JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles | Behavioral/Cognitive

Brain Responses to Naturalistic Language During Movie-Watching as Markers of Current and Future Cognitive Function in Older Adults

https://doi.org/10.1523/JNEUROSCI.2139-25.2026

Received: 20 November 2025
Revised: 6 October 2026
Accepted: 5 November 2026

Copyright © 2026 the authors

This Early Release article has been peer reviewed and accepted, but has not been through the composition and copyediting processes.The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Brain Responses to Naturalistic Language During Movie-Watching as Markers of Current and Future Cognitive Function in Older Adults

Running Title: Naturalistic Language and Cognitive Aging

Xianmin Gong[1,2,#,*], Hongwei Zhang[3,#], Yuejiao Wang[3,#], Patrick C. M. Wong[4,5,6], Helene H. Fung[2],
Bonnie Lam[7], Vincent C. T. Mok[7,8,9], Timothy C. Y. Kwok[7,10], Jean Woo[7,11],
Ka Ho Wong[3], Man Wai Mak[12], Helen Meng[1,3,*]

[1] Stanley Ho Big Data Decision Analytics Research Centre, The Chinese University of Hong Kong, Hong Kong SAR, China

[2] Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China

[3] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR, China

[4] Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong SAR, China

[5] Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China

[6] Department of Otorhinolaryngology, Head and Neck Surgery, The Chinese University of Hong Kong, Hong Kong SAR, China

[7] Division of Neurology, Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong SAR, China

[8] Gerald Choa Neuroscience Centre, Margaret K. L. Cheung Research Centre for Management of Parkinsonism, Therese Pei Fong Chow Research Centre for Prevention of Dementia, Lui Che Woo Institute of Innovative Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China

[9] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China

[10] Jockey Club Centre for Osteoporosis Care and Control, The Chinese University of Hong

JNeurosci Accepted Manuscript

Kong, Hong Kong SAR, China

[11] Jockey Club Institute of Aging, The Chinese University of Hong Kong, Hong Kong SAR, China

[12] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China

[#] These authors (Xianmin Gong, Hongwei Zhang, and Yuejiao Wang) contributed equally to this work.

* Correspondence authors: Xianmin Gong (E-mail: xianmingong@cuhk.edu.hk), Helen Meng (E-mail: hmmeng@cuhk.edu.hk)

Number of pages: 32

Number of figures/ tables: 6 figures and 4 tables in the main text.

Number of words for abstract/ introduction/ discussion: 231/ 631/ 1300

Conflict of interest statement: The authors have submitted patent applications containing similar content to the United States Patent and Trademark Office (Application No.: 18/814,269) and the China National Intellectual Property Administration (Application No.: PCT/CN2024/117702). All authors declare no other conflict of interest.

Acknowledgements: 1) This research is part of the research project entitled “Research and Development of Artificial Intelligence in Extraction and Identification of Spoken Language Biomarkers for Screening and Monitoring of Neurocognitive Disorders” (Grant No.: T45-407/19-N) funded by Theme-based Research Scheme, Research Grants Council, University Grants Committee, Hong Kong SAR, China; 2) AI tools (Claude and Copilot) were used to check spelling and grammar in this manuscript.

JNeurosci Accepted Manuscript

Abstract

Monitoring global cognitive function in older adults is crucial for timely detection and intervention of cognitive impairment. While cognitive tests and traditional magnetic resonance imaging have been widely used, more sensitive methods are needed. Conventionally, language production rather than comprehension has been emphasized for monitoring cognitive decline, as comprehension appears preserved in controlled laboratory paradigms. However, naturalistic language comprehension engages multiple aging-sensitive cognitive systems, making language-related brain activations during naturalistic tasks potentially sensitive markers of global cognitive function. This study investigated whether language-related activations during naturalistic movie-watching functional magnetic resonance imaging (mw-fMRI) could effectively monitor older adults' global cognitive function as assessed by the Montreal Cognitive Assessment (MoCA). We recruited 97 (43 females and 54 males) non-demented participants aged 60 years or older to undergo MRI scans and complete the MoCA at baseline; 50 (28 females and 22 males) completed follow-up assessments after two years. Language comprehension during movie watching elicited activations in core frontotemporal and extended language networks that significantly correlated with MoCA scores. Using nested cross-validation, support vector classifiers demonstrated that mw-fMRI activation features effectively classified baseline cognitive status (area under the curve [AUC] = 0.811 ± 0.019) and forecasted two-year cognitive change (AUC = 0.763 ± 0.036), outperforming demographics and traditional MRI features. We localized predictive activations and discussed potential compensatory mechanisms underlying these patterns. Our findings highlight naturalistic language comprehension during mw-fMRI as a biologically interpretable, ecologically valid tool for monitoring global cognitive function in older adults.

JNeurosci Accepted Manuscript

Key words: movie-watching fMRI; language comprehension; neurocognitive disorder; naturalistic fMRI; machine learning

Significance Statement

Detection of cognitive decline in older adults is critical for timely intervention, yet current methods have limitations. Conventional laboratory language tasks suggest comprehension remains intact with aging, leading researchers to focus on language production for monitoring decline. However, real-world language comprehension (e.g., understanding conversations in movies) requires coordinating multiple cognitive processes that are vulnerable to aging. We find that brain activations during naturalistic movie watching effectively detect current cognitive status and forecast future cognitive changes, outperforming traditional brain imaging approaches. This passive, engaging task requires no explicit responses, making it feasible for older adults and clinical populations. Our findings establish ecologically valid neural markers that could help identify individuals at risk for cognitive impairment before they meet diagnostic criteria for dementia.

JNeurosci Accepted Manuscript

Introduction

Aging is accompanied by decline in global cognitive function spanning domains such as memory, attention, perceptual-motor ability, executive function, language, and social cognition (Sachdev et al., 2014). Abnormal decline can progress to neurocognitive disorders such as Alzheimer's disease, underscoring the need for markers that assess current cognitive status and forecast future changes to enable timely intervention.

Standardized cognitive tests such as the Montreal Cognitive Assessment (MoCA) detect cognitive impairment once present (Nasreddine et al., 2005). However, they show limited prognostic value (Arevalo-Rodriguez et al., 2021). Neuroimaging offers potential for detecting brain alterations that precede overt cognitive decline, yet common magnetic resonance imaging (MRI) approaches have constraints (Frisoni et al., 2010). Structural MRI (sMRI) quantifies anatomy correlated with cognition, but anatomy does not fully determine function (Oschwald et al., 2019). Resting-state functional MRI (rs-fMRI) captures intrinsic connectivity related to cognition but is not directly evoked by cognitive processes. It is vulnerable to mind wandering that compromises reliability (Campbell and Schacter, 2017; Ranasinghe and Mapa, 2024). Traditional task-based fMRI directly engages cognitive processes but typically targets single domains. It often requires extensive repetitions and overt responses, limiting feasibility for comprehensive cognitive assessment within a single session (Kemik et al., 2024).

Movie-watching fMRI (mw-fMRI) involving naturalistic language comprehension may address these limitations. Conventionally, language production rather than comprehension has been emphasized for monitoring cognitive decline, as aging impacts production while comprehension appears relatively preserved in controlled laboratory paradigms (Shafto and Tyler, 2014; Shankar et al., 2025). However, naturalistic studies challenge this view by revealing more apparent age-related deficits in language comprehension under conditions approximating everyday listening (Schneider et al., 2002). Naturalistic language comprehension differs fundamentally from controlled word or sentence tasks. It requires

JNeurosci Accepted Manuscript

integrating information across speakers over time, making pragmatic inferences and predictions, managing distractions, and processing emotional, social, and perspective-taking demands. It depends on dynamic coordination between language functions and other cognitive processes (e.g., attention, working memory, control, social cognition) that may be sensitive to aging (Wingfield and Stine-Morrow, 2000; Schneider et al., 2002; Presacco et al., 2016).

Supporting this, studies show that speech during mw-fMRI activates the core frontotemporal language network (the left inferior frontal gyrus, superior temporal gyrus, and middle temporal gyrus) subserving basic linguistic processes, as well as extended language networks spanning cortical and subcortical regions subserving multiple other cognitive processes (Tie et al., 2015; Bottenhorn et al., 2018; Tikka et al., 2018). This coactivation enables simultaneous assessment of multiple cognitive functions. Moreover, passive movie viewing minimizes task demands and reduces head motion, making it well suited for older adults and clinical populations (Bottenhorn et al., 2018; Tikka et al., 2018; Häusler and Hanke, 2021). These features make mw-fMRI a promising candidate for scalable, ecologically valid neural markers of current and future cognitive status.

In the present study, we focused on language-related brain activations during movie watching as interpretable markers of global cognition. Activations were modeled using stimulus-locked speech regressors and prespecified contrasts within a general linear modeling framework. This approach differs from the data-driven methods typically applied to naturalistic fMRI, such as inter-subject correlation and independent component analysis. Although these methods can reveal synchronization and network structure, the resulting markers often lack clear cognitive interpretability. In contrast, our strategy yields neural markers with greater mechanistic and clinical relevance (Gruskin and Patel, 2022; Rajimehr et al., 2024).

We hypothesized that (i) global cognition would be significantly associated with language-related activations within core and extended language networks during mw-fMRI, and (ii) these activations would effectively classify concurrent global cognition and forecast

JNeurosci Accepted Manuscript

longitudinal cognitive changes. By combining ecological validity with interpretable neural markers, we aimed to establish sensitive, mechanistically grounded indices for monitoring global cognitive function. Furthermore, by examining language comprehension in naturalistic rather than controlled laboratory settings, we sought to advance existing research on age-related changes in language processing.

Materials and Methods

Participants

We recruited a cohort of Hong Kong Chinese older adults aged 60 or above (see Table 1). Inclusion criteria were being a native Cantonese speaker capable of reading traditional Chinese characters, having normal or correct-to-normal visual acuity, right-handedness, and eligibility for MRI scanning. Participants were excluded if they had hearing difficulties, dementia as determined by established screening criteria for the Hong Kong population, a history of depression or other neurological diseases, or contraindications for MRI. The dementia screening criteria were: (i) a score ≥ 2 on the Eight-item Informant Interview to Differentiate Aging and Dementia (AD8), an informant-reported measure of cognitive decline (Galvin et al., 2005), and (ii) a score that fell two or more standard deviations below the normative mean on the Hong Kong List Learning Test (HKLLT), a verbal learning and memory test (Chan and Kwok, 1999). The study was approved by the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (Joint CUHK-NTEC CREC; protocol number: 2019.317). The study was performed in accordance with the ethical standards of the Declaration of Helsinki. All participants provided informed consent before the study and were fully debriefed afterward.

During their baseline visit (years 2021-2024), participants underwent MRI scans (including sMRI, rs-fMRI and mw-fMRI) and completed the MoCA. A follow-up visit was scheduled two years after the baseline visit (years 2023-2026), during which participants completed the MoCA and AD8 without MRI scans. To control potential environmental confounds, the same test administrators conducted all cognitive assessments in the same

JNeurosci Accepted Manuscript

testing environment at baseline and follow-up visits. At the time of data analysis, 97 participants (43 females and 54 males) had completed the baseline assessments, of whom 50 (28 females and 22 males) had completed the follow-up assessments. Comparisons between participants who had completed the follow-up visits and those who had not showed no significant differences in sex ratio, education, age, baseline MoCA score, baseline AD8 score and baseline HKLLT score (all p values > 0.1), indicating that the subsample of 50 participants was representative of the whole sample.

Statistical Analysis

The primary objectives of this study were to investigate the association between global cognition and language comprehension, as well as the effectiveness of language-related brain activations during mw-fMRI for monitoring global cognition. To achieve these objectives, we formulated three interrelated key questions: (**Q1**) Can mw-fMRI effectively capture language-related activations in the language brain networks of older adults? (**Q2**) Do these activations reflect concurrent global cognitive function and signal its longitudinal changes? (**Q3**) Can these activations serve as neural markers to classify global cognitive status and forecast its changes, potentially outperforming sMRI and rs-fMRI features?

Global cognition was assessed using the MoCA, which is well-established for assessing global cognitive function (Nasreddine et al., 2005). MoCA scores range from 0 to 30, with a higher score indicating better global cognitive function. We assessed participants' global cognitive function using their MoCA scores at baseline (M_B) and follow-up (M_F). Longitudinal cognitive change (ΔM) was calculated as the difference between the follow-up and baseline MoCA scores ($\Delta M = M_F - M_B$). More positive ΔM values indicated better cognitive preservation, while more negative ΔM values indicated more severe cognitive decline.

