MR imaging signatures of brain age and disease over the lifespan derived from a Deep Brain Network based on an international, diverse population of 14,468 individuals

Supplementary Materials

S.1. LifespanCN racial demographics

Suppl. table 1. Racial breakdown of the LifespanCN training set.

**Note: not all participants had available race and ethnicity information.*

	Number of participants per racial group					
Dataset Name	***Asian***	***Black***	***White***	***Multi-Racial***	***Other***	***N/A****
ADC	0	22	45	3	0	9
AIBL	0	0	0	0	0	446
BLSA-1.5T	0	2	20	0	0	68
BLSA-3T	0	21	90	0	0	841
CARDIA	0	291	428	0	0	0
PAC-JHU	0	1	94	0	0	0
PAC-WASH	0	0	224	0	23	0
PAC-WISC	0	2	124	0	1	0
PING	0	0	0	0	0	398
PNC	0	521	568	0	164	143
SHIP	0	0	0	0	0	2,739

S.2. Comparison of alternative network architectures

Suppl. table 2. Comparison of alternative network architectures. Performance on brain age prediction using LifespanCN dataset and network parameters.

	InceptionResNet-v2	**DenseNet169**	**VGG16**	**ResNet50**
MAE	**3.702**	3.795	4.319	3.850
Correlation	**0.978**	0.977	0.970	0.975
# parameters	55M	14.3M	138M	25M
Depth	N/A (200)	169	16	50

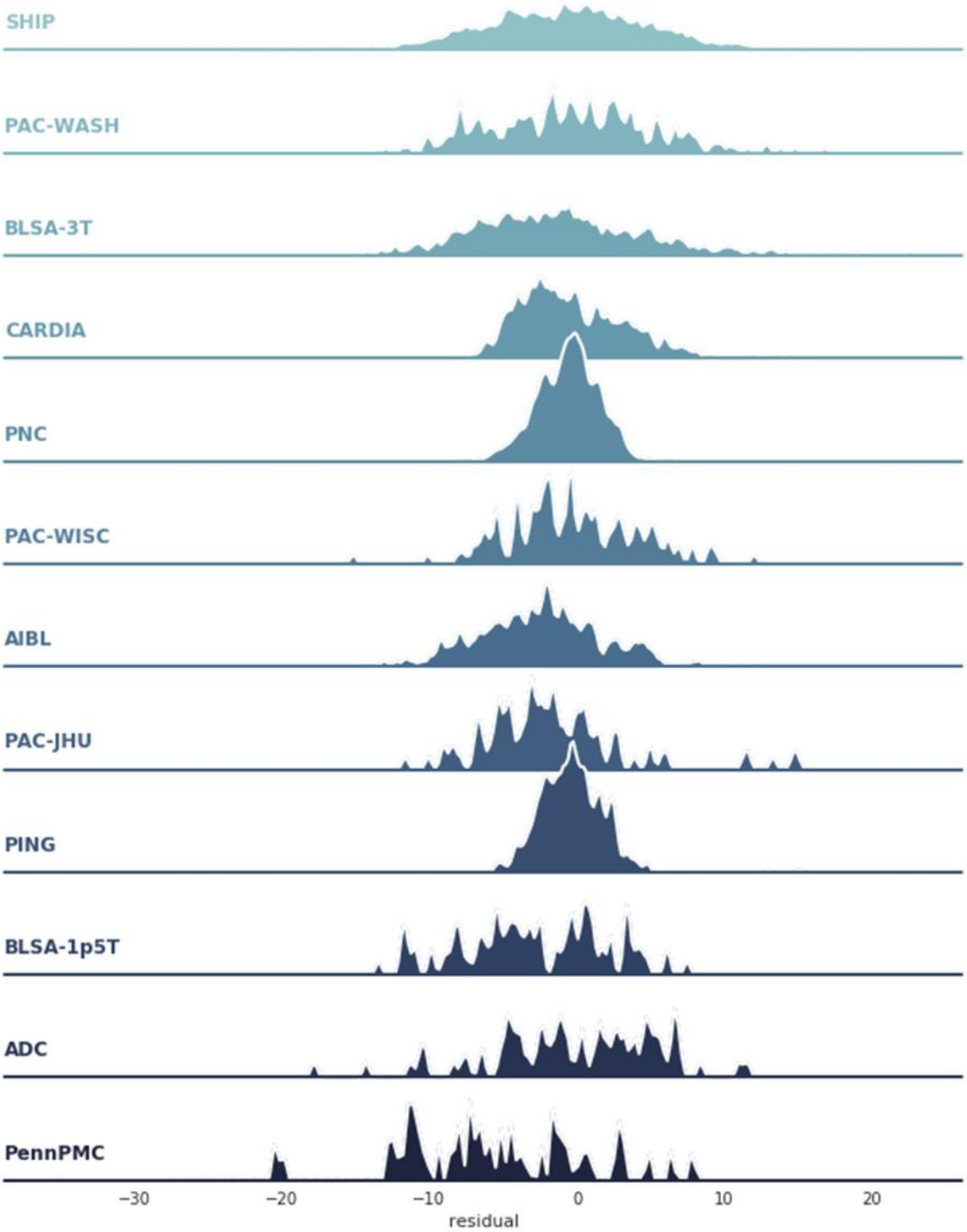
S.3. DeepBrainNet Brain Age prediction accuracy in different folds

