Research Paper

An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures

Nikhil Bhagwat, MSc; Jon Pipitone, MSc; Aristotle N. Voineskos, MD, PhD; M. Mallar Chakravarty, PhD; Alzheimer's Disease Neuroimaging Initiative

Published online on Feb. 5, 2019; subject to revision

Background: The development of diagnostic and prognostic tools for Alzheimer disease is complicated by substantial clinical heterogeneity in prodromal stages. Many neuroimaging studies have focused on case–control classification and predicting conversion from mild cognitive impairment to Alzheimer disease, but predicting scores from clinical assessments (such as the Alzheimer's Disease Assessment Scale or the Mini Mental State Examination) using MRI data has received less attention. Predicting clinical scores can be crucial in providing a nuanced prognosis and inferring symptomatic severity. **Methods:** We predicted clinical scores at the individual level using a novel anatomically partitioned artificial neural network (APANN) model. The model combined input from 2 structural MRI measures relevant to the neurodegenerative patterns observed in Alzheimer disease: hippocampal segmentations and cortical thickness. We evaluated the performance of the APANN model with 10 rounds of 10-fold cross-validation in 3 experiments, using cohorts from the Alzheimer's Disease Neuroimaging Initiative (ADNI): ADNI1, ADNI2 and ADNI1 + 2. **Results:** Pearson correlation and root mean square error between the actual and predicted scores on the Alzheimer's Disease Assessment Scale (ADNI1: $r = 0.60$; ADNI2: $r = 0.68$; ADNI1 + 2: $r = 0.63$) and Mini Mental State Examination (ADNI1: $r = 0.52$; ADNI2: $r = 0.55$; ADNI1 + 2: $r = 0.55$) showed that APANN can accurately infer clinical severity from MRI data. **Limitations:** To rigorously validate the model, we focused primarily on large cross-sectional baseline data sets with only proof-of-concept longitudinal results. **Conclusion:** The APANN provides a highly robust and scalable framework for predicting clinical severity at the individual level using high-dimensional, multimodal neuroimaging data.

Introduction

Machine-learning methods have been used extensively to distinguish people with Alzheimer disease and its prodromes from healthy controls.[1–5] However, predicting symptomatic severity at the individual level remains a challenging problem that may be more intimately related to personalized care and prognosis. Prediction is confounded by the substantial pathophysiological and clinical heterogeneity observed in prodromal stages such as mild cognitive impairment (MCI) or significant memory concern.[6–11] Although much is known about the temporal and neuroanatomical specificity regarding the aggregation of amyloid plaques and neurofibrillary tangles and resulting downstream neurodegeneration,[12] little is known about the variations in brain anatomy associated with these processes and how they inform cognitive impairment related to Alzheimer disease. Understanding the complex pathophysiological processes that characterize the varying clinical presentations is essential for biomarker development and early detection.[13–15] Furthermore, neuroanatomically informed prediction of clinical performance is an important step toward biomarker assessment and the development of assistive tools for prognosis and treatment planning.

As a structural biomarker, the hippocampus has long been associated with the pathophysiology of Alzheimer disease and related impairment.[1,16–21] However, measures of hippocampal volume lack the sensitivity to act as stand-alone biomarkers.[22–26] To achieve nuanced characterization of disease states, studies have explored hippocampal subfield-based biomarkers[23,27,28] and other neurodegeneration indicators, such as cortical atrophy quantified by cortical thickness.[19,21,29–32] Nevertheless, no characteristic localized patterns of atrophy have been associated with prodromal disease states or symptomatic severity levels, which are likely to be heavily influenced by cognitive reserve.[8,32] This motivates approaches that incorporate multiple, distributed phenotypes to predict clinical severity in service of robust diagnostic and prognostic applications.

Correspondence to: N. Bhagwat, Cerebral Imaging Centre, Douglas Mental Health University Institute, 6875 Lasalle Blvd, Montreal, QC H4H 1R3; nikhil153@gmail.com

Submitted Feb. 1, 2018; Revised Apr. 19, 2018; Accepted Aug. 1, 2018

DOI: 10.1503/jpn.180016

Previously, computational approaches using neuroimaging measures in the context of Alzheimer disease have focused on predicting diagnosis in cross-sectional data sets,[2–5] or conversion from MCI to Alzheimer disease in longitudinal analyses.[33–35] However, clinicians are more likely to treat symptoms based on the results of structured assessments rather than on a specific diagnosis. In this work, we focused on predicting clinical scores of disease severity (i.e., Alzheimer's Disease Assessment Scale [ADAS-13],[36] Mini Mental State Examination [MMSE][37]) directly from neuroimaging data.[38,39] Such neuroanatomically informed prediction of clinical performance at baseline and at future time points — particularly in people with MCI or significant memory concern — can help clinicians manage the clinical heterogeneity and make accurate diagnostic and prognostic decisions. Although our ultimate clinical goal is to provide longitudinal prognosis, in this report we focused primarily on a thorough validation of data sets from a single time point (baseline), an important first step in model development for longitudinal tasks. We also performed a proof-of-concept analysis to verify the ability of the proposed model to provide longitudinal prediction.

For this prediction task, we proposed an anatomically partitioned artificial neural network (APANN) model. Artificial neural networks (ANNs) and related deep-learning approaches have delivered state-of-the-art performance in classification and prediction problems for computer vision, speech recognition, natural language processing and other domains.[40–45] The ANNs provide highly flexible computational frameworks that can be used to extract latent features corresponding to the hierarchical structural and functional organization of the brain and are well suited for problems with high dimensional data, unlike more standard models.[41,43] To this end, the primary objective of this study was to assess whether ANN models could accurately predict ADAS-13 and MMSE clinical scores using T_1-weighted brain MRI data. In a larger context, we aim to build an ANN-based computational framework that can process high dimensional, distributed structural changes captured by multiple phenotypic measures to make prognostic predictions.

We designed, trained and tested our model using participants from 2 Alzheimer's Disease Neuroimaging Initiative (ADNI) cohorts. We used a combination of high dimensional (> 30 000) features derived from 2 neuroanatomical measures in the T_1-weighted images: hippocampal segmentation and cortical thickness. We generated these measures using MAGeT Brain and CIVET pipelines (see Methods), respectively. We present a model with an innovative modular design that enables the analysis of this high dimensional, multimodal input. It also allows for inclusion of new input modalities without having to retrain the entire model, and it offers simultaneous prediction of multiple clinical scores (e.g., ADAS-13 and MMSE). Given the high dimensionality of the input data, we have addressed the need for large training examples by introducing a novel data augmentation method. The method presented in this paper is not limited solely to the prediction of severity in Alzheimer disease; it can be applied to train a variety of deep-learning models that use high dimensional neuroimaging data to tackle many diagnostic and prognostic questions.

Methods

Data sets

We used baseline data from participants in the ADNI1 (n = 818) and ADNI2 (n = 788) databases[46] (http://adni.loni.usc.edu). After exclusions based on quality control of the image preprocessing outputs, the final number of participants we used was 669 from ADNI1 and 690 from ADNI2 (see Table 1 for demographic details).

Our objective was to predict MMSE and ADAS-13 scores. The MMSE is one of the most widely used cognitive assessments for the diagnosis of Alzheimer disease and related dementias;[47,48] its scores range from 0 to 30, with lower scores indicating greater cognitive impairment. The ADAS-13 is a modified version of the ADAS-cog assessment, and it has a maximum score of 85. Although ADAS-13 has some overlap with the MMSE, it also includes components that target memory, language and praxis. In contrast to the MMSE,

Table 1: Data set demographics for ADNI1 and ADNI2 cohorts*

Parameter	ADNI1 (n = 669)	ADNI2 (n = 690)
Acquisition	Scanner: 1.5 T Voxel sizes: 1.2 mm × 1.25 mm × 1.25 mm	Scanner: 3.0 T Voxel sizes: 1.2 mm × 1 mm × 1 mm
Diagnosis, no.	Cognitively healthy: 198 Late mild cognitive impairment: 326 Alzheimer disease: 145	Cognitively healthy: 179 Significant memory concern: 77 Early mild cognitive impairment: 162 Late mild cognitive impairment: 149 Alzheimer disease: 123
Sex, no.	Male: 377 Female: 292	Male: 361 Female: 329
Age, yr	75.0 ± 6.7	72.6 ± 7.2
Education, yr	15.5 ± 3.1	16.3 ± 2.6
ADAS-13 score	18.4 ± 9.2 (1.0, 54.7)	16.1 ± 10.14 (1.0, 52.0)
MMSE score	26.7 ± 2.7 (18.0, 30.0)	27.5 ± 2.7 (19.0, 30.0)

ADAS-13 = Alzheimer's Disease Assessment Scale; MMSE = Mini Mental State Examination; SD = standard deviation.
*Findings are presented as mean ± SD (minimum, maximum) unless otherwise specified.

higher scores on the ADAS-13 indicate greater cognitive impairment.

We pooled participants from all diagnostic categories to build models for the entire spectrum of clinical performance. We did not use diagnostic grouping, because we modelled Alzheimer disease progression on a continuum, a method that has been shown to be useful in other studies of Alzheimer disease progress.[49,50]

MRI processing

First, we preprocessed the MRIs using the bpipe pipeline (https://github.com/CobraLab/minc-bpipe-library/), consisting of N4-correction,[51] neck cropping to improve linear registration and BEaST brain extraction.[52] We then used the preprocessed data to extract hippocampal segmentations and cortical thickness measures, referred to as input modalities in this work. We performed computations using the GPC supercomputer at the SciNet HPC Consortium.[53]

Hippocampal segmentation

We produced hippocampal segmentations of T_1-weighted MRIs using the MAGeT brain pipeline.[24,54] Briefly, this pipeline began with 5 manually segmented, high-resolution 3 T T_1-weighted images,[55] each registered nonlinearly to 15 ADNI images selected at random (known as the template library). Then, each image in the template library was registered in a nonlinear fashion to all images in the ADNI data sets, and the segmentations from each atlas were warped via the template library transformations to each ADNI image. This process resulted in 75 (no. atlas × no. templates) candidate segmentations for each image, which were fused into a single segmentation using voxel-wise majority voting.

