evaluated alongside the effectiveness of different models in adjusting predicted scores.

While all models achieved similar discriminative performance (comparable AUCs), notable differences emerged in calibration, stability, and sensitivity to age-related confounding. PMDN consistently achieved higher balanced accuracy and F1-scores, with lower variability across cross-validation folds and the most reliable probability estimates (lowest Brier scores). In contrast, the baseline model relied on chronological age, and CF-Net reduced age dependence only marginally with considerable variability. PMDN substantially attenuated age prediction correlations, particularly in CN and MCI groups, indicating that explicit residualization more effectively removes confound-related variance and allows the model to focus on disease-relevant features.

An important observation is that age-related prediction bias was not equally attenuated across diagnostic groups. While PMDN substantially reduced the association between age and predicted AD logits in CN and MCI subjects, age-related associations persisted in the AD group. This may reflect the greater overlap between age-related brain changes and disease-related neurodegeneration in individuals with AD, making it more difficult to disentangle age-related and disease-specific signals in the presence of established pathology. Notably, PMDN nevertheless showed lower variability in the age-logit association for the AD group compared with the baseline and CF-Net models, suggesting more consistent behavior across evaluation folds.

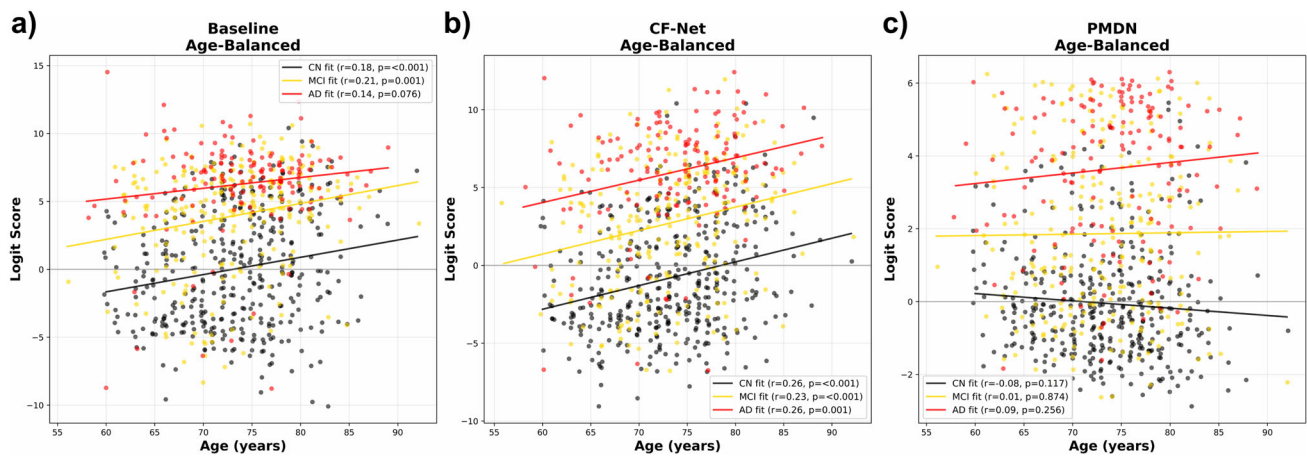


Fig. 2 Age versus predicted AD logits for median-fold models on the age-balanced test set. Panels show scatter plots for **a** Baseline, **b** CF-Net, and **c** PMDN. Points are color-coded by diagnostic group (CN: black, MCI: gold, AD: red), with fitted regression lines and corresponding correlation statistics

Table 4 Comparison of the calibration scores for the AIBL + DELCODE dataset

Model	Accuracy	F1-score	Log-Loss	Brier	ECE	MCE
Logit	0.77 ± 0.02	0.73 ± 0.03	0.54 ± 0.02	0.18 ± 0.01	0.13 ± 0.02	0.29 ± 0.05
Probit	0.77 ± 0.02	0.72 ± 0.03	0.54 ± 0.02	0.18 ± 0.01	0.13 ± 0.02	0.29 ± 0.05
Isotonic	0.77 ± 0.02	0.73 ± 0.03	0.50 ± 0.03	0.16 ± 0.01	0.08 ± 0.02	0.20 ± 0.05
BBQ	0.76 ± 0.03	0.71 ± 0.04	0.53 ± 0.04	0.17 ± 0.02	0.07 ± 0.01	0.19 ± 0.05
CNN	0.77 ± 0.02	0.73 ± 0.03	1.26 ± 0.08	0.28 ± 0.02	0.28 ± 0.03	0.64 ± 0.07

Values are reported as mean ± standard deviation (mean ± std)

