Supplementary Materials

Table S1. Scanner information for 9668 scans in ADNI dataset.

Vendor	Scanner Model	Field Strength	Number of Scans
GE	Discovery MR750	3.0T	886
	Discovery MR750w	3.0T	123
	Genesis Signa	1.5T	248
		3.0T	6
	Signa Excite	1.5T	892
		3.0T	6
	Signa HDx	1.5T	463
		3.0T	36
	Signa HDxt	1.5T	208
		3.0T	419
	Signa Premier	3.0T	33
	Signa UHP	3.0T	1
Philips	Achieva dStream	3.0T	88
	Ingenia	3.0T	180
	Ingenia Elition X	3.0T	8
	Achieva	1.5T	73
		3.0T	541
	Gemini	3.0T	32
	Gyroscan Intera	1.5T	12
	Gyroscan NT	1.5T	2
	Ingenuity	3.0T	18
	Intera	1.5T	333
		3.0T	217
	Intera Achieva	1.5T	6
SIEMENS	Allegra	3.0T	32
	Avanto	1.5T	391
	Biograph_mMR	3.0T	19
	Espree	1.5T	24
	Numaris4	1.5T	2
	Prisma	3.0T	191
	Prisma_fit/Magnetom Prisma_fit	3.0T	466
	Skyra	3.0T	406
	Skyra_fit	3.0T	17
	Sonata	1.5T	379
	SonataVision	1.5T	31
	Symphony/Symphony Tim	1.5T	671
	Trio/TrioTim	3.0T	1487
	Verio	3.0T	718
	Skyra|DicomCleaner	3.0T	3

Table S2. Scanner information for 940 scans in AIBL dataset.

Vendor	Scanner Model	Field Strength	Number of Scans
SIEMENS	Avanto	1.5T	245
	Trio/TrioTim	3.0T	622
	Verio	3.0T	73

Table S3. Scanner information for 1453 scans in MACC dataset.

Vendor	Scanner Model	Field Strength	Number of Scans
SIEMENS	Prisma	3.0T	85
	Trio/TrioTim	3.0T	1368

Table S4. Scanner information for 2519 scans in OASIS dataset.

Vendor	Scanner Model	Field Strength	Number of Scans
SIEMENS	Avanto	1.5T	2
	Biograph_mMR	3.0T	812
	Magnetom Vida	3.0T	305
	Prisma_fit/Magnetom Prisma_fit	3.0T	1
	Sonata	1.5T	40
	Trio/TrioTim	3.0T	1359

Table S5. Our study used five regional brain volumes (Hippocampus, Fusiform, MidTemp, Ventricles, WholeBrain), computed by summing the volumes of various FreeSurfer regions of interest. Note that the 6th anatomical volume (intracranial volume) was directly provided by FreeSurfer.

Regional volumetric feature	Brain regions used to compute regional volumes
Hippocampus	Left-Hippocampus, Right-Hippocampus
Fusiform	lh_fusiform_volume, rh_fusiform_volume
MidTemp	lh_middletemporal_volume, rh_middletemporal_volume
Ventricles	Left-Inf-Lat-Vent, Left-Lateral-Ventricle, Right-Inf-Lat-Vent, Right-Lateral-Ventricle
WholeBrain	WM-hypointensities, Left-Cerebellum-Cortex, Left-Cerebellum-White-Matter, Left-Thalamus-Proper, Left-Caudate, Left-Putamen, Left-Pallidum, Left-Hippocampus, Left-Amygdala, Left-Accumbens-area, Left-VentralDC, Right-Cerebellum-Cortex, Right-Cerebellum-White-Matter, Right-Thalamus-Proper, Right-Caudate, Right-Putamen, Right-Pallidum, Right-Hippocampus, Right-Amygdala, Right-Accumbens-area, Right-VentralDC, lhCortexVol, rhCortexVol, lhCerebralWhiteMatterVol, rhCerebralWhiteMatterVol

Table S6. Complete set of L2C-XGBw and L2C-XGBnw features and their corresponding original features.