Cortical thickness measures

We input the preprocessed images into the CIVET pipeline[29,56–59] to estimate cortical thickness at 40 962 vertices per hemisphere, which could then be grouped by region of interest (ROI) based on a surface atlas.

Anatomically partitioned artificial neural network

Artificial neural networks are a biologically inspired family of graphical machine-learning models that can perform prediction tasks using high dimensional input (Fig. 1A). These ANN models can be designed to contain multiple hidden layers, which hierarchically encode latent features that inform the objective task. The neuron connections represent a set of weights for the preceding input values, which are then combined and filtered with a nonlinear function. In neuroimaging, a few variants of ANN models (such as autoencoders and restricted Boltzmann machines) have been investigated for classification and prediction tasks.[43,60] The model used in the current study differs significantly from these approaches in both design and implementation.

From a design perspective, we leveraged the hierarchical structure of ANNs to build a modular (Fig. 1B) architecture that was capable of multimodal input integration (Fig. 1C) and multitask predictions (Fig. 1D). We achieved the following objectives in 3 stages (Fig. 1E). Stage I consisted of anatomically partitioned modules (2 hidden layers per module) that extracted features from individual anatomic input sources (hippocampus and cortical surface). These individual anatomic features served as input to stage II, where they were combined at a higher layer in the hidden-layer hierarchy. Finally, we used these integrated features to perform multiple tasks simultaneously; these task-specific hidden layers were represented by the higher layers in stage III (4 hidden layers total). This APANN mitigated overfitting by reducing the number of model parameters compared with classical, fully connected hidden-layer architectures. It also allowed for independent pretraining of each input source in a single branch. These individual pretrained branches could then be used to train stage II to integrate features efficiently.

Empirical distributions

The input dimensionality of MRI data greatly exceeds the available number of samples, leaving machine-learning models susceptible to overfitting.[14,30] This necessitates the critical step of feature engineering: the transformation of high dimensional raw input to a meaningful and computationally manageable feature space.[61] Techniques for addressing high dimensionality include downsampling, handcrafting features based on biological priors (e.g., atlases), principal component analysis and others. One can also increase the sample size by adding transformed data (e.g., linear transformations, image patches) to deal with the high dimensionality. In this study, we used a novel data augmentation method that leveraged the MRI preprocessing pipelines to produce a set of empirical samples for both the hippocampal and cortical thickness input modalities in place of a single point estimate per participant. This boost in training sample size made it feasible to train these models with a large parameter space and helped prevent overfitting by exposing the model to a large set of possible variations in anatomic input associated with a given severity level. Adding linear and nonlinear transformations of original input data is a common practice in machine learning.[42,44] In computer vision applications, this typically means translation, rotation or dropping of certain pixels to capture a larger set of commonly encountered variations in input features to which the classifier should be invariant. In structural MRI data, we were more interested in modelling the joint voxel distribution of anatomic segmentations than in achieving high translational invariance, because the location of anatomic structures is relatively consistent across individuals. Thus, the empirical samples that were generated as part of the common segmentation and cortical surface extraction pipelines helped train the model to be invariant to the methodologically driven perturbations of input values. In turn, this mitigated overfitting and helped the model learn anatomic patterns relevant to clinical performance.

For the hippocampal inputs, the empirical samples referred to a set of "candidate segmentations" generated

from a multi-atlas segmentation pipeline (Fig. 2A)[24,54] that model the underlying joint label distribution over the set of voxels for a given participant. For the cortical thickness inputs, the empirical samples referred to cortical thickness values from a set of vertices belonging to a given cortical ROI (Fig. 2B). In traditional approaches, these samples are usually fused to produce a point estimate of the feature.[3,32] We have detailed the sample-generation process for both input types below.

Hippocampal segmentation

We produced 75 candidate segmentations and 1 fused segmentation for each participant via the MAGeT brain pipeline.[24] We segmented the ADNI1 and ADNI2 data sets using 2 separate template libraries of 15 images for each cohort. These candidate segmentations were binary masks of the left and right hippocampal voxels.

We rigidly aligned candidate segmentations to a common space (a participant chosen at random from the ADNI1 data set) to maximize anatomic correspondence across participants. We split each segmentation into left and right hemispheres and aligned both rigidly to this common space using the ANTS registration toolkit.[62]

To remove outlier segmentations resulting from misregistration or poor segmentation, we computed the Dice κ between rigidly aligned candidate segmentations and the

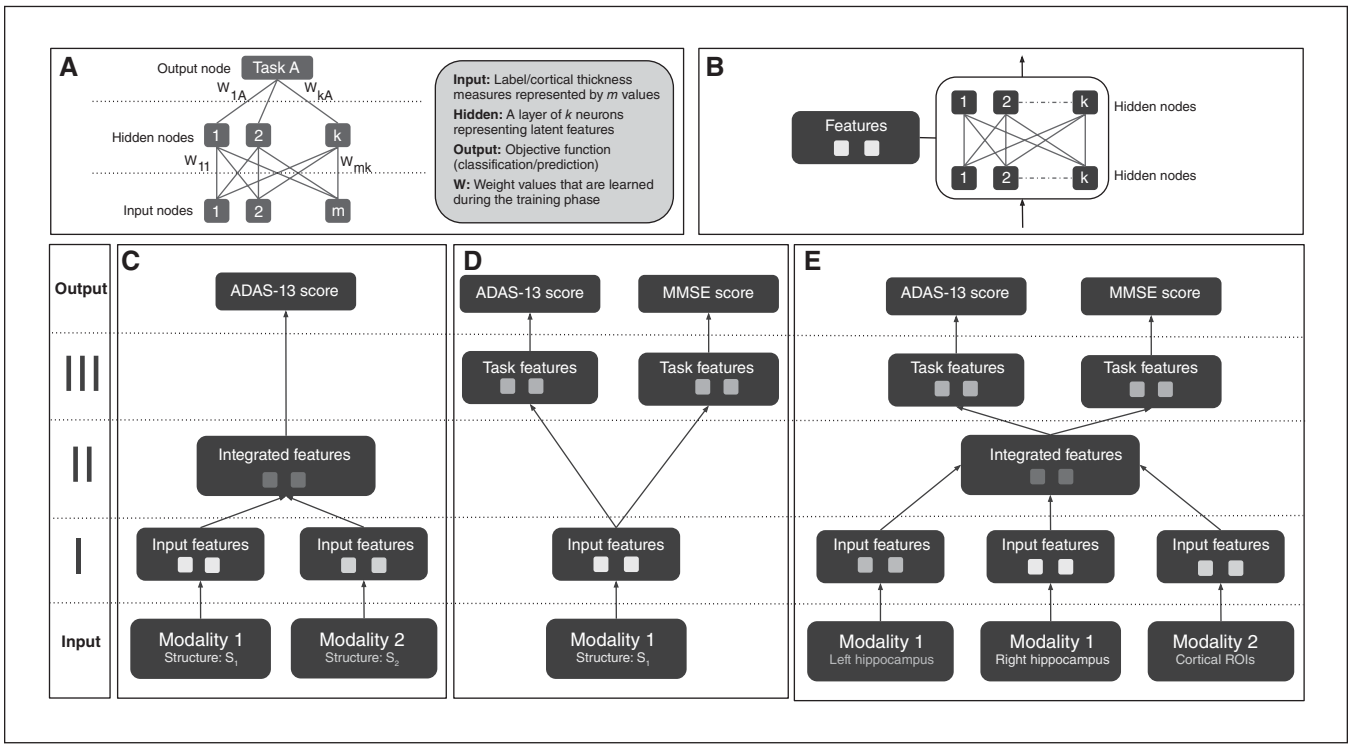

Fig. 1: (A) Structure of a generic ANN model. A neural net may consist of multiple hidden layers that encode a hierarchical set of features from input, informative of the prediction/classification task at hand. The connections between layers represent the model weights, which are updated via backpropagation based on loss function associated with the task. (B) A single feature module consisting of multiple hidden layers. This is a building block of the APANN architecture, which facilitates pretraining of individual branches per input modality. (C) A multimodal ANN with a single output task. This design consists of stage I and stage II feature modules. Stage I modules learn features from each modality that are combined in the stage II feature module. Only single-task performance is used to update the weights of the model in this architecture. (D) A multi-task ANN with a single input modality. This design consists of stage I and stage III feature modules. The stage I module learns individual features from a given modality, which are then fed into task-specific feature modules connected to the output nodes for joint prediction of the 2 tasks (ADAS-13 and MMSE score prediction). Prediction performance from both tasks is used to update the weights of the stage I feature module. Left hippocampal, right hippocampal and cortical thickness input modalities are trained separately using this design to learn input feature modules from each modality. (E) The proposed multimodal, multitask APANN model comprising anatomic partitioning. This design consists of stage I, stage II and stage III feature modules. Stage I consists of pretrained feature modules from each modality. These input features are fed into stage II to learn integrated features, which in turn are fed into the task-specific feature modules in stage III. The stage III modules are connected to the output nodes for joint prediction of the 2 tasks (ADAS-13 and MMSE score prediction). Prediction performance from both tasks is used to update the weights of the stage I and stage II feature modules. The partitioned architecture reduces the number of model parameters, which along with the pretrained feature modules helps mitigate overfitting issues. Input data dimensionality is as follows: 16 086 (left hippocampal), 16 471 (right hippocampal) and 686 (cortical thickness). For details regarding hyperparameters (number of hidden nodes, learning policies, weight regularization etc.) of APANN, see Table 2. ADAS-13 = Alzheimer's Disease Assessment Scale; ANN = artificial neural network; APANN = anatomically partitioned artificial neural network; MMSE = Mini Mental State Examination; ROI = region of interest.

preselected common space segmentation, and then excluded any candidate segmentations with a Dice κ of less than 1 standard deviation from the mean over all participants.