Bold values indicate the best performance for each respective metric

The differences between PMDN and CF-Net stem from their underlying optimization strategies. PMDN applies a deterministic GLM-based correction to remove variance explained by confounders, producing stable and consistent feature representations. Adversarial training relies on a min-max optimization with competing objectives, which may lead to incomplete convergence and residual confounding. This instability explains CF-Net's variable performance and less effective bias mitigation. Clinically, these findings are especially relevant for MCI populations, where age- and disease-related structural changes overlap. Models that rely on age as a proxy risk overestimating disease probability in older individuals, while PMDN enforces age-invariant predictions and provides better calibrated probability estimates. Notably, PMDN's more conservative predictions yield improved calibration (Brier = 0.22 ± 0.01) at the expense of larger absolute prediction errors, whereas CF-Net favors more confident but less consistently calibrated predictions (Brier = 0.26).

Post hoc calibration primarily improved the reliability of predicted probabilities rather than the model's discriminative ability, as classification performance was largely preserved across methods. Monotonic approaches (logit, probit, isotonic regression) produced smooth and reliable calibration

curves, whereas nonmonotonic methods like BBQ introduced irregularities. Parametric methods are stable but less flexible, while nonparametric methods capture local variations at the cost of increased variability. Evaluation of calibration should include metrics such as ECE and MCE alongside visual inspection.

Despite these promising results, several limitations should be considered. First, the analysis focused on age as a single confounder, whereas other factors such as sex, scanner variability, and site effects may also influence model predictions. Second, the dataset size remains limited relative to the high dimensionality of neuroimaging data, which may affect the robustness and generalizability of deep learning networks. Third, adversarial training may benefit from longer or more carefully tuned optimization, and its performance here may not represent its full potential. Future work should extend these approaches to multiconfounder settings, larger and more diverse cohorts, and disease-focused regional inputs, such as hippocampal or medial temporal lobe measures, as a complementary strategy to whole-brain voxel-wise representations. Such anatomically informed inputs may encourage greater reliance on clinically established disease-related anatomical markers and may help reduce the influence of widespread age-related structural changes.

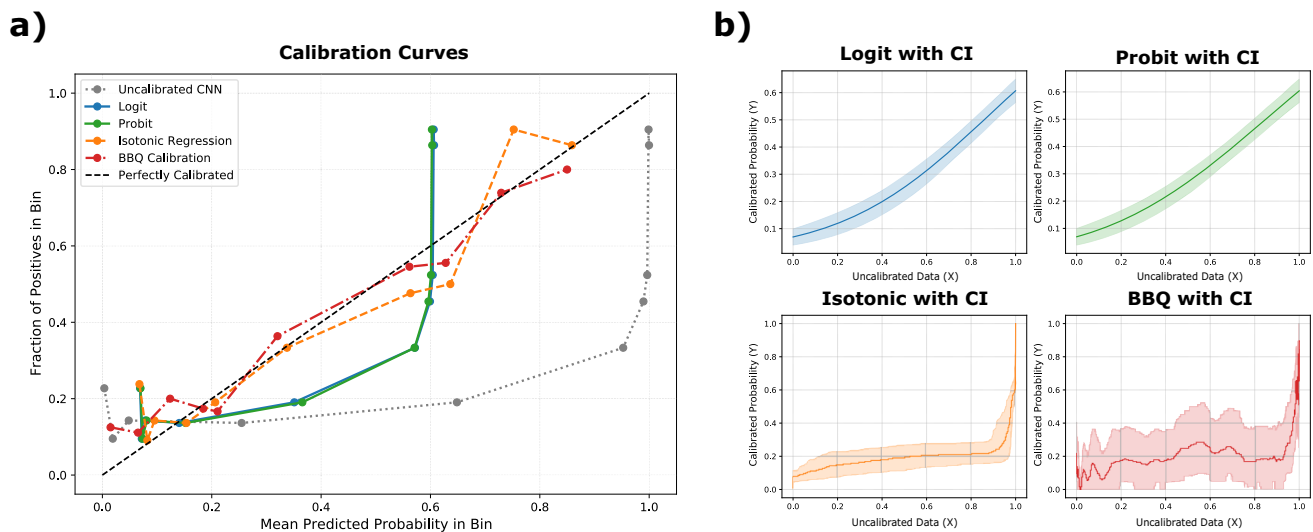


Fig. 3 Calibration plots: **a** Reliability of the various calibration methods. **b** Comparison of the calibrated probability estimates and confidence intervals

Combining disease-focused regional representations with explicit confounder modeling may therefore provide an additional strategy for improving age robustness while preserving disease-relevant information.