Suppl. table 3. Prediction accuracy in brain age prediction using LifespanCN dataset in different folds of cross-validation.

InceptionResNet-v2				
	Slices Combined		Individual Slices	
	MAE	Correlation	MAE	Correlation
CV1	3.681	0.977	4.530	0.961
CV2	3.561	0.977	4.372	0.965
CV3	3.664	0.978	4.263	0.967
CV4	3.635	0.977	4.384	0.965
CV5	3.969	0.979	4.407	0.965
All	3.702	0.978	4.391	0.964

S.4. DeepBrainNet Brain Age deltas per site

Suppl. figure 1. Distribution of the Brain Age deltas per site.



S.5. Mean absolute error and mean error by gender

Suppl. table 4. Summary statistics of brain age deltas for males and females for prediction using the LifespanCN set.

	Mean absolute error	Mean error
Male	3.68	0.03
Female	3.72	-0.31

S.6. Correlation of deltas with gender-specific models

Suppl. table 5. Correlation of deltas obtained from gender-specific models to deltas obtained from mixed-gender model. These gender-specific models were trained with LifespanCN set separated by gender (with 5-fold cross validation).

Model	Delta Correlation
Female only	0.968
Male only	0.977

S.7. Effect of preprocessing on Brain Age predictions

Suppl. table 6. Brain age prediction accuracy for minimally preprocessed versus preprocessed images. The table presents MAE for the LifespanCN predictions (with 5-fold cross-validation) and predictions on out sample dataset.

Processing	LifespanCN cross-validated	SHIP out of sample
Minimally preprocessed (Skull-stripping + linear alignment)	3.702	4.120
Additional preprocessing (Bias correction + histogram normalization)	3.698	4.106