To further compact the bounding box of all candidate segmentations, we excluded voxels with low information density by keeping only structural voxels present in at least 25% of candidate segmentations across the ADNI1 and ADNI2 data sets. After filtering operations, the 3-dimensional volumes were flattened into a 1-dimensional vector of included voxels per candidate segmentation.

Upon completion of this process, the vectorized voxels represented the hippocampal input for the APANN model. The lengths of the input vectors were 16086 for the left hippocampus and 16471 for the right.

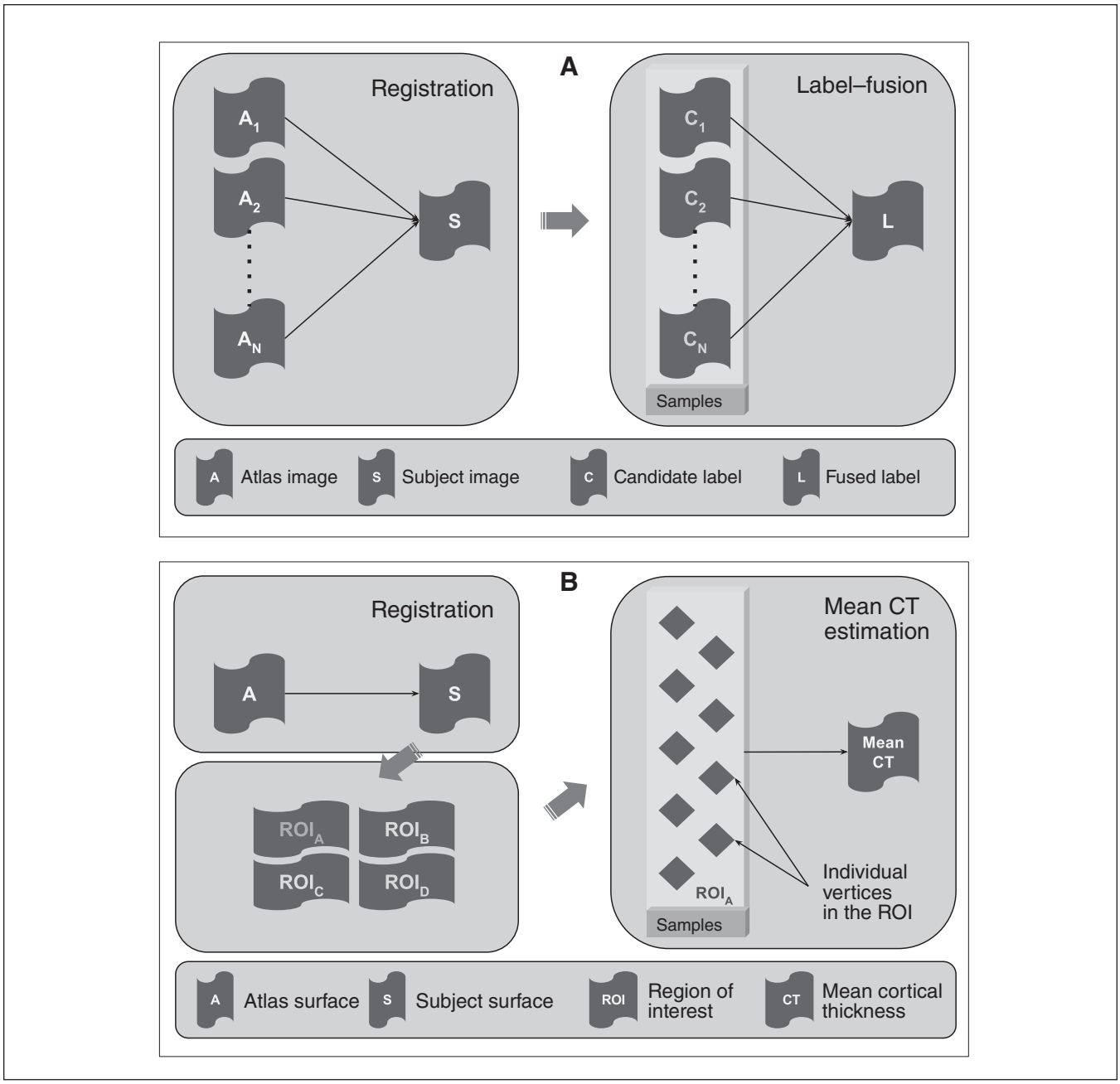

Fig. 2: (A) Schematic of a multi-atlas segmentation pipeline depicting registration and label-fusion stages. The box highlights the candidate labels derived from different atlases that were treated as empirical samples in the context of structural labels. These labels are usually fused into a single label that serves as a point-estimate mask of a given structure. (B) Schematic of a cortical thickness estimation pipeline comprising surface registration, parcellation and average thickness estimation. The box highlights the individual vertices in a given region of interest, which are treated as empirical samples in the context of the cortical thickness measure. The thickness values of these vertices are usually averaged out to estimate mean thickness over a region of interest. CT = cortical thickness; ROI = region of interest.

Cortical thickness

Preprocessing with CIVET produces cortical thickness values at 40 962 vertices per hemisphere. We assigned these cortical vertices to unique ROIs based on a predefined atlas. We created a custom atlas (Fig. 3) consisting of 686 ROIs, maintaining bilateral symmetry (343 ROIs per hemisphere) using data-driven parcellation based on spectral clustering (http://scikit-learn.org/stable/modules/generated/sklearn.cluster.spectral_clustering.html). Spectral clustering allows for the creation of ROIs with a similar number of vertices, which is desirable for unbiased sampling of vertices to estimate cortical thickness. Also, work by Khundrakpam and colleagues[63] suggests that increasing the spatial resolution of a cortical parcellation may improve predictive performance, further supporting the use of this data-driven atlas over neuroanatomically derived parcellations.[64,65] During implementation, we used the connectivity information from the cortical mesh of the template as the adjacency matrix. Upon generating sets of vertices per ROI, we treated each vertex as a sample from a distribution that characterized the thickness of that ROI. Thus, the cortical thickness features for each individual could be characterized by a distribution of thickness values per ROI, instead of the mean thickness values computed as point estimates (Fig. 2B).

Standardization across modalities

The independent empirical sampling processes for hippocampal and cortical thickness inputs necessitated a standardization step, which is described in Appendix 1, available at jpn.ca/180016.

Training procedure

The training procedure consisted of 2 parts: training individual branches per input modality and fine-tuning the unified model consisting of pretrained branches and additional integrated and task-specific feature layers. In the first part, we trained separate models independently using individual hippocampal and cortical thickness modalities (Fig. 1D). We trained the model to jointly predict both tasks (ADAS-13 scores and MMSE scores). At the end of this training procedure, we obtained the set of weights for the hidden layers in stage I for each input branch. We then extended the model with stage II and III hidden layers and further trained it to learn integrated and task-specific feature layers (Fig. 1E). We used both tasks in this training procedure as well. For both parts, we determined the hyperparameters of the model (Table 2) using an inner cross-validation loop. The code using Caffe toolbox (http://caffe.berkeleyvision.org/) for the APANN design and training is available at https://github.com/CobraLab/NI-ML/tree/master/projects/APANN. The computational resource requirements are provided in Appendix 1.

Performance validation

We compared the performance of the APANN model separately for prediction of MMSE and ADAS-13 scores. We conducted 3 experiments to compare the performance of each cohort separately and together: ADNI1, ADNI2 and ADNI1 + 2. The latter was an effort to evaluate model robustness in a context of multicohort, multisite studies, which is becoming increasingly prevalent in the field. In each experiment, we compared the performance of the 2 inputs separately and together: hippocampal input, cortical thickness input and a combined hippocampal + cortical thickness input. We used Pearson correlation (r) and root mean square error (RMSE) values between true and predicted clinical scores as our