For the research question **Q1**, we collected brain data (see the Brain Data Acquisition section below). Voxel-wise mixed-effects general linear modeling (see the Extraction of Brain Features section below) was then performed to examine whether regions within the core and

JNeurosci Accepted Manuscript

extended language networks were significantly activated by movie speech (monologues and dialogues) during the mw-fMRI task. For **Q2**, we analyzed correlations between these language-related activations and both M_B and ΔM. For **Q3**, we conducted two binary classification analyses using supervised machine learning to compare the performance of mw-fMRI against traditional MRI approaches (sMRI and rs-fMRI) in classifying baseline global cognitive status and forecasting global cognitive changes. Although MoCA scores are continuous, we primarily focused on binary classification, with regression analyses reported as supplementary results. This approach was adopted not only to facilitate comparison across imaging feature modalities and with prior studies, but also to enhance clinical interpretability. In the present study, a baseline MoCA score of 20 was used to split the sample. After applying the recommended education adjustment of adding one point for individuals with ≤6 years of education, this threshold corresponded exactly to the locally validated MoCA screening cutoff for cognitive impairment, including mild cognitive impairment and dementia (Yeung et al., 2014). Importantly, this cutoff serves as a screening reference, not as a diagnostic criterion (Yeung et al., 2014). Because no participant in the present study met other criteria for dementia, individuals with lower MoCA scores were likely to have mild cognitive impairment or to be at elevated risk for cognitive impairment. Thus, the baseline classification was intended to distinguish non-demented participants with relatively higher versus lower global cognitive function, with the latter group representing individuals at greater risk for cognitive impairment.

Specifically, we first stratified participants into groups based on median splits. For baseline status, we used the median (= 20) of M_B across participants to divide them into the M_{high} group (with $M_B > 20$) and the M_{low} group (with $M_B \leq 20$). For cognitive changes, we used the median (= 1) of ΔM across participants to divide them into the ΔM_{high} group (with $\Delta M > 1$) and the ΔM_{low} group (with $\Delta M \leq 1$). Notably, consistent with prior studies (Caballero et al., 2021; Cacciamani et al., 2023), not all participants exhibited cognitive decline over two years (as median $\Delta M = 1$). We then examined how well language-related activations during mw-fMRI classified these groups compared to sMRI and rs-fMRI features (see the Supervised

JNeurosci Accepted Manuscript

Machine Learning Analyses section below). Demographics of these groups are shown in Table 1.

[Insert Table 1 around here]

Brain Data Acquisition

All participants underwent whole-brain sMRI, rs-fMRI and mw-fMRI scans during the baseline assessment session. The scans were performed using a Siemens MAGNETOM Prisma 3 Tesla MRI Scanner, equipped with a 64-channel head/neck coil.

sMRI. A high-resolution T1-weighted anatomical image was acquired for each participant using a magnetization-prepared rapid gradient-echo (MPRAGE) sequence: repetition time (TR) = 1800 ms, echo time (TE) = 2.53 ms, flip angle = 8°, voxel size = 0.8 × 0.8 × 0.8 mm^3, matrix size = 272 × 272, field of view (FoV) = 240 × 240 mm^2, number of slices = 208 and total scan duration = 264 seconds.

rs-fMRI. Participants remained resting with their eyes closed during the rs-fMRI scan. T2-weighted blood-oxygen-level-dependent (BOLD) signals were collected using a multiband (factor = 6) gradient-echo planar imaging (EPI) sequence: TR = 900 ms, TE = 24 ms, flip angle = 90°, voxel size = 2 × 2 × 2 mm^3, matrix size = 104 × 104, FoV = 206 × 206 mm^2 and number of slices = 72. For each participant, 560 images were acquired within 514 seconds.

mw-fMRI. Using the same multiband EPI sequence, 736 fMRI images were acquired for each participant when they were watching a 662-second Cantonese movie clip about everyday interactions in a Hong Kong family (https://www.youtube.com/watch?v=KJDQDyJ9Xsg). The movie clip contained a mixture of speech segments (monologues or dialogues) and non-speech segments (Wang et al., 2024a, 2024b). Visual frames (with subtitles) were projected onto a mirror inside the MRI scanner, while the auditory stream was delivered to participants through MRI-compatible earphones.

Language comprehension performance. After the scans, participants completed a post-movie test of 18 multiple-choice questions related to speech content in the movie clip.

JNeurosci Accepted Manuscript

Post-movie test scores ranged from 0 to 18, with a higher score indicating better language comprehension performance. All participants confirmed that it was their first time watching the movie and could see and hear it clearly.

Brain Data Preprocessing

sMRI. SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) and CAT12 (Gaser et al., 2022) toolkits in MATLAB were used to preprocess the data. A standard pipeline was followed, including denoising, bias correction, normalization to the Montreal Neurological Institute (MNI) space, and segmentation into gray matter, white matter and cerebrospinal fluid. Data quality was assessed both manually and using the image quality rating indices generated by the CAT12 toolkit.

rs-fMRI. The data were preprocessed using SPM12 and DPARSF (Yan and Zang, 2010) toolkits in MATLAB. A standard pipeline was followed, including discarding the initial 10 volumes, realignment to correct for head motion, co-registration to sMRI, normalization to the MNI space using DARTEL, regressing out nuisance covariates (Friston's 24 motion parameters, white matter signal, cerebrospinal fluid signal and global signal), detrending, and smoothing with a 6 mm full-width-at-half-maximum isotropic Gaussian kernel.

mw-fMRI. The data were preprocessed using SPM12. A standard pipeline was followed, including discarding the first nine volumes, realignment, co-registration, segmentation, normalization to the MNI space, and spatial smoothing with an isotropic 6 mm full-width-at-half-maximum isotropic Gaussian kernel.

Extraction of Brain Features for Classification Analyses

sMRI. We computed the volumes of gray matter, white matter and cerebrospinal fluid for each participant across 274 regions of interest (ROIs), including 210 cortical areas and 36 subcortical areas from the Brainnetome Atlas (Fan et al., 2016) and 28 cerebellar areas from the SUIT atlas (Diedrichsen, 2006). These 822 ROI-level volumetric measures (n = 274 * 3) served as sMRI features for our subsequent machine learning analyses. As the hippocampal and parahippocampal gyri are particularly sensitive to cognitive aging (Kaye et al., 1997), their volumetric measures (n = 16 * 3) served as another sMRI feature set for

JNeurosci Accepted Manuscript

297 machine learning analyses.

298 ***rs-fMRI.*** We calculated the voxel-wise amplitude of low-frequency fluctuations (ALFF),
299 which has been reported to be effective in detecting and predicting cognitive decline and
300 impairment (de Vos et al., 2018). ALFF quantifies total power within the low-frequency band,
301 where authentic fMRI signals are typically observed. Emerging evidence indicates that
302 authentic fMRI signals can extend beyond 0.1 Hz (the conventionally defined upper limit of
303 the low-frequency band), potentially reaching up to 0.5 Hz or even 0.8 Hz (Boubela et al.,
304 2013; Chen and Glover, 2015). Per the Nyquist theorem, our sampling rate (TR = 0.9 s)
305 limits resolvable signals to 0.55 Hz (= 1 / (2 × TR)). Therefore, the low-frequency band was
306 defined as 0.01-0.6 Hz rather than the conventionally used range of 0.01-0.1 Hz (Chen and
307 Glover, 2015). The current study found that the broader band of 0.01-0.6 Hz significantly
308 improved performance in classifying M_B and forecasting ΔM compared to the 0.01-0.1 Hz
309 band. Consistent with previous studies (de Vos et al., 2018), we calculated the within-
310 participant standardized ALFF (mALFF) by dividing each participant's voxel-wise ALFF
311 values by that participant's mean ALFF across the whole brain. For our subsequent machine
312 learning analyses, the resulting voxel-level mALFF values within the gray matter were used
313 as rs-fMRI features (number of features = 183,255). In addition, we extracted regional
314 homogeneity and functional connectivity features using independent component analysis (de
315 Vos et al., 2018). However, these alternative features underperformed compared to mALFF
316 and were not pursued further.

317 ***mw-fMRI.*** To extract language-related mw-fMRI features, we performed a voxel-wise
318 two-level mixed-effects analysis. At the first level (within-participant level), a general linear
319 model was applied to estimate voxel-wise activations for each participant. Specifically, we
320 regressed the fMRI time course in each voxel on event regressors representing speech
321 (monologue and dialogue) and non-speech movie segments. Head motion parameters were
322 included as nuisance regressors, including six head motion regressors (derived from the
323 realignment procedure during preprocessing) and an additional regressor representing head
324 motion spikes, defined as framewise displacement exceeding 1 mm (Cao et al., 2016). All

JNeurosci Accepted Manuscript

regressors were convolved with the canonical hemodynamic response function without temporal and spatial derivatives. A high-pass filter (cutoff period = 128 s) and a first-order autoregressive model were applied to remove low-frequency drift and control for temporal autocorrelation. Then, voxel-wise *t*-tests were performed to contrast the regression coefficients of speech events versus non-speech events within the gray matter, resulting in a contrast brain map and a corresponding *t*-map for each participant. The contrast brain map contained voxel-wise differences in the regression coefficients between speech and non-speech events, while the *t*-map contained voxel-wise statistical *t*-values indicating the significance of these differences. A voxel with a higher contrast value and a higher *t*-value was more strongly activated by speech (vs. non-speech), indicating greater involvement in language processing during mw-fMRI. The voxel-wise *t*-values in each participant's *t*-value brain map served as mw-fMRI language-related activation features for our subsequent machine-learning analyses. Importantly, the first-level analysis was conducted entirely within each participant's own data, with no information shared across participants.

At the second-level (group-level) analysis, participants' first-level contrast maps were submitted to voxel-wise one-sample *t*-tests to define inclusion masks that constrained the language-related mw-fMRI features used in the subsequent supervised machine learning analyses (see below). Performing this analysis on the full sample would allow test participants to influence feature definition, resulting in data leakage. To avoid this problem, the second-level analysis was nested within the machine-learning cross-validation procedure and was repeated independently within each outer loop: the mask was generated using only participants in the corresponding training set for that loop. Specifically, in each outer loop, a group-level *t*-map was generated from the training participants, with higher *t*-values indicating stronger language-related activation. This *t*-map was thresholded (family-wise error FWE corrected $p < 0.05$ and cluster size > 15 voxels) to create a mask of language-related voxels. The resulting mask was then applied to the first-level *t*-maps of both the training and the held-out test participants within the same fold, ensuring that test participants made no contribution to mask definition. Consequently, a fold-specific mask was derived for

JNeurosci Accepted Manuscript

each outer fold, with mask size varying slightly across folds. For descriptive and visualization purposes only, we also report the mask derived from the full sample (n = 7,884 voxels; Figure 2), but this whole-sample mask was not used for model training or evaluation.

Supervised machine learning analyses: Classifying and forecasting cognition

The extracted brain features (sMRI, rs-fMRI and mw-fMRI features) and demographics (age, sex and education) were used for two binary machine-learning tasks: (i) Task 1 was to classify global cognitive status at baseline (M_{high} versus M_{low}), and (ii) Task 2 was to forecast global cognitive changes over two years (ΔM_{high} versus ΔM_{low}). Research on machine-learning-based classification, particularly in neuroscience, has shown that the nested k-fold cross-validation method offers greater statistical confidence and less biased performance estimates compared to other techniques such as single holdout, k-fold, or train-validation-test (Poldrack et al., 2020; Ghasemzadeh et al., 2024). We employed 10 runs of nested 10-fold cross-validation for both tasks, and used the average AUC on the outer test sets to evaluate task performance (Figure 1). To ensure a fair comparison, identical feature selection procedures and cross-validation procedures were applied across all three types of MRI features. As described above, before feature selection, mw-fMRI features were constrained to language-related regions using a group-level inclusion mask derived from second-level analyses of the training participants within each outer loop. Thus, no information from the held-out test data entered the definition of the mw-fMRI feature space.

Chance-level performance was estimated using random permutation tests. For these tests, we used 100 global label permutations. In each permutation, the class labels were shuffled, and the entire classification pipeline was repeated identically, including fold-specific mask estimation, feature selection, inner-loop hyperparameter tuning, and outer-loop testing. This ensured that the permutation results reflected the full analysis pipeline rather than only the final classifier-fitting step.

[Insert Figure 1 around here]

Feature selection. To reduce feature dimensions and mitigate the infamous “curse of dimensionality,” a consistent two-step feature selection process was applied separately to

JNeurosci Accepted Manuscript

the three types of brain features (Figure 1). In the first step, we retained features that significantly correlated with the outcome variable (i.e., M_B or ΔM) at a significance level of $p < 0.01$. This step was to ensure that relevant features were selected for further analysis. In the second step, an L1-penalty was applied to eliminate redundant or minimally contributing features using the *SelectFromModel* function from the *scikit-learn* toolkit in Python. Specifically, the retained features from the first step were input into a linear support vector classifier (SVC) with L1-regularization, where the regularization parameter C_1 was treated as a hyperparameter. Only features with coefficients exceeding the mean coefficient of the linear SVC were retained. It is important to note that the coefficients of this linear SVC were used solely for L1-penalty-based feature selection. This SVC model itself was a by-product and did not contribute directly to the subsequent classification tasks.