Conclusion

In this study, we investigated two approaches for mitigating age-related bias, PMDN and CF-Net, and additionally evaluated post hoc calibration methods to improve the reliability of model predictions. PMDN achieved comparable or improved discriminative performance relative to the baseline and CF-Net models, while demonstrating greater stability across evaluations. Importantly, PMDN substantially attenuated age-related associations in cognitively normal and mild cognitive impairment subjects, although residual age-related associations persisted in the AD group. In contrast, CF-Net provided less consistent attenuation of age-related associations and exhibited greater variability. Post hoc calibration further improved the reliability of predicted probabilities while largely preserving classification performance, with isotonic regression providing the most favorable calibration results. Further validation in larger and more diverse cohorts, with multiple confounders and anatomically informed disease-focused representations, will be important to assess their potential for clinical translation.

Acknowledgements The study data were provided by the DELCODE study group of the Clinical Research Unit of the German Center for Neurodegenerative Diseases (DZNE). Details and participating sites can be found at <https://dzne.de/en/research/studies/clinical-studies/delcode>. Data collection and sharing were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the

National Institute of Biomedical Imaging and Bioengineering, and generous support of other industry partners. A complete listing of ADNI investigators can be found at https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Data were provided by the AIBL study group. AIBL researchers are listed at <https://aibl.csiro.au>.

Funding Open Access funding enabled and organized by Projekt DEAL. The German Research Foundation (DFG) provided funding for the study to Martin Dyrba (project no. 454834942) and Peter Dechent (project no. 374011584).