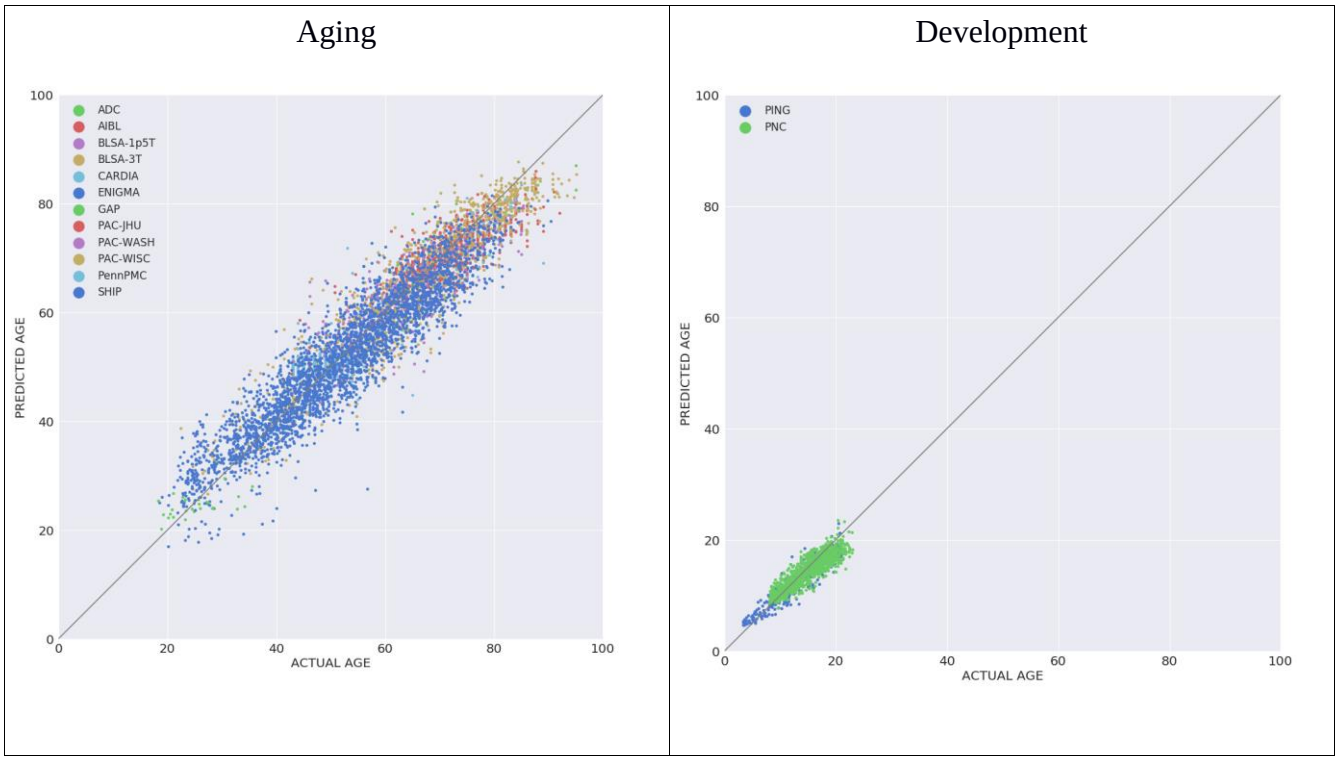
S.8. Correlation of deltas obtained from age-specific models

Suppl. table 7. Correlation of deltas obtained from age-specific models to deltas obtained from full sample model. The development set contains 2 studies from LifespanCN and has subjects between 3 and 22 years old. The aging set contains 12 studies from LifespanCN and has subjects between 18 and 95 years old. These models were trained with 5-fold cross-validation.

Model	Delta Correlation
Developmental set	0.973
Aging set delta correlation	0.954

S.9. Predictions from age-specific models

Suppl. figure 2. Plot of predictions obtained from age-specific models described in Suppl. Table 7.