Model estimation and evaluation. To estimate model performance and ensure robustness, we implemented 10 runs of nested 10-fold cross-validation (10 outer iterations per run) using the *RepeatedStratifiedKFold* function from the *scikit-learn* toolkit. Different random seeds were applied across runs to enhance robustness. Nested cross-validation involves two loops: an outer loop for performance evaluation and an inner loop for hyperparameter tuning, ensuring unbiased estimates of model generalization and fairness in comparing different brain features (Poldrack et al., 2020; Ghasemzadeh et al., 2024). Stratification was used in both the outer and inner loops to preserve comparable class proportions across folds, particularly for the smaller forecasting cohort, and to avoid highly imbalanced or single-class test folds.

In each outer iteration of each run, participants' raw features (before feature selection) were split into an outer training set (90% of participants) and an outer test set (10% of participants). We further implemented an inner cross-validation loop with five iterations nested within each outer iteration. In the inner loop, the outer training set was split into an inner training set (80% of the outer training set) and an inner validation set (20% of the outer training set). The inner training set underwent the two-step feature selection process (with hyperparameter C_1), and the selected features were used to train another linear SVC model

JNeurosci Accepted Manuscript

for validation, with a regularization parameter C_2 treated as a hyperparameter. Candidate values for each hyperparameter included 0.1, 1, 5 and 10. For a given combination of hyperparameters (C_1, C_2), the fitted model was evaluated on the inner validation set using the AUC metric. In this way, the inner cross-validation loop was repeated for each candidate combination of hyperparameters (C_1, C_2), and the combination yielding the highest average AUC across the five inner iterations was designated as the optimal set of hyperparameters [C_1^*, C_2^*]. This [C_1^*, C_2^*] combination was then passed to the outer iteration.

Using the optimal set of hyperparameters [C_1^*, C_2^*], the outer training set underwent the previously described two-step feature selection, and a new linear SVC model was trained for the selected features. Its performance was then evaluated on the outer test set by AUC. For each model, the ten outer-fold AUCs were averaged within each cross-validation repeat, and the reported mean and standard deviation were then calculated across the ten repeat-level mean AUCs. It is important to highlight that hyperparameter tuning was only conducted using the data of inner iterations, which did not overlap with the outer test sets. This approach ensures that nested k-fold cross-validation provides an unbiased estimate of the model's generalization performance without depending on an external validation dataset (Poldrack et al., 2020; Ghasemzadeh et al., 2024).

Feature specificity analyses. To evaluate whether the predictive performance of mw-fMRI features was specific to language-related activations, we conducted three sets of control analyses. (1) We assessed model performance using voxel-wise activation features sampled from outside the mw-fMRI language regions. A set of non-language voxels matched in size to the primary language-feature set (n = 7,884 voxels) was randomly selected, and the corresponding voxel-wise t-values were submitted to the same feature-selection and nested cross-validation procedures used in the main analyses. (2) Given that head motion during fMRI may be associated with cognitive function in older adults (Haller et al., 2014; Geerligs et al., 2017; Hausman et al., 2022), we examined whether motion-related variables could account for the observed predictive effects/ Specifically, we evaluated the classification performance of models using the six realignment parameters alone, as well as

JNeurosci Accepted Manuscript

models combining these motion parameters with the mw-fMRI features. (3) To determine whether the predictive effects of mw-fMRI reflected generic functional connectivity information rather than specific language-related activation patterns, we performed additional classification analyses using functional connectivity features derived from mw-fMRI data. Functional connectivity was estimated among 116 AAL atlas regions and submitted to the same nested cross-validation procedures used in the main analyses. Together, these control analyses tested whether language-related mw-fMRI activation features provided predictive information beyond that captured by non-language activations, head motion, or whole-brain functional connectivity. Evidence for specificity of language-related activations would be supported if language-related mw-fMRI features demonstrated superior classification performance relative to these alternative feature sets.

Localization of predictive mw-fMRI features. To identify reliable predictive features from our models and increase the models' interpretability, we analyzed the spatial distribution and functional significance of selected mw-fMRI features (i.e., language-related activations in brain voxels, with each voxel corresponding to one feature). These analyses were conducted separately for the classification and forecasting tasks.

For each voxel, we calculated its selection probability as the ratio between the number of times it was selected during nested cross-validation and the total number of iterations. This voxel selection frequency provided an initial measure of each voxel's predictive importance. Given the spatial nature of fMRI data and the fact that brain functions are typically organized in distinct anatomical regions, we identified clusters of predictive voxels. As typically defined in the literature, a cluster refers to a group of spatially continuous voxels within anatomical regions (delineated by the AAL3 atlas). To minimize the impact of sporadic selections while maintaining sensitivity, we employed a minimum cluster size threshold of 5 voxels (Woo et al., 2014; Ge et al., 2025). This threshold helped eliminate isolated voxel selections that might represent noise rather than meaningful predictive features.

For each identified cluster, we computed a cluster-wise selection frequency defined as the proportion of cross-validation iterations where the cluster contained at least one selected

JNeurosci Accepted Manuscript

voxel. This metric allowed us to assess the consistency of feature selection across different data partitions during the classification analyses. To characterize each cluster's spatial properties, we identified its peak coordinate (in MNI space) as the location of the voxel with the highest selection frequency within the cluster. To evaluate the functional relevance of these clusters, we implemented two correlation analyses: (i) We calculated the mean correlation between activations of selected voxels within each cluster and our primary outcome measures (M_B or ΔM); (ii) We computed correlations with language performance scores from the post-movie test to establish the clusters' relationship with language processing. These analyses provided complementary evidence for both the predictive utility and functional significance of the identified clusters.

Code Availability

Data analysis code is available at https://github.com/hwzzhang/mwfmri_analysis.

Results

Language-related Activations in mw-fMRI

To examine whether mw-fMRI captures language-related activations in the language brain networks of older adults (research question **Q1**), we mapped brain activations significantly associated with language comprehension during mw-fMRI. While data-driven approaches (e.g., inter-subject correlation and ICA) have provided valuable insights into brain function during naturalistic movie watching (Gruskin and Patel, 2022; Rajimehr et al., 2024), they often yield results that lack clear cognitive interpretability. To enhance interpretability, we performed a voxel-wise general linear model analysis to identify gray matter voxels that were more strongly activated during movie segments with speech than those without. This analysis resulted in the group-level statistical *t*-map, where higher *t*-values indicated stronger language-related activations across participants. Brain regions showing significant language-related activations are displayed in Figure 2 Part A. To facilitate cross-study comparisons, we labeled all brain regions (Figure 2 Part D) using the widely adopted Automated Anatomical Labeling 3 (AAL3) atlas (Rolls et al., 2020).

JNeurosci Accepted Manuscript

To validate the relevance of these activations (Figure 2 Part A) to language network engagement, we first compared them qualitatively with established language networks documented in the literature (Price, 2012; Friederici and Gierhan, 2013; Deldar et al., 2020). The activated regions during mw-fMRI encompassed the well-documented core frontotemporal language network responsible for language processing, including the left inferior frontal gyrus (AAL3 labels: Frontal_Inf_Oper_L, Frontal_Inf_Orb_L and Frontal_Inf_Tri_L), superior temporal gyrus (Temporal_Sup_L) and middle temporal gyrus (Temporal_Mid_L; Price, 2012; Friederici and Gierhan, 2013; Deldar et al., 2020). Significant activations during mw-fMRI were also observed in the extended language network supporting (but not directly responsible for) language processing, including multiple cortical and subcortical regions (Price, 2012; Friederici and Gierhan, 2013; Deldar et al., 2020). Hereafter, we refer to these activated regions as the **mw-fMRI language regions** for conciseness.

To provide a more rigorous quantitative evaluation of language-network engagement, we compared the mw-fMRI language regions with the Language Network Atlas (LanA), a validated probabilistic functional atlas derived from over 800 individuals (Lipkin et al., 2022). LanA assigns each brain voxel a probability value (P), reflecting its probability of engagement in language processing. We defined two reference networks from the LanA using different probability thresholds: a central LanA network with a threshold of $P > 0.5$ (Figure 2 Part B) and an extended LanA network with a threshold of $P > 0.2$ (Figure 2 Part C). Since the number of voxels in the reference LanA networks varies with probability thresholds, our comparisons focused on voxel distribution patterns rather than absolute voxel counts. The comparisons showed that the mw-fMRI language regions covered all AAL3-defined brain regions (albeit not precisely the same voxels) within the central LanA network (Parts B and D in Figure 2), with voxel distributions across brain regions showing significant consistency between the two ($r = 0.61$, $p < 0.001$). Furthermore, the mw-fMRI language regions exhibited substantial overlap with the extended LanA network in both anatomical coverage and voxel distribution patterns ($r = 0.80$, $p < 0.001$; Parts C and D in

JNeurosci Accepted Manuscript

Figure 2).

To assess whether these findings depended on the selected LanA probability thresholds, we further recalculated the mw-fMRI–LanA correspondence across 11 thresholds ranging from $P > 0.1$ (lower threshold Ps would result in inclusion of overly nonspecific voxels) to $P > 0.6$ (higher threshold Ps would result in overly sparse networks). Across these thresholds (Figure 3), the regional voxel-distribution correlations remained consistently high ($r = 0.65$-0.80, all $p \leq 9.5*10^{-21}$). Spatial-overlap measures showed a similar robust pattern, with the Dice coefficient remaining stable across intermediate thresholds and the proportion of LanA voxels covered by mw-fMRI regions increasing as the LanA threshold became more stringent. These results demonstrate the effectiveness of mw-fMRI in capturing language-related activations in older adults.

[Insert Figures 2 & 3 around here]

Correlations Between mw-fMRI and Global Cognition

To test whether language comprehension during mw-fMRI relates to global cognitive function (research question **Q2**), we examined the correlations between cognitive measures (M_B and ΔM) and two types of language-related measures from mw-fMRI: language comprehension performance at the behavioral level and language-related activations at the brain level.

Correlational analyses (Figure 4 Part A) revealed a strong positive correlation between language performance and M_B ($r = 0.57$, $p < 0.001$), confirming the link between language processing ability and global cognitive function. However, neither M_B nor language performance showed a significant correlation with ΔM ($p > 0.05$), suggesting the limitations of behavioral cognitive assessments in forecasting cognitive changes.

At the brain level, we examined the correlations between participants' language-related activation strength (in each voxel of the mw-fMRI language regions) and both M_B (Figure 4 Part B) and ΔM (Figure 4 Part C). Only voxels with significant correlations are displayed ($p < 0.05$)—note that the threshold p was uncorrected to provide a more comprehensive view of the distribution of significant voxels. Voxels across widespread mw-

JNeurosci Accepted Manuscript

fMRI language regions showed significant positive correlations with M_B. These regions spanned the core frontotemporal language network (including the left inferior frontal gyrus, middle temporal gyrus and superior temporal gyrus) and the extended language network (such as the right middle temporal gyrus, superior temporal gyrus and cerebellar areas). This result indicated that participants with better global cognitive function engaged these voxels more strongly during language processing in the mw-fMRI task.

Fewer voxels showed significant correlations with ΔM than with M_B, suggesting that language-related activations more strongly reflect concurrent cognitive status than forecasting cognitive changes. Notably, the pattern of correlations with ΔM varied systematically across regions: Negative correlations were predominant in the core frontotemporal language network, while positive correlations emerged in the extended language network (including the right precentral gyrus, lingual areas and cerebellum). This dissociation suggests distinct relationships between language network components and future cognitive decline: Stronger activations in the extended network during mw-fMRI were associated with less decline, whereas stronger activations in the core network were linked to more decline.

[Insert Figure 4 around here]

Classifying Global Cognition and Forecasting Its Changes

To examine whether language-related mw-fMRI features serve as neural markers for monitoring global cognition more effectively compared to conventional MRI methods (sMRI and rs-fMRI) in monitoring global cognitive function (research question **Q3**), we evaluated the performance of demographics and each type of MRI features on two binary classification tasks: (i) distinguishing baseline global cognitive status (i.e., M_{high} versus M_{low}) and (ii) forecasting global cognitive changes (i.e., ΔM_{high} versus ΔM_{low}). AUC, along with other metrics (accuracy, specificity, and sensitivity), was used to evaluate model performance. For each model, the ten outer-fold metrics (e.g., AUCs) were averaged within each cross-validation repeat, and the reported mean and standard deviation were then calculated across the ten repeat-level mean metrics. AUCs were compared across feature sets using

JNeurosci Accepted Manuscript

matched paired permutation contrasts, as all feature sets were evaluated using the same participants, cross-validation folds, and random seeds. This paired approach preserves the dependency structure across models and directly tests whether one feature set outperformed another under identical resampling conditions.