Original feature	**L2C features**
Demographic and time-related features	APOE4, is_male, educ, marital_status, current_age, month_since_baseline,
Clinical diagnosis (categorical: CN, MCI, AD)	mr_dx, time_since_mr_dx, best_dx, time_since_best_dx, worst_dx, time_since_worst_dx, milder, time_since_milder,
MMSE score (numeric, ordinal: 0-30)	mr_MMSE, time_since_mr_MMSE, mr_change_MMSE, low_MMSE, time_since_low_MMSE, high_MMSE, time_since_high_MMSE,
CDR_GLOBAL score (numeric, ordinal: 0, 0.5, 1, 2, 3)	mr_CDR, time_since_mr_CDR, mr_change_CDR, low_CDR, time_since_low_CDR, high_CDR, time_since_high_CDR,
Ventricle volume (numeric, continuous)	mr_Ventricles, time_since_mr_Ventricles, mr_change_Ventricles, low_Ventricles, time_since_low_Ventricles, high_Ventricles, time_since_high_Ventricles,
Fusiform volume (numeric, continuous)	mr_Fusiform, time_since_mr_Fusiform, mr_change_Fusiform, low_Fusiform, time_since_low_Fusiform, high_Fusiform, time_since_high_Fusiform,
WholeBrain volume (numeric, continuous)	mr_WholeBrain, time_since_mr_WholeBrain, mr_change_WholeBrain, low_WholeBrain, time_since_low_WholeBrain, high_WholeBrain, time_since_high_WholeBrain,
Hippocampus volume (numeric, continuous)	mr_Hippocampus, time_since_mr_Hippocampus, mr_change_Hippocampus, low_Hippocampus, time_since_low_Hippocampus, high_Hippocampus, time_since_high_Hippocampus,
Middle temporal volume (numeric, continuous)	mr_MidTemp, time_since_mr_MidTemp, mr_change_MidTemp, low_MidTemp, time_since_low_MidTemp, high_MidTemp, time_since_high_MidTemp,
Intracranial volume (numeric, continuous)	mr_ICV, time_since_mr_ICV, mr_change_ICV, low_ICV, time_since_low_ICV, high_ICV, time_since_high_ICV

Table S7. L2C-FNN input vector dimensionality. For discrete features, the calculation is performed after one-hot encoding, including an additional class for missing data or unknown.

Category	L2C features	Feature dimension
Discrete features (dimensions are after one-hot encoding, including "unknown" class)	apoe	4
	is_male	3
	marital_status	3
	mr_CDR	6
	high_CDR	6
	low_CDR	6
	mr_dx	4
	best_dx	4
	worst_dx	4
	milder	3
MRI features	mr_Ventricles, time_since_mr_Ventricles, mr_change_Ventricles, low_Ventricles, time_since_low_Ventricles, high_Ventricles, time_since_high_Ventricles, mr_Fusiform, time_since_mr_Fusiform, mr_change_Fusiform, low_Fusiform, time_since_low_Fusiform, high_Fusiform, time_since_high_Fusiform, mr_WholeBrain, time_since_mr_WholeBrain, mr_change_WholeBrain, low_WholeBrain, time_since_low_WholeBrain, high_WholeBrain, time_since_high_WholeBrain, mr_Hippocampus, time_since_mr_Hippocampus, mr_change_Hippocampus, low_Hippocampus, time_since_low_Hippocampus, high_Hippocampus, time_since_high_Hippocampus, mr_MidTemp, time_since_mr_MidTemp, mr_change_MidTemp, low_MidTemp, time_since_low_MidTemp, high_MidTemp, time_since_high_MidTemp, mr_ICV, time_since_mr_ICV, mr_change_ICV, low_ICV, time_since_low_ICV, high_ICV, time_since_high_ICV	42
Cognitive features	mr_MMSE, time_since_mr_MMSE, mr_change_MMSE, low_MMSE, time_since_low_MMSE, high_MMSE, time_since_high_MMSE, time_since_mr_CDR, time_since_low_CDR, time_since_high_CDR,	10
Diagnostic features	time_since_mr_dx, time_since_best_dx, time_since_worst_dx	3
Demographic and time-dependent features	baseline education level, current_age, month_since_baseline	3