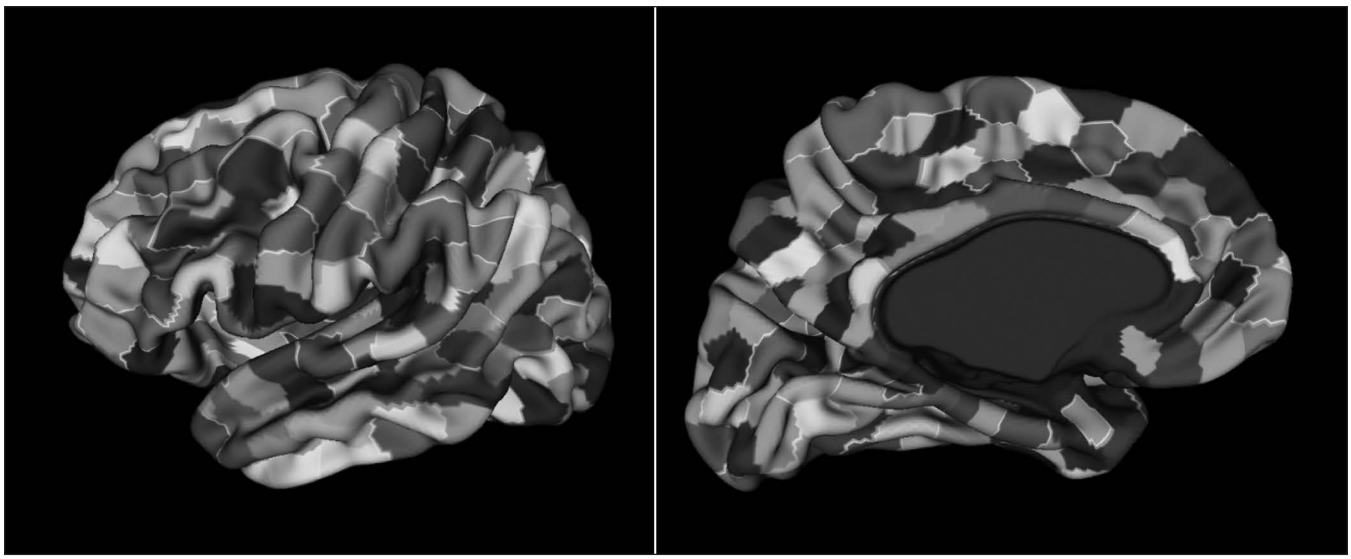

Fig. 3: A custom cortical surface parcellation (atlas) made up of 686 regions of interest, each consisting of a roughly equal number of vertices. We obtained the parcellations using a triangular surface mesh obtained from a CIVET model. The vertices of the mesh were grouped based on spatial proximity using a spectral clustering method (http://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html). Bilateral symmetry within the vertices of the hemispheres was preserved. The atlas was propagated to each participant to obtain thickness samples per region of interest.

performance metrics. We evaluated all experiments using 10 rounds of a 10-fold nested cross-validation procedure. The outer folds were created by dividing the participant pool into 10 nonoverlapping subsets. During each run, we chose 9 of 10 subsets as a training set and evaluated performance on the subset that was held back. During model training, we created 3 inner folds by further dividing the training set under consideration to determine the optimal combination of hyperparameters (e.g., number of hidden nodes) using a grid search. Then, we stratified the outer folds to maintain a similar ratio of ADNI1 and ADNI2 participants in each fold. We compared the performance of APANN in all experiments against 3 commonly used machine-learning models: linear regression with lasso, support vector machine and random forest. The results are provided in Appendix 1.

Our secondary, proof-of-concept analysis consisted of a longitudinal experiment to predict clinical scores at baseline and 1 year simultaneously, using only baseline MRI data. This was in an effort to demonstrate the applicability of APANN from a clinical standpoint, where the end goal was to predict a person's future diagnostic and/or prognostic states. We limited our analysis to the ADAS-13 scale (because its larger score range offered better sensitivity to longitudinal changes) and to the individual ADNI1 and ADNI2 cohorts. Because of missing data, the number of participants for this experiment dropped to 553 for ADNI1 and 590 for ADNI2.

Results

The mean correlation (r) and RMSE performance values for all 3 experiments with 3 input modality configurations are summarized in Figure 4, Table 3 and Table 4. Scatter plots for predicted and actual ADAS-13 and MMSE scores are shown in Figure 5. We generated scatter plots using scores from all test subsets in a randomly chosen round of a 10-fold run. Results for the longitudinal experiment are shown in Figure 6. Individual results for each experiment are detailed below. Comparisons with other models are provided in Appendix 1. Briefly, results from all 3 experiments indicated that the APANN model offered better predictive performance with hippocampal inputs. The cortical thickness input, when used independently, did not offer improvement. However, the combined hippocampal + cortical thickness input offered significantly higher performance improvement over reference models across all 3 experiments.

Experiment 1: ADNI1 cohort

The combined hippocampal + cortical thickness input provided the best results for ADAS-13 prediction (r = 0.60, RMSE = 7.11). We observed similar trends for MMSE prediction with the combined hippocampal + cortical thickness input (r = 0.52, RMSE = 2.25). The hippocampal input alone yielded findings of r = 0.53, RMSE = 7.56 for ADAS-13 score prediction and r = 0.40, RMSE = 2.41 for MMSE. The cortical thickness input alone yielded findings of r = 0.51, RMSE = 7.67 for ADAS-13 score prediction and r = 0.50, RMSE = 2.29 for MMSE.

Experiment 2: ADNI2 cohort

Similar to experiment 1, the combined hippocampal + cortical thickness input provided the best results for ADAS-13 prediction (r = 0.68, RMSE = 7.17). We observed similar trends for MMSE prediction with the combined hippocampal + cortical thickness input (r = 0.55, RMSE = 2.25). The hippocampal input alone yielded findings of r = 0.52, RMSE = 8.32 for ADAS-13 score prediction and r = 0.40, RMSE = 2.51 for MMSE. The cortical thickness input alone yielded findings of r = 0.63, RMSE = 7.58 for ADAS-13 score prediction and r = 0.52, RMSE = 2.31 for MMSE.

Table 2: Hyperparameter search space for the 4 models*

Model	Hyperparameters
Linear regression with lasso	L1-penalty: 0.001 to 1 (with increments of 0.01)
Support vector regression	Kernel: {linear, rbf}, C: [0.001, 0.01, 1, 10, 100]
Random forest regression	N_estimators: 10 to 210 (with increments of 25) min_sample_split: [2, 4, 6, 8]
APANN	Fixed hyperparameters Network architecture Stage I (input features): 2 hidden layers with equal nodes in each layer Stage II (integrated features): 1 hidden layer Stage III (task features): 1 hidden layer Activation nonlinearity: ReLU Tunable hyperparameters Stage I number of hidden nodes: [25, 50, 100, 200] Stage II number of hidden nodes: [25, 50] Stage III number of hidden nodes: [25, 50] Learning rate: [1e-6, 1e-5, 1e-4] Learning policy: [Nesterov, Adagrad] Weight_decay: [1e-4,1e-3,1e-2] Dropout rate: [0, 0.25, 0.5] (only for stage I)

APANN = anatomically partitioned artificial neural network; ReLU = rectified linear unit.
*We performed a grid search of the hyperparameters using a nested inner loop for each cross-validation round. For the APANN model, the fixed hyperparameters refer to a broader network of design choices that remained identical for all cross-validation rounds. The tunable hyperparameters for APANN were optimized for each fold.

Experiment 3: ADNI1 + 2 cohort

Similar to experiments 1 and 2, the combined hippocampal + cortical thickness input provided the best results for ADAS-13 prediction (r = 0.63, RMSE = 7.32). We observed similar trends for MMSE prediction with the combined hippocampal + cortical thickness input (r = 0.55, RMSE = 2.25). The hippocampal input alone yielded findings of r = 0.54, RMSE = 7.99 for ADAS-13 score prediction and r = 0.45, RMSE = 2.42 for MMSE. The cortical thickness input alone yielded findings of r = 0.57, RMSE = 7.79 for ADAS-13 score prediction and r = 0.50, RMSE = 2.37 for MMSE.

A further analysis of results in this experiment stratified by participant-cohort membership (ADNI1 v. ADNI2) showed that APANN had a smaller performance bias toward any particular cohort (i.e., models performing well on only a single cohort) than other models (see Appendix 1).

Longitudinal prediction

Similar to experiments 1 to 3, the combined hippocampal + cortical thickness input provided the best results (ADNI1: r = 0.58, RMSE = 7.1 for baseline and r = 0.59, RMSE = 9.08 for 1-year score prediction; ADNI2: r = 0.64, RMSE = 7.07 for baseline and r = 0.65, RMSE = 9.07 for 1-year score prediction). The hippocampal input alone yielded better performance than the cortical thickness input alone for baseline and 1-year score prediction in the ADNI1 cohort. The cortical thickness input alone yielded better performance than the hippocampal input alone for baseline and 1-year score prediction in the ADNI2 cohort.