S.10. Generalization of gender specific models to opposite gender

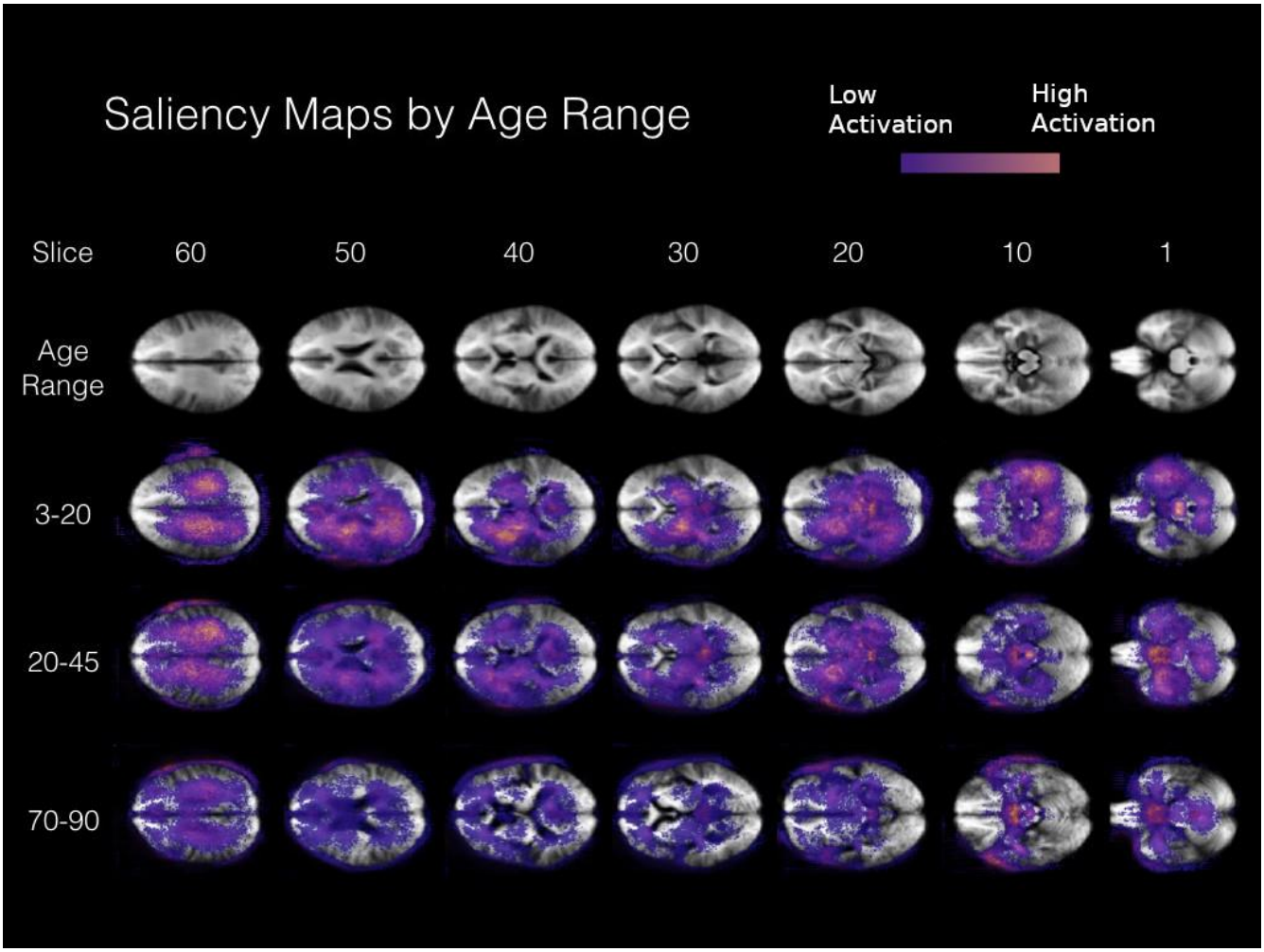
Suppl. table 8. Correlation of deltas obtained from opposite-sex models to deltas obtained from full sample model

Training	Testing	MAE	Delta Correlation
Male	Female	3.898	0.948
Female	Male	3.798	0.955

S.11. Visualization of model activation

Suppl. figure 3. Saliency maps were computed by method described by [Simonyan 2014]. The computed gradients of each sample within the respective age range were averaged. Higher activation represents the importance of a region in prediction.

Reference: https://arxiv.org/pdf/1312.6034v2.pdf



S.12. Robustness of small sample experiments

Suppl. table 9. Transfer learning experiments with sample sizes equal to or less than 100 were repeated to test robustness.

Experiment	N	Accuracy		AUC	
		DeepBrainNet	ImageNet	DeepBrainNet	ImageNet
AD	100	0.78	0.67	0.84	0.71
MCI	100	0.64	0.62	0.68	0.64
SCZ	100	0.73	0.61	0.80	0.70
	50	0.64**	0.59	0.72**	0.51

* Scores reported are the average of 2 runs

** At N=50 for SCZ, only one of the runs converged to the scores listed – Other run converged to baseline (52% - Accuracy)

S.13. Mean Error of predictions for 3 levels of regularization

Suppl. table 10. The mean absolute error, mean error, and Cohens'd effect sizes between disease and control for the 3 levels of fit shown in Figure 2.