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Ethical Approval The studies were approved by the respective neuroimaging initiatives internal review boards of each of the participating study sites. See <https://adni.loni.usc.edu> and <https://aibl.org.au> for details. All initiatives met common ethical standards in the collection of the data such as the Declaration of Helsinki. Analysis of the data was approved by the internal review board of the Rostock University Medical Center, reference number A 2020-0182.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Rashid AB, Kausik MAK (2024) AI revolutionizing industries worldwide: a comprehensive overview of its diverse applications. *Hybrid Adv* 7:100277. <https://doi.org/10.1016/j.hybadv.2024.100277>
2. Zhang J, Che Y, Liu R, Wang Z, Liu W (2025) Deep learning-driven multi-omics analysis: enhancing cancer diagnostics and therapeutics. *Brief Bioinform* 26(4):440. <https://doi.org/10.1093/bib/bbaf440>
3. Gadhave DG, Sugandhi VV, Jha SK, Nangare SN, Gupta G, Singh SK, Dua K, Cho H, Hansbro PM, Paudel KR (2024) Neurodegenerative disorders: mechanisms of degeneration and therapeutic approaches with their clinical relevance. *Ageing Res Rev* 99:102357. <https://doi.org/10.1016/j.arr.2024.102357>
4. Zhao L (2025) Advances in functional magnetic resonance imaging-based brain function mapping: a deep learning perspective. *Psychoradiology* 5:007. <https://doi.org/10.1093/psyrad/kkaf007>
5. Yao Z, Wang H, Yan W, Wang Z, Zhang W, Wang Z, Zhang G (2023) Artificial intelligence-based diagnosis of Alzheimer's disease with brain MRI images. *Eur J Radiol* 165:110934. <https://doi.org/10.1016/j.ejrad.2023.110934>
6. Alsubaie MG, Luo S, Shaikat K (2024) Alzheimer's disease detection using deep learning on neuroimaging: a systematic review. *Mach Learn Knowl Ext* 6(1):464–505. <https://doi.org/10.3390/make6010024>
7. Dhar T, Dey N, Borra S, Sherratt RS (2023) Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Trans Technol Soc* 4(1):68–75. <https://doi.org/10.1109/TTS.2023.3234203>
8. Anagnostopoulos C, Schaar M, Collet J-P, Vecino R (2025) Real-world data and causal machine learning to enhance drug development. *Ther Innov Regul Sci*. <https://doi.org/10.1007/s43441-025-00849-0>
9. Dinsdale NK, Bluemke E, Smith SM, Arya Z, Vidaurre D, Jenkinson M, Namburete AIL (2021) Learning patterns of the ageing brain in MRI using deep convolutional networks. *Neuroimage* 224:117401. <https://doi.org/10.1016/j.neuroimage.2020.117401>
10. Mehra S, Raut G, Purkayastha RD, Vishvakarma SK, Biasizzo A (2023) An empirical evaluation of enhanced performance softmax function in deep learning. *IEEE Access* 11:34912–34924. <https://doi.org/10.1109/ACCESS.2023.3265327>
11. Pearce, T., Brintrup, A., Zhu, J.: (2021) Understanding softmax confidence and uncertainty. *CoRR* [arXiv:2106.04972](https://arxiv.org/abs/2106.04972)
12. Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, Lautenschlager NT, Lenzo N, Martins RN, Maruff P, Masters C, Milner A, Pike K, Rowe C, Savage G, Szoëke C, Taddei K, Villemagne V, Woodward M, Ames D (2009) 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* 21(4):672–687. <https://doi.org/10.1017/S1041610209009405>
13. Jessen F, Spottke A, Boecker H, Brosseron F, Buerger K, Catak C, Fliessbach K, Franke C, Fuentes Casañ M, Heneka M, Janowitz D, Kilimann I, Laske C, Menne F, Nestor P, Peters O, Priller J, Pross V, Ramirez A, Duzel E (2018) Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (delcode). *Alzheimer's Res Ther*. <https://doi.org/10.1186/s13195-017-0314-2>
14. Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell L, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W. (2008) The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging* 27(4):685–691. <https://doi.org/10.1002/jmri.21049>
15. George A, Kuzniecky R, Rusinek H, Pardoe HR (2020) Human Epilepsy Project Investigators: Standardized brain MRI acquisition protocols improve statistical power in multicenter quantitative morphometry studies. *J Neuroimaging* 30(1):126–133. <https://doi.org/10.1111/jon.12673>
16. Zhao Q, Adeli E, Pohl K (2020) Training confounder-free deep learning models for medical applications. *Nat Commun* 11. <https://doi.org/10.1038/s41467-020-19784-9>
17. Vento, A., Zhao, Q., Paul, R., Pohl, K.M., Adeli, E.: A penalty approach for normalizing feature distributions to build confounder-free models. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 387–397. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_37
18. Lu M, Zhao Q, Zhang J, Pohl KM, Fei-Fei L, Niebles JC, Adeli E. Metadata normalization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021 Jun 20 (pp. 10912-10922)*. IEEE
19. Martensen OH, Strauß T, Ramedani M, Dyrba M.: Comparison of post-hoc calibration methods for neural network likelihood scores. In: *Bildverarbeitung Für die Medizin 2026*, pp. 443–449. Springer, Wiesbaden (2026). https://doi.org/10.1007/978-3-658-51100-5_87
20. Rainio O, Teuvo J, Klén R (2024) Evaluation metrics and statistical tests for machine learning. *Sci Rep* 14. <https://doi.org/10.1038/s41598-024-56706-x>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Majid Ramedani¹ · Jaya C. Terli¹ · Ole H. Martensen² · Tobias Strauß² · Katharina Buerger^{3,4} · Peter Dechent⁵ · Emrah Düzel^{6,7} · Klaus Fliessbach^{8,9} · Maria Gemenetzi^{10,11} · Wenzel Glanz⁶ · Julian Hellmann-Regen^{11,12,13} · Stefan Hetzer¹⁴ · Enise I. Incesoy^{6,7,15} · Daniel Janowitz^{3,4} · Marie T. Kronmüller⁹ · Christoph Laske^{16,17} · Falk Lüsebrink⁶ · Robert Perneczky^{3,18,19,20} · Oliver Peters^{11,12,13} · Lukas Preis¹² · Josef Priller^{10,11,21,22,23} · Boris-Stephan Rauchmann^{18,24,25} · Ayda Rostamzadeh²⁶ · Anja Schneider^{8,9} · Björn H. Schott^{27,28,29} · Sebastian Sodenkamp^{16,30} · Annika Spottke^{9,31} · Eike J. Spruth^{10,11} · Sophia Stöcklein^{3,32} · Jens Wiltfang^{27,28,33} · Frank Jessen^{9,26,34} · Stefan J. Teipel^{1,35} · Martin Dyrba¹ 