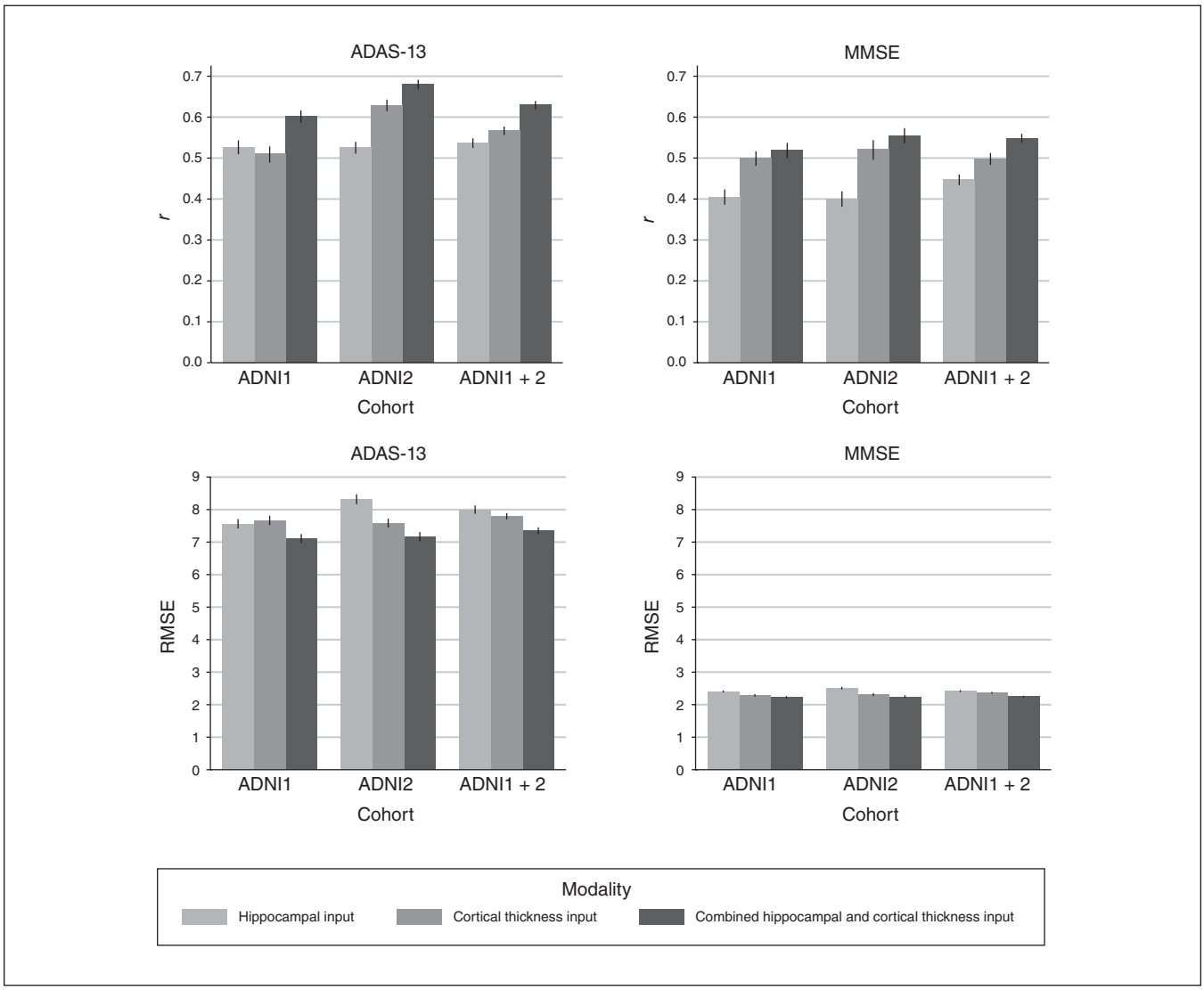

Fig. 4: Performance of anatomically partitioned artificial neural network subject to individual and combined input modalities. The Pearson r and RMSE values were averaged over 10 rounds of 10 folds. All models were trained with a nested inner loop that searched for optimal hyperparameters. ADAS-13 = Alzheimer's Disease Assessment Scale; ADNI = Alzheimer's Disease Neuroimaging Initiative; MMSE = Mini Mental State Examination; RMSE = root mean square error.

Discussion

We have presented an ANN model for the prediction of cognitive scores in Alzheimer disease using high dimensional structural MRI data. We showed that information from voxel-level hippocampal segmentations and highly granular cortical parcellations can be leveraged to infer cognitive performance and clinical severity at the level of the individual. This ability of the APANN model to predict ADAS-13 and MMSE and scores based on structural MRI features may prove to be valuable from a clinical perspective in helping to build prognostic tools. Our proof-of-concept longitudinal experiment showed that APANN could successfully predict future scores (at 1 year) from baseline MRI data. The results comparing APANN to several other models are provided in Appendix 1. These findings highlighted the performance gains offered by using high dimensional features as inputs. In the sections that follow, we discuss the performance of the APANN model in terms of clinical scale, input modalities, data sets and related literature.

Table 3: Prediction performance for ADAS-13 scores*

	Hippocampal input		Cortical thickness input		Combined hippocampal and cortical thickness input	
Model	*r*	RMSE	*r*	RMSE	*r*	RMSE
ADNI1						
Linear regression with lasso	0.22 ± 0.11	8.72 ± 0.81	0.56 ± 0.08	7.44 ± 0.72	0.56 ± 0.08	7.42 ± 0.74
Support vector regression	0.23 ± 0.11	8.70 ± 0.85	0.52 ± 0.08	7.68 ± 0.76	0.53 ± 0.08	7.62 ± 0.78
Random forest regression	0.15 ± 0.10	9.27 ± 0.80	0.54 ± 0.08	7.55 ± 0.76	0.54 ± 0.08	7.51 ± 0.77
APANN	0.53 ± 0.09	7.56 ± 0.76	0.51 ± 0.10	7.67 ± 0.76	0.60 ± 0.08	7.11 ± 0.72
ADNI2						
Linear regression with lasso	0.14 ± 0.11	9.69 ± 0.70	0.61 ± 0.07	7.77 ± 0.71	0.61 ± 0.07	7.78 ± 0.71
Support vector regression	0.21 ± 0.10	9.75 ± 0.79	0.63 ± 0.07	7.65 ± 0.68	0.63 ± 0.07	7.66 ± 0.70
Random forest regression	0.24 ± 0.09	9.77 ± 0.76	0.58 ± 0.07	7.97 ± 0.65	0.58 ± 0.08	7.97 ± 0.67
APANN	0.52 ± 0.07	8.32 ± 0.79	0.63 ± 0.07	7.58 ± 0.71	0.68 ± 0.06	7.17 ± 0.71
ADNI1 + 2						
Linear regression with lasso	0.12 ± 0.08	9.37 ± 0.50	0.58 ± 0.06	7.71 ± 0.48	0.58 ± 0.06	7.71 ± 0.48
Support vector regression	0.18 ± 0.07	9.39 ± 0.54	0.59 ± 0.05	7.65 ± 0.42	0.59 ± 0.05	7.65 ± 0.42
Random forest regression	0.18 ± 0.09	9.63 ± 0.61	0.57 ± 0.05	7.76 ± 0.46	0.57 ± 0.05	7.75 ± 0.46
APANN	0.54 ± 0.06	7.99 ± 0.59	0.57 ± 0.05	7.79 ± 0.51	0.63 ± 0.05	7.32 ± 0.53

APANN = anatomically partitioned artificial neural network; RMSE = root mean squared error; SD = standard deviation.
*Findings are presented as mean ± SD.

Table 4: Prediction performance for MMSE scores*

	Hippocampal input		Cortical thickness input		Combined hippocampal and cortical thickness input	
Model	*r*	RMSE	*r*	RMSE	*r*	RMSE
ADNI1						
Linear regression with lasso	0.23 ± 0.12	2.54 ± 0.18	0.49 ± 0.08	2.28 ± 0.17	0.50 ± 0.08	2.27 ± 0.17
Support vector regression	0.25 ± 0.12	2.59 ± 0.19	0.48 ± 0.07	2.31 ± 0.16	0.50 ± 0.07	2.28 ± 0.16
Random forest regression	0.22 ± 0.11	2.63 ± 0.21	0.48 ± 0.08	2.30 ± 0.17	0.49 ± 0.08	2.28 ± 0.17
APANN	0.40 ± 0.09	2.41 ± 0.15	0.50 ± 0.09	2.29 ± 0.20	0.52 ± 0.08	2.23 ± 0.17
ADNI2						
Linear regression with lasso	0.19 ± 0.12	2.64 ± 0.19	0.46 ± 0.08	2.39 ± 0.19	0.47 ± 0.08	2.39 ± 0.19
Support vector regression	0.28 ± 0.14	2.72 ± 0.24	0.52 ± 0.07	2.32 ± 0.18	0.54 ± 0.07	2.30 ± 0.18
Random forest regression	0.25 ± 0.12	2.67 ± 0.24	0.50 ± 0.09	2.33 ± 0.17	0.51 ± 0.08	2.31 ± 0.17
APANN	0.40 ± 0.09	2.51 ± 0.21	0.52 ± 0.12	2.31 ± 0.25	0.55 ± 0.10	2.25 ± 0.21
ADNI1 + 2						
Linear regression with lasso	0.15 ± 0.08	2.64 ± 0.12	0.50 ± 0.07	2.31 ± 0.13	0.50 ± 0.07	2.31 ± 0.13
Support vector regression	0.22 ± 0.07	2.71 ± 0.13	0.52 ± 0.07	2.31 ± 0.13	0.52 ± 0.07	2.30 ± 0.13
Random forest regression	0.17 ± 0.08	2.74 ± 0.14	0.50 ± 0.07	2.31 ± 0.14	0.50 ± 0.07	2.31 ± 0.14
APANN	0.45 ± 0.06	2.42 ± 0.14	0.50 ± 0.07	2.37 ± 0.15	0.55 ± 0.06	2.25 ± 0.12

APANN = anatomically partitioned artificial neural network; RMSE = root mean squared error; SD = standard deviation.
*Findings are presented as mean ± SD.

Clinical scale comparisons

Performance comparisons between clinical scales based on correlation values indicated that predicting MMSE scores was more challenging across all inputs and cohorts. This disparity between performances may have been due to the higher sensitivity of the ADAS-13 assessment, reflected in its comparatively larger scoring range, which improved its association with the structural measures.

Input modality comparisons

The results from all 3 experiments indicated that the APANN model offered better predictive performance with the combined hippocampal + cortical thickness input. The use of cortical thickness outperformed hippocampal segmentation in all 3 experiments for both scales, except in the ADNI1 cohort for ADAS-13 prediction, where the hippocampal segmentation input showed a slightly higher performance. This finding highlighted the importance of incorporating multiple phenotypes for biomarker development that are indicative of cognitive performance. The ability of the APANN model to handle multimodal input is crucial for building clinical tools to leverage disparate MRI, clinical and genetic markers.