Classifying M_{high} versus M_{low}. As shown in Table 2 and Figure 5 (Part A), mw-fMRI features achieved higher classification performance (AUC = 0.811 ± 0.019) than demographic variables (age, sex and education; AUC = 0.758 ± 0.027), sMRI (AUC = 0.705 ± 0.038) and rs-fMRI (AUC = 0.718 ± 0.034; all $p < 0.001$). All features exceeded chance-level performance (all $p < 0.001$), with sMRI and rs-fMRI performance approximating previous studies using comparable methods (see Supplementary Information). Combining mw-fMRI with M_B yielded a higher AUC (0.793 ± 0.035), but this increase was not statistically significant; and combining mw-fMRI with other feature sets also did not significantly improve performance (all $p > 0.05$). These results demonstrate the superiority of mw-fMRI in classifying older adults' concurrent global cognitive status.

Forecasting ΔM_{high} versus ΔM_{low}. As shown in Table 3 and Figure 5 (Part B), mw-fMRI (AUC = 0.763 ± 0.036) outperformed rs-fMRI (AUC = 0.702 ± 0.068; $p < 0.05$), and both outperformed other features (demographics, baseline MoCA M_B, and sMRI), as well as exceeded chance-level performance (all $p < 0.001$). Demographic variables (AUC = 0.566 ± 0.065) and sMRI (AUC = 0.363 ± 0.052) performed no better than chance level. Combining mw-fMRI with other features did not significantly improve its performance (all $p > 0.05$). These results demonstrate the superiority of mw-fMRI over both established cognitive tests (such as the MoCA) and traditional MRI (sMRI and rs-fMRI) in forecasting global cognitive changes in older adults.

[Insert Table 2, Table 3, and & Figure 5 around here]

Specificity Analyses

Performance of voxels outside mw-fMRI language regions. To verify that the mw-fMRI features are specifically related to the mw-fMRI language regions, we conducted control analyses using a similar number of randomly sampled voxel-wise *t*-values (n = 7,884

JNeurosci Accepted Manuscript

voxels) from outside the mw-fMRI language regions. Applying the same feature selection and cross-validation procedures, these control features yielded significantly lower performance in both the classification (AUC = 0.699 ± 0.041; $p < 0.01$) and forecasting tasks (AUC = 0.456 ± 0.090; $p < 0.001$). These results demonstrate that the predictive power of mw-fMRI stemmed specifically from language-related activations, rather than nonspecific brain activity.

Potential confounding effects of head motion*.* Because head motion during fMRI has been shown to correlate with cognitive function in older adults (Haller et al., 2014; Geerligs et al., 2017; Hausman et al., 2022), we examined its potential confounding effects. First, we compared motion levels between mw-fMRI and rs-fMRI (Table S4; Figure S2, left panel). Framewise displacement (FD; Power et al., 2015), an overall motion index derived from the six realignment parameters, was greater during mw-fMRI (mean = 0.252 mm) than during rs-fMRI (mean = 0.220 mm). FD during both scans was negatively correlated with M_B (r = -0.255 for mw-fMRI FD and r = -0.355 for rs-fMRI FD), but neither was significantly correlated with ΔM (ps > 0.10; see Table S5). We further examined whether the six head-motion parameters contributed to the classification performance of fMRI features. The results (Table S6; Figure S2, right panel) showed that both head motion of mw-fMRI (AUC = 0.487 ± 0.044 and 0.493 ± 0.039 for classifying M_B and ΔM) and that of rs-fMRI (AUC = 0.553 ± 0.044 and 0.340 ± 0.057 for classifying M_B and ΔM) yielded only modest classification performance. After combining with head motion parameters, the performance of neither mw-fMRI (AUC = 0.809 ± 0.012 and 0.707 ± 0.051 for classifying M_B and ΔM) nor rs-fMRI (AUC = 0.716 ± 0.027 and 0.685 ± 0.071 for classifying M_B and ΔM) was improved. mw-fMRI outperformed rs-fMRI regardless of whether head motion was included in the models. These results demonstrate that the predictive power of mw-fMRI was not attributable to head motion, supporting the specificity of mw-fMRI language-related features.

Classification performance based on mw-fMRI functional connectivity. To further assess the specificity of the classification performance of language-related mw-fMRI features, we performed additional classification analysis using functional connectivity

JNeurosci Accepted Manuscript

features, which have been associated with trait-like cognitive and emotional phenotypes (Finn and Bandettini, 2021). Functional connectivity was estimated separately from preprocessed mw-fMRI and rs-fMRI data using the 116-region AAL atlas. Regional time series were extracted by averaging the signal across voxels within each atlas-defined region after resampling the atlas to the functional grid using nearest-neighbor interpolation. Regional time series were linearly detrended and band-pass filtered using a fifth-order Butterworth filter (0.01–0.10 Hz). Nuisance effects were regressed using Friston's 24 motion parameters, mean white-matter and cerebrospinal-fluid signals, mean global signal, and spike regressors for volumes with framewise displacement (FD) exceeding 0.5 mm. Functional connectivity was quantified as the Pearson correlation between each pair of denoised regional time series. The 6,670 unique off-diagonal elements of each 116*116 connectivity matrix were Fisher *z*-transformed, yielding a 97*6,670 participant-by-edge feature matrix.

The mw-fMRI and rs-fMRI feature matrices were entered separately into classification analyses using the same 10-run nested cross-validation framework described above (Figure 1). The results showed that classification performance based on functional connectivity was near chance. For mw-fMRI connectivity, the AUC was 0.480 $\pm$ 0.073 for M_B classification and 0.560 $\pm$ 0.089 for ΔM prediction. For rs-fMRI connectivity, the AUC was 0.599 $\pm$ 0.046 for M_B classification and 0.405 $\pm$ 0.043 for ΔM prediction. These values were lower than those obtained using language-related mw-fMRI features and rs-fMRI mALFF features, both of which exceeded an average AUC of 0.7 (Tables 2 and 3). Thus, the superior performance of language-related mw-fMRI features was unlikely to reflect generic functional connectivity information, supporting the specificity of mw-fMRI language-related features.

Spatial Distribution of Predictive mw-fMRI Features

The top-selected clusters (selection frequencies ≥ 74% and 37% for the classification and forecasting tasks, respectively) are visualized in brain maps in Figure 6 (Parts A and C). These frequency thresholds were chosen based on the "elbow" criterion, a commonly used

JNeurosci Accepted Manuscript

heuristic in determining the number of clusters and features to be selected (Tibshirani and Walther, 2005). Specifically, we observed substantially lower selection frequencies for the remaining clusters, suggesting that they contributed significantly less to the classification performance. The peak coordinates (the location of the most frequently selected voxel within each cluster) and selection frequencies of the top-selected clusters are detailed in Figure 6 (Parts B and D). We then investigated each cluster's contribution to classification by computing the average correlations between its selected voxels and the outcome measures (M_B or ΔM). To understand each cluster's functional relevance, we further examined the average correlations between the selected voxels in the cluster and language performance, measured in terms of the post-movie test scores.

Features for classifying baseline cognitive status. For classifying M_{high} versus M_{low}, as shown in Figure 6 (Parts A and B), the most frequently selected clusters were from the core frontotemporal (i.e., the left middle and superior temporal gyri) and extended language networks (such as the right middle and superior temporal gyri, right temporal pole and right cerebellum). These clusters consistently exhibited positive correlations with both the M_B and language performance (post-movie test), indicating that stronger activations were related to better baseline global cognitive function and language processing abilities.

Features for forecasting cognitive changes. For forecasting ΔM_{high} versus ΔM_{low}, the most frequently selected clusters showed different correlation patterns, as shown in Figure 6 (Parts C and D). We observed that: (i) Within the core frontotemporal language network, the predictive clusters (i.e., the left inferior frontal gyrus and middle temporal gyrus) showed negative correlations with ΔM (r = -.40 to -.49), but weak positive correlations with language performance (r = .10 to .17). This pattern indicates that stronger activations in these language-related brain regions contributed minimally to language performance, but forecasted greater cognitive decline. (ii) In contrast, within the extended language network, the most predictive clusters (i.e., the right precentral gyrus) showed a positive correlation with ΔM (r = 0.45) and a weak positive correlation with language performance (r = 0.11), indicating that stronger activations in these clusters enhanced language processing and

JNeurosci Accepted Manuscript

forecasted less cognitive decline. These results align with our correlational analyses reported above in Figure 4 Part C, suggesting a functional dissociation in cognitive forecasting. More specifically, heightened activations in the core language network during mw-fMRI signaled impending cognitive decline, while those in the extended language network signaled cognitive preservation. This dissociation could be interpreted through the neural compensation mechanism, to be elaborated in the Discussion section.

[Insert Figure 6 around here]

Supplementary Regression Analyses: Treating MoCA Scores as Continuous Outcomes

As supplementary analyses, we further evaluated whether mw-fMRI features predicted continuous MoCA outcomes using machine-learning regression. The analytic procedures were similar to those used in the binary classification analyses, except that the categorical group labels were replaced with continuous outcomes: baseline MoCA score (M_B) and longitudinal change in MoCA score (ΔM). Specifically, we used the same feature sets and nested cross-validation workflow as in the classification analyses, but implemented regression models, including support vector regression (SVR), Ridge regression, and Elastic Net regression. Model performance was evaluated using mean absolute error (*MAE*) and R^2, where lower *MAE* indicates smaller prediction error and higher R^2 indicates better prediction.

The results are displayed in Table 4. For the prediction of M_B, mw-fMRI features yielded the best performance when Ridge regression was implemented (*MAE* = 2.992, R^2 = 0.110). This performance was comparable to that of sMRI features (*MAE* = 3.048, R^2 = 0.114), slightly better than that of rs-fMRI features (*MAE* = 3.128, R^2 = 0.055), but lower than that of demographics (*MAE* = 2.776, R^2 = 0.223). For the prediction of ΔM, mw-fMRI features showed limited predictive performance, with R^2 near or below zero. Adding demographics, baseline MoCA, sMRI, or rs-fMRI features to mw-fMRI did not substantially improve prediction of either M_B or ΔM. These findings suggest that mw-fMRI features were useful for classifying clinically interpretable cognitive-status groups, while their ability to predict exact continuous MoCA scores or longitudinal changes was limited.

JNeurosci Accepted Manuscript

Discussion

This study demonstrates that global cognitive function modulates brain activations associated with language comprehension during naturalistic movie-watching fMRI (mw-fMRI), and these activations serve as effective neural markers for monitoring (classifying and forecasting) global cognitive function in non-demented older adults. Compared to traditional MRI approaches (sMRI and rs-fMRI), mw-fMRI uniquely captures brain activation patterns during the natural cognitive processing of language, providing more direct and ecologically valid measures of cognitive functions. Specifically, our findings reveal that: (i) mw-fMRI effectively engaged both the core frontotemporal and extended language networks; (ii) language-related activations in these networks correlated with global cognition (M_B) and its changes over two years (ΔM); and (iii) these activations outperformed sMRI and rs-fMRI features in classifying global cognitive status (M_{high} versus M_{low}) and forecasting its changes (ΔM_{high} versus ΔM_{low}).

Our mw-fMRI results revealed significant language-related activation in both core and extended language networks of older adults, consistent with previous studies in younger populations (Tie et al., 2015; Tikka et al., 2018; Häusler and Hanke, 2021; Yao et al., 2022; Rajimehr et al., 2024). The core network, anchored by the left inferior frontal gyrus (the Broca's area), middle temporal gyrus and superior temporal gyrus (the Wernicke's area), is predominantly responsible for core linguistic processes such as semantic and syntactic processing (Price, 2012; Friederici and Gierhan, 2013; Deldar et al., 2020). Beyond this core network, we observed robust activations in the extended language network, comprising the right-hemispheric homologues of the core frontotemporal regions, multiple additional cortical regions on both hemispheres and subcortical structures — particularly the cerebellum. The right-hemispheric homologues process paralinguistic features critical for natural communication, such as visuospatial context, emotional prosody and pragmatic aspects of language (Price, 2012; Friederici and Gierhan, 2013; Deldar et al., 2020). While not directly engaged in linguistic processing, the additional cortical areas provide crucial cognitive

JNeurosci Accepted Manuscript

support (such as attention, working memory and executive control mechanisms) for natural language processing. The cerebellar involvement may reflect its role in the temporal sequencing of linguistic elements, articulatory planning and semantic prediction during natural language comprehension (LeBel and D'Mello, 2023). These distributed activations reflect the complex brain architecture underlying natural language processing, where multiple cognitive operations unfold dynamically and simultaneously. This complexity of natural language processing mirrors the integrative nature of global cognitive function, providing a neurobiological basis for using language-related activations in mw-fMRI to monitor global cognitive function.