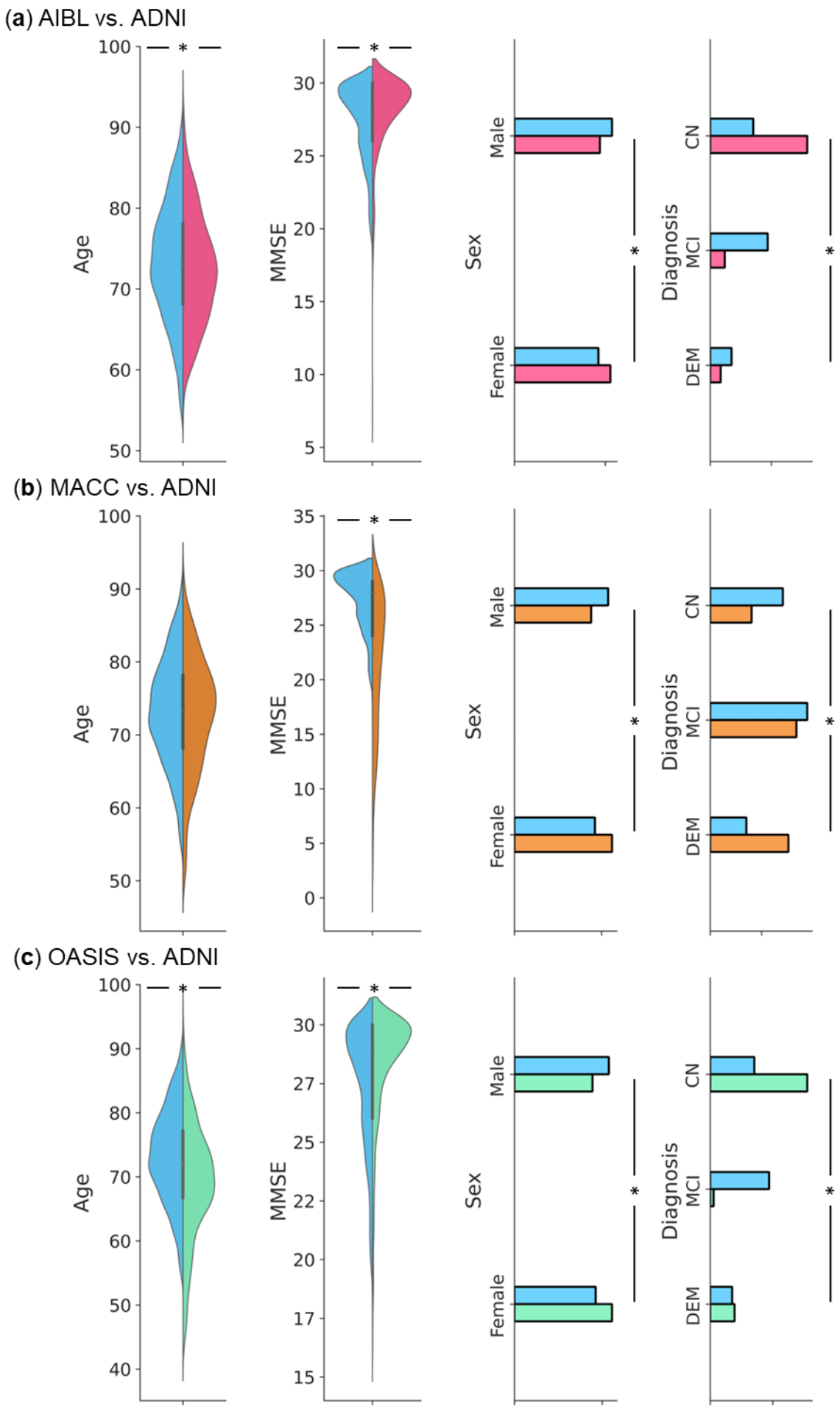

(a) AIBL vs. ADNI
(b) MACC vs. ADNI
(c) OASIS vs. ADNI
Age
MMSE
Sex
Male
Female
Diagnosis
CN
MCI
DEM
*

Figure S1. Baseline age, baseline MMSE, sex and baseline clinical diagnosis distribution differences between ADNI and external test set. (a) Distributions of age, sex, MMSE and clinical diagnosis for ADNI (blue) and AIBL (pink). (b) Distributions of age, sex, MMSE and clinical diagnosis for ADNI (blue) and MACC (yellow). (c) Distributions of age, sex, MMSE and clinical diagnosis for ADNI (blue) and OASIS (green). To test for differences in mean age and mean MMSE between ADNI and the external dataset, a permutation test was used. To test for differences in distributions of sex and diagnosis between ADNI and the external dataset, the chi square test was used. * indicates statistical significance after correcting for multiple comparisons with false discovery rate (FDR) $q < 0.05$. See Table 2 in the main text for the full set of statistical tests.