Data set comparisons

Between experiments 1 and 2, we observed that the ADNI2 cohort yielded better performance than the ADNI1 cohort across all models. This may have been because of the differences in acquisition protocols, because ADNI2 images were acquired at a higher field strength with better resolution. Such an improvement in image acquisition would likely provide superior quality segmentations and cortical thickness measures.[66] In experiment 3, we combined data from the ADNI1 and ADNI2 cohorts. Pooling data from different data

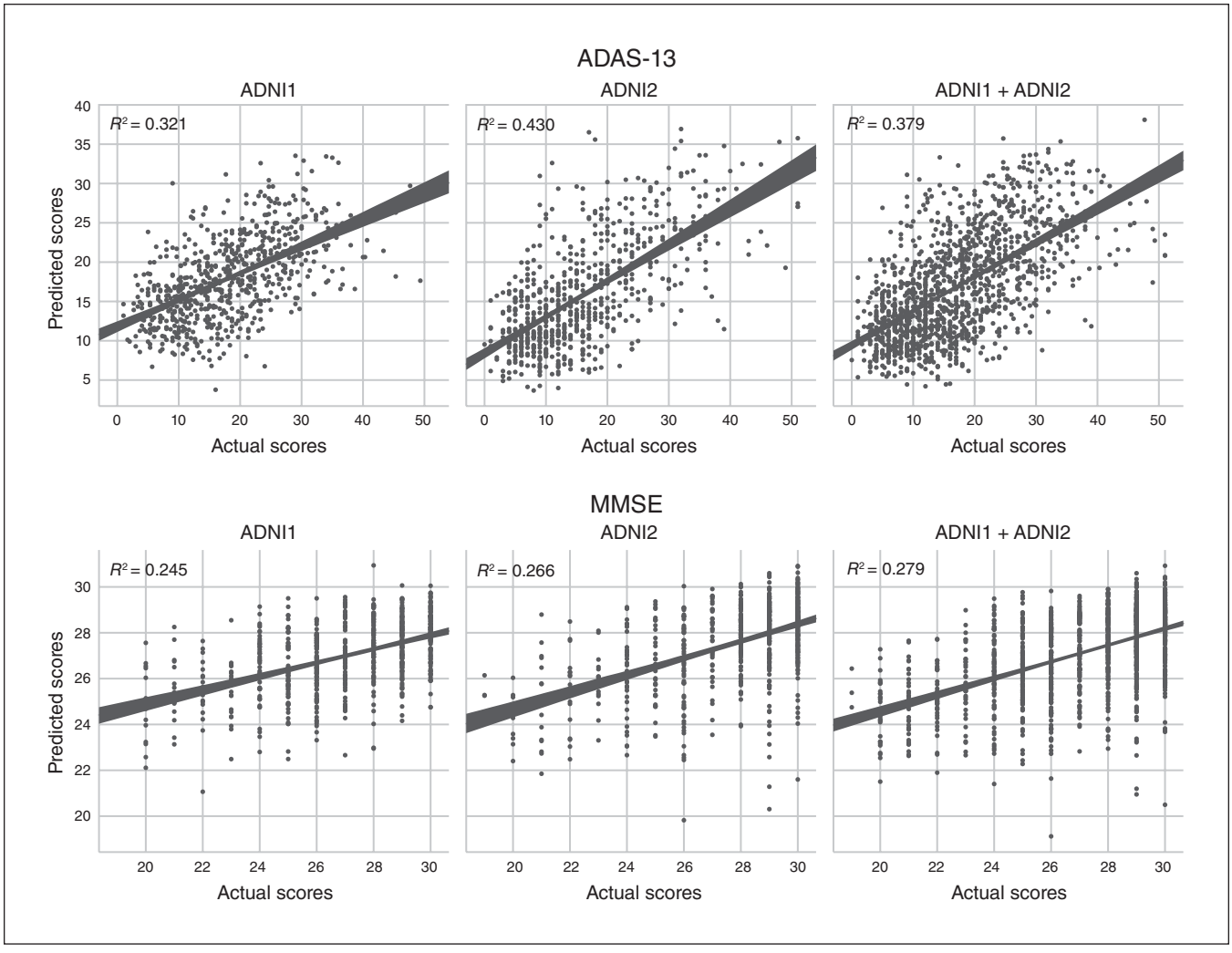

Fig. 5: Scatter plots for predicted and actual ADAS-13 and MMSE scores for 3 cohorts (ADNI1, ADNI2 and ADNI1 + 2). Scatter plots were generated by concatenating scores from all the test subsets of a randomly chosen 10-fold validation run. ADAS-13 = Alzheimer's Disease Assessment Scale; ADNI = Alzheimer's Disease Neuroimaging Initiative; MMSE = Mini Mental State Examination.

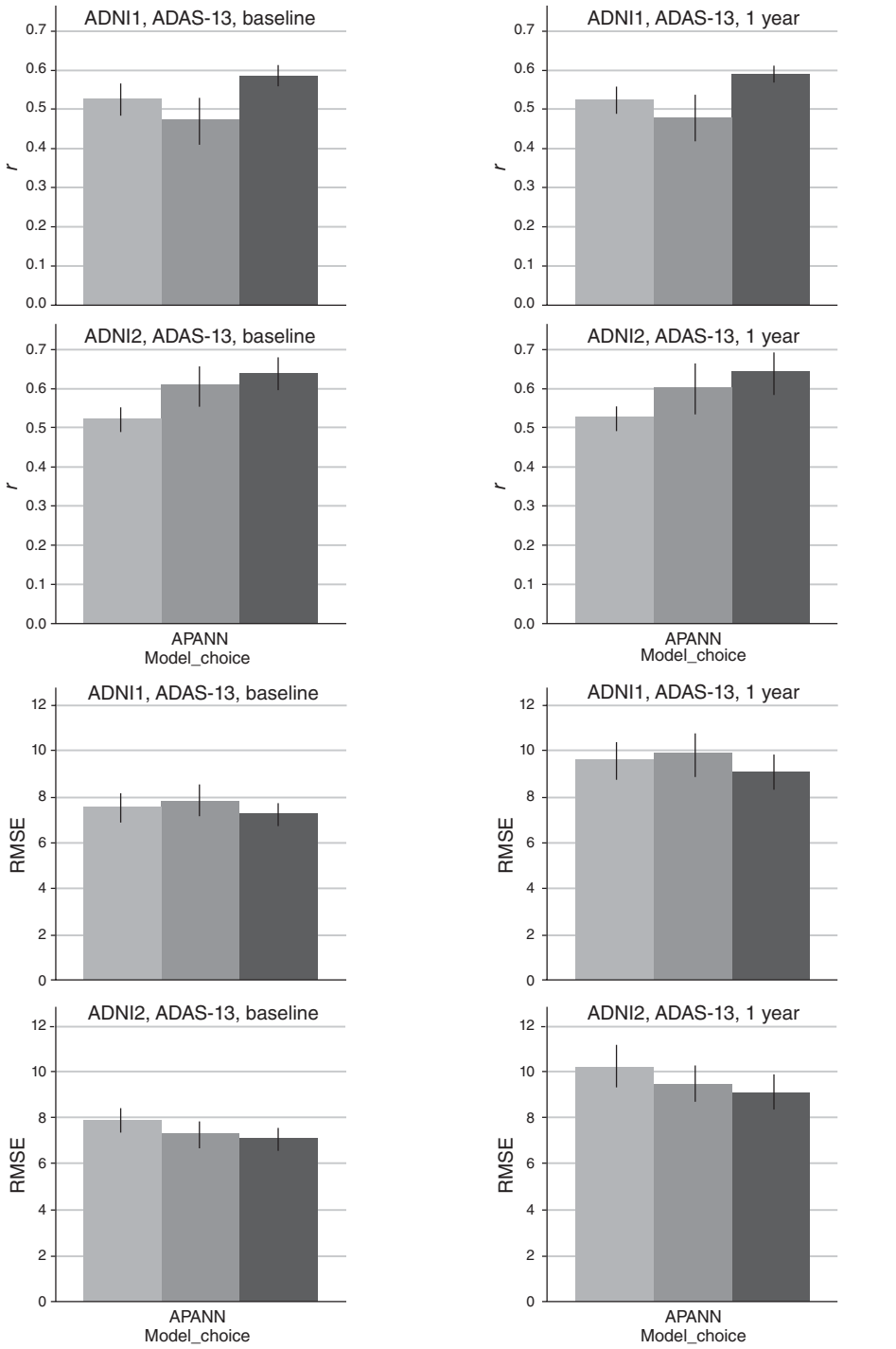

Fig. 6: Simultaneous predictions of ADAS-13 scores at baseline and 1 year. The top 2 rows show the Pearson *r* values based on predicted and actual ADAS-13 scores over 10-fold cross-validation for ADNI1 and ADNI2 cohorts respectively. The bottom 2 rows show the RMSE between predicted and actual ADAS-13 scores for ADNI1 and ADNI2 cohorts, respectively. The left column shows performance at baseline, and the right column shows performance at 1 year. Models were trained separately for each input. ADAS-13 = Alzheimer's Disease Assessment Scale; ADNI = Alzheimer's Disease Neuroimaging Initiative; APANN = anatomically partitioned artificial neural network; RMSE = root mean square error.

sets is increasingly important to verify the generalizability of the model in a larger population that extends beyond a single study. Interestingly, experiment 3 outperformed experiment 1, but underperformed compared with experiment 2. This was partially expected because of substantial differences in the individual feature distributions (e.g., hippocampal segmentations) resulting from differences in the acquisition protocols. In such cases, it becomes imperative to build models that are invariant to data set–specific biases resulting from nonuniform data-collection practices. The results from experiment 3 showed that APANN offered consistent performance that was comparable to that of experiments 1 and 2, and it had low data set–specific bias compared with other models (see Appendix 1). We speculate that models incorporating high dimensional, multimodal input were less susceptible to multicohort and multisite study-design artifacts, a characteristic that is desirable for the development of clinical tools in practical settings.