Majid Ramedani
majid.ramedani@dzne.de

Jaya C. Terli
jayachandra.terli@dzne.de

Ole H. Martensen
ole.martensen@uni-rostock.de

Tobias Strauß
tobias.strauss@uni-rostock.de

Katharina Buerger
katharina.buerger@med.uni-muenchen.de

Peter Dechent
peter.dechent@med.uni-goettingen.de

Emrah Düzel
emrah.duezel@dzne.de

Klaus Fliessbach
Klaus.Fliessbach@ukbonn.de

Maria Gemenetzi
maria.gemenetzi@charite.de

Wenzel Glanz
Wenzel.Glanz@dzne.de

Julian Hellmann-Regen
julian.hellmann@charite.de

Stefan Hetzer
stefan.hetzer@charite.de

Enise I. Incesoy
enise.incesoy@dzne.de

Daniel Janowitz
Daniel.Janowitz@med.uni-muenchen.de

Marie T. Kronmüller
MarieTheres.Kronmueller@dzne.de

Christoph Laske
Christoph.Laske@med.uni-tuebingen.de

Falk Lüsebrink
Falk.Luesebrink@dzne.de

Robert Perneczky
Robert.Perneczky@med.uni-muenchen.de

Oliver Peters
oliver.peters@charite.de

Lukas Preis
lukas.preis@charite.de

Josef Priller
josef.priller@charite.de

Boris-Stephan Rauchmann
boris.rauchmann@med.uni-muenchen.de

Ayda Rostamzadeh
ayda.rostamzadeh@uk-koeln.de

Anja Schneider
anja.schneider@dzne.de

Björn H. Schott
Bjoern-Hendrik.Schott@dzne.de

Sebastian Sodenkamp
sebastian.sodenkamp@med.uni-tuebingen.de

Annika Spottke
annika.spottke@dzne.de

Eike J. Spruth
eike.spruth@charite.de

Sophia Stöcklein
Sophia.Stoecklein@med.uni-muenchen.de

Jens Wiltfang
jens.wiltfang@med.uni-goettingen.de

Frank Jessen
frank.jessen@uk-koeln.de

Stefan J. Teipel
stefan.teipel@med.uni-rostock.de

- ¹ German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany
- ² Department of Mathematics, University of Rostock, Rostock, Germany
- ³ German Center for Neurodegenerative Diseases (DZNE), Munich, Germany
- ⁴ Institute for Stroke and Dementia Research (ISD), Ludwig Maximilian University Munich, Munich, Germany
- ⁵ MR-Research in Neurosciences, Department of Cognitive Neurology, University Medical Center Goettingen, Goettingen, Germany
- ⁶ German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany
- ⁷ Institute of Cognitive Neurology and Dementia Research (IKND), Otto-von-Guericke University Magdeburg, Magdeburg, Germany
- ⁸ Department of Old Age Psychiatry and Cognitive Disorders, University Hospital Bonn, Bonn, Germany
- ⁹ German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

- ¹⁰ Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany
- ¹¹ German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany
- ¹² Department of Psychiatry and Neurosciences, Charité – Universitätsmedizin Berlin, Berlin, Germany
- ¹³ Experimental and Clinical Research Center (ECRC), Charité – Universitätsmedizin Berlin, Berlin, Germany
- ¹⁴ Berlin Center for Advanced Neuroimaging, Charité – Universitätsmedizin Berlin, Berlin, Germany
- ¹⁵ Department for Psychiatry and Psychotherapy, University Clinic Magdeburg, Magdeburg, Germany
- ¹⁶ German Center for Neurodegenerative Diseases (DZNE), Tuebingen, Germany
- ¹⁷ Section for Dementia Research, Hertie Institute for Clinical Brain Research and Department of Psychiatry and Psychotherapy, University of Tuebingen, Tuebingen, Germany
- ¹⁸ Department of Psychiatry and Psychotherapy, Ludwig Maximilian University Munich, Munich, Germany
- ¹⁹ Munich Cluster for Systems Neurology (SyNergy), Munich, Germany
- ²⁰ Ageing Epidemiology Research Unit (AGE), School of Public Health, Imperial College London, London, UK
- ²¹ Department of Psychiatry and Psychotherapy, Technical University of Munich, Munich, Germany
- ²² German Center for Mental Health (DZPG), Munich, Germany
- ²³ UK Dementia Research Institute, University of Edinburgh, Edinburgh, UK
- ²⁴ Department of Neuroradiology, Ludwig Maximilian University Munich, Munich, Germany
- ²⁵ Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK
- ²⁶ Department of Psychiatry and Psychotherapy, University of Cologne, Cologne, Germany
- ²⁷ Department of Psychiatry and Psychotherapy, University Medical Center Goettingen, Goettingen, Germany
- ²⁸ German Center for Neurodegenerative Diseases (DZNE), Goettingen, Germany
- ²⁹ Department of Psychiatry and Psychotherapy, Otto von Guericke University, Magdeburg, Germany
- ³⁰ Department of Psychiatry and Psychotherapy, University of Tuebingen, Tuebingen, Germany
- ³¹ Department of Neurology, University Hospital Bonn, Bonn, Germany
- ³² Department of Radiology, Ludwig Maximilian University Munich, Munich, Germany
- ³³ Department of Medical Sciences, University of Aveiro, Aveiro, Portugal
- ³⁴ Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany
- ³⁵ Department of Psychosomatic Medicine, Rostock University Medical Center, Rostock, Germany