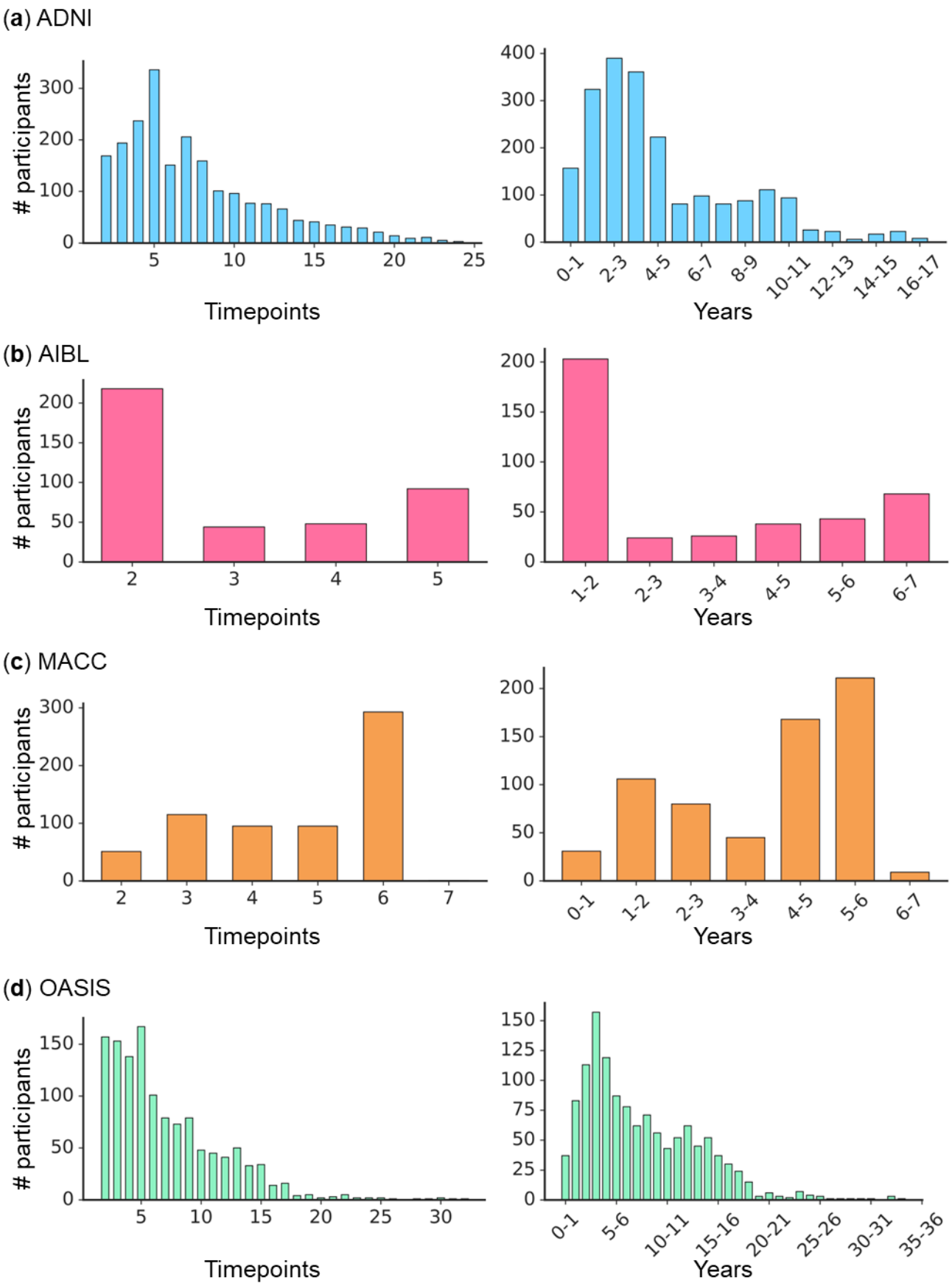

(a) ADNI
participants
Timepoints
Years
(b) AIBL
participants
Timepoints
Years
(c) MACC
participants
Timepoints
Years
(d) OASIS
participants
Timepoints
Years

Figure S2. Left: Distribution of the number of timepoints for all participants in each dataset. **Right:** Distribution of the number of years between baseline and the last observation for all participants in each dataset. Note that year 0-1 means $0 < t \leq 12$, where t is the interval between baseline and the last observed timepoint; year 1-2 means $12 < t \leq 24$, and so on. (a) ADNI. (b) AIBL. (c) MACC. (d) OASIS.