Longitudinal analysis

Consistent with the first 3 experiments, the combined hippocampal segmentation + cortical thickness input offered the best performance for 1-year score prediction, with similar correlation results but higher RMSE. This finding suggests that uncertainty is likely to increase with a larger time span for longitudinal tasks (1 year v. 2 years v. 5 years), making predictions more challenging. As well, further consideration is needed of cases in which information from multiple time points (baseline + 1 year) is used to generate subsequent (2 years +) performance prediction. Missing time points become an increasingly important barrier to such tasks. Nevertheless, APANN showed promising results for investigating more sophisticated longitudinal predictions.

Related work

Prediction of clinical scores is a relatively underexplored task. For a fair comparison, we have limited our discussion to 2 recent studies involving baseline prediction with MRI features by Stonington and colleagues[38] and Zhang and colleagues.[39] Both works used structural MRI from the ADNI1 baseline data set to predict MMSE and ADAS-cog scores (which uses 11 of the 13 subscales of ADAS-13; http://adni.loni.usc.edu/data-samples/data-faq/). The ADAS-cog and ADAS-13 scores are strongly correlated ($r > 0.9$ for the ADNI1 and ADNI2 cohorts considered in this manuscript). Stonington and colleagues[38] used relevance vector regression models with a sample size of 586; correlation values were $r = 0.48$ (MMSE) and $r = 0.57$ (ADAS-cog). Zhang and colleagues[39] proposed a computational framework called Multi-Modal Multi-Task (M3T) that offers multitask feature selection and multimodal support vector machines for regression and classification tasks. With only MRI-based features, M3T achieved correlations of $r = 0.50$ (MMSE) and $r = 0.60$ (ADAS-cog) with a sample size of 186. In comparison, the APANN model offered correlations of $r = 0.52$ (MMSE), and $r = 0.60$ (ADAS-13) with a much larger cohort (669 ADNI1 participants). Although APANN offered similar performance for the ADNI1 data set, it had several key advantages over the other models. In contrast to M3T, which implemented 2 separate stages for feature extraction and regression (or classification) tasks, APANN provided a unified model that performed seamless feature extraction and multitask prediction using multimodal input. From a scalability perspective, APANN was capable of handling high dimensional input and extending to incorporate new modalities without retraining the entire model. In contrast, M3T had 93 magnetic resonance atlas-based features[64] with a total of 189 multimodal (MRI, FDG-PET and cerebrospinal fluid) features.[39] Moreover, with APANN we replicated performance in the ADNI2 cohort and demonstrated an improved correlation performance of $r = 0.55$ (MMSE) and $r = 0.68$ (ADAS-13) with 690 participants, further validating the model's generalizability.

Other recent works have addressed clinical score prediction using sparse Bayesian learning[67] and graph-guided feature selection,[68] with 98 and 93 imaging features, respectively. Both works reported strong performance in Alzheimer disease and cognitively normal groups, but performance degraded after inclusion of people with MCI. For example, Yu and colleagues[68] reported correlations of $r = 0.745$ (MMSE) and $r = 0.74$ (ADAS-cog) for specific subsets of Alzheimer disease/cognitively normal participants, but performance degraded to $r = 0.382$ (MMSE) and $r = 0.472$ (ADAS-cog) for a subset of MCI/cognitively normal participants. Clinically, the prognosis of people with MCI is of high interest. Predicting their cognitive performance is crucial for early interventions, potential lifestyle changes and treatment planning. To the best of our knowledge, APANN is the first work to tackle high input dimensionality (> 30 000 features), validated across the continuum from healthy controls to patients with Alzheimer disease, in multiple cohorts with site and scanner differences. Such validation is increasingly important with the availability of newer and larger data sets, such as the UK biobank (www.ukbiobank.ac.uk/about-biobank-uk/).

Clinical translation

The ultimate clinical goal of this work is to provide longitudinal prognosis and to predict individuals' future clinical states. The rigorously validated APANN provides a computational platform for a variety of longitudinal tasks, such as the 1-year ADAS-13 prediction task investigated in the proof-of-concept experiment. We envision the APANN model applied to the MRI data of people at risk from prodromal stages (MCI, significant memory concern etc.) and even early stages of Alzheimer disease to predict their future clinical scores and other clinical-state proxies. The ability of the APANN model to capture relevant subtle neuroanatomical changes from high dimensional, multimodal MRI data can be leveraged to provide nuanced diagnosis and prognosis for various symptom subdomains, assisting or verifying clinicians' decision-making. Having a clear prognosis can help with early intervention, clinical trial recruitment and caregiver arrangements.

Limitations

In this work we applied APANN primarily to cross-sectional data sets and a proof-of-concept longitudinal data set. From a clinical perspective, it is crucial to note that the use of a specific clinical or cognitive test is subjective, contingent on availability and associated with its own set of biases. Further, similar to the clinical diagnosis that uses several sources of information to create a composite of the patient's clinical profile, we envision the proposed MRI-based prediction framework as another assistive instrument that will be interpreted in the larger context of an overall clinical picture. We acknowledge that the cross-sectional experiments in this work were a first step toward building assistive MRI-based models. We believe that the design flexibility of APANN can be used for handling multimodal input and multiple scale predictions that could minimize modality-specific and scale-specific biases, respectively.

Large-scale models such as APANN that are subjected to high dimensional input require significant computational resources. Thus, we have limited the scope of this work to classical ANNs as a prototypical example to demonstrate the feasibility of large-scale analysis with structural neuroimaging data. Nevertheless, the training regimens discussed in this work should motivate further development of state-of-the-art neural network architectures, such as 3-dimensional convolutional networks, toward various neuroimaging applications. Another common drawback of models with deep architectures is the lack of interpretability of the model parameters compared with simpler models; this prohibits localizing most predictive brain regions. In our view, this limitation is a model design trade-off that in turn allows for the capture of distributed changes that are often present in the heterogeneous atrophy patterns of Alzheimer disease prodromes. The computational flexibility of ANNs allow us to model the collective impact of these atrophy patterns and predict clinical performance more accurately.

Conclusion

The presented APANN model, together with empirical sampling procedures, offers a sophisticated machine-learning framework for high dimensional, multimodal structural neuroimaging analysis. By going beyond low-dimensional, anatomic prior-based feature sets, we can build more sensitive models capable of capturing the subtle neuroanatomical changes associated with cognitive symptoms in Alzheimer disease. The results validate the strong predictive performance of the APANN model across 2 independent cohorts, as well as its robustness when these 2 cohorts were combined. From clinical standpoint, these attributes make APANN a promising approach for building diagnostic and prognostic tools that would help identify people at risk and provide clinical-trajectory assessments, facilitating early intervention and treatment planning.

Acknowledgements: N. Bhagwat receives support from the Alzheimer Society of Canada. A. Voineskos is funded by the Canadian Institutes of Health Research, the Ontario Mental Health Foundation, the Brain and Behavior Research Foundation and the National Institute of Mental Health (R01MH099167 and R01MH102324). M. Chakravarty is funded by the Weston Brain Institute, the Alzheimer Society of Canada, the Michael J. Fox Foundation for Parkinson's Research, the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada and Fondation de Recherches Santé Québec. Computations were performed on the GPC supercomputer at the SciNet HPC Consortium and the Kimel Family Translational Imaging-Genetics Research (TIGR) Lab computing cluster. SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada, the Government of Ontario, the Ontario Research Fund Research Excellence Program and the University of Toronto. The TIGR Lab cluster is funded by the Canada Foundation for Innovation Research Hospital Fund. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904), and ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from Abbott; the Alzheimer's Association; the Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research Development LLC.; Johnson & Johnson Pharmaceutical Research Development LLC; Medpace Inc.; Merck & Co. Inc.; Meso Scale Diagnostics LLC; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research provides funds to support ADNI clinical sites in Canada. Private-sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is Rev March 26, 2012, coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. The ADNI data are disseminated by the Laboratory for Neuro-Imaging at the University of California, Los Angeles. This research was also supported by National Institutes of Health grants P30 AG010129 and K01 AG030514.

Affiliations: From the Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ont. (Bhagwat, Chakravarty); the Cerebral Imaging Centre, Douglas Mental Health University Institute, Verdun, Que. (Bhagwat, Chakravarty); the Kimel Family Translational Imaging-Genetics Research Lab, Research Imaging Centre, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ont. (Bhagwat, Pipitone, Voineskos); the Department of Psychiatry, University of Toronto, Toronto, Ont. (Voineskos); and the Department of Psychiatry, McGill University, Montreal, Que. (Chakravarty), Canada.

Competing interests: None declared.

Contributors: N. Bhagwat, J. Pipitone and M. Chakravarty designed the study. Data were collected by the Alzheimer's Disease Neuroimaging Initiative, and all authors participated in data analysis. N. Bhagwat and J. Pipitone wrote the article, which all authors reviewed. All authors approved the final version to be published and can certify that no other individuals not listed as authors have made substantial contributions to the paper.