	Mean Absolute Error	Mean Error	Cohens'd	Cohens'd 95% CI
AD	4.81	-1.94	1.19	0.98, 1.40
	5.47	-2.04	1.26	1.04, 1.48
	7.14	-4.78	1.16	0.95, 1.37
MCI	4.14	-1.62	0.52	0.32, 0.72
	5.15	-2.44	0.62	0.42, 0.82
	6.26	-4.66	0.55	0.35, 0.75
Schizophrenia	4.18	0.75	0.63	0.49, 0.77
	7.67	1.68	0.79	0.65, 0.93
	9.48	2.64	0.76	0.62, 0.90
Depression	4.18	0.73	0.11	0.01, 0.21
	7.27	-4.11	0.12	0.02, 0.22
	8.56	-5.41	0.09	-0.01, 0.19

S.14. Mixed-effects model testing for different model fits

Suppl. table 11. Results of a mixed-effects model used to determine which model fit captures the most differentiation in the residual values of controls versus disease subjects.

Group	Comparison	Fit with most separation	p-value
AD	Middle vs. Tight	Middle fit	0.021
	Middle vs. Loose	Middle fit	0.017
	Loose vs. Tight	Tight fit	0.311
MCI	Middle vs. Tight	Middle fit	0.043
	Middle vs. Loose	Middle fit	0.039
	Loose vs. Tight	Loose fit	0.550
Schizophrenia	Middle vs. Tight	Middle fit	0.074
	Middle vs. Loose	Middle fit	0.078
	Loose vs. Tight	Loose fit	0.105
Depression	Middle vs. Tight	Middle fit	0.550
	Middle vs. Loose	Middle fit	0.472
	Loose vs. Tight	Tight fit	0.508

We conduct appropriate testing to examine whether the brain age gap values differentiate disease (e.g., AD, MCI, Schizophrenia or Depression) and controls subjects, and whether such discrimination differ by the chose models (loose, middle and tight). Hence, we are testing the difference (by models) of the difference (by diagnosis) in brain age gaps.

We represent the brain age gap for subject i under model m (m=1,2,3) as $R_{i,m}$. D_i is the binary disease indicator with value 1 for diseased subjects and 0 for controls.

$$R_{i,m} = c_m + b_m D_i * m + a_i + noise$$

For a specific model m, c_m represents the average brain age gap among controls, and b_m represents the degree of differentiation in brain age gap comparing diseased versus controls. a_i is the random intercept that quantifies the subject-specific deviation of brain age gap from the population average. a_i is shared across all three model fits for each subject i, and takes care of the possible within-subject correlation in brain age gap values.

We further illustrate the above model under one particular scenario. That is, comparing loose fit (m=1) and middle fit (m=2) models in terms of their differentiation in AD versus controls.

Under the loose fit (m=1),

if subject i is a control subject (D_i=0), then

$$R_{i,1,D_i=0} = c_1 + a_i + noise$$

if subject i is an AD subject (D_i=1), then

$$R_{i,1,D_i=1} = c_1 + b_1 + a_i + noise$$

The average discrimination between AD vs. control under loose fit model is then the difference between the above b_1.

While using the middle fit model (m=2),

if subject i is a control subject, then

$$R_{i,2,D_i=0} = c_2 + a_i + noise$$

if subject i is an AD subject, then

$$R_{i,2,D_i=1} = c_2 + b_2 + a_i + noise$$

The average discrimination between AD vs. control under middle fit model is then the difference between the above b_2.

Hence to comparing middle fit and loose fit model, we will be testing the contrast of their discriminations under the null hypothesis that H_0: $b_2 - b_1 = 0$.

The significance of the differential discrimination from the mixed effects models were determined based on likelihood ratio (LRT) tests of the fixed effects.

The mixed effects model is used because the model-specific brain age gaps $R_{i,m}$ are generated from the same subject's data. Hence for any pairwise comparison such as middle fit vs. tight fit, the data might be correlated within subject. Mixed effects models with the subject-specific random intercept a_i are known to provide valid inference for correlated outcome data.