Indeed, our analyses demonstrated that language-related activations in mw-fMRI significantly correlated with global cognitive function and its longitudinal changes over two years. Furthermore, mw-fMRI language-related activations outperformed the widely used sMRI and rs-fMRI features in classifying global cognitive status and forecasting its changes. Our findings include: (i) In classifying cognitive status, mw-fMRI achieved significantly higher performance (AUC = 0.811 ± 0.019) than sMRI (AUC = 0.705 ± 0.030) and rs-fMRI (AUC = 0.718 ± 0.034). A systematic review (Figure S1 and Table S1) of studies using similar algorithms found that AUC values ranged from 0.54 to 0.90 (mean 0.70) for rs-fMRI features (Table S2) and from 0.65 to 0.84 (mean 0.73) for sMRI features (Table S3). These values were comparable to those obtained in the present study. Additional details are provided in the Supplementary Information. (ii) In forecasting cognitive change, mw-fMRI also demonstrated better performance (AUC = 0.763 ± 0.036) than sMRI (AUC = 0.363 ± 0.052) and rs-fMRI (AUC = 0.702 ± 0.068). Notably, neither sMRI nor the baseline MoCA score (M_B) performed above chance level in forecasting cognitive changes, suggesting the limited sensitivity of structural brain measures (sMRI) and behavioral cognitive assessments (e.g., the MoCA) in forecasting subtle cognitive changes over the relatively short two-year duration in our study. The superior performance of mw-fMRI highlights its practical utility in monitoring cognitive function.

To better understand the brain mechanisms underlying these results, we further

JNeurosci Accepted Manuscript

localized the most predictive mw-fMRI features (i.e., language-related activations) and examined their contributions to the classification and forecasting tasks. The results revealed distinct contribution patterns that could be interpreted using the **neural compensation hypothesis** (Cabeza et al., 2002; Park and Reuter-Lorenz, 2009). This hypothesis posits that aging brains may recruit additional resources to enhance task performance through hyperactivation in the core network or engagement of extended networks (Cabeza et al., 2002; Park and Reuter-Lorenz, 2009). Successful compensation (i.e., effective hyperactivation) results in enhanced performance, while failed compensation (i.e., ineffective hyperactivation) brings no performance benefits. Our results demonstrated the occurrence of successful and failed compensation in different classification tasks and brain regions. In classifying the baseline cognitive status, selected activation features positively correlated with both global cognitive function and language task performance (the post-movie test score), suggesting successful compensation. In forecasting cognitive changes, selected features showed modest correlations (with *r* around 0.1) with language task performance, suggesting failed or ineffective compensation.

Moreover, the abovementioned failed or ineffective compensation was largely linked to greater cognitive decline. Similar patterns of ineffective hyperactivation preceding cognitive decline have been reported previously (Bonnet et al., 2010; Hantke et al., 2013). Such compensatory failure becomes increasingly prevalent with advancing age and cognitive impairment (Park and Reuter-Lorenz, 2009; Hantke et al., 2013). These results suggest that cognitive changes are signaled not only by activation magnitude, but also by how effectively these activations support task performance. The differential predictive roles of the core versus extended networks highlight that flexible recruitment of alternative neural resources to support functioning is crucial for cognitive preservation during aging.

While mw-fMRI showed promising performance in distinguishing binary cognitive-status and cognitive-change groups, its performance on predicting continuous MoCA scores was more limited, particularly for longitudinal change. This discrepancy likely reflects the greater difficulty of predicting fine-grained individual variation in continuous cognitive

JNeurosci Accepted Manuscript

outcomes. In addition, regression performance might have been constrained by characteristics of the present sample, such as the relatively restricted range of MoCA scores, the small magnitude of longitudinal change over two years, and the modest sample size. Therefore, the regression results should not be interpreted as contradicting the classification findings, but rather as indicating that mw-fMRI may be more suited for clinically interpretable classification than for precise prediction of continuous cognitive scores.

Several limitations in this study should be acknowledged. First, we focused specifically on language-related activations in mw-fMRI. Future work could examine activations in other brain networks to more comprehensively understand the utility of mw-fMRI in monitoring cognitive function. Second, our classification analyses employed linear support vector classifiers with rigorous nested cross-validation. Alternative approaches, such as deep learning, might improve performance with more data and warrant further investigation. Third, the present movie-watching paradigm included subtitles, and therefore should be interpreted as an audiovisual narrative-viewing task with concurrent spoken- and written-language input rather than as a purely visual movie-watching task. Although subtitles reflect a common naturalistic viewing condition and may support comprehension, especially in older adults, the current design cannot determine whether the observed prediction of cognitive performance was driven primarily by audiovisual narrative processing, spoken-language comprehension, written-language processing, or their combination. Lastly, as we only recruited non-demented older adults, future research should investigate the potential of mw-fMRI in monitoring dementia and other neurocognitive disorders.

Conclusion

This study advances our understanding of age-related changes in language processing by examining naturalistic rather than controlled laboratory comprehension, revealing critical links between global cognitive function and real-world language processing in older adults. While cognitive tests like the MoCA effectively assess current cognitive status, they have limited predictive power for future decline (Kandiah et al., 2014; Arevalo-

JNeurosci Accepted Manuscript

Rodriguez et al., 2021). Our findings show that mw-fMRI outperforms traditional MRI methods (sMRI and rs-fMRI) in both classifying concurrent cognitive status and forecasting future cognitive changes. By revealing how naturalistic language comprehension engages multiple cognitive systems and how compensatory neural mechanisms relate to cognitive preservation, this work establishes mw-fMRI as a biologically interpretable and ecologically valid neuroimaging marker. These results provide a foundation for further validation and expansion of mw-fMRI applications for monitoring cognitive impairment.

References

Arevalo-Rodriguez I, Smailagic N, Roqué-Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A, Pedraza OL, Cosp XB, Cullum S (2021) Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). Cochrane Database Syst Rev.

Bonnet MC, Allard M, Dilharreguy B, Deloire M, Petry KG, Brochet B (2010) Cognitive compensation failure in multiple sclerosis. Neurology 75:1241–1248.

Bottenhorn KL, Flannery JS, Boeving ER, Riedel MC, Eickhoff SB, Sutherland MT, Laird AR (2018) Cooperating yet distinct brain networks engaged during naturalistic paradigms: A meta-analysis of functional MRI results. Netw Neurosci 3:27–48.

Boubela RN, Kalcher K, Huf W, Kronnerwetter C, Filzmoser P, Moser E (2013) Beyond noise: using temporal ICA to extract meaningful information from high-frequency fMRI signal fluctuations during rest. Front Hum Neurosci 7:168.

Caballero HS, McFall GP, Wiebe SA, Dixon RA (2021) Integrating three characteristics of executive function in non-demented aging: Trajectories, classification, and biomarker predictors. J Int Neuropsychol Soc 27:158–171.

Cabeza R, Anderson ND, Locantore JK, McIntosh AR (2002) Aging gracefully: compensatory brain activity in high-performing older adults. Neuroimage 17:1394–1402.

Cacciamani F, Bercu A, Bouteloup V, Grasset L, Planche V, Chêne G, Dufouil C, Group

JNeurosci Accepted Manuscript

MCS (2023) Understanding factors associated with the trajectory of subjective cognitive complaints in groups with similar objective cognitive trajectories. Alzheimers Res Ther 15:205.

Campbell KL, Schacter DL (2017) Ageing and the resting state: is cognition obsolete? Lang Cogn Neurosci 32:661–668.

Cao W, Cao X, Hou C, Li T, Cheng Y, Jiang L, Luo C, Li C, Yao D (2016) Effects of cognitive training on resting-state functional connectivity of default mode, salience, and central executive networks. Front Aging Neurosci 8:70.

Chan AS, Kwok I (1999) Hong Kong List Learning Test (HKLLT): manual and preliminary norms. Hong Kong: Department of Psychology, The Chinese University of Hong Kong.

Chen JE, Glover GH (2015) BOLD fractional contribution to resting-state functional connectivity above 0.1 Hz. Neuroimage 107:207–218.

de Vos F, Koini M, Schouten TM, Seiler S, van der Grond J, Lechner A, Schmidt R, de Rooij M, Rombouts SARB (2018) A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. Neuroimage 167:62–72.

Deldar Z, Gevers-Montoro C, Khatibi A, Ghazi-Saidi L (2020) The interaction between language and working memory: a systematic review of fMRI studies in the past two decades. AIMS Neurosci 8:1–32.

Diedrichsen J (2006) A spatially unbiased atlas template of the human cerebellum. Neuroimage 33:127–138.

Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, Yang Z, Chu C, Xie S, Laird AR (2016) The human brainnetome atlas: a new brain atlas based on connectional architecture. Cereb Cortex 26:3508–3526.

Finn ES, Bandettini PA (2021) Movie-watching outperforms rest for functional connectivity-based prediction of behavior. Neuroimage 235:117963.

Friederici AD, Gierhan SME (2013) The language network. Curr Opin Neurobiol 23:250–254.

Frisoni GB, Fox NC, Jack Jr CR, Scheltens P, Thompson PM (2010) The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 6:67–77.

JNeurosci Accepted Manuscript

Galvin JE, Roe CM, Powlishta KK, Coats MA, Muich SJ, Grant E, Miller JP, Storandt M, Morris JC (2005) The AD8: a brief informant interview to detect dementia. Neurology 65:559–564.

Gaser C, Dahnke R, Thompson PM, Kurth F, Luders E, Initiative ADN (2022) CAT–A computational anatomy toolbox for the analysis of structural MRI data. biorxiv:2006–2022.

Ge S, Xiang Y, Hua X, Wang Z, Hu Q, Guo Y, Huang J, Zhao C, Wu J, Wang X (2025) The characteristics of brain function alterations in patients with chronic prostatitis/chronic pelvic pain syndrome across varying symptom severities evaluated by NIH-CPSI. Front Neurosci 19:1511654.

Geerligs L, Tsvetanov KA, Henson RN (2017) Challenges in measuring individual differences in functional connectivity using fMRI: the case of healthy aging. Hum Brain Mapp 38:4125–4156.

Ghasemzadeh H, Hillman RE, Mehta DD (2024) Toward generalizable machine learning models in speech, language, and hearing sciences: Estimating sample size and reducing overfitting. J Speech, Lang Hear Res 67:753–781.

Gruskin DC, Patel GH (2022) Brain connectivity at rest predicts individual differences in normative activity during movie watching. Neuroimage 253:119100.

Haller S, Monsch AU, Richiardi J, Barkhof F, Kressig RW, Radue EW (2014) Head motion parameters in fMRI differ between patients with mild cognitive impairment and Alzheimer disease versus elderly control subjects. Brain Topogr 27:801–807.

Hantke N, Nielson KA, Woodard JL, Breting LMG, Butts A, Seidenberg M, Smith JC, Durgerian S, Lancaster M, Matthews M (2013) Comparison of semantic and episodic memory BOLD fMRI activation in predicting cognitive decline in older adults. J Int Neuropsychol Soc 19:11–21.

Häusler CO, Hanke M (2021) A studyforrest extension, an annotation of spoken language in the German dubbed movie “Forrest Gump” and its audio-description. F1000Research 10.