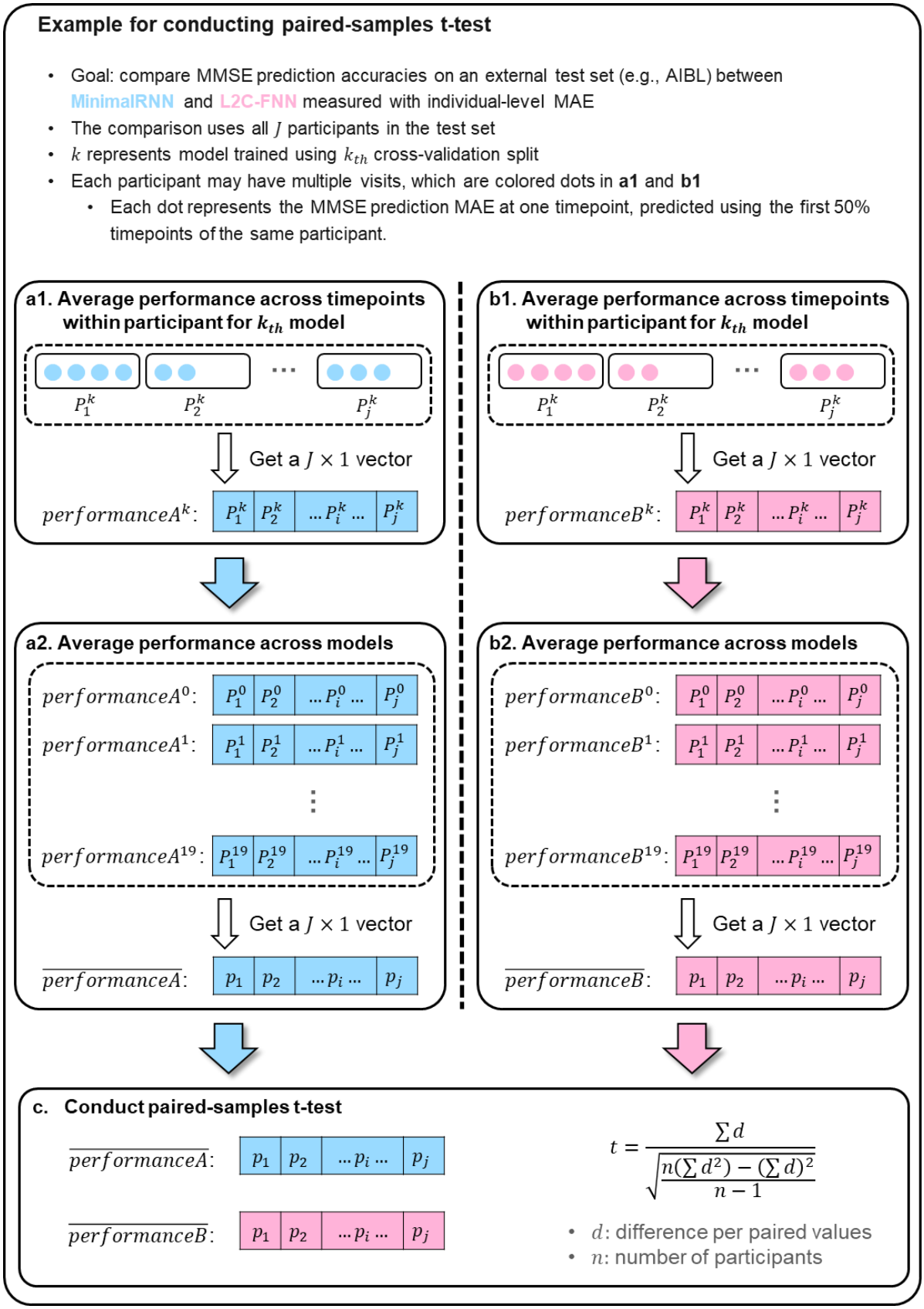


Figure S3. Illustration of paired-samples t-test for comparing performance measured with individual-level metrics (e.g., MMSE MAE) of MinimalRNN and L2C-FNN on an external test set. (a1) For a given model, we averaged the MMSE MAE within each

participant across all timepoints for MinimalRNN. (a2) Averaging the participant-level MMSE MAE obtained in (a1) across the 20 models. (b1 & b2) Same as a1 and a2 but for L2C-FNN. (c) Conduct paired-samples t-test to obtain p value.

Example for conducting permutation test

- Goal: compare clinical diagnosis accuracies on an external test set (e.g., AIBL) between MinimalRNN and L2C-FNN measured with group-level metrics (e.g., mAUC)
- The comparison uses all J participants in the test set
- k represents model trained using k_{th} cross-validation split
- Each participant may have multiple visits, which are colored triangles in **a** and **b**
  - Each triangle represents the clinical diagnosis prediction probabilities at one timepoint, predicted using the first 50% timepoints of the same participant.