References

1. Gerardin E, Chételat G, Chupin M, et al. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *Neuroimage* 2009;47:1476-86.
2. Chupin M, Gérardin E, Cuingnet R, et al. Fully automatic hippocampus segmentation and classification in Alzheimer's

disease and mild cognitive impairment applied on data from ADNI. *Hippocampus* 2009;19:579-87.
3. Cuingnet R, Gerardin E, Tessieras J, et al. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuroimage* 2011;56:766-81.
4. Zhang D, Wang Y, Zhou L, et al.; Alzheimer's Disease Neuroimaging Initiative. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 2011;55:856-67.
5. Casanova R, Hsu F-C, Sink KM, et al. Alzheimer's disease risk assessment using large-scale machine learning methods. *PLoS One* 2013;8:e77949.
6. Murray ME, Graff-Radford NR. Ross OA, Petersen RC, Duara R, Dickson DW. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol* 2011;10:785-96.
7. Coupé P, Manjón JV, Fonov V, et al. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. *Neuroimage* 2011;54:940-54.
8. Stern Y. Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurol* 2012;11:1006-12. 9.
Whitwell JL, Dickson DW, Murray ME, et al. Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study. *Lancet Neurol* 2012;11:868-77.
10. Lam B, Masellis M, Freedman M, et al. Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. *Alzheimers Res Ther* 2013;5:1.
11. Eskildsen SF, Coupe P, Garcia-Lorenzo D, et al. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *Neuroimage* 2013;65:511-21.
12. Braak H, Braak E. Frequency of stages of Alzheimer-related lesions in different age categories. *Neurobiol Aging* 1997;18:351-7.
13. Ferreira D, Verhagen C, Hernández-Cabrera JA, et al. Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Sci Rep* 2017;7:46263.
14. Atluri G, Padmanabhan K, Fang G, et al. Complex biomarker discovery in neuroimaging data: finding a needle in a haystack. *Neuroimage Clin* 2013;3:123-31.
15. Scheltens NME, Galindo-Garre F, Pijnenburg YAL, et al. The identification of cognitive subtypes in Alzheimer's disease dementia using latent class analysis. *J Neurol Neurosurg Psychiatry* 2016;87:235-43.
16. Jack CR Jr, Petersen RC, Xu YC, et al. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology* 1999;52:1397-403.
17. Duchesne S, Caroli A, Geroldi C, et al. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *Neuroimage* 2009;47:1363-70.
18. Frisoni GB, Fox NC, Jack CR Jr, et al. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol* 2010;6:67-77.
19. Sabuncu MR, Desikan RS, Sepulcre J, et al. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Arch Neurol* 2011;68:1040-8.
20. Coupé P, Eskildsen SF, Manjón JV, et al. Alzheimer's Disease Neuroimaging Initiative. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. *Neuroimage* 2012;59:3736-47.
21. Eskildsen SF, Coupe P, Fonov VS, et al. Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. *Neurobiol Aging* 2015;36:S23-31.
22. Perrin RJ, Fagan AM, Holtzman DM. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. *Nature* 2009;461:916.
23. La Joie R, Perrotin A, de La Sayette V, et al. Hippocampal subfield volumetry in mild cognitive impairment, Alzheimer's disease and semantic dementia. *Neuroimage Clin* 2013;3:155-62.
24. Pipitone J, Park MTM, Winterburn J, et al. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *Neuroimage* 2014; 101:494-512.
25. Bhagwat N, Pipitone J, Winterburn JL, et al. Manual-protocol inspired technique for improving automated MR image segmentation during label fusion. *Front Neurosci* 2016;10:325.
26. Sankar T, Park MTM, Jawa T, et al. Your algorithm might think the hippocampus grows in Alzheimer's disease: caveats of longitudinal automated hippocampal volumetry. *Hum Brain Mapp* 2017;38:2875-96.
27. Mueller SG, Weiner MW. Selective effect of age, Apo e4, and Alzheimer's disease on hippocampal subfields. *Hippocampus* 2009;19:558-64.
28. Amaral RSC, Park MTM, Devenyi GA, et al. Manual segmentation of the fornix, fimbria, and alveus on high-resolution 3T MRI: application via fully-automated mapping of the human memory circuit white and grey matter in healthy and pathological aging. *Neuroimage* 2018;170:132-50.
29. Lerch JP, Pruessner JC, Zijdenbos A, et al. Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. *Cereb Cortex* 2005;15:995-1001.
30. Klöppel S, Stonnington CM, Chu C, et al. Automatic classification of MR scans in Alzheimer's disease. *Brain* 2008;131:681-9.
31. Lerch JP, Pruessner J, Zijdenbos AP, et al. Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiol Aging* 2008;29:23-30.
32. Querbes O, Aubry F, Pariente J, et al. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain* 2009;132:2036-47.
33. Davatzikos C, Bhatt P, Shaw LM, et al. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol Aging* 2011;32:2322.e19-27.
34. Coupé P, Eskildsen SF, Manjón JV, et al. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. *Neuroimage Clin* 2012;1:141-52.
35. Moradi E, Pepe A, Gaser C, et al.; Alzheimer's Disease Neuroimaging Initiative. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* 2015;104:398-412.
36. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. *Am J Psychiatry* 1984;141:1356-64.
37. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975;12:189-98.
38. Stonnington CM, Chu C, Klöppel S, et al. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *Neuroimage* 2010;51:1405-13.
39. Zhang D, Shen D; Alzheimer's Disease Neuroimaging Initiative. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* 2012;59:895-907.
40. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;313:504-7.
41. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 2013;35:1798-828.
42. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. *Advances in neural information processing systems 25.* Red Hook (NY): Curran Associates, Inc.; 2012:1097-105.
43. Plis SM, Hjelm DR, Salakhutdinov R, et al. Deep learning for neuroimaging: a validation study. *Front Neurosci* 2014;8:229.
44. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *MICCAI* 2015;9351:234-41.
45. Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models. In: Xing EP, Jebara T, editors. *Proceedings of the 31st International Conference on Machine Learning;* 2014 Jun. 22–24; Beijing, China. *J Machine Learning Res* 2014;32:595-603.
46. Wyman BT, Harvey DJ, Crawford K, et al. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimers Dement* 2013;9:332-7.
47. O'Bryant SE, Humphreys JD, Smith GE, et al. Detecting dementia with the Mini-Mental State Examination in highly educated individuals. *Arch Neurol* 2008;65:963-7.
48. Sheehan B. Assessment scales in dementia. *Ther Adv Neurol Disord* 2012;5:349-58.
49. Jack CR Jr, Knopman DS, Jagust WJ, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol* 2013;12:207-16.

50. Iturria-Medina Y, Sotero RC, Toussaint PJ, et al.; Alzheimer's Disease Neuroimaging Initiative. Early role of vascular dysregulation on late-onset Alzheimer's disease based on multifactorial data-driven analysis. *Nat Commun* 2016;7:11934.
51. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* 2010;29:1310-20.
52. Eskildsen SF, Coupé P, Fonov V, et al. BEaST: brain extraction based on nonlocal segmentation technique. *Neuroimage* 2012; 59:2362-73.
53. Loken C, Gruner D, Groer L, et al. SciNet: lessons learned from building a power-efficient top-20 system and data centre. *J Phys Conf Ser* 2010:256;012026.
54. Chakravarty MM, Steadman P, van Eede MC, et al. Performing label-fusion-based segmentation using multiple automatically generated templates. *Hum Brain Mapp* 2013;34:2635-54.
55. Winterburn JL, Pruessner JC, Chavez S, et al. A novel in vivo atlas of human hippocampal subfields using high-resolution 3T magnetic resonance imaging. *Neuroimage* 2013;74:254-65.
56. Collins DL, Neelin P, Peters TM, et al. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *J Comput Assist Tomogr* 1994;18:192-205.
57. MacDonald D, Kabani N, Avis D, et al. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. *Neuroimage* 2000;12:340-56.
58. Kim JS, Singh V, Lee JK, et al. Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *Neuroimage* 2005;27:210-21.
59. Ad-Dab'bagh Y, Lyttelton O, Muehlboeck JS, et al. The CIVET image-processing environment: a fully automated comprehensive pipeline for anatomical neuroimaging research. In: Corbetta M, Nichols T, Pietrini P, editors. *Proceedings of the 12th Annual Meeting of the Organization for Human Brain Mapping*; 2006 Jun. 11–15; Florence, Italy. *Neuroimage* 2006;31(Suppl 1):2266.
60. Suk H-I, Lee S-W, Shen D; Alzheimer's Disease Neuroimaging Initiative. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct Funct* 2015;220:841-59.
61. Mwangi B, Tian TS, Soares JC. A review of feature reduction techniques in neuroimaging. *Neuroinformatics* 2014;12:229-44.
62. Avants BB, Epstein CL, Grossman M, et al. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal* 2008;12:26-41.
63. Khundrakpam BS, Tohka J, Evans AC; Brain Development Cooperative Group. Prediction of brain maturity based on cortical thickness at different spatial resolutions. *Neuroimage* 2015;111:350-9.
64. Kabani NJ. 3D anatomical atlas of the human brain. *Neuroimage*. 1998;7:P-0717.
65. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 2002;15:273-89.
66. Chow N, Hwang KS, Hurtz S, et al. Comparing 3T and 1.5T MRI for mapping hippocampal atrophy in the Alzheimer's Disease Neuroimaging Initiative. *AJNR Am J Neuroradiol* 2015;36:653-60.
67. Wan J, Zhang Z, Yan J, et al. Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2012 Jun. 16-21; Providence (RI). Red Hook (NY): Curran Associates, Inc.; 2012:940-7.
68. Yu G, Liu Y, Shen D. Graph-guided joint prediction of class label and clinical scores for the Alzheimer's disease. *Brain Struct Funct* 2016;221:3787-801.