JNeurosci Accepted Manuscript

908 Hausman HK, Hardcastle C, Kraft JN, Evangelista ND, Boutzoukas EM, O'Shea A, Albizu A,
909 Langer K, Van Etten EJ, Bharadwaj PK (2022) The association between head motion
910 during functional magnetic resonance imaging and executive functioning in older adults.
911 Neuroimage: Reports 2:100085.
912 Kandiah N, Zhang A, Cenina AR, Au WL, Nadkarni N, Tan LC (2014) Montreal Cognitive
913 Assessment for the screening and prediction of cognitive decline in early Parkinson's
914 disease. Parkinsonism Relat Disord 20:1145–1148.
915 Kaye JA, Swihart T, Howieson D, Dame A, Moore MM, Karnos T, Camicioli R, Ball M, Oken
916 B, Sexton G (1997) Volume loss of the hippocampus and temporal lobe in healthy
917 elderly persons destined to develop dementia. Neurology 48:1297–1304.
918 Kemik K, Ada E, Çavuşoğlu B, Aykaç C, Savaş DDE, Yener G (2024) Detecting language
919 network alterations in mild cognitive impairment using task-based fMRI and resting-
920 state fMRI: A comparative study. Brain Behav 14:e3518.
921 LeBel A, D'Mello AM (2023) A seat at the (language) table: incorporating the cerebellum into
922 frameworks for language processing. Curr Opin Behav Sci 53:101310.
923 Lipkin B, Tuckute G, Affourtit J, Small H, Mineroff Z, Kean H, Jouravlev O, Rakocevic L,
924 Pritchett B, Siegelman M (2022) Probabilistic atlas for the language network based on
925 precision fMRI data from 800 individuals. Sci Data 9:529.
926 Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings
927 JL, Chertkow H (2005) The Montreal Cognitive Assessment, MoCA: a brief screening
928 tool for mild cognitive impairment. J Am Geriatr Soc 53:695–699.
929 Oschwald J, Guye S, Liem F, Rast P, Willis S, Röcke C, Jäncke L, Martin M, Mérillat S
930 (2019) Brain structure and cognitive ability in healthy aging: a review on longitudinal
931 correlated change. Rev Neurosci 31:1–57.
932 Park DC, Reuter-Lorenz P (2009) The adaptive brain: aging and neurocognitive scaffolding.
933 Annu Rev Psychol 60:173–196.
934 Poldrack RA, Huckins G, Varoquaux G (2020) Establishment of best practices for evidence
935 for prediction: a review. JAMA Psychiatry 77:534–540.

JNeurosci Accepted Manuscript

936 Power JD, Schlaggar BL, Petersen SE (2015) Recent progress and outstanding issues in
937 motion correction in resting state fMRI. Neuroimage 105:536–551.
938 Presacco A, Simon JZ, Anderson S (2016) Evidence of degraded representation of speech
939 in noise, in the aging midbrain and cortex. J Neurophysiol 116:2346–2355.
940 Price CJ (2012) A review and synthesis of the first 20 years of PET and fMRI studies of
941 heard speech, spoken language and reading. Neuroimage 62:816–847.
942 Rajimehr R, Xu H, Farahani A, Kornblith S, Duncan J, Desimone R (2024) Functional
943 architecture of cerebral cortex during naturalistic movie-watching. Neuron:Advanced
944 online publication.
945 Ranasinghe P, Mapa MST (2024) Functional connectivity and cognitive decline: A review of
946 rs-fMRI, EEG, MEG, and graph theory approaches in aging and dementia. Explor Med
947 5:797–821.
948 Rolls ET, Huang C-C, Lin C-P, Feng J, Joliot M (2020) Automated anatomical labelling atlas
949 3. Neuroimage 206:116189.
950 Sachdev PS, Blacker D, Blazer DG, Ganguli M, Jeste D V, Paulsen JS, Petersen RC (2014)
951 Classifying neurocognitive disorders: the DSM-5 approach. Nat Rev Neurol 10:634–
952 642.
953 Schneider BA, Daneman M, Pichora-Fuller MK (2002) Listening in aging adults: from
954 discourse comprehension to psychoacoustics. Can J Exp Psychol 56:139–152.
955 Shafto MA, Tyler LK (2014) Language in the aging brain: the network dynamics of cognitive
956 decline and preservation. Science (80-) 346:583–587.
957 Shankar R, Bundele A, Mukhopadhyay A (2025) A systematic review of natural language
958 processing techniques for early detection of cognitive impairment. Mayo Clin Proc Digit
959 Heal 3:100205.
960 Tibshirani R, Walther G (2005) Cluster validation by prediction strength. J Comput Graph
961 Stat 14:511–528.
962 Tie Y, Rigolo L, Ozdemir Ovalioglu A, Olubiyi O, Doolin KL, Mukundan Jr S, Golby AJ (2015)
963 A new paradigm for individual subject language mapping: movie-watching fMRI. J

JNeurosci Accepted Manuscript

Neuroimaging 25:710–720.

Tikka P, Kauttonen J, Hlushchuk Y (2018) Narrative comprehension beyond language: Common brain networks activated by a movie and its script. PLoS One 13:e0200134.

Wang Y, Gong X, Meng L, Wu X, Meng H (2024a) Large Language Model-based FMRI Encoding of Language Functions for Subjects with Neurocognitive Disorder. Proc Interspeech 2024:1485–1489.

Wang Y, Gong X, Wu X, Wong P, Fung HH, Mak MW, Meng H (2024b) Naturalistic Language-Related Movie-Watching fMRI Task for Detecting Neurocognitive Decline and Disorder. In: 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp 31–35. IEEE.

Wingfield A, Stine-Morrow EAL (2000) Language and speech. In: The handbook of aging and cognition, 2nd ed. (Craik FIM, Salthouse TA, eds), pp 359–416. Lawrence Erlbaum Associates Publishers.

Woo C-W, Krishnan A, Wager TD (2014) Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. Neuroimage 91:412–419.

Yan C, Zang Y (2010) DPARSF: a MATLAB toolbox for" pipeline" data analysis of resting-state fMRI. Front Syst Neurosci 4:13.

Yao S, Rigolo L, Yang F, Vangel MG, Wang H, Golby AJ, Liebenthal E, Tie Y (2022) Movie-watching fMRI for presurgical language mapping in patients with brain tumour. J Neurol Neurosurg Psychiatry 93:220–221.

Yeung PY, Wong LL, Chan CC, Leung JL, Yung CY (2014) A validation study of the Hong Kong version of Montreal Cognitive Assessment (HK-MoCA) in Chinese older adults in Hong Kong. Hong Kong Med J 20:504–510.

JNeurosci Accepted Manuscript

Table Legends

Table 1. Demographic information of participants.

MoCA: Montreal Cognitive Assessment. AD8: Eight-item Informant Interview to Differentiate Aging and Dementia. HKLLT: Hong Kong List Learning Test (*Z*-scores based on the Hong Kong norm were used). M_B*:* the baseline MoCA score. ΔM: longitudinal change in the MoCA score. M_{high} group versus M_{low} group: M_B above versus below the median. ΔM_{high} group versus ΔM_{low} group: ΔM above versus below the median. N/A: not applicable.

Table 2. Performance on classifying baseline global cognitive status (M_{high} versus M_{low}).

JNeurosci Accepted Manuscript

Sample size n = 97. M_{high} versus M_{low}: M_B above versus below the median. AUC: area under the curve.

* sMRI features include only hippocampal and parahippocampal volumes, as whole-brain volumetric measures performed significantly worse (AUC = 0.634 ± 0.054).

† The performance achieved by mw-fMRI features and the best performance achieved by feature combinations are highlighted in bold. Performance was measured using the mean (± *SD*) of the evaluation metrics, calculated from the outer test sets across all iterations of nested cross-validation.

Table 3. Performance on forecasting global cognitive changes (ΔM_{high} versus ΔM_{low}).

Sample size n = 50. M_B: the baseline MoCA score. ΔM: longitudinal change in the MoCA score. ΔM_{high} versus ΔM_{low}: ΔM above versus below the median. N/A: not applicable. AUC: area under the curve.

* sMRI features include only hippocampal and parahippocampal volumes; whole-brain volumetric measures performed similarly (AUC = 0.355 ± 0.088).

† The performance achieved by mw-fMRI features and the best performance achieved by feature combinations are highlighted in bold. Performance was measured using the mean (± *SD*) of the evaluation metrics, calculated from the outer test sets across all iterations of nested cross-validation.

Table 4. Machine-learning regression performance for predicting baseline global cognitive function (M_B) and longitudinal global cognitive change (ΔM).

M_B: baseline MoCA score. ΔM: longitudinal change in the MoCA score. SVR: support vector regression. MAE: mean absolute error.

JNeurosci Accepted Manuscript

Figure Legends

Figure 1. Flowchart showing the procedures for data collection, feature selection, and supervised machine learning analysis. GM/WM/CSF: gray matter/white matter/cerebrospinal fluid. mALFF: standardized amplitude of low-frequency fluctuation. MoCA: Montreal Cognitive Assessment. M_B: the baseline MoCA score. ΔM: longitudinal change in the MoCA score. Task 1 is to classify M_{high} versus M_{low}: baseline MoCA score above versus below the median. Task 2 is to forecast ΔM_{high} versus ΔM_{low}: ΔM above versus below the median. SVC: support vector classifier. CV: cross-validation. C_1 and C_2 are hyperparameters in the feature selection module and the SVC model, respectively, with C_1^* and C_2^* representing their optimal values. k is the index of inner cross-validation, and l is the index of outer cross-validation. AUC: area under the curve.

Figure 2. Significant language-related activations in the mw-fMRI task (Part A) and their spatial overlap with established language networks (Parts B to D). Inf: inferior. L: left. LanA: a validated probabilistic functional atlas of the language network. Med: medial. Mid: middle. Oper: opercular. Orb: orbital. Post: posterior. R: right. Sup: superior. Supp: supplementary. Tri: triangular.

Figure 3. Robustness of spatial correspondence between mw-fMRI language-related regions and the LanA atlas across probability thresholds. The overlap between mw-fMRI language-related regions and the LanA atlas was evaluated across 11 LanA probability thresholds ranging from $P > 0.10$to $P > 0.60$. **Left:** AAL3 regional voxel-distribution correlations between mw-fMRI regions and thresholded LanA masks across probability thresholds. **Right:** Spatial overlap quantified by the Dice coefficient and the proportion of LanA voxels covered by mw-fMRI regions across probability thresholds. Dashed vertical lines indicate the two thresholds used in the main analysis: $P > 0.20$, defining the extended LanA network, and $P > 0.50$, defining the central LanA network.

JNeurosci Accepted Manuscript

Figure 4. Correlations between cognitive measures (M_B and ΔM) and language-related measures during mw-fMRI (including language performance and language-related activations within the mw-fMRI language regions). M_B: the MoCA score at baseline. ΔM: longitudinal change in the MoCA score.

Figure 5. Receiver operating characteristic (ROC) curves for classifying baseline cognitive status (Part A) and forecasting longitudinal cognitive change (Part B).

Figure 6. Spatial distributions of mw-fMRI feature clusters and cluster-wise correlations for classifying baseline cognitive status (Parts A and B) and forecasting cognitive changes (Parts C and D). M_B: the baseline MoCA score. ΔM: longitudinal change in the MoCA score. Corr.: correlation. 95% CI: 95% confidence interval.

JNeurosci Accepted Manuscript

Table 1. Demographic information of participants

Measures	**Statistics for different samples**			
	M_{high} group	**M_{low} group**	**Group comparison**	**Whole sample**
Sample size	45	52	N/A	97
Sex (f/m)	14 / 31	29 / 23	$\chi^2 = 5.94$, $p = 0.01$	43 / 54
Age	69.56 ± 5.58	73.10 ± 5.93	$t = 3.01$, $p = 0.003$	71.45 ± 6.01
Education	9.84 ± 3.81	6.35 ± 3.07	$t = -4.99$, $p < 0.001$	7.97 ± 3.84
AD8	2.77 ± 2.17	3.98 ± 2.32	$t = 2.64$, $p = 0.01$	3.44 ± 2.38
HKLLT	-0.80 ± 0.56	-0.80 ± 0.72	$t = 0.01$, $p = 0.99$	-0.80 ± 0.65
M_B	23.69 ± 2.18	17.12 ± 2.97	$t = -6.57$, $p < 0.001$	20.16 ± 4.21
M_B median	N/A	N/A	N/A	20
	ΔM_{high} group	**ΔM_{low} group**	**Group comparison**	**Whole sample**
Sample size	25	25	N/A	50
Sex (f/m)	17 / 8	11 / 14	$\chi^2 = 2.92$, $p = 0.09$	28 / 22
Age	70.04 ± 6.12	70.68 ± 5.95	$t = 0.37$, $p = 0.71$	70.36 ± 6.05
Education	7.64 ± 2.65	7.52 ± 3.03	$t = -0.14$, $p = 0.88$	7.58 ± 2.85
AD8	3.88 ± 2.18	2.88 ± 2.43	$t = -1.53$, $p = 0.13$	3.38 ± 2.36
HKLLT	-0.55 ± 0.79	-0.74 ± 0.46	$t = -1.07$, $p = 0.29$	-0.65 ± 0.64
M_B	22.60 ± 4.41	20.96 ± 4.12	$t = -1.36$, $p = 0.18$	21.78 ± 4.35
ΔM	3.56 ± 0.98	-0.88 ± 1.51	$t = -12.33$, $p < 0.001$	1.34 ± 2.56
ΔM median	N/A	N/A	N/A	1

MoCA: Montreal Cognitive Assessment. AD8: Eight-item Informant Interview to Differentiate Aging and Dementia. HKLLT: Hong Kong List Learning Test (*Z*-scores based on the Hong Kong norm were used). M_B: the baseline MoCA score. ΔM: longitudinal change in the MoCA score. M_{high} group versus M_{low} group: M_B above versus below the median. ΔM_{high} group versus ΔM_{low} group: ΔM above versus below the median. N/A: not applicable.