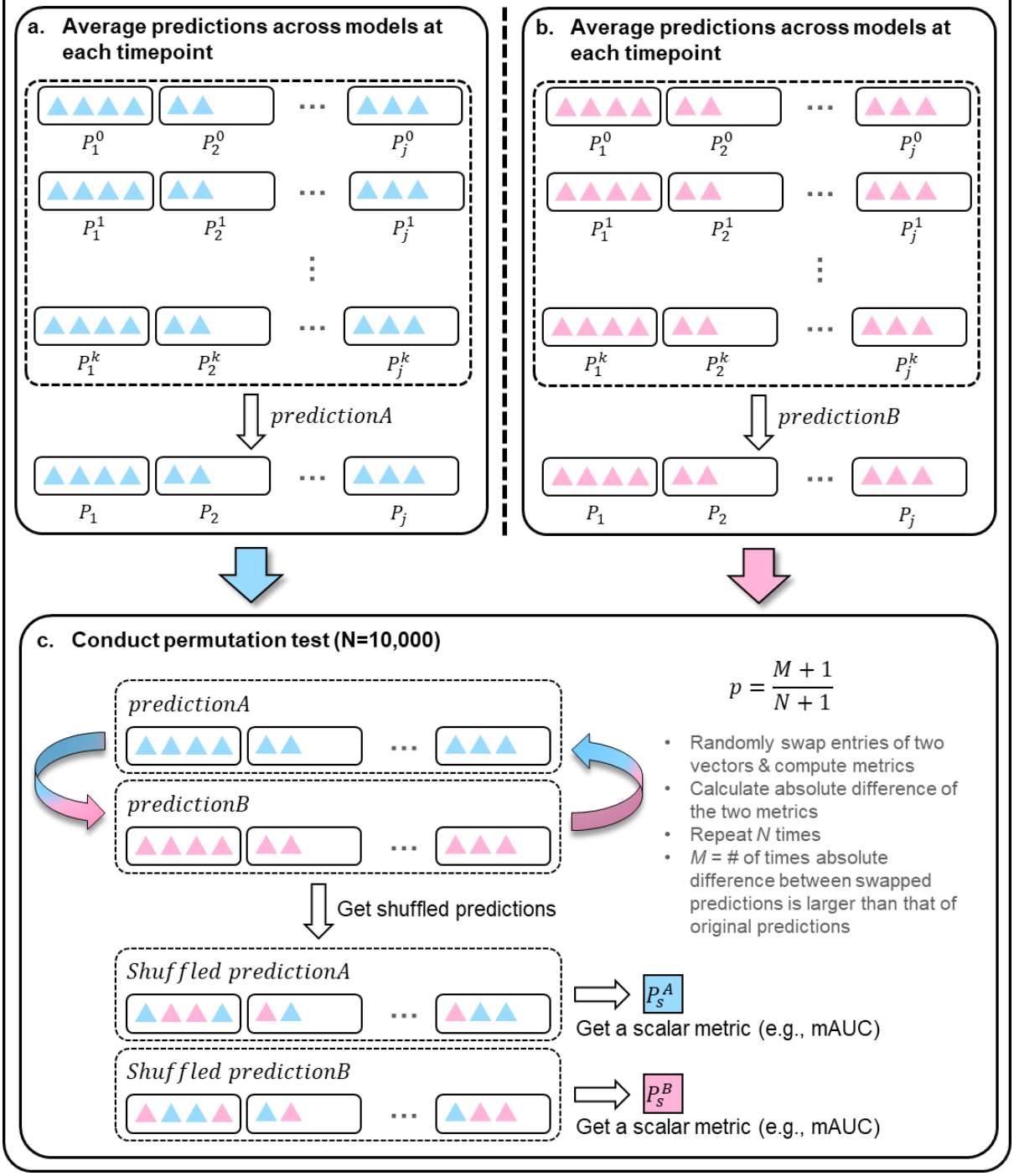
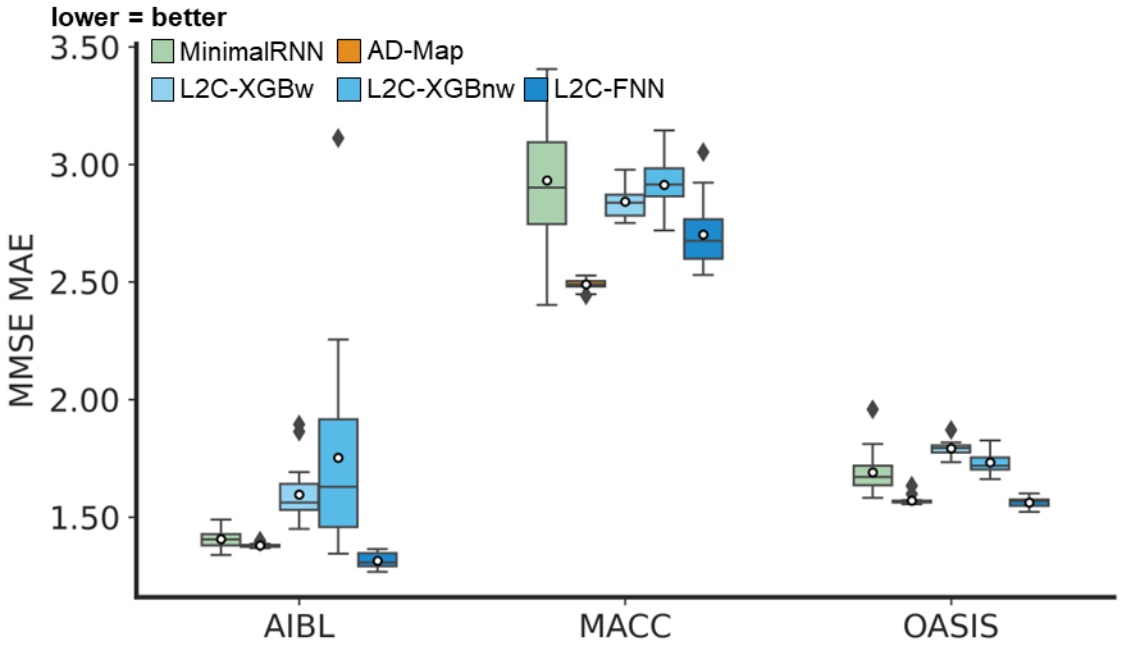


Figure S4. Illustration of permutation test for comparing performance measured with group-level metric (e.g., diagnostic prediction mAUC) of MinimalRNN and L2C-FNN on external test set. (a) For each participant, we averaged the predictions at each timepoint across 20 models for MinimalRNN. (b) Same as A but for L2C-FNN. (c) Permute 10,000 times and compute group-level metric (e.g., mAUC) on permuted predictions to obtain p value

(**a**)



(**b1**)

AIBL MMSE	Minimal RNN	AD-Map	L2C-XGBw	L2C-XGBnw	L2C-FNN
MinimalRNN		n.s.	**	***	*
AD-Map	n.s.		n.s.	**	n.s.
L2C-XGBw	**	n.s.		**	***
L2C-XGBnw	***	**	**		***
L2C-FNN	*	n.s.	***	***	

(**b2**)

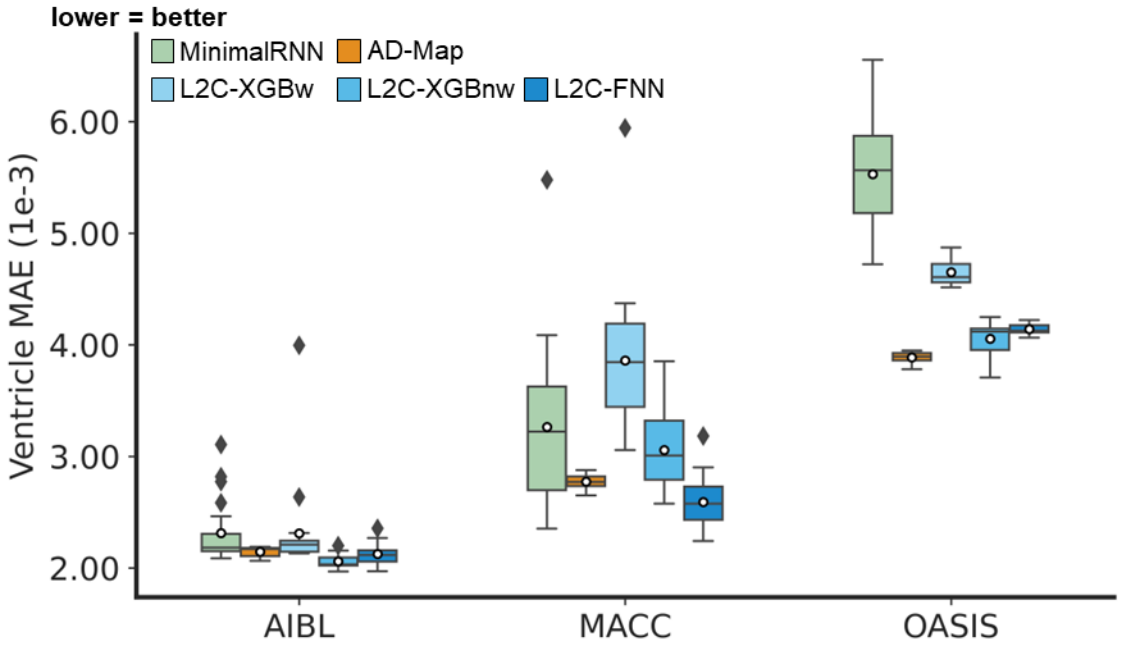
MACC MMSE	Minimal RNN	AD-Map	L2C-XGBw	L2C-XGBnw	L2C-FNN
MinimalRNN		***	n.s.	n.s.	**
AD-Map	***		***	***	**
L2C-XGBw	n.s.	***		n.s.	*
L2C-XGBnw	n.s.	***	n.s.		**
L2C-FNN	**	**	*	**	

(**b3**)

OASIS MMSE	Minimal RNN	AD-Map	L2C-XGBw	L2C-XGBnw	L2C-FNN
MinimalRNN		***	**	n.s.	***
AD-Map	***		***	***	n.s.
L2C-XGBw	**	***		*	***
L2C-XGBnw	n.s.	***	*		***
L2C-FNN	***	n.s.	***	***	

Figure S5. Cross-cohort MMSE prediction error (MAE) on three external test datasets for AD-only experiments (setting other dementia diagnosis to NaN). (a) Boxplots display the variability across 20 trained models (from ADNI) for MMSE prediction assessed using MAE. The x-axis denotes the test dataset used for evaluation. (b) Statistical significance in the prediction error between all models. Each row shows the statistical difference between a model and all other models. For example, the first row of each 5 x 5 table corresponds to the statistical difference between MinimalRNN and the other models – green indicates that MinimalRNN performs better, while red indicates that MinimalRNN performs worse. Therefore, the colors are always flipped between red and green across the diagonal. "*" indicates $p < 0.05$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "**" indicates $p < 0.001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "***" indicates $p < 0.00001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "n.s." indicates no statistical significance ($p \geqslant 0.05$) or did not survive FDR correction.

(a)



(b1)

AIBL Ventricle	Minimal RNN	AD-Map	L2C-XGBw	L2C-XGBnw	L2C-FNN
MinimalRNN		n.s.	n.s.	*	*
AD-Map	n.s.		n.s.	n.s.	n.s.
L2C-XGBw	n.s.	n.s.		*	n.s.
L2C-XGBnw	*	n.s.	*		n.s.
L2C-FNN	*	n.s.	n.s.	n.s.	

(b2)