JNeurosci Accepted Manuscript

Table 2. Performance on classifying baseline global cognitive status (M_{high} versus M_{low}).

Features	AUC	Accuracy	Sensitivity	Specificity	Random permutation (AUC)
Demographics (Dem.)	0.758±0.036	0.684±0.023	0.789±0.037	0.566±0.033	0.527±0.063
sMRI*	0.705±0.031	0.645±0.033	0.837±0.048	0.430±0.065	0.524±0.044
rs-fMRI	0.718±0.036	0.673±0.042	0.701±0.054	0.642±0.063	0.492±0.041
mw-fMRI†	0.811±0.019	0.722±0.020	0.787±0.036	0.645±0.036	0.502±0.073
mw-fMRI + Dem.	0.809±0.023	0.724±0.023	0.777±0.023	0.661±0.047	0.498±0.080
mw-fMRI + sMRI*	0.809±0.024	0.732±0.019	0.790±0.035	0.659±0.030	0.503±0.076
mw-fMRI + rs-fMRI	0.793±0.026	0.715±0.023	0.738±0.021	0.691±0.048	0.491±0.073
All imaging + Dem.†	0.817±0.022	0.740±0.030	0.767±0.040	0.707±0.057	0.491±0.074

Sample size n = 97. M_{high} versus M_{low}: M_B above versus below the median. AUC: area under the curve.

* sMRI features include only hippocampal and parahippocampal volumes, as whole-brain volumetric measures performed significantly worse (AUC = 0.634 ± 0.054).

† The performance achieved by mw-fMRI features and the best performance achieved by feature combinations. Performance was measured using the mean (± *SD*) of the evaluation metrics, calculated from the outer test sets across all iterations of nested cross-validation.

JNeurosci Accepted Manuscript

Table 3. Performance on forecasting global cognitive changes (ΔM_{high} versus ΔM_{low}).

Features	AUC	Accuracy	Sensitivity	Specificity	Random permutation (AUC)
Demographics (Dem.)	0.566±0.069	0.582±0.036	0.562±0.074	0.628±0.078	0.452±0.053
M_B	0.680±0.025	0.400±0.000	0.500±0.000	0.500±0.000	0.491±0.118
sMRI*	0.363± 0.054	0.416±0.034	0.400±0.081	0.528±0.087	0.518±0.076
rs-fMRI	0.702±0.071	0.628±0.041	0.615±0.071	0.630±0.069	0.510±0.093
mw-fMRI†	0.763±0.036	0.702±0.052	0.683±0.087	0.745±0.070	0.492±0.115
mw-fMRI + Dem.	0.762±0.048	0.656±0.062	0.647±0.055	0.692±0.100	0.490±0.108
mw-fMRI + M_B†	0.793±0.035	0.716±0.053	0.677±0.089	0.782±0.052	0.567±0.097
mw-fMRI + sMRI	0.710±0.052	0.648±0.040	0.608±0.042	0.710±0.083	0.490±0.109
mw-fMRI + rs-fMRI	0.762±0.069	0.678±0.051	0.667±0.072	0.683±0.094	0.485±0.092
All imaging + Dem.	0.723±0.066	0.642±0.056	0.620±0.090	0.673±0.078	0.480±0.094

Sample size n = 50. M_B: baseline MoCA score. ΔM: longitudinal change in the MoCA score. ΔM_{high} versus ΔM_{low}: ΔM above versus below the median. N/A: not applicable. AUC: area under the curve.

* sMRI features include only hippocampal and parahippocampal volumes; whole-brain volumetric measures performed similarly (AUC = 0.355 ± 0.088).

† The performance achieved by mw-fMRI features and the best performance achieved by feature combinations. Performance was measured using the mean (± *SD*) of the evaluation metrics, calculated from the outer test sets across all iterations of nested cross-validation.

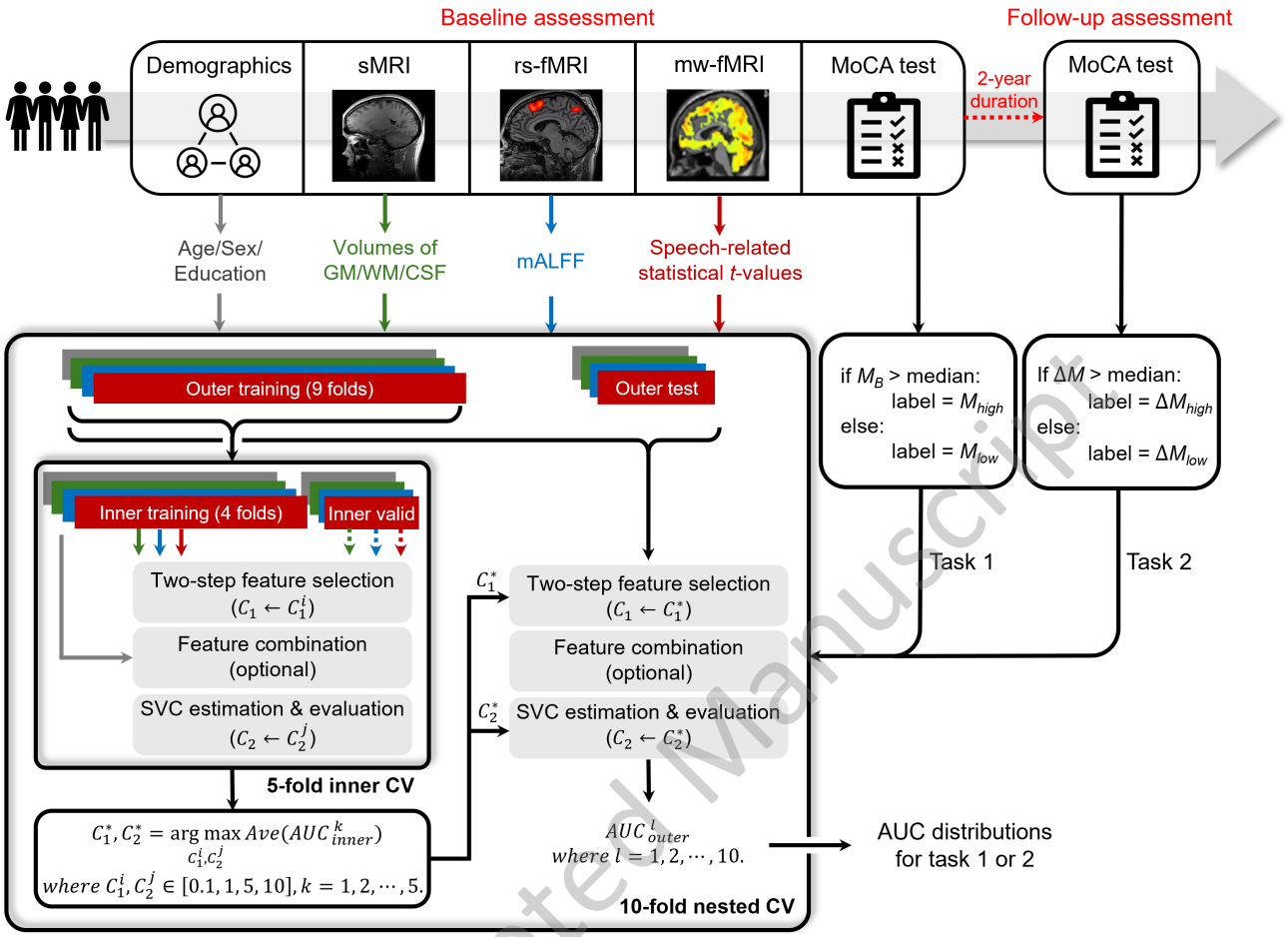
JNeurosci Accepted Manuscript

Table 4. Machine-learning regression performance for predicting baseline global cognitive function (M_B) and longitudinal global cognitive change (ΔM)

Features	Predicting M_B (n = 97)			Predicting ΔM (n = 50)		
	SVR (MAE / R^2)	Ridge (MAE / R^2)	Elastic Net (MAE / R^2)	SVR (MAE / R^2)	Ridge (MAE / R^2)	Elastic Net (MAE / R^2)
Demographics (Dem.)	2.825 / 0.201	2.776 / 0.223	2.780 / 0.221	2.288 / −0.172	2.282 / −0.079	2.279 / −0.077
M_B	-	-	-	2.115 / 0.020	2.125 / 0.020	2.128 / 0.020
sMRI	3.086 / 0.099	3.048 / 0.114	3.068 / 0.103	2.332 / −0.132	2.295 / −0.088	2.301 / −0.093
rs-fMRI	3.155 / 0.044	3.128 / 0.055	3.112 / 0.062	2.193 / −0.113	2.185 / −0.115	1.965 / 0.109
mw-fMRI	3.059 / 0.089	2.992 / 0.110	3.134 / 0.051	2.218 / −0.008	2.236 / −0.024	2.210 / −0.002
mw-fMRI + Dem.	3.067 / 0.086	2.993 / 0.109	3.063 / 0.083	2.230 / −0.023	2.218 / −0.009	2.212 / −0.005
mw-fMRI + M_B	-	-	-	2.206 / −0.026	2.230 / −0.018	2.224 / −0.021
mw-fMRI + sMRI	3.062 / 0.087	2.996 / 0.108	3.123 / 0.043	2.204 / −0.003	2.219 / −0.015	2.207 / −0.003
mw-fMRI + rs-fMRI	3.074 / 0.092	3.080 / 0.086	3.094 / 0.068	2.194 / −0.100	2.170 / −0.096	1.967 / 0.113
All imaging + Dem.	3.074 / 0.092	3.073 / 0.089	3.076 / 0.074	2.198 / −0.103	2.177 / −0.104	1.987 / 0.101

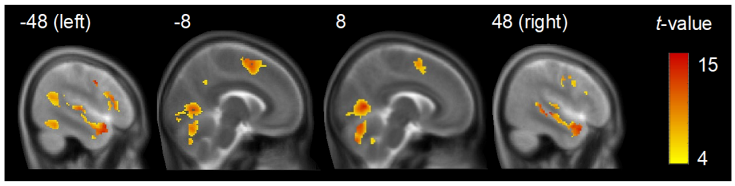
M_B: baseline MoCA score. ΔM: longitudinal change in the MoCA score. SVR: support vector regression. MAE: mean absolute error.

JNeurosci Accepted Manuscript

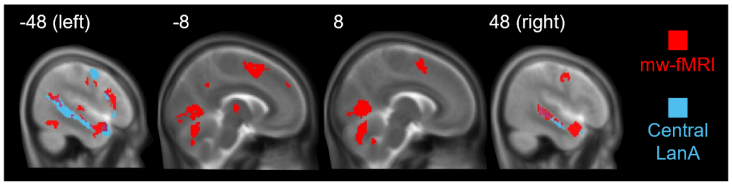


JNeurosci Accepted Manuscript

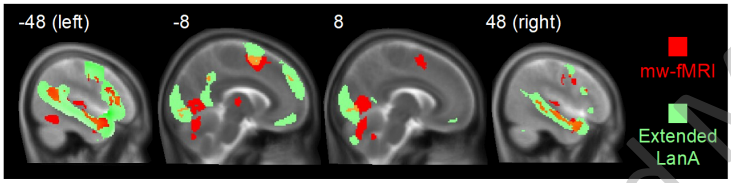
A The mw-fMRI language regions (significantly activated by speech events during movie watching)

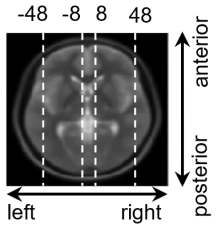


B Spatial overlap between the mw-fMRI language regions and the central LanA network



C Spatial overlap between the mw-fMRI language regions and the extended LanA network





Positions of the sagittal slices displayed in parts A to C

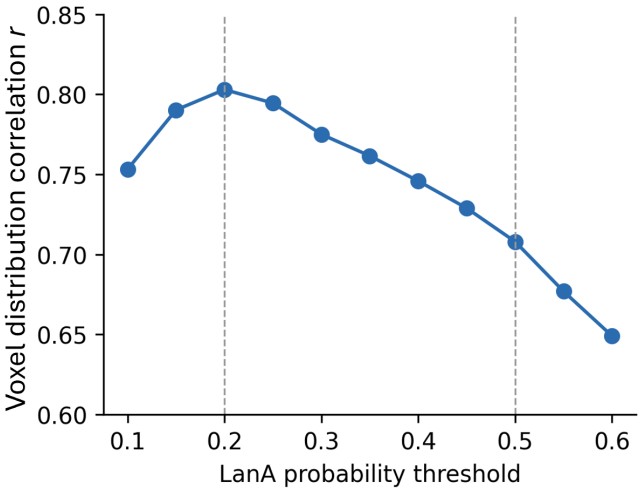
D Voxel distributions in the mw-fMRI language regions and the LanA networks