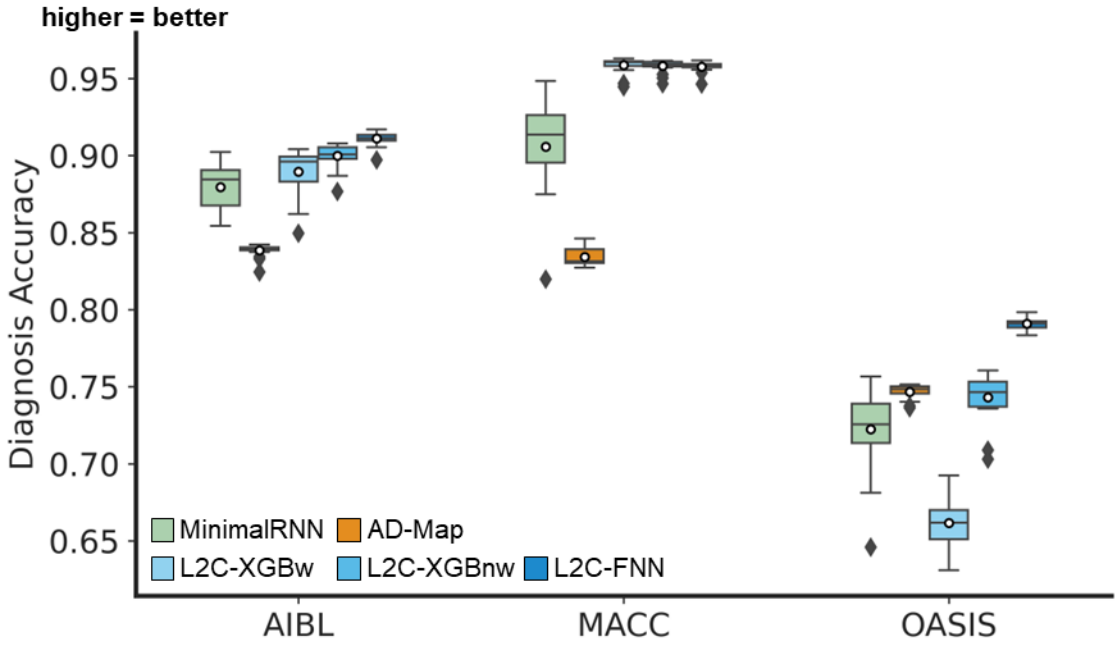
MACC Ventricle	Minimal RNN	AD-Map	L2C-XGBw	L2C-XGBnw	L2C-FNN
MinimalRNN		**	***	*	***
AD-Map	**		***	n.s.	n.s.
L2C-XGBw	***	***		***	***
L2C-XGBnw	*	n.s.	***		***
L2C-FNN	***	n.s.	***	***	

(b3)

OASIS Ventricle	Minimal RNN	AD-Map	L2C-XGBw	L2C-XGBnw	L2C-FNN
MinimalRNN		***	*	***	***
AD-Map	***		***	n.s.	n.s.
L2C-XGBw	*	***		***	***
L2C-XGBnw	***	n.s.	***		n.s.
L2C-FNN	***	n.s.	***	n.s.	

Figure S6. Cross-cohort ventricle volume prediction error (MAE) on three external test datasets for AD-only experiments (setting other dementia diagnosis to NaN). (a) Boxplots display the variability across 20 trained models (from ADNI) for ventricle volume prediction assessed using MAE. The x-axis denotes the test dataset used for evaluation. (b) Statistical significance in the prediction error between all models. Each row shows the statistical difference between a model and all other models. For example, the first row of each 5 x 5 table corresponds to the statistical difference between MinimalRNN and other models – green indicates that MinimalRNN performs better, while red indicates that MinimalRNN performs worse. Therefore, the colors are always flipped between red and green across the diagonal. "*" indicates $p < 0.05$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "**" indicates $p < 0.001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "***" indicates $p < 0.00001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "n.s." indicates no statistical significance ($p \geqslant 0.05$) or did not survive FDR correction.

(a)



(b1)

AIBL Diagnosis	Minimal RNN	AD-Map	L2C-XGBw	L2C-XGBnw	L2C-FNN
MinimalRNN		n.s.	n.s.	n.s.	n.s.
AD-Map	n.s.		*	*	**
L2C-XGBw	n.s.	*		n.s.	n.s.
L2C-XGBnw	n.s.	*	n.s.		n.s.
L2C-FNN	n.s.	**	n.s.	n.s.	

(b2)

MACC Diagnosis	Minimal RNN	AD-Map	L2C-XGBw	L2C-XGBnw	L2C-FNN
MinimalRNN		**	*	*	*
AD-Map	**		**	**	**
L2C-XGBw	*	**		n.s.	n.s.
L2C-XGBnw	*	**	n.s.		n.s.
L2C-FNN	*	**	n.s.	n.s.	

(b3)

OASIS Diagnosis	Minimal RNN	AD-Map	L2C-XGBw	L2C-XGBnw	L2C-FNN
MinimalRNN		n.s.	**	n.s.	n.s.
AD-Map	n.s.		**	n.s.	n.s.
L2C-XGBw	**	**		**	**
L2C-XGBnw	n.s.	n.s.	**		**
L2C-FNN	n.s.	n.s.	**	**	

Figure S7. Cross-cohort clinical diagnosis prediction accuracy (mAUC) on three external test datasets for AD-only experiments (setting other dementia diagnosis to NaN). (a) Boxplots display the variability across 20 trained models (from ADNI) for clinical diagnosis prediction assessed using mAUC. The x-axis denotes the test dataset used for evaluation. (b) Statistical significance in the prediction error between all models. Each row shows the statistical difference between a model and all other models. For example, the first row of each 5 x 5 table corresponds to the statistical difference between MinimalRNN and other models – green indicates that MinimalRNN performs better, while red indicates that MinimalRNN performs worse. Therefore, the colors are always flipped between red and green across the diagonal. "*" indicates $p < 0.05$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "**" indicates $p < 0.001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "***" indicates $p < 0.00001$ and statistical significance after multiple comparisons correction (FDR $q < 0.05$). "n.s." indicates no statistical significance ($p \geqslant 0.05$) or did not survive FDR correction.