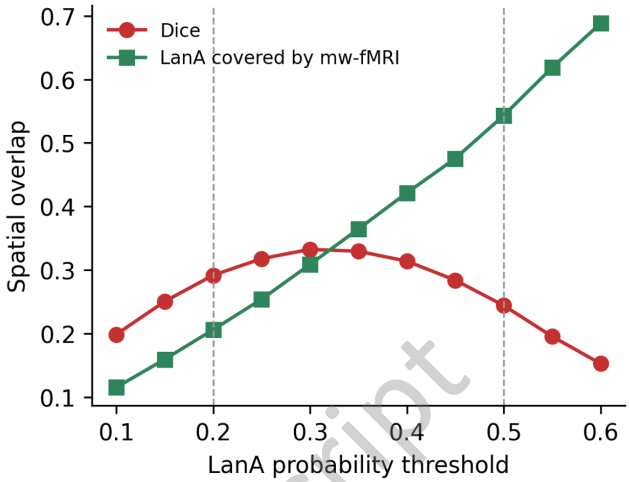
	AAL3 labels of brain regions	Number of brain voxels: mw-fMRI regions	Central LanA	Extended LanA
Frontal lobe	Frontal_Inf_Oper_L / R	116	24	676
	Frontal_Inf_Orb_L / R	20	0	604
	Frontal_Inf_Tri_L / R	269	79	1843
	Frontal_Med_Orb_L / R	0	0	170
	Frontal_Mid_L / R	191	0	222
	Frontal_Sup_L	24	0	385
	Frontal_Sup_Medial_L	49	0	680
	Precentral_L / R	548	221	1408
	Rectus_L / R	0	0	452
	Supp_Motor_Area_L / R	329	55	868
Temporal lobe	Fusiform_L / R	525	0	79
	Parahippocampal_L	0	0	49
	Heschl_L / R	75	0	0
	Temporal_Inf_L	149	0	192
	Temporal_Mid_L / R	2110	2034	5568
	Temporal_Pole_Mid_L / R	100	65	554
	Temporal_Pole_Sup_L / R	400	216	1101
	Temporal_Sup_L / R	1522	400	2324
Parietal lobe	Angular_L / R	114	0	421
	Parietal_Inf_L	36	0	15
	Parietal_Sup_L / R	100	0	0
	Postcentral_L / R	93	0	227
	Precuneus_L / R	91	0	522
	SupraMarginal_L	16	0	402
Occipital lobe	Calcarine_L / R	340	0	1391
	Cuneus_L / R	0	0	514
	Lingual_L / R	927	0	906
	Occipital_Mid_L	190	0	505
	Occipital_Sup_R	36	0	209
Subcortical areas	Cerebellum	2648	62	2204
	Caudate_R	25	0	0
	Cingulum_Mid_L	0	0	86
	Cingulum_Post_L	31	0	127
	Putamen_L	109	0	31
	Thalamus_L	61	0	61

Color intensity indicates percentage:

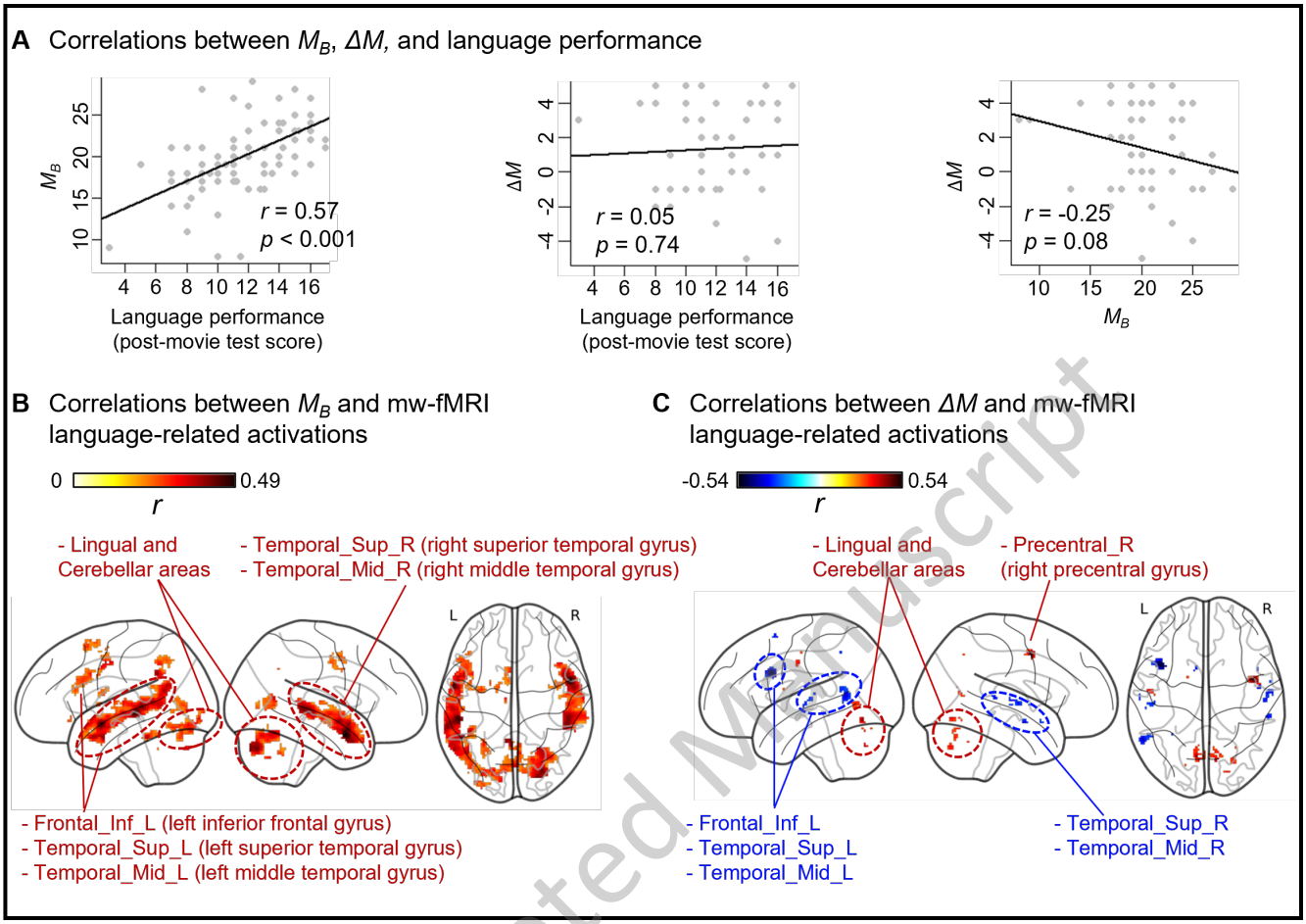
0% 100% 0% 100% 0% 100%

JNeurosci Accepted Manuscript



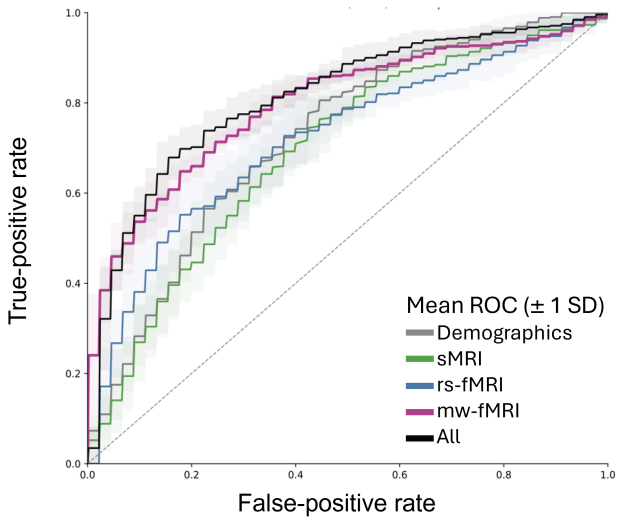


JNeurosci Accepted Manuscript

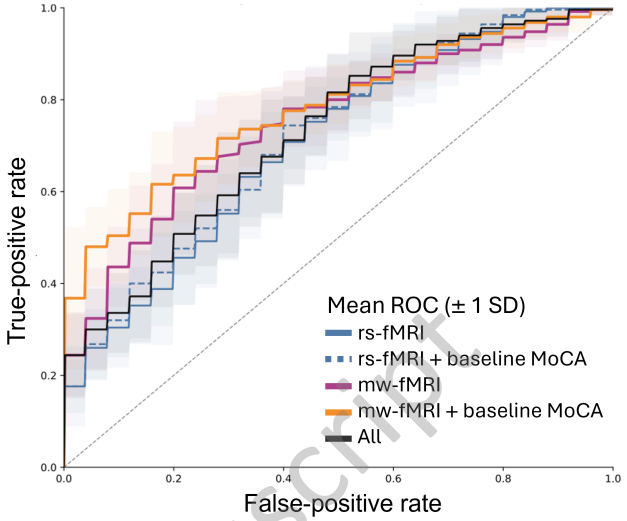


JNeurosci Accepted Manuscript

A. ROC curves for classifying baseline cognitive status



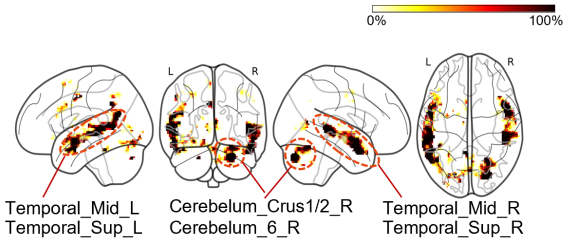
B. ROC curves for forecasting cognitive change



JNeurosci Accepted Manuscript

Sources of mw-fMRI features selected for classifying M_{high} versus M_{low}

A Selection frequency of features in Step-1 feature selection



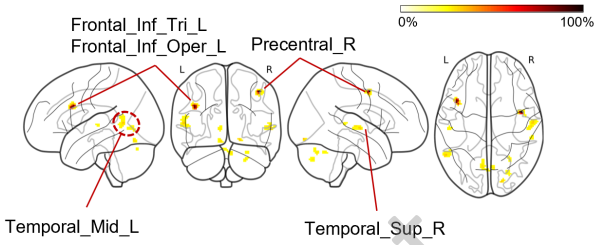
B Primary feature sources in Step-2 feature selection

AAL3 label	Frequency of selection	Correlation coefficients: mean [5-95th percentile] Corr. with M_B	Corr. with post-movie test
Temporal_Mid_L* (left middle temporal)	100%	.34 [.32, .37]	.27 [.23, .30]
Temporal_Mid_R (right middle temporal)	100%	.36 [.32, .40]	.28 [.23, .34]
Cerebelum_6_R (right cerebellum, area 6)	100%	.33 [.29, .36]	.27 [.22, .33]
Temporal_Sup_R* (right superior temporal)	95%	.32 [.30, .35]	.24 [.18, .31]
Temporal_Sup_L (left superior temporal)	87%	.37 [.28, .47]	.26 [.16, .34]
Cerebelum_Crus1_R (right cerebellum, crus 1)	86%	.33 [.28, .38]	.31 [.26, .38]
Cerebelum_Crus2_R (right cerebellum, crus 2)	74%	.30 [.28, .34]	.23 [.17, .29]

** Including the superior temporal pole area*

Sources of features selected for forecasting ΔM_{high} versus ΔM_{low}

C Selection frequency of features in Step-1 feature selection



D Primary feature sources in Step-2 feature selection

AAL3 label	Frequency of selection	Correlation coefficients: mean [5-95th percentile] Corr. with ΔM	Corr. with M_B	Corr. with post-movie test
Precentral_R (right precentral)	94%	.45 [.41, .50]	.05 [-.04, .14]	.11 [.03, .22]
Frontal_Inf_Tri_L (left inferior frontal, triangular part)	61%	-.45 [-.51, -.41]	.20 [.13, .28]	.17 [.07, .27]
Frontal_Inf_Oper_L (left inferior frontal, opercular part)	52%	-.49 [-.58, -.41]	.22 [.14, .31]	.10 [.01, .20]
Temporal_Sup_R (right superior temporal)	38%	-.40 [-.42, -.38]	.24 [.15, .32]	.16 [.08, .23]
Temporal_Mid_L (left middle temporal)	37%	-.40 [-.44, -.38]	.33 [.27, .37]	.12 [.05, .19]

JNeurosci Accepted Manuscript